CALCULUS LECTURE NOTES

VAUGHN CLIMENHAGA

SOME CONTEXT. These notes were written to supplement a two-semester calculus course designed for math majors at the University of Houston, which I taught for several years. Much of the material and the approach in these notes is drawn from the textbooks by Spivak and by Stewart, to which I am deeply indebted.

CONTENTS

Date: February 25, 2026.

Part I. Functions, limits, and continuity

Lecture 1 — Sets of numbers

The material in this lecture corresponds to Chapters 1 and 2 of Spivak's book. For full details of the construction of the real numbers – which we do not give in this course! – see Chapters 28–30 of Spivak.

Consider the following numbers.

$$0 \qquad 1 \qquad 2 \qquad -1 \qquad \frac{1}{2} \qquad \sqrt{2} \qquad \pi \qquad e \qquad i$$

Now imagine that you meet an alien who knows mathematics and has learned English but does not know any of our notation for numbers, and in particular has a different way of writing all the numbers in the list above. How would you describe those numbers to the alien in such a way that you can be sure you are both talking about the same thing?

A reasonable description of 0 might be "the number that represents nothing". Another way of describing it would be via the property that adding 0 to any number does not change it: formally, we write

$$\text{for every number } x, \text{ we have } x + 0 = x.$$

Observe that 0 is the only number with this property, so this describes 0 uniquely.

Now you might describe 1 as "the smallest positive whole number", and 2 as "the whole number that comes after 1", or "the smallest whole number larger than 1". For -1 we might say "it is the only number x with the property that $x + 1 = 0$"; recall that we already described 0 and 1. Similarly, $\frac{1}{2}$ can be described as "the only number x with the property that $2x = 1$".

We might try to do the same thing with $\sqrt{2}$ and say that it is "the only number x with the property that $x^2 = 2$". But this doesn't quite work, since $x = -\sqrt{2}$ also has this property. So we need to add the word "positive" to that description; $\sqrt{2}$ is the only *positive* number x such that $x^2 = 2$.

What about π? We might describe π as "the ratio of the circumference of a circle to its diameter", or as "the area of a circle with radius 1". But how do we know that these values are the same as each other, or that they don't change from one circle to another? And what exactly do we mean by "circumference" and "area"? There is no trouble with talking about perimeter and area for polygons with straight edges, but for a shape like a circle that involves curves, these concepts are not so straightforward; as we will see later on, the way to define them for the circle is to *approximate* the circle with polygons.[1]

You have probably encountered the number $e \approx 2.71828\ldots$ as "the natural logarithmic base". But what does that mean? In fact it is not so easy to give a suitable description of this number, and we will have to postpone this until later in the course.

[1] One may also wonder if there is a way to describe π without invoking geometry; after all, $\sqrt{2}$ also has an important geometric description as "the length of the diagonal of a square with side length 1", but as we saw above can also be described in a more direct way.

Finally, the number i can be described as "the number x with the property that $x^2 = -1$". But two problems arise at this point. First, there is no real number with this property; i is an *imaginary* number (though in general we will prefer the terminology "complex number"), and so we really need to start talking about what exactly we mean by the word "number". The second problem is the opposite of the first; if we accept i as a "number", then $-i$ *also* has the property that $(-i)^2 = -1$, and so "the number x with the property that $x^2 = -1$" describes both i and $-i$. We will ignore this second problem for now,[2] and pause briefly to identify several morals of the preceding discussion.

(1) It is important to give *precise definitions* of things. But in your previous mathematics courses, it has probably happened that you encountered and used some mathematical objects (such as the number e) without having a completely satisfactory definition. We will attempt to move away from this practice, and to put everything on a more careful foundation.
(2) It is reasonable to define a number, or other mathematical object, in terms of its properties – "the only number x with the property that…" – however, when we do so we must consider the questions of *existence* (are there any numbers x with this property?) and of *uniqueness* (does this property uniquely determine the number?), which play a fundamental role in many places in mathematics.
(3) When dealing with general real numbers, it is often useful (if not essential) to use a process of *approximation*. We will explore this point in more detail very soon.

The previous discussion concerned numbers. But what is a number? By "number" we generally mean an element of one (or more) of the following sets.

- $\mathbb{N} = \{1, 2, 3, \dots\}$ is the set of *natural numbers*. Some people include 0 in $\mathbb{N}$ as well.
- $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ is the set of *integers*. The notation stands for the German word *Zahlen* (numbers).
- $\mathbb{Q} = \{p/q : p \in \mathbb{Z}, q \in \mathbb{N}\}$ is the set of *rational numbers*. The notation stands for "quotient". Recall that the "set builder" notation used here means "the set of all possible quotients p/q formed by choosing $p \in \mathbb{Z}$ and $q \in \mathbb{N}$".[3]
- $\mathbb{R}$ is the set of *real numbers*.
- $\mathbb{C} = \{x + iy : x, y \in \mathbb{R}\}$ is the set of *complex numbers*. Here i is the *imaginary* square root of -1.

These sets are nested:

$$\mathbb{N} \subsetneqq \mathbb{Z} \subsetneqq \mathbb{Q} \subsetneqq \mathbb{R} \subsetneqq \mathbb{C}. \tag{1.1}$$

Observe that we didn't say anything about what the set of real numbers actually is. You may have heard $\mathbb{R}$ described as the set of all points on the number line; this amounts to a geometric interpretation where each positive real number should correspond to the

[2]Basically the resolution of this second problem is that we just pick one of the numbers x with that property, call it i, and get on with our life. Maybe the alien picked the other one, but it turns out not to matter; formally, the map $i \mapsto -i$ gives an involution of the set of complex numbers that preserves everything we're interested in. But you don't need to worry about this here…

[3]The notation "$p \in \mathbb{Z}$" means "p is an element of the set of integers", or more concisely, "p is in the set of integers", or even more concisely, "p is an integer".

length of some line segment in Euclidean space. This approach is reasonable but some difficulties arise when we try to interpret arithmetic through this lens. For the sum of two numbers x and y, we could put together line segments with those lengths and with the same direction to make line segments with lengths $x + y$ and $x - y$. But it is not so clear how to do multiplication or division; how can we use two line segments with lengths x and y to produce line segments with lengths xy and x/y?[4]

Taking a more symbolic approach, we may observe that every real number x has a decimal representation

$$x = \lfloor x \rfloor + 0.a_1a_2a_3a_4\cdots, \tag{1.2}$$

where $\lfloor x \rfloor$ means "the greatest integer that is less than or equal to x", and each a_i is a digit from the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

Since we have rules for adding, subtracting, multiplying, and dividing decimal representations, this seems to give us the tools that we need to work with real numbers. However, there are a few ambiguities and questions that we would do well to keep in mind.

- The rules for arithmetic operations with decimal representations are defined for *finite* decimal representations, not infinite ones. For example, in the algorithm for addition, we start at the right-most decimal place and then work to the left. But if the decimal representation is infinite, there is no right-most decimal place. So what do we do?
- A single real number may have more than one decimal representation: for example, $1 = 0.99999\cdots$. So it is not quite accurate to say that "$\mathbb{R}$ is the set of all decimal representations".
- Why is $\mathbb{Q}$ not enough? That is, is $\mathbb{R} \setminus \mathbb{Q} = \{x \in \mathbb{R} : x \notin \mathbb{Q}\}$ nonempty? We talk about "the square root of 2" (for example) as the unique real number $x > 0$ such that $x^2 = 2$, but how do we know that such a real number exists? And if we know that it exists, how do we know that it is not actually a rational number?

For the first issue raised above, it is instructive to consider how we might compute a decimal expansion for $x + y$ in the following cases.

(1) $x = 0.123123123\ldots$ (which we abbreviate as $x = 0.\overline{123}$) and $y = 0.\overline{456}$.
(2) $x = 0.\overline{456}$ and $y = 0.\overline{789}$.
(3) $x = 0.\overline{142857}$ and $y = 0.\overline{3}$.
(4) $x = 0.959559559559559\ldots$ and $y = 0.\overline{4}$.

In the first example, there is no issue; we simply add separately in each decimal place, there are no carries, and we obtain

$$\begin{array}{r} 0.123123123\ldots \\ +0.456456456\ldots \\ \hline = 0.579579579\ldots \end{array}$$

[4]Of course we can use them to make a rectangle with *area* xy, but this is not the same thing.

In the second example, life gets a little more complicated, since the sum in every decimal place is ≥ 10, so every decimal place has a carry:

$$\begin{array}{r} {\scriptstyle 1\ 1\,1\,1\,1\,1\,1\,1\,1\,1} \\ 0.456456456\ldots \\ +0.789789789\ldots \\ \hline = 1.246246246\ldots \end{array}$$

In the third example, we have carries in some places but not in others:

$$\begin{array}{r} {\scriptstyle 1\ \ 1\ \ \ \ 1\ \ 1} \\ 0.142857142857\ldots \\ +0.333333333333\ldots \\ \hline = 0.476190476190\ldots \end{array}$$

Our implicit strategy here has been to start at the left and work to the right – which is fine because the decimal expansion *has* a leftmost position – and at each step to check whether or not there is a carry by looking at the next digit and determining whether or not the sum will be ≥ 10. So far this has worked just fine, because we can check this by looking only one digit ahead. But the fourth example requires a little more care: when we add $9 + 4$ we clearly get a carry in the previous position, but what happens when we add $5 + 4$? In this case we need to look one digit further and see if *that* digit sum is ≥ 10, in which case we would end up with $5 + 4 + 1 = 10$. For the values of x and y that are given, we actually end up with a carry in every position – but we sometimes need to look very far ahead to confirm this!

Another strategy that works for the first three examples is to observe that x and y have repeating decimal expansions and thus actually represent rational numbers. In the third example above, we have $x = \frac{1}{7}$ and $y = \frac{1}{3}$, so $x + y = \frac{3}{21} + \frac{7}{21} = \frac{10}{21}$, and with some more work you can see that this agrees with the decimal answer we got. This strategy, though, does not work for the fourth example, because in that instance x cannot be represented as a fraction.

One final strategy, which works equally well for all choices of x and y, is the following: for each $n \in \mathbb{N}$, truncate x and y to the first n digits of their decimal expansions, which can be easily added by starting at the right and working left. As n gets larger and larger, so that we write down more and more digits of x and y, this should give us a better and better approximation to the true value of $x + y$. Here is what the first few steps of this procedure look like for the fourth example above:

$$\begin{array}{r} {\scriptstyle 1} \\ 0.9 \\ +0.4 \\ \hline = 1.3 \end{array} \qquad \begin{array}{r} {\scriptstyle 1} \\ 0.95 \\ +0.44 \\ \hline = 1.39 \end{array} \qquad \begin{array}{r} {\scriptstyle 1\ 1\,1} \\ 0.959 \\ +0.444 \\ \hline = 1.403 \end{array} \qquad \begin{array}{r} {\scriptstyle 1\ 1\,1} \\ 0.9595 \\ +0.4444 \\ \hline = 1.4039 \end{array} \qquad \begin{array}{r} {\scriptstyle 1\ 1\,1} \\ 0.95955 \\ +0.44444 \\ \hline = 1.40399 \end{array} \qquad \begin{array}{r} {\scriptstyle 1\ 1\,1\,1\,1\,1} \\ 0.959559 \\ +0.444444 \\ \hline = 1.404003 \end{array}$$

We see that as we take more and more digits in our computation, the sum is getting closer and closer to $1.4040040004\ldots$, which is the value of $x + y$.

We can take away a few lessons from this example.

(1) There may be multiple approaches to solving a particular kind of problem.
(2) Some of those approaches may work in more cases than others.
(3) To reiterate a point already made earlier, in working with real numbers the idea of *approximation* – by rational numbers or otherwise – is very powerful. It may be the case that it is not clear how to do something for a particular real number

x, but that we do know how to do it for some values of y that approximate x, and that we can use this to get closer and closer to the actual answer for x itself.

This language of "approximate" and "closer and closer" will be made more precise soon, when we discuss *limits*.

For the second issue raised above (multiple decimal representations), we point out that in fact this also arises in our definition of $\mathbb{Q}$, since we may have $p/q = a/b$ for some $p, a \in \mathbb{Z}$ and $q, b \in \mathbb{N}$ even if $p \neq a$ and $q \neq b$. In that case we can guarantee a unique representation by requiring that p, q have no common factors except 1 (they are *relatively prime*); we can do a similar thing for real numbers by requiring that our decimal representations do not terminate in an infinite sequence of 9s.

For the third issue above, we will defer a discussion of why $x^2 = 2$ has a solution in $\mathbb{R}$, and present the proof that it does *not* have a solution in $\mathbb{Q}$.

Theorem 1.1. *There is no rational number x such that $x^2 = 2$.*

Proof. We use the technique of *proof by contradiction*: that is, we start by assuming that there *is* a rational number x with $x^2 = 2$, then show that this would imply a statement that we know to be false. This means that the original assumption must be false.

To this end, suppose that $p, q \in \mathbb{N}$ are relatively prime and that $(p/q)^2 = 2$. Then $p^2 = 2q^2$. Now we need the following fact, whose proof we leave to the reader.

Exercise 1.2. An integer $n \in \mathbb{Z}$ is even if and only if its square n^2 is even.

Since $p^2 = 2q^2$ is even, Exercise 1.2 implies that p is even as well; thus $p = 2a$ for some $a \in \mathbb{N}$, and we get

$$2q^2 = p^2 = (2a)^2 = 4a^2 \quad \Rightarrow \quad q^2 = 2a^2.$$

Thus q^2 is even, and again Exercise 1.2 implies that q is even. But then p, q are both even, contradicting our assumption that they are relatively prime. This contradiction shows that no such p, q exist, which completes the proof of the theorem. □

The result of Theorem 1.1 is usually stated as "The square root of 2 is irrational". In order to state it this way, though, we need to show that there *is* a square root of 2 in the real numbers, which we defer until later.

Lecture 2 Functions: Review of basic concepts

This lecture corresponds to §1.1 of Stewart's book and Chapter 3 of Spivak.

2.1. Functions and their graphs

A *function* f from a set X to a set Y is a rule that assigns to each element $x \in X$ an element $f(x) \in Y$. The set X is the *domain* of f, and the set Y is the *codomain* or *target space*. The *range* of f is

$$f(X) := \{f(x) : x \in X\} \subset Y,$$

where we use "$A := B$" to mean that A is defined as equal to B. Note that we use the notation $A \subset B$ to mean "A is a subset of B", without requiring that $A \neq B$; some

authors use $A \subseteq B$ instead. If we need to specify that A is a *proper* subset of B, we will write $A \subsetneq B$, as in (1.1).

There are several ways to define a function.

- We can use a formula, such as $f(x) = x^2$. In this case the domain is usually understood to be the set of all real numbers for which the formula also gives a real number, although sometimes a smaller domain is explicitly specified.
- We can list every element of X and then say which element of Y it is mapped to by f. In this case the domain is explicitly specified.
- We can give a verbal description or an algorithm that defines the function. For example, consider the function $f\colon \mathbb{N} \to \mathbb{N}$ defined by taking $f(n)$ to be the nth prime number when the primes are listed in increasing order.
- We can define a function *piecewise* by partitioning the domain X into subsets and using one of the above methods to define f on each of these. For example, the "Collatz function" $f\colon \mathbb{N} \to \mathbb{N}$ is defined by

$$f(n) = \begin{cases} 3n+1 & \text{if } n \text{ is odd,} \\ n/2 & \text{if } n \text{ is even.} \end{cases}$$

We are used to drawing the graph of a function from the real line to itself as a curve in the plane $\mathbb{R}^2$, which consists of all the points (x, y) for which $y = f(x)$. More generally, given two sets X and Y we can form the *direct product*

$$X \times Y := \{(x, y) : x \in X, y \in Y\}, \tag{2.1}$$

and then the graph of a function $f\colon X \to Y$ is defined to be

$$\operatorname{graph}(f) := \{(x, f(x)) : x \in X\} \subset X \times Y. \tag{2.2}$$

Remark 2.1. The difference between the *ordered pair* (x, y) in (2.1) and the set $\{x, y\}$ is two-fold: firstly, the elements of the ordered pair are allowed to be the same (if X, Y have any elements in common), and secondly, in the ordered pair we keep track of the order in which the elements appear, so that (x, y) is distinct from (y, x) if $x \neq y$, whereas $\{x, y\} = \{y, x\}$. Notice that the same notation (x, y) can refer both to an ordered pair and to the open interval with endpoints x and y (when x, y are real numbers), so one must always be alert to the context in which it occurs to see which is meant.

Remark 2.2. Formally, if we take set theory as the foundation of mathematics, then we should define ordered pairs in terms of sets: this can be done by declaring (x, y) to be the set $\{\{x\}, \{x, y\}\}$ (recall that an element of a set could be a set in its own right); then (x, x) is represented by $\{\{x\}, \{x\}\} = \{\{x\}\}$, and we see that $(x, y) = \{\{x\}, \{x, y\}\}$ and $(y, x) = \{\{y\}, \{x, y\}\}$ are distinct from each other when $y \neq x$.[5] For practical purposes we do not bother with this level of detail, however, and continue to simply work with (x, y) as we always have.

Recall the vertical line test, which says that a set $\Gamma \subset \mathbb{R}^2$ is the graph of a function (on some domain in $\mathbb{R}$) if and only if every vertical line intersects Γ at most once. A more precise version says that Γ is the graph of a function from X to $\mathbb{R}$ if and only if every vertical line whose x-coordinate is in X intersects Γ exactly once. The vertical

[5] See `https://en.wikipedia.org/wiki/Ordered_pair#Kuratowski's_definition` for more.

line through $(x, 0)$ is the set $\{(x, y) : y \in \mathbb{R}\}$, and so we can formulate a version of the vertical line test that works for general sets X and Y.

Vertical Line Test. *Given two sets X and Y, and a subset $\Gamma \subset X \times Y$, the following are equivalent.*

(1) There is a function $f\colon X \to Y$ such that $\Gamma = \operatorname{graph}(f)$.
(2) For all $x \in X$, we have $\#(\Gamma \cap \{(x, y) : y \in Y\}) = 1$.
(3) For all $x \in X$, there exists a unique $y \in Y$ such that $(x, y) \in \Gamma$.

Remark 2.3. If A is a set, the notation $\#A$ denotes the number of elements in A. Thus the equation in the second item above is just the statement that each "vertical line" intersects Γ in exactly one point.

The terms "for all" and "there exists" that we used above are important enough that we give them their own notation: we write $\forall$ to mean "for all", and $\exists$ to mean "there exists". Thus the third item above could be written as

$$\forall x \in X \ \exists \text{ a unique } y \in Y \text{ s.t. } (x, y) \in \Gamma, \tag{2.3}$$

where we also use the abbreviation "s.t." for "such that".

2.2. Injectivity, surjectivity, and bijectivity

Definition 2.4. Consider a function $f\colon X \to Y$. Given $x \in X$, the *image* of x under f is $f(x) \in Y$. If $f(x) = y$, we say that x *is a pre-image of* y. The function f is *one-to-one* (1-1) if every $y \in Y$ has at most one pre-image; such a function is also called *injective*. In this case we can define an inverse function $f^{-1}\colon \operatorname{range}(f) \to X$ by the condition that $f^{-1}(y)$ is the unique $x \in X$ such that $f(x) = y$; that is, the unique pre-image of y.

Remark 2.5. Do not confuse the *inverse* f^{-1} with the *reciprocal* $1/f$. The reason for the power-type notation in the inverse is that if $Y = X$, so that f maps the set X to itself, then for each $n \in \mathbb{N}$ we can define

$$f^n = \overbrace{f \circ f \circ \cdots \circ f}^{n \text{ times}}$$

to be the nth iterate of f under *composition*, and these functions have the property that[6]

$$f^m \circ f^n = f^{m+n}. \tag{2.4}$$

The identity function $\mathrm{Id}(x) = x$ has the property that $\mathrm{Id} \circ f^n = f^n$, so it makes sense to write $f^0 = \mathrm{Id}$. Then if (2.4) is to hold for negative integers as well, we should have

$$f^{-1} \circ f = f^{-1} \circ f^1 = f^{-1+1} = f^0 = \mathrm{Id};$$

in other words, f^{-1} should be the inverse function for f.

The following are all equivalent to the definition of injectivity.

(1) For every $y \in Y$, $\#\{x \in X : f(x) = y\} \leq 1$.
(2) For every $y \in Y$, $\#(\operatorname{graph}(f) \cap \{(x, y) : x \in X\}) \leq 1$.
(3) If $x_1, x_2 \in X$ are such that $f(x_1) = f(x_2)$, then $x_1 = x_2$.

[6]With this notation in mind, observe that $f^2(x)$, $f(x)^2$, and $f(x^2)$ all mean different things. Unfortunately there is a common inconsistency in notation regarding trigonometric functions (which we will also be guilty of): the notation $\sin^2(x)$ most commonly means $(\sin(x))^2$ rather than $\sin(\sin(x))$.

The first of these just restates the definition. The second is a mildly more complicated version of the first, which has a geometric interpretation called the *horizontal line test*: the set $\{(x, y) : x \in X\}$ represents the horizontal line with second coordinate y, so f is injective if every horizontal line meets its graph in at most one point. It is a (short) exercise to prove that the third condition is equivalent to the other two.

Definition 2.6. A function $f\colon X \to Y$ is *onto* (or *surjective*) if every $y \in Y$ has a pre-image in X; in other words, for every $y \in Y$ there exists $x \in X$ such that $f(x) = y$. If f is both 1-1 and onto (both injective and surjective), then it is called a *bijection*.

Using the "quantifiers"[7] $\forall$ and $\exists$, the definition of surjectivity can be rewritten as

$$\forall y \in Y \ \exists x \in X \text{ s.t. } f(x) = y. \tag{2.5}$$

Remark 2.7. Whenever you see the symbols $\forall$ and $\exists$, it is helpful to interpret the statement containing them by thinking of a game between you and an adversary. Your goal is to verify the truth of the statement, and your adversary's goal is to make the statement be false. Each $\forall$ represents a turn taken by your adversary, in which they make a choice over which you have no control. Each $\exists$ represents a turn that you take, in which you get to control the choice. The overall statement is true if you have a winning strategy: that is, if no matter what your adversary does, you can make your choices in such a way that the innermost statement is true.

Consider the specific example given in (2.5); suppose that $X = Y = \mathbb{R}$ and that $f(x) = 2x+1$. Then on the first turn of the game, your adversary picks a real number y, over which you have no control. You want to choose a number x such that $y = f(x) = 2x + 1$; some simple algebra suggests that you should pick $x = (y - 1)/2$. Indeed, if you make this pick, then $f(x) = 2(y - 1)/2 + 1 = y$, and thus no matter what number y your adversary picks (with the $\forall$), you can always make a choice of x (with the $\exists$) such that the innermost statement, "$f(x) = y$", is true. Thus the entire statement in (2.5) is true, and indeed, this function is onto.

On the other hand, if $f(x) = x^2$, then your adversary, after a moment's thought, will realize that if they choose $y = -1$ on their turn, then no matter what choice of x you make, you will have $f(x) = x^2 \geq 0 > -1 = y$, and in particular, $f(x) \neq y$. Thus your adversary can force the innermost statement to be false, and thus the entire statement in (2.5) is false. And indeed, this function is not onto.

When we discuss limits in a few lectures, we will encounter more complicated expressions involving $\forall$ and $\exists$.

Lecture 3 — Examples of functions

This lecture corresponds to §§1.2–1.5 in Stewart and Chapter 4 in Spivak.

[7] In logic, $\forall$ is called the *universal quantifier* and $\exists$ is called the *existential quantifier*.

3.1. Polynomials and rational functions

We will need to study functions whose domain and range lie in $\mathbb{R}$. One fundamental class of examples is the set of *polynomials*: a function $f\colon \mathbb{R} \to \mathbb{R}$ is a polynomial if there is a nonnegative integer n (called the *degree* of f and sometimes written $\deg f$) and real numbers $a_0, a_1, \ldots, a_n$ (the *coefficients* of f) such that

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n. \tag{3.1}$$

This can also be written using summation notation:

$$f(x) = \sum_{k=0}^{n} a_k x^k.$$

Low-degree polynomials are given specific names:

A polynomial of degree …	…is called …
0	constant
1	linear
2	quadratic
3	cubic
4	quartic
5	quintic

If f is a polynomial, then a number $r \in \mathbb{R}$ such that $f(r) = 0$ is called a *root* of the polynomial. You can use polynomial long division to prove the following.

Exercise 3.1. Prove that $r \in \mathbb{R}$ is a root of a polynomial f if and only if there is a polynomial g such that $f(x) = (x - r)g(x)$ for all $x \in \mathbb{R}$. Show that in this case, $\deg g = (\deg f) - 1$.

Exercise 3.2. Use Exercise 3.1 to show that the number of roots of a non-constant polynomial f is at most $\deg f$.

Remark 3.3. When f is linear, it is easy to find the unique root. When f is quadratic, we have the *quadratic formula* that produces the roots (if they exist) in terms of the coefficients. There is a corresponding formula to find the roots of cubic polynomials, but it is rather more complicated. There is even a formula to solve quartic equations, but it is nothing short of horrific to write out in full (it would take several pages to do so). It turns out that there is no formula to solve quintic equations in general. By this we do not mean that "no formula is known". Rather, we mean that it can be *proved* that no such formula exists, so that the fact that we do not know such a formula is not a reflection of our own ignorance, but rather a fundamental fact about the mathematical universe. The proof of this fact uses what is called *Galois theory*, which is well beyond the scope of this course.

Definition 3.4. If f, g are polynomials, then $r(x) := f(x)/g(x)$ is called a *rational function*. A polynomial $p(x)$ is a *factor* of a polynomial $f(x)$ if there is a polynomial $q(x)$ such that $f(x) = p(x)q(x)$; then polynomials f and g *have no common factor* if there is no non-constant polynomial p that is a factor of both f and g. If f, g are polynomials with no common factors, then the *degree* of the rational function f/g is $\max(\deg f, \deg g)$.

If the word "polynomial" is replaced with the word "integer" in the previous definition, then this turns into the description of the rational numbers. The rules for doing arithmetic with rational functions are completely analogous to those for doing arithmetic with rational numbers: for multiplication we simply multiply the numerators and denominators of the two functions, while for addition and subtraction we must first put everything over a common denominator. Thus for the rational functions $\frac{1}{x}$ and $\frac{1}{x+1}$ we have

$$\frac{1}{x} - \frac{1}{x+1} = \frac{x+1}{x(x+1)} - \frac{x}{x(x+1)} = \frac{1}{x(x+1)}. \tag{3.2}$$

3.2. Trigonometric functions

We define the sine and cosine functions as follows. Given $t \in \mathbb{R}$, let $P(t) \in \mathbb{R}^2$ be the point on the unit circle obtained by starting at the point $(1, 0)$ and moving counterclockwise until a total arc length of t has been reached. Then $\cos(t)$ is the x-coordinate of $P(t)$, and $\sin(t)$ is the y-coordinate of $P(t)$.

There are four more standard trigonometric functions, defined as:[8]

$$\sec x = \frac{1}{\cos x}, \quad \tan x = \frac{\sin x}{\cos x}, \quad \csc x = \frac{1}{\sin x}, \quad \cot x = \frac{\cos x}{\sin x}.$$

Because $P(t)$ lies on the unit circle $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$, we have

$$\cos^2 t + \sin^2 t = 1 \text{ for all } t \in \mathbb{R}.$$

Two other fundamental trigonometric identities that we will need later on are the formulas for sine and cosine of sums of angles:

$$\sin(x + y) = \sin x \cos y + \cos x \sin y, \tag{3.3}$$

$$\cos(x + y) = \cos x \cos y - \sin x \sin y. \tag{3.4}$$

For the time being we omit the proofs of these identities, which can be given by elementary geometric arguments.

Remark 3.5. We said "sums of angles" even though no angles appeared in the discussion so far. In the case when $0 < t < \pi/2$, we can consider the triangle with vertices at $O = (0, 0)$, $P = (\cos t, \sin t)$, and $Q = (\cos t, 0)$; then $t = \measuredangle(POQ)$ and $\cos t$, $\sin t$ give the lengths of the sides PQ and OP, respectively.

There is a weak point in these definitions, though. What exactly do we mean by "arc length along the circle"? To make this a little more concrete: given a point (x, y) in the first quadrant (so $x > 0$ and $y > 0$), how do we compute the arc length along the circle from $(0, 0)$ to (x, y)? If we replace the arc between these two points with a straight line, then length is easy to compute; the length of the straight line between two points (x_1, y_1) and (x_2, y_2) is just the distance between those points, which is given by the Pythagorean distance formula $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. We can get a better approximation by picking a few points on the arc, connecting them with straight lines, and adding up the lengths of those line segments. Intuitively, it seems reasonable to say

[8]Yes, we're using a different variable (x) here than we did in the paragraph above (t). This should not bother you. The identity of a function is not affected by what we call the variable we are feeding into it. *That which we call a rose*, and so on and so forth.

that the length of the arc is somehow given by this approximation procedure, provided we take enough points. But this will take some work to make precise, and until we do that work we need to regard our definitions of the trigonometric functions as provisional, since they rely on a notion of arc length that we have not really nailed down yet.

3.3. Exponential functions

Given $a > 0$, we would like to define an *exponential* function $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = a^x$. But what does a^x mean? When x is a natural number it is clear enough: $a^1 = a$, $a^2 = a \cdot a$, $a^3 = a \cdot a \cdot a$, and so on: a^x means the product of a with itself x times.

What about negative values of x? Or non-integer values? The first thing to observe is that when x, y are natural numbers, we clearly have

$$a^{x+y} = a^x a^y. \tag{3.5}$$

We want to define a^x for more general values of x in such a way that (3.5) continues to hold. To this end, we first note that whatever a^0 is, it should have the property that $a^0 a^x = a^{0+x} = a^x$ for all $x \in \mathbb{N}$; this is only possible if $a^0 = 1$, so we define

$$a^0 := 1. \tag{3.6}$$

Now if x is a negative integer, then $x = -n$ for some $n \in \mathbb{N}$, and in order for (3.5) to hold with $y = n$, we must have

$$a^{-n} = \frac{a^{x+y}}{a^y} = \frac{a^{-n+n}}{a^n} = \frac{a^0}{a^n} = \frac{1}{a^n}.$$

Thus we define

$$a^{-n} := \frac{1}{a^n} \tag{3.7}$$

for every $n \in \mathbb{N}$. Now we have defined a^x for every $x \in \mathbb{Z}$. But what about non-integer values? Again, (3.5) seems to tell us what to do. For example, whatever $a^{1/2}$ is, using (3.5) with $x = y = 1/2$ tells us that

$$a = a^1 = a^{1/2+1/2} = a^{1/2} a^{1/2} = (a^{1/2})^2 \quad \Rightarrow \quad a^{1/2} = \sqrt{a}.$$

More generally, if p/q is any rational number, then iterating (3.5) q times gives

$$(a^{p/q})^q = a^{\frac{p}{q} \cdot q} = a^p,$$

and thus $a^{p/q}$ must be *defined* to be the qth root of a^p. We are nearly there – we have defined a^x whenever $x \in \mathbb{Q}$ – but two questions remain to be addressed.

(1) Why does a^p always have a qth root when $p, q \in \mathbb{N}$, and why should it be unique?
(2) What are we to make of a^x when x is an irrational number?

The first question will be addressed when we study the intermediate value theorem. For the second question, we start by observing that in order to describe an irrational number such as π, we can use a sequence of increasingly accurate rational approximations: $\pi \approx 3.14$, then $\pi \approx 3.14159$, then $\pi \approx 3.1415926535$, and so on. Since a^x was defined whenever $x \in \mathbb{Q}$, we know what is meant by $a^{3.14}$, $a^{3.14159}$, etc., and it would be reasonable to expect that these are "increasingly accurate approximations" to a^π. To make this precise requires the notion of *limit*, which we start discussing in the next lecture.

In the meantime, we observe that if $a > 1$, then the function $\mathbb{N} \to \mathbb{R}$ defined by $f(x) = a^x$ also has the property that it is *strictly increasing*: if $x, y \in \mathbb{N}$ satisfy $x < y$, then $f(x) = a^x < a^y = f(y)$. If $a < 1$, then it is *strictly decreasing*. (If $a = 1$, then $f(x) = 1$ for all x, and the function is not so exciting.) Just as we want to define $f\colon \mathbb{R} \to \mathbb{R}$ in such a way that (3.5) is preserved, we would also like to preserve this property of being strictly increasing (if $a > 1$) or decreasing (if $a < 1$). If and when we can do this, we will have a 1-1 function $f\colon \mathbb{R} \to \mathbb{R}$. We will eventually prove that the range of this function is $(0, \infty)$, and thus there is an inverse function $f^{-1}\colon (0, \infty) \to \mathbb{R}$, which is called the *logarithm with base* a, and denoted $\log_a$. Observe that given $x, y \in (0, \infty)$, if we write $s = \log_a x$ and $t = \log_a y$, then (3.5) gives

$$a^{s+t} = a^s a^t = xy,$$

and thus $s + t = \log_a(xy)$. In other words, $\log_a$ satisfies the identity

$$\log_a(xy) = \log_a(x) + \log_a(y). \tag{3.8}$$

So far we have no reason to prefer one value of $a > 0$ over another. Eventually we will see that there is a *natural logarithmic base* $e \approx 2.71828\ldots$, also called *Euler's constant*; however, the motivation for this number needs to wait until we discuss derivatives.

Lecture 4 — Limits, intuitively

This lecture corresponds to §§2.1–2.2 in Stewart, and the beginning of Chapter 5 in Spivak.

We have now encountered several situations in which there is some quantity that we want to compute but cannot do so directly, while at the same time we can compute approximations to this quantity. We briefly describe these, together with some new ones.

(1) *Arithmetic with decimal expansions.* If we want to add (or multiply) the decimal expansions for x and y, we can take the first n digits of each, add those using the normal rules for addition, and then let n get larger and larger to get a better and better approximation for $x + y$.
(2) *Exponential functions.* If x is an irrational number and we want to make sense of the expression 2^x, we can approximate x with rational numbers p/q, for which $2^{p/q}$ is defined as the qth root of 2^p (which is itself 2 multiplied by itself p times).
(3) *Arc length.* The number π is the circumference of a circle with diameter 1. Here "circumference" means the length traveled if we go once around the circle, and this can be approximated by drawing a regular polygon with a large number of sides. The perimeter of this polygon is something we can calculate using Pythagoras' formula.
(4) *Area.* The number π is also the area of a circle with radius 1. Here "area" is something we can make sense of for rectangles – where it is width times height – and so we can imagine covering the circle by a large number of small rectangles, then adding up their areas to get an approximate value for the area of the circle.

As we use smaller and smaller rectangles in this procedure, our approximation should get better and better.

(5) *Instantaneous velocity.* Suppose I throw a ball straight up into the air, and $f(t)$ represents its height at time t. Then the *average velocity* of the ball between time t and time $t+h$ is given by (total distance traveled) / (time elapsed), which is $\frac{f(t+h)-f(t)}{h}$. As h gets smaller and smaller, this gives a better and better approximation to the *instantaneous velocity* of the ball at time t.

(6) *Tangent lines.* The graph of a function $y = f(x)$ gives a curve in $\mathbb{R}^2$. The tangent line to this curve at a given point $(a, f(a))$ is the line through this point that "goes in the same direction as the curve". But what does "direction of the curve" mean? If $(b, f(b))$ is a nearby point on the curve, then the *secant line* corresponding to a and b is the line passing through $(a, f(a))$ and $(b, f(b))$, which has slope given by $\frac{\text{rise}}{\text{run}} = \frac{f(b)-f(a)}{b-a}$. As b gets closer and closer to a, this secant line gives a better and better approximation to the tangent line to f at $(a, f(a))$. For example, if $f(x) = x^2$ and $a = 2$, then the slope of the secant line is

$$\frac{b^2-4}{b-2} = \frac{(b-2)(b+2)}{b-2} = b+2;$$

note that this makes sense as long as $b \neq 2$, but if $b = 2$ then the initial expression is no longer defined. Nevertheless we can use the final expression to deduce that as b gets closer and closer to 2, the slope of the secant line gets closer and closer to 4, so the slope of the tangent line is 4, and the equation of the tangent line is $y - 4 = 4(x-2)$, which simplifies to $y = 4x - 4$.

If you have taken some calculus before, you may recognize the first two of these as instances of *continuity*, the next two as instances of *integrals*, and the last two as instances of *derivatives*. For now, we focus on the common thread between all of them, which is the idea of a *limit*.

Definition 4.1. Given a real-valued function f, a real number a in the domain of f, and a real number L, we say that L *is the limit of* f *at* a, and write $\lim_{x\to a} f(x) = L$, if we can make the values of $f(x)$ be arbitrarily close to L by taking x to be sufficiently close (but not equal) to a. In this case we also write "$f(x) \to L$ as $x \to a$", or sometimes $f(x) \xrightarrow{x\to a} L$.

Note that the value $f(a)$ does not affect the limit. We will make Definition 4.1 more precise in a couple lectures. For now we just mention that we will also have occasion to talk about the limit of a sequence: if $x_1, x_2, x_3, \ldots$ is a sequence of real numbers, we say that L *is the limit of* x_n *(as* $n \to \infty$*)*, and write $\lim_{n\to\infty} x_n = L$, if we can make the values of x_n be arbitrarily close to L by taking n to be sufficiently large. The first three examples in the list above can be interpreted as the limit of a sequence (sequence of rational approximations, sequence of polygons, sequence of coverings by rectangles), while the last two can be interpreted as the limit of a function.

Lecture 5 Computing limits

This lecture corresponds to §2.3 in Stewart, and the end of Chapter 5 in Spivak.

5.1. Algebraic tricks and numerical approximations

Sometimes we can use some algebraic simplifications to compute limits: for example, the tangent line computation in the previous lecture went as follows:

$$\lim_{x\to 2}\frac{x^2-4}{x-2}=\lim_{x\to 2}\frac{(x-2)(x+2)}{x-2}=\lim_{x\to 2}(x+2),$$

and it is not much of a stretch to convince yourself that $x+2$ approaches 4 as x approaches 2. (We will give a more complete justification of this in the next few lectures.) What about the limit of the reciprocal quantity? We can write

$$\lim_{x\to 2}\frac{x-2}{x^2-4}=\lim_{x\to 2}\frac{x-2}{(x-2)(x+2)}=\lim_{x\to 2}\frac{1}{x+2},$$

and it seems reasonable to expect that this approaches $\frac{1}{4}$, but how can we be sure of this? (Recall, after all, that in the definition of limit we never actually make the substitution $x=2$, we merely choose values of x that are extremely close to 2.) We might try gathering some numerical evidence, choosing numbers like $x=2.000001$ and seeing what happens. But we should be wary, because strange things can happen if we blindly rely on a calculator or computer. For example, suppose we want to compute

$$\lim_{t\to 0} f(t) \text{ for } f(t)=\frac{\sqrt{t^2+1}-1}{t^2}.$$

When $t=.01$, a numerical computation shows that

$$f(.01)=0.499988\ldots,$$

which suggests that the limit is $\frac{1}{2}$. (Though it's worth pausing for a moment and thinking about whether we should really always expect the limit to be a "nice" number.) When $t=.001$, a similar computation gives $f(.001)=0.499999875\ldots$, lending further credence to the conjecture. But if we keep going, then for some extremely small t, perhaps .00001, perhaps 10^{-10} (it depends on the details of which calculator or computer you use), the computer will start returning the answer 0. This is because it only stores some finite number of digits, and eventually the numerator of $f(t)$ gets stored as 0. On the other hand, we can use some algebra to observe that

$$\frac{\sqrt{t^2+1}-1}{t^2}=\frac{\sqrt{t^2+1}-1}{t^2}\frac{\sqrt{t^2+1}+1}{\sqrt{t^2+1}+1}=\frac{(t^2+1)-1}{t^2(\sqrt{t^2+1}+1)}=\frac{1}{\sqrt{t^2+1}+1}, \tag{5.1}$$

and it once again seems very reasonable to expect that this approaches $\frac{1}{\sqrt{1}+1}=\frac{1}{2}$.

Remark 5.1. The algebraic trick we have used several times in this and the previous lecture is worth remembering: a^2-b^2 factors as $(a-b)(a+b)$, and an expression containing something of the form $a-b$ (or $a+b$) can sometimes be simplified by multiplying both numerator and denominator by the *conjugate* expression, so that for example we get $\sqrt{A}-\sqrt{B}=\frac{A-B}{\sqrt{A}+\sqrt{B}}$.

In each of the examples above, we reached a point where we concluded with "it seems reasonable to expect that" the limit is given by substituting the limiting value of t (or x, or whatever the independent variable is) into the expression. We will start justifying this shortly. First one more cautionary tale is in order.

Example 5.2. Define $f\colon (0,\infty) \to \mathbb{R}$ by $f(x) = \sin\frac{1}{x}$. Then for $x = \frac{1}{n\pi}$ we have $f(x) = \sin\frac{1}{x} = \sin n\pi = 0$, but it is *not* true that $f(x) \to 0$ as $x \to 0$. Indeed, if $x = \frac{2}{n\pi}$, then $f(x) = \sin\frac{n\pi}{2}$, which is equal to ± 1 whenever n is odd. Thus in this case the limit *does not exist*. A similar example with sequences (which is easier to state) is $x_n = (-1)^n$, so that x_n is 1 when n is even and -1 when n is odd.

5.2. The limit laws

Our primary tools for computing limits will be algebraic manipulations of the sort described above, together with the following set of *limit laws*.

Theorem 5.3 (Limit laws)**.** *Let f, g be functions defined around a point $a \in \mathbb{R}$, and suppose that $\lim_{x\to a} f(x)$ and $\lim_{x\to a} g(x)$ exist. Let c be any real number. Then the following limits all exist and are given by the values shown.*

(1) $\lim\limits_{x\to a}\big(f(x)+g(x)\big) = \lim\limits_{x\to a} f(x) + \lim\limits_{x\to a} g(x)$

(2) $\lim\limits_{x\to a}\big(f(x)-g(x)\big) = \lim\limits_{x\to a} f(x) - \lim\limits_{x\to a} g(x)$

(3) $\lim_{x\to a}\big(cf(x)\big) = c\lim_{x\to a} f(x)$

(4) $\lim\limits_{x\to a}\big(f(x)g(x)\big) = \Big(\lim\limits_{x\to a} f(x)\Big)\Big(\lim\limits_{x\to a} g(x)\Big)$

(5) $\lim\limits_{x\to a}\dfrac{f(x)}{g(x)} = \dfrac{\lim_{x\to a} f(x)}{\lim_{x\to a} g(x)}$ *provided* $\lim\limits_{x\to a} g(x) \neq 0$

(6) $\lim\limits_{x\to a}\Big(\big(f(x)\big)^n\Big) = \Big(\lim\limits_{x\to a} f(x)\Big)^n$

(7) $\lim\limits_{x\to a} c = c$

(8) $\lim\limits_{x\to a} x = a$

(9) $\lim\limits_{x\to a} x^n = a^n$ *for every* $n \in \mathbb{Z}$

(10) $\lim\limits_{x\to a} \sqrt[n]{x} = \sqrt[n]{a}$ *for every odd* $n \in \mathbb{Z}$ *(works for even n also if $a > 0$)*

(11) $\lim\limits_{x\to a} \sqrt[n]{f(x)} = \sqrt[n]{\lim\limits_{x\to a} f(x)}$ *for every odd n (works for even n also if* $\lim\limits_{x\to a} f(x) > 0$*)*

Laws 1, 4, 5, 7, and 8 will be proved later, after we have given the precise definition of a limit. The remaining laws follow from these four:

- Law 3 follows from Laws 4 and 7 by putting $g(x) = c$.
- Law 2 follows from Laws 1 and 3 by writing $f - g = f + (-1)g$ and putting $c = -1$.
- Law 6 is proved by iterating Law 4.
- Law 9 follows from Laws 6 and 8.
- Law 11 follows from Law 6.
- Law 10 follows from Laws 8 and 11.

Exercise 5.4. Write down the details of the proofs of Laws 2, 3, 6, 9, 10, and 11 using Laws 1, 4, 5, 7, and 8 as suggested in the list above.

As an example of the limit laws in action, we can justify the examples from the start of this lecture. For the first two we note that

$$\lim_{x\to 2}(x+2) \overbrace{=}^{\text{Law 2}} \lim_{x\to 2} x + \lim_{x\to 2} 2 \overbrace{=}^{\text{Laws 8 and 7}} 2+2=4,$$

$$\lim_{x\to 2} \frac{1}{x+2} \overbrace{=}^{\text{Law 5 (and 7)}} \frac{1}{\lim_{x\to 2}(x+2)} = \frac{1}{4}.$$

For the third, we have

$$\begin{aligned}
\lim_{t\to 0} \frac{\sqrt{t^2+1}-1}{t^2} &= \lim_{t\to 0} \frac{1}{\sqrt{t^2+1}+1} && \text{algebra from (5.1)} \\
&= \frac{\lim_{t\to 0} 1}{\lim_{t\to 0}(\sqrt{t^2+1}+1)} && \text{by Law 5} \\
&= \frac{1}{(\lim_{t\to 0}\sqrt{t^2+1})+1} && \text{by Laws 1 and 7} \\
&= \frac{1}{\sqrt{(\lim_{t\to 0} t^2)+1}+1} && \text{by Laws 11, 1, and 7} \\
&= \frac{1}{\sqrt{(\lim_{t\to 0} t)^2+1}+1} && \text{by Law 9} \\
&= \frac{1}{\sqrt{0^2+1}+1} && \text{by Law 8} \\
&= \frac{1}{2}.
\end{aligned}$$

In practice we eventually carry out these steps without writing each of them explicitly, but when you are first encountering limits it is important to understand why the overall computation works.

We occasionally work with *one-sided limits*. For example, we say that L is the *left-hand limit* of f at a if we can make $f(x)$ arbitrarily close to L by taking x sufficiently close to a *and to the left of* a *(that is,* $x<a$*)*. In this case we write

$$\lim_{x\to a^-} f(x) = L.$$

The *right-hand limit* is defined analogously and written $\lim_{x\to a^+}$.

Example 5.5. Let $f(x)=x$ for $x\le 0$ and $f(x)=x+1$ for $x>0$. Then

$$\lim_{x\to 0^-} f(x) = 0 \text{ and } \lim_{x\to 0^+} f(x) = 1.$$

Finally, we mention *infinite limits*. We say that $\lim_{x\to a} f(x)=\infty$ if $f(x)$ can be made arbitrarily large by taking x sufficiently close to (but not equal to) a. We define $\lim_{x\to a} f(x)=-\infty$ analogously, and make the obvious definitions for infinite one-sided limits.

Example 5.6. Let $f(x)=\frac{1}{x}$ for $x\neq 0$. Then

$$\lim_{x\to 0^-} f(x) = -\infty \text{ and } \lim_{x\to 0^+} f(x) = \infty.$$

Definition 5.7. If any of the one-sided or two-sided limits of f at a is ∞ or $-\infty$, then we say that $x = a$ is a *vertical asymptote* of $y = f(x)$.

Lecture 6 — Limits, rigorously

This lecture corresponds to §2.4 in Stewart and the middle of Chapter 5 in Spivak.

6.1. The definition at the heart of calculus

To prove the limit laws, we need the formal definition of limits, which uses the quantifier notation we introduced earlier.

Definition 6.1. Let f be a function that is defined on an open interval containing a, except possibly at a itself. Then $L \in \mathbb{R}$ is the limit of f as $x \to a$ if[9]

$$\forall \varepsilon > 0,\ \exists \delta > 0 \text{ s.t. if } 0 < |x - a| < \delta, \text{ then } |f(x) - L| < \varepsilon. \tag{6.1}$$

In this case we write $L = \lim_{x \to a} f(x)$.

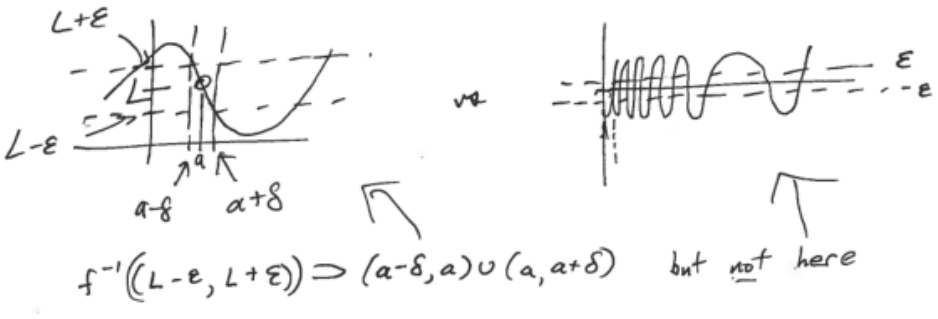


FIGURE 1. The ε-δ definition of limit

Figure 1 shows how the definition can be interpreted in terms of a graph: our adversary (recall Remark 2.7) chooses ε, which determines a horizontal strip with height 2ε, and we must choose δ such that the part of the graph lying inside the corresponding vertical strip (with width 2δ) is contained inside the horizontal strip chosen by our adversary. Observe that a smaller choice of ε will generally require a smaller choice of δ, so it is crucial that δ is allowed to depend on ε.

Exercise 6.2. Use the definition of limit to prove that $\lim_{x \to a} f(x) = \lim_{h \to 0} f(a + h)$, where by this equality we mean (as in the limit laws) that if one of the limits exists, then so does the other one, and in this case they are equal.

[9]Here ϵ and δ are the Greek letters 'epsilon' and 'delta'; one can imagine that they stand for "error" and "displacement", respectively. If you do not yet know the Greek alphabet, you should learn it; mathematicians tend to run out of letters if they are restricted to one alphabet, and so it is useful to have another one handy.

6.2. Some examples

Example 6.3. Earlier we claimed that $\lim_{x\to 2}(x+2) = 4$. To prove this directly from the definition, observe that we are putting $f(x) = x+2$, $a = 2$, and $L = 4$. Suppose our adversary chooses $\varepsilon > 0$. We want to choose x close enough to 2 that we are guaranteed to have $|(x+2) - 4| < \varepsilon$. Observe that this 'error term' can be written as

$$|(x+2) - 4| = |x - 2|,$$

and thus we have $|(x+2) - 4| < \varepsilon$ if and only if $|x-2| < \varepsilon$. Let $\delta = \varepsilon$. Then if x satisfies $0 < |x-2| < \delta$, we must have $|x-2| < \varepsilon$, and therefore $|f(x) - 4| < \varepsilon$, which proves that $\lim_{x\to 2}(x+2) = 4$.

Example 6.4. Based on the previous example and the limit law for multiplication, we expect to find that $\lim_{x\to 2}(x+2)^2 = 16$. So we put $f(x) = (x+2)^2$, $a = 2$, and $L = 16$, then we check the definition. The error term $|f(x) - L|$ can be written as

$$|(x+2)^2 - 16| = |x^2 + 4x + 4 - 16| = |x^2 + 4x - 12| = |(x-2)(x+6)|.$$

Thus once our adversary has chosen $\varepsilon > 0$, we have[10]

$$|f(x) - L| < \varepsilon \quad \Leftrightarrow \quad |(x-2)(x+6)| < \varepsilon.$$

Our goal is to choose $\delta > 0$ such that

$$\text{if } 0 < |x-2| < \delta, \text{ then } |(x-2)(x+6)| < \varepsilon. \tag{6.2}$$

It is tempting to look at this and decide that we should make δ be equal to $\varepsilon/|x+6|$ – after all, if $\delta = \varepsilon/|x+6|$ and $|x-2| < \delta$, then the bound we want follows immediately.

The problem with this is that δ is not allowed to depend on x. Remember the order of events in (6.1): first our adversary chooses ε, then we choose δ, and only after δ is chosen do we start checking the error estimate for various values of x.

Thus we must proceed in a different way, and address (6.2) in two steps. First we will require that $\delta \le 1$, so that we have

$$\text{if } 0 < |x-2| < \delta, \text{ then } 1 \le 2 - \delta < x < 2 + \delta \le 3, \text{ so } 7 < x + 6 < 9. \tag{6.3}$$

Using this bound, we deduce that

$$\text{if } 0 < |x-2| < \delta, \text{ then } |(x-2)(x+6)| < 9\delta. \tag{6.4}$$

Thus if we also have $\delta \le \varepsilon/9$, then we can use (6.4) to conclude that (6.2) is true. We set $\delta = \min(1, \varepsilon/9)$, and conclude that

$$0 < |x-2| < \delta \;\Rightarrow\; |x+6| < 9 \;\Rightarrow\; |f(x) - L| = |(x-2)(x+6)| < 9\delta \le \varepsilon.$$

This proves that $\lim_{x\to 2}(x+2)^2 = 16$.

When working through examples like these, it is very important to always keep in mind the distinction between a statement that you *are trying to prove is true*, such as (6.2), and a statement that you *have already proved is true*, such as (6.3) and (6.4). You will need to write down both sorts; remember which is which.

[10]The notation $P \Leftrightarrow Q$ means that statements P and Q are equivalent: if P is true then Q is true as well, and vice versa. In this case we often say that "P is true if and only if Q is true", and abbreviate the written form to "P is true iff Q is true".

Example 6.5. To prove that $\lim_{x\to 2} \frac{1}{x+2} = \frac{1}{4}$, observe that we are putting $f(x) = \frac{1}{x+2}$, $a = 2$, and $L = \frac{1}{4}$. The error term $|f(x) - L|$ can be written as

$$\Big|\frac{1}{x+2} - \frac{1}{4}\Big| = \Big|\frac{4-(x+2)}{4(x+2)}\Big| = \frac{|2-x|}{4|x+2|}.$$

Once our adversary chooses $\varepsilon > 0$, our goal is to choose $\delta > 0$ such that

$$\text{if } 0 < |x-2| < \delta, \text{ then } \frac{|2-x|}{4|x+2|} < \varepsilon.$$

To get some control on the denominator, we use the same trick as in the previous example and first assume that $\delta \leq 1$, so that if $0 < |x-2| < \delta$, then $1 < x < 3$, and thus $3 < |x+2| < 5$. It follows that for every such x we have

$$|f(x) - L| = \frac{|x-2|}{4|x+2|} \leq \frac{\delta}{12}.$$

Thus we put $\delta = \min(1, 12\varepsilon)$, and conclude that

$$0 < |x-2| < \delta \;\Rightarrow\; \frac{1}{|x+2|} < \frac{1}{3} \;\Rightarrow\; |f(x) - L| = \frac{|x-2|}{4|x+2|} < \frac{\delta}{12} \leq \varepsilon,$$

which proves that $\lim_{x\to 2} \frac{1}{x+2} = \frac{1}{4}$.

Already these examples involve enough calculations that you can imagine how much worse it would be to prove that $\lim_{t\to 0} \frac{\sqrt{t^2+1}-1}{t^2} = \frac{1}{2}$ directly from the definition. This illustrates the power of the limit laws, which we will prove soon.

6.3. Another formulation, using sequences

The formal definition of the limit of a function has a natural analogue for sequences.

Definition 6.6. Given a sequence x_n of real numbers, we say that L is the limit of x_n as $n \to \infty$, and write $L = \lim_{n\to\infty} x_n$ (or sometimes $x_n \to L$) if the following is true: for every $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $n \geq N$, we have $|x_n - L| < \varepsilon$.

Notice that the role of δ here is replaced by N, because we are interested in what happens as n becomes very large.

Exercise 6.7. Prove that $\lim_{n\to\infty} \frac{1}{n} = 0$.

It is useful to bear in mind the following consequence of the definition of limit.

Proposition 6.8. *If $\lim_{x\to a} f(x) = L$ and x_n is a sequence converging to a such that $x_n \neq a$ for all n, then $\lim_{n\to\infty} f(x_n) = L$.*

Proof. For every $\epsilon > 0$, the definition of limit gives $\delta > 0$ such that if $0 < |x-a| < \delta$, then $|f(x) - L| < \epsilon$. Similarly, the definition of limit of a sequence gives $N \in \mathbb{N}$ such that $|x_n - a| < \delta$ for all $n \geq N$. Since $x_n \neq a$ gives $0 < |x_n - a|$, we conclude that $|f(x_n) - L| < \epsilon$ for all $n \geq N$. By the definition of limit of a sequence, this means that $\lim_{n\to\infty} f(x_n) = L$. □

In fact the converse of this is true as well.

Proposition 6.9. *If* $\lim_{x\to a} f(x) \neq L$*, then there exists a sequence* $x_n \to a$ *such that* $x_n \neq a$ *for all* n *and* $\lim_{n\to\infty} f(x_n) \neq L$*.*

Proof. Since $\lim_{x\to a} f(x) \neq L$, our adversary has a winning move, and can choose $\varepsilon > 0$ such that no matter what $\delta > 0$ we choose, there is x with $0 < |x - a| < \delta$ such that $|f(x) - L| \geq \varepsilon$. In particular, given $n \in \mathbb{N}$ we can consider what happens when $\delta = \frac{1}{n}$; the previous sentence guarantees that there exists x_n with $0 < |x_n - a| < \frac{1}{n}$ such that $|f(x_n) - L| \geq \varepsilon$. Now it is a straightforward exercise to show that $x_n \to a$ and $f(x_n) \not\to L$. □

Combining Propositions 6.8 and 6.9 gives the following useful criterion.

Corollary 6.10. $\lim_{x\to a} f(x) = L$ *if and only if* $\lim_{x_n\to a} f(x_n) = L$ *for every sequence* $x_n \to a$ *with* $x_n \neq a$*.*

6.4. A limit that does not exist

Let us briefly look at an example of how to use the definition to show that a limit does *not* exist. Let $f(x) = \sin(\frac{1}{x})$ for all $x > 0$; see the right-hand graph in Figure 1. To show that $\lim_{x\to 0^+} f(x)$ does not exist, we must show that no matter what value of $L \in \mathbb{R}$ we choose for the putative limit, our adversary can win the game by choosing a value of $\varepsilon > 0$ that makes (6.1) fail, demonstrating that $\lim_{x\to 0^+} f(x) \neq L$. Indeed, if we choose a value of L and then our adversary chooses $\varepsilon = \frac{1}{2}$, in order to win the game (which we could do if L is the limit) we would need to find $\delta > 0$ such that

$$\text{if } 0 < x < \delta, \text{ then } |f(x) - L| < \tfrac{1}{2}. \tag{6.5}$$

If we could do this, then for every $x, y \in (0, \delta)$, we would have

$$|f(x) - f(y)| = |(f(x) - L) + (L - f(y))| \leq |f(x) - L| + |f(y) - L| < \tfrac{1}{2} + \tfrac{1}{2} = 1. \tag{6.6}$$

Observe that for every $n \in \mathbb{N}$, the points

$$x_n := \frac{1}{(2n + \frac{1}{2})\pi} \quad \text{and} \quad y_n := \frac{1}{(2n - \frac{1}{2})\pi}$$

have the property that

$$f(x_n) = \sin(2\pi n + \tfrac{\pi}{2}) = \sin\tfrac{\pi}{2} = 1 \quad \text{and} \quad f(y_n) = \sin(2\pi n - \tfrac{\pi}{2}) = \sin(-\tfrac{\pi}{2}) = -1.$$

We want to choose n large enough that $x_n, y_n \in (0, \delta)$. Since $x_n < y_n$ it is enough to guarantee that $\frac{1}{(2n-\frac{1}{2})\pi} < \delta$, or equivalently, $2n - \frac{1}{2} > \frac{1}{\delta\pi}$. Thus by choosing $n \in \mathbb{N}$ with $n > \frac{1}{2\delta\pi} + \frac{1}{4}$, we obtain two points $x_n, y_n \in (0, \delta)$ for which

$$|f(x_n) - f(y_n)| = |1 - (-1)| = 2 > 1,$$

so that (6.6) is false. Since this happens no matter what $\delta > 0$ we choose, we conclude that we cannot win the game, and thus $\lim_{x\to 0^+} \sin(\frac{1}{x})$ does not exist.

Remark 6.11. Notice that the argument in the last part of this discussion, about choosing n large enough that $x_n, y_n \in (0, \delta)$, is exactly the proof that $\lim_{n\to\infty} y_n = 0$.

Remark 6.12. Instead of working directly with the definition of limit, we could observe that $x_n \to 0$ and $y_n \to 0$, with $x_n, y_n \neq 0$, but $\lim_{n\to\infty} f(x_n) \neq \lim_{n\to\infty} f(y_n)$, so no matter what value of L we choose, it is impossible to have both $L = \lim_{n\to\infty} f(x_n)$ and $L = \lim_{n\to\infty} f(x_n)$. By Corollary 6.10, this implies that $\lim_{x\to 0} f(x)$ does not exist.

Lecture 7 — Proving the limit laws

This lecture corresponds to §2.4 and Appendix F in Stewart, and the end of Chapter 5 in Spivak.

As promised earlier, we now prove the limit laws from Theorem 5.3. It is enough to prove Laws 1, 4, 5, 7, and 8; the others are consequences of these, as explained there.

Exercise 7.1. Prove Limit Laws 7 and 8 using the definition of limit by observing that for Law 7, you can choose any $\delta > 0$ that you like,[11] and for Law 8, you can choose $\delta = \varepsilon$.

Proposition 7.2 (Limit Law 1)**.** *If* $\lim_{x\to a} f(x) = L$ *and* $\lim_{x\to a} g(x) = M$*, then* $\lim_{x\to a}(f(x) + g(x)) = L + M$*.*

Proof. The error bound we wish to control can be estimated as follows:

$$|(f(x) + g(x)) - (L + M)| = |(f(x) - L) + (g(x) - M)| \le \underbrace{|f(x) - L|}_{\text{I}} + \underbrace{|g(x) - M|}_{\text{II}}.$$

We can control I using the fact that $\lim_{x\to a} f(x) = L$, and II using the fact that $\lim_{x\to a} g(x) = M$. Indeed, once our adversary chooses $\varepsilon > 0$, then

(1) we can choose $\delta_1 > 0$ such that $0 < |x - a| < \delta_1$ implies $|f(x) - L| < \frac{\varepsilon}{2}$, and
(2) we can choose $\delta_2 > 0$ such that $0 < |x - a| < \delta_2$ implies $|g(x) - M| < \frac{\varepsilon}{2}$.

Let $\delta = \min(\delta_1, \delta_2)$. Then for every $0 < |x - a| < \delta$, we have

$$|(f(x) + g(x)) - (L + M)| \le |f(x) - L| + |g(x) - M| < \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} = \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this proves that $\lim_{x\to a}(f(x) + g(x)) = L + M$, and completes the proof of Limit Law 1. □

We go next to Law 4.

Proposition 7.3 (Limit Law 4)**.** *If* $\lim_{x\to a} f(x) = L$ *and* $\lim_{x\to a} g(x) = M$*, then* $\lim_{x\to a}(f(x)g(x)) = LM$*.*

Proof. Now the error term we must control is

$$\begin{aligned} |f(x)g(x) - LM| &= |f(x)g(x) - Lg(x) + Lg(x) - LM| \\ &\le \underbrace{|g(x)||f(x) - L|}_{\text{I}} + \underbrace{|L||g(x) - M|}_{\text{II}}, \end{aligned}$$

where we use the trick of adding and subtracting the same expression in order to gain some control over what we are dealing with. (This trick will appear many times.) As in the previous proof, we control I and II using the limits of f and g, respectively; once our adversary has chosen $\varepsilon > 0$, we want to make each of them $< \varepsilon/2$. Start with II. Since $\lim_{x\to a} g(x) = M$, there is $\delta_1 > 0$ such that

$$\text{if } 0 < |x - a| < \delta_1, \text{ then } |g(x) - M| < \frac{\varepsilon}{2|L| + 1}. \tag{7.1}$$

[11] This is the one and only case in which δ does not depend on ε.

This further implies that

$$|L||g(x) - M| < \frac{|L|\varepsilon}{2|L| + 1} < \frac{\varepsilon}{2},$$

which controls II as desired.[12] But what about I? To control this, we first use one more time the fact that $\lim_{x\to a} g(x) = M$, to choose $\delta_2 > 0$ such that

$$\text{if } 0 < |x - a| < \delta_2, \text{ then } |g(x) - M| < 1, \text{ so } |g(x)| < |M| + 1. \tag{7.2}$$

Then since $\lim_{x\to a} f(x) = L$, there is $\delta_3 > 0$ such that

$$\text{if } 0 < |x - a| < \delta_3, \text{ then } |f(x) - L| < \frac{\varepsilon}{2(|M| + 1)}. \tag{7.3}$$

Finally, we set $\delta = \min(\delta_1, \delta_2, \delta_3)$; then whenever $0 < |x - a| < \delta$, the inequalities in (7.1), (7.2), and (7.3) are all true, and we have

$$\begin{aligned} |f(x)g(x) - LM| &\le |g(x)||f(x) - L| + |L||g(x) - M| \\ &\le (|M| + 1)\frac{\varepsilon}{2(|M| + 1)} + |L|\frac{\varepsilon}{2|L| + 1} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, this proves that $\lim_{x\to a} f(x)g(x) = LM$, as claimed. □

Finally, we prove Law 5, starting with the special case when $f(x) = 1$.

Proposition 7.4. *If* $\lim_{x\to a} g(x) = M \neq 0$, *then* $\lim_{x\to a} \frac{1}{g(x)} = \frac{1}{M}$.

Proof. Suppose our adversary chooses $\varepsilon > 0$. The error term that we must control is

$$\Big|\frac{1}{g(x)} - \frac{1}{M}\Big| = \frac{|M - g(x)|}{|g(x)M|}. \tag{7.4}$$

The numerator becomes small when $x \approx a$, but what if the denominator also becomes small? We need to get a lower bound on $|g(x)|$, which we do by choosing $\delta_1 > 0$ small enough that

$$\text{if } 0 < |x - a| < \delta_1, \text{ then } |g(x) - M| < \frac{|M|}{2}, \text{ so } |g(x)| > |M| - \frac{|M|}{2} = \frac{|M|}{2}.$$

In this case we have $\frac{1}{|g(x)|} < \frac{2}{|M|}$, so the error term in (7.4) can be estimated as

$$\Big|\frac{1}{g(x)} - \frac{1}{M}\Big| = \frac{|g(x) - M|}{|M|}\frac{1}{|g(x)|} < \frac{2|g(x) - M|}{|M|^2}.$$

Once more using the fact that $\lim_{x\to a} g(x) = M$, there is $\delta_2 > 0$ such that

$$\text{if } 0 < |x - a| < \delta_2, \text{ then } |g(x) - M| < \frac{\varepsilon|M|^2}{2}.$$

Let $\delta = \min(\delta_1, \delta_2)$; then for every x with $0 < |x - a| < \delta$, both of the above estimates hold, and we have

$$\Big|\frac{1}{g(x)} - \frac{1}{M}\Big| < \frac{2}{|M|^2}|g(x) - M| < \frac{2}{|M|^2}\frac{\varepsilon|M|^2}{2} = \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this proves that $\lim_{x\to a} \frac{1}{g(x)} = \frac{1}{M}$. □

[12] The reason we use $2|L| + 1$ in the denominator, and not $2|L|$, is that we might have $L = 0$.

The general case of Law 5 follows from this proposition together with Law 4: if $\lim_{x\to a} f(x) = L$ and $\lim_{x\to a} g(x) = M$, then

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} f(x) \cdot \frac{1}{g(x)} = \Big(\lim_{x\to a} f(x)\Big)\Big(\lim_{x\to a} \frac{1}{g(x)}\Big) = L \cdot \frac{1}{M} = \frac{L}{M},$$

where the second equality uses Law 4 (limit of products) and the third equality uses Proposition 7.4.

Lecture 8 Theorems about limits

This lecture corresponds to §2.3 and Appendix F in Stewart, and parts of Chapter 5 in Spivak.

8.1. Direct substitution

Some limits can be computed via the method of "direct substitution".

Proposition 8.1. *If f is a polynomial, then $\lim_{x\to a} f(x) = f(a)$ for every $a \in \mathbb{R}$.*

Proof. We prove this by induction in the degree of f. If $\deg f = 0$ then $f(x) = c$ is a constant, so the claim is true by Law 7. If the claim is true for polynomials of degree $n-1$, and f is a polynomial of degree n, then $f(x) = xg(x) + c$, where $c \in \mathbb{R}$ and g is a polynomial of degree $n-1$, so we have

$$\begin{aligned}
\lim_{x\to a} f(x) &= \lim_{x\to a}(xg(x)) + \lim_{x\to a} c && \text{by Law 1}\\
&= \Big(\lim_{x\to a} x\Big)\Big(\lim_{x\to a} g(x)\Big) + c && \text{by Laws 4 and 7}\\
&= ag(a) + c && \text{by Law 8 and the inductive hypothesis}\\
&= f(a).
\end{aligned}$$

Thus the result holds for every n by induction. □

Proposition 8.2. *If f is a rational function and a is in the domain of f, then $\lim_{x\to a} f(x) = f(a)$.*

Proof. Let g, h be polynomials such that $f(x) = g(x)/h(x)$ for all x where $h(x) \neq 0$. By Proposition 8.1, we have

$$\lim_{x\to a} g(x) = g(a) \quad \text{and} \quad \lim_{x\to a} h(x) = h(a) \neq 0,$$

so Limit Law 5 gives

$$\lim_{x\to a} f(x) = \frac{\lim_{x\to a} g(x)}{\lim_{x\to a} h(x)} = \frac{g(a)}{h(a)} = f(a).$$ □

8.2. Two-sided and one-sided limits

Now we relate two-sided and one-sided limits.

Theorem 8.3. $\lim_{x\to a} f(x) = L$ *if and only if* $\lim_{x\to a^-} f(x) = L = \lim_{x\to a^+} f(x)$.

Proof. ($\Rightarrow$):[13] If $\lim_{x\to a} f(x) = L$, then for every $\varepsilon > 0$ there is $\delta > 0$ such that

$$\text{if } 0 < |x - a| < \delta, \text{ then } |f(x) - L| < \varepsilon.$$

In particular, $x \in (a, a+\delta) \Rightarrow |f(x) - L| < \varepsilon$, and since $\varepsilon > 0$ was arbitrary this implies that $\lim_{x\to a^+} f(x) = L$. Similarly, $x \in (a-\delta, a) \Rightarrow |f(x) - L| < \varepsilon$, and thus $\lim_{x\to a^-} f(x) = L$.

($\Leftarrow$): If both of the one-sided limits exist and are equal to L, then for every $\varepsilon > 0$ there are $\delta_1, \delta_2 > 0$ such that

$$\text{if } x \in (a, a+\delta_1), \text{ then } |f(x) - L| < \varepsilon, \text{ and}$$
$$\text{if } x \in (a-\delta_2, a), \text{ then } |f(x) - L| < \varepsilon.$$

Taking $\delta = \min(\delta_1, \delta_2)$, we see that for every x with $0 < |x-a| < \delta$, we have $x \in (a, a+\delta_1)$ or $x \in (a-\delta_2, a)$, and in either case we get $|f(x) - L| < \varepsilon$. Thus for every $\varepsilon > 0$ we can produce the required $\delta > 0$, which shows that $\lim_{x\to a} f(x) = L$. □

Example 8.4. Consider $\lim_{x\to 0} |x|$. For every $x > 0$ we have $|x| = x$, so

$$\lim_{x\to 0^+} |x| = \lim_{x\to 0^+} x = 0 \text{ by Limit Law 8.}$$

For every $x < 0$ we have $|x| = -x$, so

$$\lim_{x\to 0^-} |x| = \lim_{x\to 0^-} (-x) = -\lim_{x\to 0^-} x = -0 = 0 \text{ by Limit Laws 3 and 8.}$$

Since the one-sided limits exist and agree, Theorem 8.3 implies that $\lim_{x\to 0} |x| = 0$.

Example 8.5. Define $f\colon \mathbb{R} \setminus \{0\} \to \mathbb{R}$ by $f(x) = x/|x|$, so $f(x) = 1$ if $x > 0$ and -1 if $x < 0$. Then

$$\lim_{x\to 0^+} f(x) = \lim_{x\to 0^+} 1 = 1 \quad \text{and} \quad \lim_{x\to 0^-} f(x) = \lim_{x\to 0^-} -1 = -1 \quad \text{by Law 7,}$$

so by Theorem 8.3, $\lim_{x\to 0} f(x)$ does not exist.

8.3. Inequalities and limits

Finally, we prove two results demonstrating that inequalities between functions can be passed to the corresponding limits.

Theorem 8.6. *If* $f(x) \le g(x)$ *for every* x*, and if both* $\lim_{x\to a} f(x)$ *and* $\lim_{x\to a} g(x)$ *exist, then* $\lim_{x\to a} f(x) \le \lim_{x\to a} g(x)$.

Proof. Let $L = \lim_{x\to a} f(x)$ and $M = \lim_{x\to a} g(x)$. We use proof by contradiction. Suppose that $M < L$. Then by Law 2, we have

$$\lim_{x\to a} \big(g(x) - f(x)\big) = M - L < 0,$$

[13] To prove an "if and only if" result, one often proves each direction separately. Here "($\Rightarrow$)" means that we are proving that the first statement (two-sided limit) implies the second (one-sided limits), and "($\Leftarrow$)" means that we are proving that the second implies the first.

and so putting $\varepsilon = L - M > 0$, the definition of limit gives $\delta_1 > 0$ such that

$$\text{if } 0 < |x - a| < \delta_1, \text{ then } 0 < |(g(x) - f(x)) - (M - L)| < \varepsilon = L - M,$$

and thus for all such x, we have

$$g(x) \leq |g(x) - f(x) - (M - L)| + f(x) + M - L < L - M + f(x) + M - L = f(x).$$

This contradicts the assumption that $f(x) < g(x)$, demonstrating that we must have $M \geq L$ after all. □

Theorem 8.7 (Squeeze Theorem)**.** *Suppose that f, g, h are functions such that $f(x) \leq g(x) \leq h(x)$ for every x, and that moreover $\lim_{x\to a} f(x) = \lim_{x\to a} h(x) = L$. Then we have $\lim_{x\to a} g(x) = L$.*

Proof. Given $\varepsilon > 0$, there are $\delta_1, \delta_2 > 0$ such that

$$\text{if } 0 < |x - a| < \delta_1, \text{ then } L - \varepsilon < f(x) < L + \varepsilon, \text{ and}$$
$$\text{if } 0 < |x - a| < \delta_2, \text{ then } L - \varepsilon < h(x) < L + \varepsilon.$$

Let $\delta = \min(\delta_1, \delta_2)$. Then if $0 < |x - a| < \delta$, the estimates on $f(x)$ and $h(x)$ both hold, so

$$L - \varepsilon < f(x) \leq g(x) \leq h(x) < L + \varepsilon,$$

which implies that $|g(x) - L| < \varepsilon$. Since $\varepsilon > 0$ was arbitrary, this proves that $\lim_{x\to a} g(x) = L$. □

Remark 8.8. Both of the preceding theorems continue to hold if we replace two-sided limits by one-sided limits. Moreover, in the hypotheses of the theorems, it is enough for the inequalities to hold for every x that is sufficiently close to a; in other words, if there is $\delta > 0$ such that the inequalities hold for all $0 < |x - a| < \delta$, then the conclusion of the theorem still holds.

Lecture 9 Continuity

This lecture corresponds to §2.5 in Stewart and Chapter 6 in Spivak.

9.1. Definition and basic examples

The 'direct substitution' property from the start of the previous lecture is important enough to study at greater length, and we make the following definition.

Definition 9.1. A function f is *continuous at a point a* if $\lim_{x\to a} f(x) = f(a)$.

Note that this definition actually requires three things to be true:

(1) $f(a)$ must be defined (a is in the domain of f);
(2) the limit $\lim_{x\to a} f(x)$ exists;
(3) the two values are equal.

If any of these three fails, then the function is not continuous at a; in this case we say that f is *discontinuous at a*.

Example 9.2. Consider the piecewise constant *Heaviside function*

$$H(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x \geq 0. \end{cases}$$

This function is continuous at a for every $a \neq 0$, but is discontinuous at 0, because the limit does not exist there.

The value of the Heaviside function "jumps" from 0 to 1 as x increases through 0. This type of behavior is important enough to give a name to.

Definition 9.3. If the left-hand and right-hand limits of f at a both exist, but take different values, then we say that f has a *jump discontinuity* at a.

Example 9.4. Consider the *floor function* $f(x) = \lfloor x \rfloor$, which takes x to the largest integer n such that $n \leq x$. This is continuous at every $x \in \mathbb{R} \setminus \mathbb{Z}$ and has a jump discontinuity at every $x \in \mathbb{Z}$.

Remark 9.5. Using the ϵ-δ definition of a limit, we see that f is continuous at a if and only if the following is true: for every $\epsilon > 0$ there exists $\delta > 0$ such that for every $|x - a| < \delta$, we have $|f(x) - f(a)| < \epsilon$.

Remark 9.6. Note that in the previous remark we wrote $|x - a| < \delta$ instead of the usual $0 < |x - a| < \delta$ that appears in the definition of the limit. The reason we can omit the first inequality is that when $|x - a| = 0$ we have $x = a$, and thus $f(x) = f(a)$, so $|f(x) - f(a)| = 0 < \epsilon$ automatically.

Remark 9.7. Recall from Exercise 6.2 that $\lim_{x \to a} f(x) = \lim_{h \to 0} f(a + h)$, and thus f is continuous at a if and only if $\lim_{h \to 0} f(a + h) = f(a)$. As in Remark 9.5 this can be written as follows: for every $\epsilon > 0$ there exists $\delta > 0$ such that for every $|h| < \delta$, we have $|f(a + h) - f(a)| < \epsilon$.

Definition 9.8. A function f is *right-continuous* at a if $\lim_{x \to a^+} f(x) = f(a)$, and *left-continuous* at a if $\lim_{x \to a^-} f(x) = f(a)$.

Recalling Theorem 8.3, we see immediately that f is continuous at a if and only if it is both left- and right-continuous at a.

Definition 9.9. A function f is *continuous on an interval* I if it is continuous at every point $a \in I$. If I contains an endpoint that is also an endpoint of the domain of f, then we interpret 'continuous' to mean either 'left-continuous' or 'right-continuous', as appropriate.

9.2. Polynomials, rational, and root functions are continuous

Limit Laws 9 and 10 say that the functions $f(x) = x^n$ and $g(x) = \sqrt[n]{x}$ are continuous on their entire domains. A generalization of this first fact is given by Proposition 8.1 from the last lecture, which says that if f is a polynomial, then $f(a) = \lim_{x \to a} f(x)$ for every $a \in \mathbb{R}$. This can be restated by saying that f is continuous at every $a \in \mathbb{R}$, or even more succinctly as "f is continuous on $\mathbb{R}$". Similarly, Proposition 8.2 says that if f is a rational function, then it is continuous on its domain (that is, the set of values of x for which the denominator is nonzero).

Remark 9.10. The domain of a rational function is always $\mathbb{R} \setminus A$, where A is a finite set. This is because if p, q are polynomials, then $f(x) = p(x)/q(x)$ is defined whenever $q(x) \neq 0$, and by Exercise 3.2, there are at most $\deg f$ values of x for which $q(x) = 0$.

Example 9.11.

(1) $f(x) = \frac{1}{x}$ is continuous at a for every $a \neq 0$ and discontinuous at 0.
(2) $f(x) = \frac{x^2-1}{x-1}$ is continuous at every $x \neq 1$ and discontinuous at 1, where it is undefined. Observe that the function $g(x) = x + 1$ is continuous at every x (including $x = 1$) and agrees with $f(x)$ everywhere that the latter is defined.

Definition 9.12. If a function f has the property that the left-hand and right-hand limits at a both exist and agree with each other, but disagree with $f(a)$ (which may or may not be defined), then we say that f has a *removable discontinuity* at a.

Remark 9.13. When we discussed the notation "f^{-1}" for the inverse function of f, we used the notation $f^n(x)$ to denote the result of *composing* f with itself n times, so

$$f^n(x) = \overbrace{f \circ f \circ \cdots \circ f}^{n \text{ times}}(x).$$

In particular, according to that convention, $f^2(x)$ would mean $f(f(x))$, and we would write $f(x)^2$ or $(f(x))^2$ to denote $f(x)f(x)$. However, with the trigonometric functions it is standard to write $\cos^2 x$ or $\cos^2(x)$ to mean $(\cos x)(\cos x)$, instead of $\cos(\cos x)$, and from now on we will use this 'power' notation to mean multiplication of trigonometric functions. The one exception is that $\cos^{-1} x$ will still mean the inverse function arccosine.

This situation of ambiguous notation, where the same way of writing things can have multiple meanings, is unfortunate but occurs from time to time in mathematics. In general you can deduce the meaning from context, but you should be aware that this possibility exists.

9.3. New continuous functions from old ones

Theorem 9.14. *If f and g are functions that are continuous at $a \in \mathbb{R}$, then the following functions are also continuous at a:*

(1) $f + g$;
(2) $f - g$;
(3) cf for every $c \in \mathbb{R}$;
(4) fg (note that this means the product $(fg)(x) = f(x)g(x)$ rather than the composition);
(5) f/g provided $g(a) \neq 0$.

Proof. These assertions follow immediately from the corresponding limit laws. □

Theorem 9.15. *If f, g are real-valued functions such that g is continuous at a and f is continuous at $g(a)$, then $f \circ g$ is continuous at a.*

Proof. Given $\epsilon > 0$, we want to produce $\delta > 0$ such that for every $|x - a| < \delta$, we have $|f(g(x)) - f(g(a))| < \epsilon$. To accomplish this, we proceed as follows.

(1) Use the fact that f is continuous at $g(a)$ to deduce that there exists $\delta_1 > 0$ such that for all y with $|y - g(a)| < \delta_1$, we have $|f(y) - f(g(a))| < \epsilon$.

(2) Use continuity of g at a to deduce that there exists $\delta > 0$ such that for all x with $|x - a| < \delta$, we have $|g(x) - g(a)| < \delta_1$.

Now given any x with $|x - a| < \delta$, the second item gives $|g(x) - g(a)| < \delta_1$, and then the first gives $|f(g(x)) - f(g(a))| < \delta$, which proves the theorem. □

Example 9.16. Because $x^2 + 1 \geq 0$ for all $x \in \mathbb{R}$ and the square root function is continuous on $[0, \infty)$, it follows from Theorem 9.15 that the function $f(x) = \sqrt{x^2 + 1}$ is continuous on $\mathbb{R}$. Note that we could also have deduced this using Limit Law 11.

9.4. Trigonometric functions are continuous

Theorem 9.17. *The sine and cosine functions are both continuous on* $\mathbb{R}$.

Proof. We start by proving that both of these functions are continuous at 0; then we use the sum-of-angles formulas (3.3) and (3.4) to deduce continuity at every $a \in \mathbb{R}$. For continuity at 0, we start by recalling the definition of cosine and sine, as illustrated in the left-hand side of Figure 2, which suggests that $\lim_{\theta \to 0} \cos\theta = 1$ and $\lim_{\theta \to 0} \sin\theta = 0$, so that cos and sin are continuous at 0.

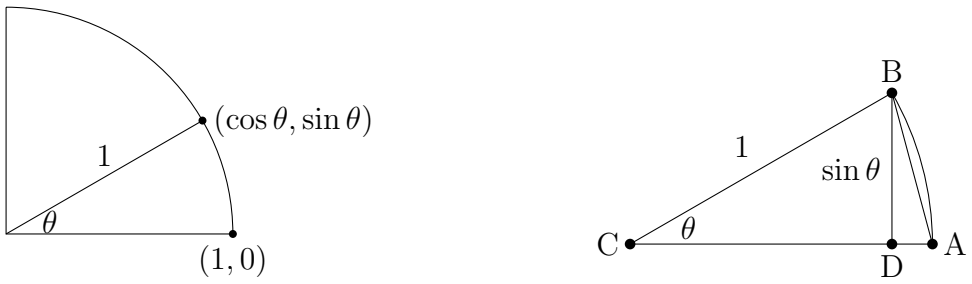


FIGURE 2. Proving continuity of sin at 0.

To prove this rigorously, we can use the right-hand side of Figure 2 to observe that

$$\sin\theta = BD < AB \leq (\text{length of arc from } A \text{ to } B) = \theta,$$

where the first inequality is because the hypotenuse of a right triangle is longer than either leg, and the second inequality is because the length of any curve between two points is at least as large as the distance between them. Thus for $\theta \in (0, \pi/2)$ we have $0 < \sin\theta < \theta$. Since $\lim_{\theta \to 0} 0 = \lim_{\theta \to 0} \theta = 0$, the squeeze theorem proves that $\lim_{\theta \to 0^+} \sin\theta = 0$. Since $\sin(-\theta) = -\sin\theta$, we conclude that $\lim_{\theta \to 0} \sin\theta = 0$, so sin is continuous at 0.

For the corresponding result for cosine, we use this fact together with the identity $1 - \cos^2\theta = \sin^2\theta$, which implies $1 - \cos\theta = \frac{\sin^2\theta}{1+\cos\theta}$, to get

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

so $\lim_{\theta \to 0} \cos\theta = 1$, which shows that cos is continuous at 0.

With these two results in hand, we can use (3.3) to deduce that for any $a \in \mathbb{R}$, the sine function is continuous at a. Indeed, from Remark 9.7 we see that it suffices to show that $\lim_{h \to 0} \sin(a + h) = \sin a$, and the sum-of-angles formula (3.3) gives

$$\lim_{h\to 0}\sin(a+h) = \lim_{h\to 0}(\sin a\cos h + \cos a\sin h) = \sin a\Big(\lim_{h\to 0}\cos h\Big) + \cos a\Big(\lim_{h\to 0}\sin h\Big)$$
$$= (\sin a)\cdot 1 + (\cos a)\cdot 0 = \sin a.$$

This proves continuity of $x \mapsto \sin x$ on $\mathbb{R}$. For cosine, we use (3.4) to get

$$\lim_{h\to 0}\cos(a+h) = \lim_{h\to 0}(\cos a\cos h - \sin a\sin h) = \cos a\Big(\lim_{h\to 0}\cos h\Big) - \sin a\Big(\lim_{h\to 0}\sin h\Big)$$
$$= (\cos a)\cdot 1 - (\sin a)\cdot 0 = \cos a,$$

and thus $x \mapsto \cos x$ is also continuous on $\mathbb{R}$. This completes the proof of Theorem 9.17. □

Exercise 9.18. Use the results proved so far to show that tan, cot, sec, and csc are all continuous on their domains.

Using Theorems 9.15 and 9.17 together with the fact that rational functions are continuous, we deduce the following.

Example 9.19.

(1) $\sin\frac{1}{x}$ is continuous at every $x \neq 0$. There is no way to extend this function to a continuous function $f\colon \mathbb{R} \to \mathbb{R}$ such that f is continuous at 0. That is, if $f\colon \mathbb{R} \to \mathbb{R}$ is any function such that $f(x) = \sin\frac{1}{x}$ for all $x \neq 0$, then f is discontinuous at 0.

(2) The function

$$f(x) := \begin{cases} x\sin\frac{1}{x} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0 \end{cases}$$

is continuous at every $x \in \mathbb{R}$. Continuity at $x \neq 0$ is a consequence of the previous part and Theorem 9.14 (product of continuous functions is continuous). For continuity at $x = 0$, we use the Squeeze Theorem, observing that $-|x| \leq x\sin\frac{1}{x} \leq |x|$ for all $x \neq 0$ and $\lim_{x\to 0} -|x| = \lim_{x\to 0}|x| = 0$.

At this point one may reasonably ask about the inverse trigonometric functions. Are they continuous? We will eventually see that they are, but our proof will require the *Intermediate Value Theorem*, which we discuss in the next lecture.

9.5. What about exponentials?

Recall that given $a > 0$, we defined $f(x) = a^x$ in the following way.

- When $x \in \mathbb{N}$, it is defined iteratively (multiply a by itself x times).
- To extend to $x \in \mathbb{Z}$, we take reciprocals: $a^{-x} = 1/a^x$.
- To extend to $x \in \mathbb{Q}$, we take roots: $a^{p/q}$ is the qth root of a^p. (There's a subtlety here, though – why do qth roots exist?)
- To extend to $x \in \mathbb{R}$, we take limits: a^x is the limit of a^{r_n} when r_n is a sequence of rational numbers approaching x.

The last step here strongly suggests that the exponential function is continuous, and indeed it is. As with the sine and cosine functions, we start by proving continuity at 0.

In fact, we start by proving something about the behavior of the exponential function when the exponent gets large.

Lemma 9.20. *For every $a \geq 1$ and $b > 1$, there exists $n \in \mathbb{N}$ such that $b^n > a$.*

Proof. Let $t = b - 1$, then $b^n = (1+t)^n \geq 1 + tn$, where the last inequality can be proved by induction or by observing that $(1+t)^n = 1 + tn + \frac{n(n-1)}{2}t^2 + \cdots + t^n \geq 1 + tn$ since all the extra terms are ≥ 0. Thus it suffices to take $n > \frac{a-1}{t}$, since this gives $b^n \geq 1 + tn > 1 + (a-1) = a$. □

Now we prove right-continuity at 0 when $a \geq 1$.

Lemma 9.21. *For every $a \geq 1$, we have $\lim_{x \to 0^+} a^x = 1$.*

Proof. Given $\epsilon > 0$, use Lemma 9.20 to find $n \in \mathbb{N}$ such that $(1+\epsilon)^n > a$, and let $\delta = 1/n$. Then taking nth roots gives $a^\delta = a^{1/n} < 1 + \epsilon$. Now for every $x \in (0, \delta)$ we have

$$1 \leq a^x \leq a^\delta < 1 + \epsilon,$$

where the second inequality uses the fact that $a^x \leq a^y$ whenever $a \geq 1$ and $x \leq y$. Because $\epsilon > 0$ was arbitrary, this completes the proof. □

Remark 9.22. Although this proof certainly establishes the desired result, it is a little bit cryptic because it gives no indication of how we decided on that particular choice of δ. The reasoning behind this choice is as follows: "Given $\epsilon > 0$, we want to find $\delta > 0$ such that every $x \in (0, \delta)$ has $|a^x - 1| < \epsilon$. Since $a^x > 1$ for all $x > 0$, this is equivalent to showing that $a^x < 1 + \epsilon$. Moreover, since $a^x < a^\delta$ for all $x \in (0, \delta)$, this is equivalent to choosing $\delta > 0$ such that $a^\delta < 1 + \epsilon$. This last inequality is equivalent to $a < (1+\epsilon)^{1/\delta}$."

Now we use a similar argument to get left-continuity.

Lemma 9.23. *For every $a \geq 1$, we have $\lim_{x \to 0^-} a^x = 1$.*

Proof. Given $\epsilon > 0$, use Lemma 9.20 to find $n \in \mathbb{N}$ such that $(\frac{1}{1-\epsilon})^n > a$, and let $\delta = 1/n$. Raising both sides to the power $-1/n$ gives $1 - \epsilon < a^{-1/n}$, and thus for every $x \in (-\delta, 0)$ we have $1 - \epsilon < a^{-\delta} \leq a^{-x} \leq 1$. □

Combining Lemmas 9.21 and 9.23 shows that $f(x) = a^x$ is continuous at 0 for every $a \geq 1$. In fact it is continuous at *every* $x \in \mathbb{R}$:

$$\lim_{h \to 0} a^{x+h} = \lim_{h \to 0} a^x a^h = a^x \lim_{h \to 0} a^h = a^x \cdot 1 = a^x.$$

Here the first equality uses the basic property of exponentials, the second equality uses Limit Law 3, and the third uses continuity at 0. Finally, we observe that for $0 < a < 1$ we have

$$\lim_{y \to x} a^y = \lim_{y \to x} (1/a)^{-y} = (1/a)^{-x} = a^x$$

for all $x \in \mathbb{R}$, where the second equality uses continuity for bases that are ≥ 1. Putting it all together, we have proved the following.

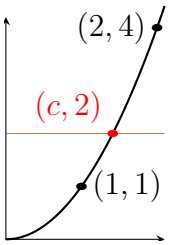
Theorem 9.24. *For every $a > 0$, the function $f(x) = a^x$ is continuous on $\mathbb{R}$.*

It is also natural to ask about the logarithm function, which is the inverse function of the exponential, and just as with the inverse trig functions, our proof of continuity will need to wait until we have discussed the Intermediate Value Theorem.

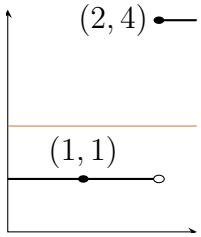
Lecture 10 Intermediate Value Theorem: Preparation

This lecture corresponds to §2.5 of Stewart and parts of Chapters 7 and 8 of Spivak; however, Stewart does not prove the theorem, and Spivak's proof differs from ours.

Now we return to a question raised in the opening lecture: Why does $\sqrt{2}$ exist? To put this a little bit more precisely, why is there a real number x with the property that $x^2 = 2$?



Informally, we might reason as follows: *The graph of the function $f(x) = x^2$ is a parabola opening upwards, as shown in the picture. The graph goes through the point $(1, 1)$, which is below the line $y = 2$, and through the point $(2, 4)$, which is above it. In order to go between these points, the graph has to cross the line $y = 2$ at some point $(c, 2)$, and then we must have $f(c) = c^2 = 2$, so $c = \sqrt{2}$.*



This reasoning is made precise by the *Intermediate Value Theorem*, which we will state in a moment. Before doing so, we observe that this reasoning must use some property of the function $f(x) = x^2$. The picture at left shows the function

$$g(x) = \begin{cases} 1 & x < 2 \\ 4 & x \geq 2 \end{cases},$$

which also has the property that its graph goes through the points $(1, 1)$ and $(2, 4)$; however, there is no value of c for which $g(c) = 2$.

This example illustrates that a function can have a discontinuity that lets it jump from one side of a line to the other without intersecting it. The Intermediate Value Theorem says that a discontinuity is the *only* way that this can happen.

Theorem 10.1 (Intermediate Value Theorem)**.** *If $f\colon [a, b] \to \mathbb{R}$ is continuous and $f(a) < r < f(b)$, then there exists a real number $c \in [a, b]$ such that $f(c) = r$.*

Remark 10.2. The IVT is not true if we replace "real number" by "rational number". Indeed, as we saw in Theorem 1.1, there is no rational number c such that $c^2 = 2$.

Remark 10.3. You may sometimes see the IVT summarized as the statement that "if you draw a curve from one side of a line to the other without lifting your pen from the paper, then you must intersect the line somewhere." And stated this way, the theorem seems obvious; how could it be otherwise? We cannot imagine how the theorem could fail, so what is there to prove?

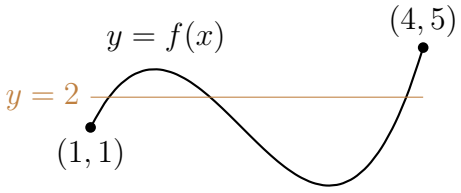
However, "we cannot imagine X happening" is quite different from saying that X never happens. After all, the early Pythagoreans could not imagine that there were two lengths whose ratio could not be expressed as a rational number,[14] but it turns out that the diagonal and side of a square form such a pair, since their ratio is $\sqrt{2}$. So until we produce a genuine proof, we must consider the possibility that perhaps there is a failure in our imagination.

Another way of thinking about this is the following: the statement involving "drawing a curve without lifting your pen from the paper" is meant to describe the graph of a

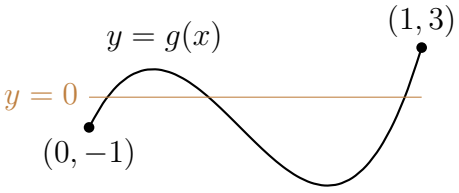
[14] A pair of such lengths are called *incommensurable*.

continuous function. However, the definitions of continuity in Definition 9.1 and Remark 9.5 are quite technical; what do they have to do with "drawing without lifting your pen"? One can interpret the IVT as providing a connection between the technical definition and the intuitive one.

Before embarking on the proof itself, we make a mild simplification. The precise values of a, b, r turn out not to be particularly important to the theorem; what is important is the relationship $f(a) < r < f(b)$. Looking at the two functions shown below, one may reasonably expect that the same argument should work for both of them, even though the values are different.[15]



$a = 1,\ b = 4,\ r = 2$



$a = 0,\ b = 1,\ r = 0$

The function g in the right-hand picture is obtained from the function f in the left-hand picture by rescaling horizontally (the interval $[1, 4]$ gets mapped to the interval $[0, 1]$) and shifting vertically (the line $y = 2$ gets moved to the line $y = 0$). In fact, suppose we have *any* real numbers $a < b$, a function $f\colon [a, b] \to \mathbb{R}$, and a real number r such that $f(a) < r < f(b)$. Then the function

$$g(x) := f(a + (b - a)x) - r$$

has the following properties:

- $g\colon [0, 1] \to \mathbb{R}$ is continuous;
- $g(0) < 0 < g(1)$; and
- given $c \in [0, 1]$, we have $g(c) = 0$ if and only if $f(a + (b - a)c) = r$.

In light of this, we will first prove the IVT under the additional assumption that $a = 0$, $b = 1$, and $r = 0$. The procedure just described shows that if we know the theorem is true in this case, then it is true for arbitrary values of a, b, r.[16]

Theorem 10.4 (IVT: simple form)**.** *If $f\colon [0, 1] \to \mathbb{R}$ is continuous and $f(0) < 0 < f(1)$, then there exists a real number $c \in [0, 1]$ such that $f(c) = 0$.*

Our proof of Theorem 10.4 uses *bisection sequences*.[17] We build two sequences $b_n, r_n \in [0, 1]$ as shown in the pictures below (where there are blue points and red points, respectively) by the following iterative procedure:

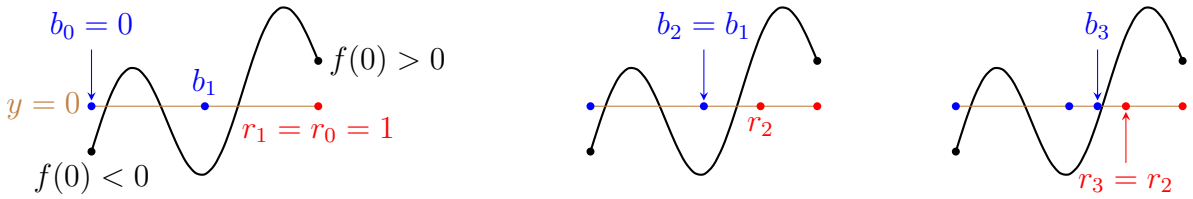
(1) start by putting $b_0 = 0$ and $r_0 = 1$;

[15] Also note that as these pictures show, there may well be more than one value of c with $f(c) = r$. To claim that there is a *unique* $c \in [a, b]$ with this property requires some extra information about f.

[16] Mathematicians often summarize this whole discussion by saying, "Without loss of generality, we assume that $a = 0, b = 1, r = 0$". This language means that we prove a specific case, which turns out to contain all the ingredients for the general case.

[17] Our proof follows an article by Stephen M. Walk entitled "The Intermediate Value Theorem is NOT Obvious—and I Am Going to Prove It to You", which appeared in *The College Mathematics Journal*, vol. 42, No. 4 (Sep. 2011), pp. 254–259.

(2) if b_n, r_n have been defined, then take $m_n = \frac{1}{2}(b_n + r_n)$ to be the midpoint of $[b_n, r_n]$, and define b_{n+1}, r_{n+1} as follows:
- if $f(m_n) = 0$, then stop (we found the desired c);
- if $f(m_n) < 0$, then let $b_{n+1} = m_n$, $r_{n+1} = r_n$;
- if $f(m_n) > 0$, then let $b_{n+1} = b_n$, $r_{n+1} = m_n$.



If the first case above ($f(m_n) = 0$) happens for some n, then we put $c = m_n$ and the theorem is proved. So we need to consider the situation where the first case never happens, and the sequences b_n, r_n go on forever. Then the following properties are immediate consequences of the definitions:

- $0 \le b_0 \le b_1 \le b_2 \le \cdots \le r_2 \le r_1 \le r_0 = 1$;
- $f(b_n) < 0 < f(r_n)$ for every $n \in \mathbb{N}$;
- $r_n - b_n = 2^{-n}$ for every $n \in \mathbb{N}$.

The pictures suggest that the root c of the equation $f(c) = 0$ should lie to the right of every b_n, and to the left of every r_n. In fact, we expect to see that $c = \lim_{n\to\infty} b_n = \lim_{n\to\infty} r_n$, where we recall that the definition of limit of a sequence was given in Definition 6.6. In the next lecture, we will complete the proof via the following steps:

- $b = \lim_{n\to\infty} b_n$ and $r = \lim_{n\to\infty} r_n$ both exist;
- $f(b) \le 0$ and $f(r) \ge 0$;
- $b = r$;
- putting $c = b = r$ gives $f(c) = 0$.

Lecture 11 IVT: Proof and consequences

The first parts of this lecture correspond to §2.5 of Stewart and parts of Chapters 7 and 8 of Spivak; however, Stewart does not prove the theorem, and Spivak's proof differs from ours. The end of this lecture corresponds to §2.6 of Stewart and parts of Chapter 5 of Spivak.

11.1. Completion of the proof

To finish proving the Intermediate Value Theorem, we need to justify the list of claims at the end of the previous lecture.

The first of these is a fundamental property of the real numbers: *every bounded nondecreasing sequence converges to a real number.* More precisely, the following is true.

Monotone Convergence Theorem for real numbers. *Let $x_1, x_2, x_3, \ldots$ be a sequence of real numbers with the following properties:*

- *the sequence is* nondecreasing, *meaning that $x_n \le x_{n+1}$ for every $n \in \mathbb{N}$;*

- *the sequence is* bounded above*, meaning that there is some* $B \in \mathbb{R}$ *such that* $x_n \le B$ *for every* $n \in \mathbb{N}$.

Then there exists some $x \in \mathbb{R}$ *such that* $x = \lim_{n\to\infty} x_n$.

Remark 11.1. We will not prove this theorem; rather, we will take it as a fundamental property of the real numbers that distinguishes them from smaller sets of numbers such as $\mathbb{Q}$. (Note that the theorem fails if we replace $\mathbb{R}$ by $\mathbb{Q}$: consider the sequence 1, 1.4, 1.41, 1.414, 1.4142, …that converges to $\sqrt{2}$, which is not in $\mathbb{Q}$.) In order to prove this theorem one needs to *construct* the real numbers, which we will not do in this course.

Remark 11.2. A number B with the property stated above is called an *upper bound* for the sequence $\{x_n\}$. The real number x produced by the theorem is in fact an upper bound for the sequence (can you prove this?), and actually satisfies the stronger property of being a *least* upper bound: in other words, every number $y < x$ is *not* an upper bound for the sequence. If you are interested in exploring the foundations of the subject, it is a worthwhile exercise to prove that the monotone convergence theorem stated above is equivalent to the *least upper bound property*, which says that every *set* of real numbers that is bounded above has a least upper bound. This least upper bound property is used in Spivak's book.

Since b_n is a nondecreasing sequence, the Monotone Convergence Theorem implies that $b := \lim_{n\to\infty} b_n$ exists. For the sequence r_n, we observe that it is nonincreasing, meaning that $r_{n+1} \le r_n$ for all n. We use the following consequence of the MCT.

Corollary 11.3. *If* $x_1 \ge x_2 \ge x_3 \ge \cdots$ *is a nonincreasing sequence of real numbers that is bounded below, then* $\lim_{n\to\infty} x_n$ *exists.*

Proof. Apply the Monotone Convergence Theorem to the nondecreasing sequence $-x_1 \le -x_2 \le -x_3 \le \cdots$. □

This corollary shows that $r := \lim_{n\to\infty} r_n$ exists, so we have proved the first claim on our list. For the second claim, we observe that $f(b_n) < 0$ for all n, and recall Theorem 8.6, which said that if $f(x) \le g(x)$ for all x, then $\lim_{x\to a} f(x) \le \lim_{x\to a} g(x)$ provided both limits exist. It is easy to prove an analogous result for sequences:

(11.1) if $x_n \le y_n$ for all n, then $\lim\limits_{n\to\infty} x_n \le \lim\limits_{n\to\infty} y_n$ provided both limits exist.

Putting $x_n = f(b_n)$ and $y_n = 0$, this gives

$$\begin{aligned} f(b) &= f\Big(\lim_{n\to\infty} b_n\Big) && \text{by definition of } b \\ &= \lim_{n\to\infty} f(b_n) && \text{by continuity of } f \\ &\le 0 && \text{by (11.1).} \end{aligned}$$

A similar argument with r_n and r gives $f(r) \ge 0$, so we have proved

$$f(b) \le 0 \le f(r). \tag{11.2}$$

Now we recall that from the definition of the sequences b_n, r_n, we have $r_n - b_n = 2^{-n}$ for all N. In particular, we have

$$r - b = \Big(\lim_{n\to\infty} r_n\Big) - \Big(\lim_{n\to\infty} b_n\Big) = \lim_{n\to\infty} (r_n - b_n) = \lim_{n\to\infty} 2^{-n},$$

where the second inequality uses Limit Law 2.

Exercise 11.4. Use Lemma 9.20 to prove that $\lim_{n\to\infty} 2^{-n} = 0$.

The exercise implies that $r - b = 0$, so $r = b$ and we can define $c = r = b$. Then (11.2) implies that $f(c) \le 0 \le f(c)$, which is only possible if $f(c) = 0$. This completes the proof of the IVT.

11.2. Applications of the IVT

Theorem 11.5. *If $f\colon [0,1] \to [0,1]$ is continuous, then there exists $x \in [0,1]$ such that $f(x) = x$; such an x is called a* fixed point *for f.*

Proof. Because f is continuous, so is $g(x) := x - f(x)$. If $f(0) = 0$ or $f(1) = 1$, then we are done. If $f(0) > 0$ and $f(1) < 1$, then $g(0) = 0 - f(0) < 0$ and $g(1) = 1 - f(1) > 0$, so the IVT applies to g on $[0,1]$ and gives $x \in (0,1)$ such that $g(x) = 0$. But this means that $x - f(x) = 0$, so $f(x) = x$. □

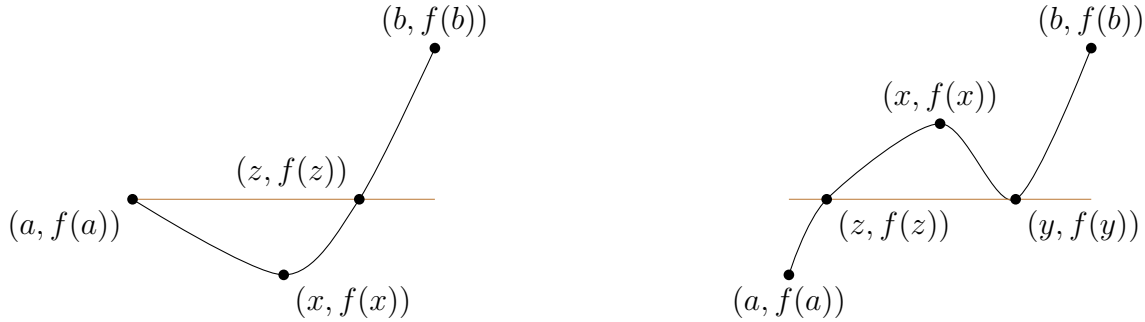
Note that the hypothesis of the preceding theorem requires that $0 \le f(x) \le 1$ for all $0 \le x \le 1$, so the theorem only applies when the image of the unit interval $[0,1]$ lies inside the unit interval.

Theorem 11.6. *If $f\colon [a,b] \to \mathbb{R}$ is 1-1 and continuous, then it is either an increasing function ($x < y$ implies $f(x) < f(y)$) or a decreasing function ($x < y$ implies $f(x) > f(y)$). In either case, its range is the closed interval I whose endpoints are $f(a)$ and $f(b)$, and the inverse function $f^{-1}\colon I \to [a,b]$ is continuous as well.*

Proof. Since f is 1-1, we either have $f(a) < f(b)$ or $f(a) > f(b)$. We give the proof in the case $f(a) < f(b)$, when f will turn out to be increasing; the proof for $f(a) > f(b)$, when f ends up being decreasing, is completely analogous.

First we prove that for every $x \in (a,b)$, we have $f(x) > f(a)$. We prove this by contradiction; we assume that there is some $x \in (a,b)$ for which $f(x) \le f(a)$, then we derive a contradiction from this, and conclude that our assumption must have been false.

Indeed, if there is $x \in (a,b)$ such that $f(x) \le f(a)$, then since f is 1-1 we have $f(x) < f(a) < f(b)$, as in the left-hand picture below. By applying the IVT to f on the interval $[x,b]$, we conclude that there is $z \in [x,b]$ such that $f(z) = f(a)$. Since $a < x$, we have $z \ne x$, which contradicts the fact that f is 1-1. Thus we have proved that $f(x) > f(a)$ for all $x \in (a,b)$.



A similar argument shows that $f(x) < f(b)$ for all $x \in (a,b)$, and we conclude that

$$f(a) < f(x) < f(b) \text{ for all } x \in (a,b). \tag{11.3}$$

In other words, range$(f) \subset [f(a), f(b)] =: I$. In fact, given any $y \in [f(a), f(b)]$, the IVT implies that there is $x \in [a, b]$ such that $f(x) = y$, and we conclude that range$(f) = I$, so that $f\colon [a, b] \to I$ is a bijection.

Now we must prove that f is increasing and that f^{-1} is continuous. To prove that it is increasing, we again proceed by contradiction. Suppose that it is *not* increasing; then there exist $x < y$ such that $f(x) > f(y)$. Because both $f(x)$ and $f(y)$ lie in I, we must have $f(a) < f(y) < f(x)$ as shown in the right-hand picture above. Applying the IVT to f on $[a, x]$ gives $z \in [a, x]$ such that $f(z) = f(y)$. Since $z \neq y$, this contradicts the assumption that f is 1-1. We conclude that f is increasing.

Finally, we show that $f^{-1}\colon I \to [a, b]$ is continuous. Given $c \in I$ and $\epsilon > 0$, let $x = f^{-1}(c)$ and consider the interval $(x - \epsilon, x + \epsilon)$. By the first part of the proof above, we see that

$$\{f(z) : z \in (x - \epsilon, x + \epsilon)\} = (f(x - \epsilon), f(x + \epsilon)). \tag{11.4}$$

Let $\delta_1 = c - f(x - \epsilon)$ and $\delta_2 = f(x + \epsilon) - c$; note that $\delta_1, \delta_2 > 0$ because f is increasing. Let $\delta = \min(\delta_1, \delta_2)$; then for every y with $0 < |y - c| < \delta$, we have

$$y \in (c - \delta, c + \delta) \subset (f(x - \epsilon), f(x + \epsilon)),$$

and by (11.4) we have $f^{-1}(y) \in (x - \epsilon, x + \epsilon)$. Rewriting this we get $|f^{-1}(y) - f^{-1}(c)| = |f^{-1}(y) - x| < \epsilon$, so f^{-1} is continuous at c. □

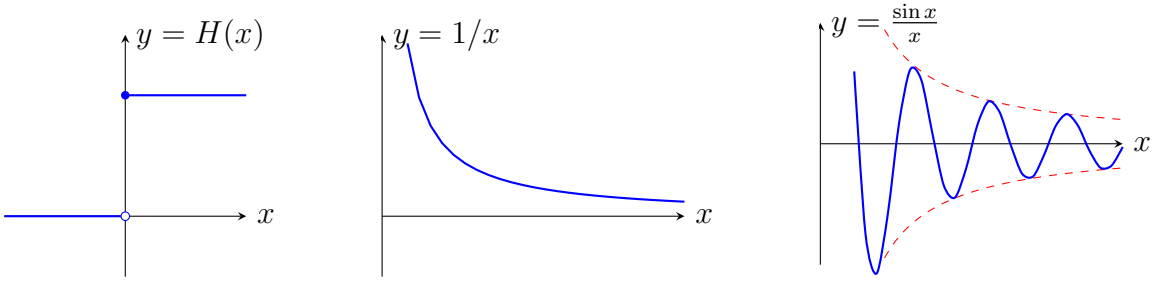
As a consequence of Theorem 11.6, we see that inverse trigonometric functions and logarithmic functions are continuous.

11.3. Limits at infinity

In addition to limits of a function f as x approaches a point $a \in \mathbb{R}$, we sometimes want to talk about "limits at infinity", where x approaches ∞ or $-\infty$. Informally, we say that $\lim_{x\to\infty} f(x) = L$ if $f(x)$ can be made arbitrarily close to L by taking x to be sufficiently large. Formally, we have the following definition, which should be compared to Definition 6.6 for the limit of a sequence.

Definition 11.7. We say that $\lim_{x\to\infty} f(x) = L$ if for every $\epsilon > 0$ there exists $R > 0$ such that for all $x > R$, we have $|f(x) - L| < \epsilon$.

There is an analogous definition for $\lim_{x\to-\infty}$, replacing $x > R$ with $x < -R$. If $\lim_{x\to\infty} f(x) = L$ or $\lim_{x\to-\infty} f(x) = L$, we say that the line $y = L$ is a *horizontal asymptote* of the graph of $y = f(x)$.



The three pictures above illustrate various ways that a function can approach a horizontal asymptote. In the first example, the Heaviside function from Example 9.2 has horizontal asymptotes at $y = 0$ and $y = 1$, and the graph eventually coincides with these

asymptotes. In the second example, the graph of $y = 1/x$ has a horizontal asymptote at $y = 0$ because $\lim_{x\to\infty} 1/x = 0$, and the graph never touches the asymptote.

In the third example, the function $\frac{\sin x}{x}$ satisfies the inequalities $-\frac{1}{x} \le \frac{\sin x}{x} \le \frac{1}{x}$ for all $x > 0$, and since $\lim_{x\to\infty} -\frac{1}{x} = \lim_{x\to\infty} \frac{1}{x} = 0$, the Squeeze Theorem implies that $\lim_{x\to\infty} \frac{\sin x}{x} = 0$, so the graph has a horizontal asymptote at $y = 0$, which it crosses infinitely often.

Example 11.8. The inverse tangent function $y = \tan^{-1} x$ has two horizontal asymptotes, one at $y = -\frac{\pi}{2}$ and one at $y = \frac{\pi}{2}$.

The limit laws work for limits at infinity as well, although we need to be careful with Laws 8–10; these need to be rewritten as

$$\lim_{x\to\infty} x = \infty, \quad \lim_{x\to\infty} \sqrt[n]{x} = \infty, \quad \lim_{x\to\infty} x^n = \infty. \tag{11.5}$$

These use the following definition.

Definition 11.9. Given a function f that is defined for all sufficiently large x, we say that $\lim_{x\to\infty} f(x) = \infty$ if for every $Y > 0$ (no matter how large) there exists $X > 0$ such that every $x > X$ has $f(x) > Y$.

We define limits involving $-\infty$ in a similar way.

Exercise 11.10. Prove (11.5) using Definition 11.9.

Theorem 11.11. *If* $\lim_{x\to\infty} f(x) = \infty$*, then* $\lim_{x\to\infty} \frac{1}{f(x)} = 0$.

Proof. Given any $\epsilon > 0$, it follows from Definition 11.9 that there is $X > 0$ such that for all $x > X$, we have $f(x) > 1/\epsilon$. This implies that $0 < 1/f(x) < \epsilon$, and since $\epsilon > 0$ was arbitrary this shows that $\lim_{x\to\infty} \frac{1}{f(x)} = 0$. □

Remark 11.12. It is tempting to think of Theorem 11.11 as just another version of Limit Law 5 by writing

$$\text{“} \lim_{x\to\infty} \frac{1}{f(x)} = \frac{1}{\lim_{x\to\infty} f(x)} = \frac{1}{\infty} = 0.\text{”}$$

However, this is not a correct application of Law 5, because “$1/\infty$” is not a well-defined expression; ∞ is not a real number and cannot actually be divided, multiplied, added, subtracted, etc. While it is true that this informal computation gives the correct answer in this instance, it should be thought of as a way of remembering Theorem 11.11 rather than as a legitimate computation. It is worth keeping in mind the following examples where naive computations along these lines lead to trouble.

(1) When $f(x) = \frac{\sin x}{x}$, we saw above that $\lim_{x\to\infty} f(x) = 0$, and it would be natural to write “$\lim_{x\to\infty} \frac{1}{f(x)} = 1/\lim_{x\to\infty} f(x) = 1/0 = \infty$”; however, **this is incorrect**. A moment's thought reveals that $\frac{1}{f(x)} = \frac{x}{\sin x}$ alternates between positive and negative values as x grows, changing sign at each vertical asymptote $x = n\pi$ ($n \in \mathbb{N}$), and so cannot go to ∞.

(2) From (11.5) we see that $\lim_{x\to\infty} x^2 = \infty$ and $\lim_{x\to\infty} x = \infty$. It is not hard to show that $\lim_{x\to\infty}(x^2 - x) = \infty$ and $\lim_{x\to\infty}(x - x^2) = -\infty$, so that any attempt to use some version of Limit Law 2 and make sense of “$\infty - \infty$” is doomed to

failure. (We will return to limits with such "indeterminate forms" later, when we study l'Hospital's rule.)

Exercise 11.13. Prove that the first example in the previous remark can be fixed by adding an extra assumption: if $\lim_{x\to\infty} f(x) = 0$ and if $f(x) > 0$ for all x, then $\lim_{x\to\infty} f(x) = \infty$.

The following exercise gives a version of Limit Laws 1, 3, and 4 for infinite limits.

Exercise 11.14. Prove that if $\lim_{x\to\infty} f(x) = \infty$ and $\lim_{x\to\infty} g(x) = \infty$, then

$$\lim_{x\to\infty} (f(x) + g(x)) = \infty \text{ and } \lim_{x\to\infty} (f(x)g(x)) = \infty.$$

Also prove that $\lim_{x\to\infty} cf(x) = \infty$ for all $c > 0$, and $\lim_{x\to\infty} cf(x) = -\infty$ for all $c < 0$.

One important consequence of Theorem 11.11 is the following.

Corollary 11.15. *For every $r > 0$, we have $\lim_{x\to\infty} \frac{1}{x^r} = 0$.*

Proof. By Theorem 11.11, it suffices to prove that $\lim_{x\to\infty} x^r = \infty$. Choosing $n \in \mathbb{N}$ sufficiently large that $\frac{1}{n} < r$, we have $x^r \geq x^{1/n} = \sqrt[n]{x}$, and $\lim_{x\to\infty} \sqrt[n]{x} = \infty$ by (11.5), so Theorem 8.6 (or rather, the equivalent theorem for limits at infinity) gives $\lim_{x\to\infty} x^r = \infty$. □

We can use Corollary 11.15 to evaluate the limit at infinity of any rational function.

Example 11.16.

$$\begin{aligned}
\lim_{x\to\infty} \frac{2x^2 + x - 5}{3x^2 - 2x + 1} &= \lim_{x\to\infty} \frac{2 + \frac{1}{x} - \frac{5}{x^2}}{3 - \frac{2}{x} + \frac{1}{x^2}} && \text{dividing top and bottom by } x^2 \\
&= \frac{2 + \lim_{x\to\infty} \frac{1}{x} - 5 \lim_{x\to\infty} \frac{1}{x^2}}{3 - 2\lim_{x\to\infty} \frac{1}{x} + \lim_{x\to\infty} \frac{1}{x^2}} && \text{by Limit Laws 5, 1, and 3} \\
&= \frac{2 + 0 - 5 \cdot 0}{3 - 2 \cdot 0 + 0} = \frac{2}{3} && \text{by Corollary 11.15.}
\end{aligned}$$

More complicated limits can sometimes be evaluated by using a little more algebraic manipulation. The following example requires our old trick of multiplying by the conjugate when we see an expression that involves adding or subtracting a square root.

Example 11.17.

$$\begin{aligned}
\lim_{x\to\infty} (\sqrt{x^2+1} - x) &= \lim_{x\to\infty} \frac{(\sqrt{x^2+1} - x)(\sqrt{x^2+1} + x)}{\sqrt{x^2+1} + x} = \lim_{x\to\infty} \frac{(x^2+1) - x^2}{\sqrt{x^2+1} + x} \\
&= \lim_{x\to\infty} \frac{1}{\sqrt{x^2+1} + x}.
\end{aligned}$$

From Exercise 11.14 we see that $\lim_{x\to\infty}(\sqrt{x^2+1}+x) = \infty$, and so Theorem 11.11 gives $\lim_{x\to\infty}(\sqrt{x^2+1} - x) = 0$.

Sometimes limits at finite values can be rephrased in terms of limits at infinity.

Example 11.18. Consider $\lim_{t\to 0^+}\left(\sqrt{\frac{1}{t}+1}-\frac{1}{\sqrt{t}}\right)$. Writing $x=1/\sqrt{t}$ so that $1/t=x^2$, we see that $x\to\infty$ as $t\to 0^+$, and thus

$$\lim_{t\to 0^+}\left(\sqrt{\frac{1}{t}+1}-\frac{1}{\sqrt{t}}\right)=\lim_{x\to\infty}(\sqrt{x^2+1}-x)=0.$$

Part II. Derivatives

Lecture 12 Derivatives

This lecture corresponds to §2.7 in Stewart and Chapter 9 in Spivak.

Recall our discussion from Lecture 4 about finding the tangent line to the graph of a function $y = f(x)$ at the point $(a, f(a))$, for $x \approx a$ we considered the *secant line* through the points $(a, f(a))$ and $(x, f(x))$, which has slope $\frac{f(x)-f(a)}{x-a}$, and then took a limit as $x \to a$ to obtain the slope of the tangent line. Equivalently, we can write $x = a + h$ and obtain the slope of the tangent line as $\lim_{h\to 0} \frac{f(a+h)-f(a)}{h}$. This procedure is fundamental to the remainder of the course, and we formalize it with the following definition.

Definition 12.1. The *derivative* of a function f at a point a in the domain of f is

$$f'(a) := \lim_{x\to a} \frac{f(x) - f(a)}{x - a} = \lim_{h\to 0} \frac{f(a + h) - f(a)}{h}, \tag{12.1}$$

if the limit exists. (The notation f' is pronounced "f prime".)

We call the function f *differentiable* at a if the limit in (12.1) exists, and *nondifferentiable* if it does not. We say that f is *differentiable on the interval* (a, b) if it is differentiable at every $c \in (a, b)$.

Remark 12.2. The two limits on the right-hand side of (12.1) are the same by Exercise 6.2, and either can be used as the definition of derivative.

Example 12.3. The function $f(x) = |x|$ is not differentiable at 0, because

$$\frac{f(x) - f(0)}{x - 0} = \frac{|x|}{x} = \begin{cases} -1 & \text{if } x < 0, \\ 1 & \text{if } x > 0, \end{cases}$$

so $\lim_{x\to 0^-} \frac{f(x)-f(0)}{x-0} = -1$ and $\lim_{x\to 0^+} \frac{f(x)-f(0)}{x-0} = 1$. Since the one-sided limits are different, the two-sided limit does not exist.

The derivative of a function has various interpretations in different applications.

- The tangent line to the curve $y = f(x)$ at the point $(a, f(a))$ has slope $f'(a)$. Thus the equation of this tangent line is $\frac{y-f(a)}{x-a} = f'(a)$. This can also be written as $y - f(a) = f'(a)(x - a)$, or $y = f(a) + f'(a)(x - a)$. In this case the *difference quotients* $\frac{f(x)-f(a)}{x-a}$ represent the slopes of the secant lines.
- A *linear function*[18] is a function $g(x) = mx + b$, where m, b are real numbers (the slope and the y-intercept, respectively); equivalently, a linear function is a polynomial with degree 1. The function $g(x) = f(a) + f'(a)(x - a)$ is the linear function that best approximates the function f near $x = a$; one way to think of this property of being the *best linear approximation* is that as we zoom in closer

[18]Technically we should call this an *affine function* and reserve the term *linear* for the case when $b = 0$ so that $g(x) = mx$, but here we will continue to use the terminology that you are probably more familiar with from previous math courses, which reflects the fact that the graph of $y = g(x)$ is a line.

and closer to the point $(a, f(a))$, the graph of $y = f(x)$ looks more and more like the tangent line $y = f(a) + f'(a)(x - a)$.

- If $f(t)$ represents the position of an object at time t – for example, $f(t)$ may represent the height of a ball that is thrown straight up into the air, or the total distance from an observer to a car traveling along a straight road – then $f'(t)$ represents the *instantaneous velocity* of the object at time t. In this case the difference quotient $\frac{f(t+h)-f(t)}{h}$ represents the average velocity of the object between time t and time $t + h$.
- More generally, if $f(t)$ represents any quantity that changes with time, then $f'(t)$ represents the instantaneous rate of change at time t, and the difference quotient represents the average rate of change from t to $t+h$. For example, $f(t)$ might represent the number of bacteria in a petri dish at time t, and then $f'(t)$ represents the rate of growth of the population.[19]
- If $f(x)$ represents the cost of producing x units of something, then $\frac{f(x+h)-f(x)}{h}$ represents the extra cost incurred by producing h more units, assuming x units were already produced. The limit $f'(x)$ is called the *marginal cost* and represents the rate at which the cost changes per extra unit when x units are being produced.

There are many other applications and interpretations besides the ones listed above. We will return to applications later. For the time being we compute a few examples to get a feel for the process.

Example 12.4. If $f(x) = c$ is a constant function, then $f'(x) = \lim_{h\to 0} \frac{c-c}{h} = 0$ at every x. Thus *a constant function has vanishing derivative*. Later we will see that the converse is true as well provided f is differentiable everywhere.

Example 12.5. More generally, if $f(x) = mx + b$ for some constants $m, b \in \mathbb{R}$, then

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{m(x+h) + b - (mx + b)}{h} = \lim_{h\to 0} \frac{mh}{h} = m.$$

This is consistent with our interpretation of $f'(x)$ as the slope of the tangent line, since in this example the graph of f itself is a straight line with slope m.

Example 12.6. Consider the function $f(x) = x^2$. The derivative of f at a is

$$\begin{aligned} f'(a) &= \lim_{h\to 0} \frac{(a+h)^2 - a^2}{h} = \lim_{h\to 0} \frac{a^2 + 2ah + h^2 - a^2}{h} \\ &= \lim_{h\to 0} \frac{2ah + h^2}{h} = \lim_{h\to 0} (2a + h) = 2a. \end{aligned}$$

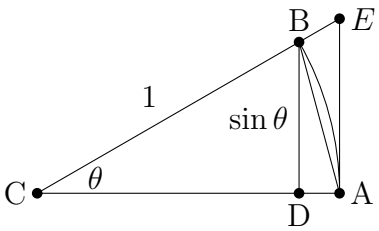
Suppose we wish to find the tangent line to the parabola $y = x^2$ at a the point $(3, 9)$. We have $f'(3) = 6$, so the tangent line has equation $y = 9 + 6(x - 3) = 9 + 6x - 18 = 6x - 9$.

[19] Strictly speaking, the number of bacteria should be an integer, so $f(t)$ would need to be either constant or discontinuous, since otherwise the intermediate value theorem would imply that at some point in time it takes a non-integer value. But it is useful to pretend that the population can be an arbitrary real number, so that we can use the tools of calculus. If the population is very large, then this fiction is generally not too disruptive.

Example 12.7. The function $f(x) = \frac{1}{x}$ has derivative

$$f'(x) = \lim_{h\to 0} \frac{\frac{1}{x+h} - \frac{1}{x}}{h} = \lim_{h\to 0} \frac{1}{h} \cdot \frac{h-(x+h)}{x(x+h)} = \lim_{h\to 0} \frac{-h}{hx(x+h)} = \lim_{h\to 0} \frac{-1}{x(x+h)} = -\frac{1}{x^2}.$$

The *normal line* to a curve $y = f(x)$ at the point $(a, f(a))$ is the line through $(a, f(a))$ that is perpendicular to the tangent line at that point. Since the tangent line has slope $f'(a)$, the normal line has slope $-\frac{1}{f'(a)}$. Suppose we wish to find the normal line to the curve $y = \frac{1}{x}$ at the point $(2, \frac{1}{2})$. The derivative is $f'(x) = -\frac{1}{4}$, so the normal line has slope 4, and the equation of the normal line is $y = \frac{1}{2} + 4(x-2)$.



Example 12.8. The derivative of the function $f(x) = \sin x$ at the point $x = 0$, if it exists, is given by

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

To compute this limit, we use the figure shown above and observe that

$$BD = \sin\theta, \quad AE = \tan\theta, \quad CB = CA = 1.$$

The triangles CBA and CEA have areas $\frac{1}{2}\sin\theta$ and $\frac{1}{2}\tan\theta$, respectively, while the circular wedge with vertices C, B, A has area $\frac{1}{2}\theta$; this wedge contains the triangle CBA and is contained in the triangle CEA, so we have

$$\sin\theta < \theta < \tan\theta$$

for all $\theta \in (0, \frac{\pi}{2})$. The first inequality gives $\frac{\sin\theta}{\theta} < 1$, and the second gives $\cos\theta < \frac{\sin\theta}{\theta}$. Putting these together gives

$$\cos\theta < \frac{\sin\theta}{\theta} < 1.$$

The left-hand and right-hand functions converge to 1 as $\theta \to 0^+$, and thus by the Squeeze Theorem we have $\lim_{\theta\to 0^+} \frac{\sin\theta}{\theta} = 1$. The sine function is even, so for $x < 0$ we have $\frac{\sin x}{x} = \frac{\sin|x|}{|x|}$, and we conclude that

$$f'(0) = \lim_{x\to 0} \frac{\sin x}{x} = 1.$$

Later we will use this to find the derivative of $f(x) = \sin x$ at any real number, not just 0.

Lecture 13 Derivative as a function

This lecture corresponds to §2.8 in Stewart and Chapter 9 in Spivak.

13.1. Domain of the derivative

Suppose the function f is differentiable on the interval (a, b); that is, the derivative $f'(x) = \lim_{h\to 0} \frac{1}{h}(f(x+h) - f(x))$ exists for every $x \in (a, b)$. Then we can interpret $f' \colon (a, b) \to \mathbb{R}$ as a function in its own right.

Remark 13.1. The domain of f' may be smaller than the domain of f. Indeed, consider the function $f(x) = \sqrt{x}$, whose domain is $[0, \infty)$. Then we have

$$\begin{aligned} f'(x) &= \lim_{h\to 0} \frac{\sqrt{x+h} - \sqrt{x}}{h} = \lim_{h\to 0} \frac{(x+h) - x}{h(\sqrt{x+h} + \sqrt{x})} = \lim_{h\to 0} \frac{h}{h(\sqrt{x+h} + \sqrt{x})} \\ &= \lim_{h\to 0} \frac{1}{\sqrt{x+h} + \sqrt{x}} = \frac{1}{2\sqrt{x}}, \end{aligned}$$

and we see that the domain of f' is $(0, \infty)$.[20]

We identify some of the common ways that differentiability can fail.

(1) If f has a discontinuity at a, then by Theorem 13.3 it is not differentiable there.
(2) If the one-sided limits in the definition of derivative exist at a but are not equal, then f is not differentiable at a, and its graph has a 'corner' there; this is what occurs for $f(x) = |x|$. In particular, this shows that continuity does not imply differentiability.
(3) If $\lim_{x\to a} \frac{f(x)-f(a)}{x-a} = \infty$ or $-\infty$, then the graph of f has "infinite slope" at a, and f is not differentiable there. We saw this occur with $f(x) = \sqrt{x}$.

This list is not comprehensive.

Exercise 13.2. Prove that the function $x \sin \frac{1}{x}$ from Example 9.19(2) is nondifferentiable at 0, but does not fit into any of the categories listed above.

It is worth noting that the domain of f' – that is, the set of points at which f is differentiable – must be contained in the set of points at which f is continuous.

Theorem 13.3. *If f is differentiable at a, then f is continuous at a.*

Proof. Since $f'(a)$ exists, we can use Limit Law 4 to conclude that

$$\begin{aligned} \lim_{x\to a}(f(x) - f(a)) &= \lim_{x\to a} \Big(\frac{f(x) - f(a)}{x - a}(x - a) \Big) \\ &= \Big(\lim_{x\to a} \frac{f(x) - f(a)}{x - a} \Big) \Big(\lim_{x\to a}(x - a) \Big) = f'(a) \cdot 0 = 0. \end{aligned}$$

Then we have

$$\lim_{x\to a} f(x) = \lim_{x\to a}(f(a) + f(x) - f(a)) = f(a) + \lim_{x\to a}(f(x) - f(a)) = f(a) + 0 = f(a),$$

[20] In addition to the fact that the formula for f' only makes sense when $x > 0$, it is also standard practice to only consider the derivative as defined at points where the *two-sided* limit makes sense; when f is only defined on one side of a point a, we usually talk only about one-sided derivatives at a.

which proves that f is continuous at a. □

As the absolute value function shows, the converse of Theorem 13.3 is false. So we summarize the situation like this: ***Every differentiable function is continuous, but not every continuous function is differentiable.***

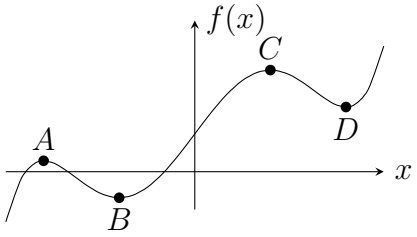
13.2. Connection between properties of f and f'

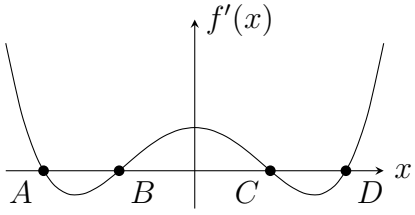
13.2.1. *Monotonicity properties*

It is useful to relate properties of the function f' to properties of the function f. Intuitively, since $f'(x)$ is the limit of the ratios $\frac{f(y)-f(x)}{y-x}$ as $y \to x$, we expect that a positive derivative, $f'(x) > 0$, corresponds to an increasing function, where $x < y$ implies $y - x > 0$ implies $f(y) - f(x) \approx f'(x)(y - x) > 0$, so $f(y) > f(x)$. We will make this precise when we study the Mean Value Theorem later on; for now, we merely observe that $f' > 0$ corresponds to regions where f is increasing, and $f' < 0$ corresponds to regions where f is decreasing.

Example 13.4. When $f(x) = x^2$, we saw in Example 12.6 that $f'(x) = 2x$. Thus $f' > 0$ on the interval $(0, \infty)$, where f is increasing, and $f' < 0$ on the interval $(-\infty, 0)$, where f is decreasing.

This qualitative relationship between f and f' is useful even when we do not have formulas for the functions. For example, in the pictures below we see that f' is positive between B and C, where the function f is increasing, and negative between A and B, and between C and D, where the function f is decreasing.





13.2.2. *Symmetry properties*

Proposition 13.5. *If f is an even function ($f(-x) = f(x)$ for all x) then f' is odd ($f'(-x) = -f'(x)$ for all x where f' exists), and vice versa.*

Proof. If f is even and is differentiable at x, then we have

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

where the first equality is the definition of derivative, and the second uses the fact that f is even. Writing $k = -h$ we can rewrite the last ratio as $\frac{f(x+k)-f(x)}{-k}$, and obtain

$$f'(-x) = \lim_{k\to 0} \frac{f(x+k) - f(x)}{-k} = -\lim_{k\to 0} \frac{f(x+k) - f(x)}{k} = -f'(x),$$

which proves that f' is odd. Conversely, if f is odd then we can make a similar computation and show that f' is even. □

Example 13.6. The function $f(x) = x^2$ is even, while its derivative $f'(x) = 2x$ (recall Example 12.6) is odd.

13.3. Notation

We will usually write the derivative of f at x as $f'(x)$, but you should be aware of other notations that are in common use. One of these, which will appear in this class, is $\frac{d}{dx}f(x)$, or $\frac{df}{dx}$. If we want to stress that this derivative is evaluated at a particular point $x = a$, we may write $\frac{d}{dx}f(x)|_{x=a}$. It is important to note that although $\frac{df}{dx}$ looks like a fraction (and one might be tempted to do things like 'cancel the ds'), it is not a fraction, and the symbols df and dx have no independent meaning. This notation comes from the idea that $\frac{df}{dx}$ is the limit of the ratios $\frac{\Delta f}{\Delta x}$, where Δx represents the change in x, and $\Delta f = f(x + \Delta x) - f(x)$ represents the corresponding change in the value of f. These ratios are of course the difference quotients $\frac{f(x+h)-f(x)}{h}$.

Remark 13.7. Rather than thinking of $\frac{df}{dx}$ as representing a fraction, it is better to think about $\frac{d}{dx}$ as representing the *operation* of differentiation, so that $\frac{d}{dx}f$ is the function that results from applying the operation of differentiation to the function f. From this point of view, $\frac{d}{dx}$ is the *differentiation operator*, and is a function whose inputs and outputs are themselves functions. We will not dwell on this point of view in this course, but it is an important one in more advanced courses.

The notation $\dot{f}$ is sometimes used for the derivative of f; this is most common when f is a function of time t, so that $\dot{f}$ represents a derivative with respect to time. In this course we will generally stick to the notation f', though. When the variable is t instead of x, we will also use the notation $\frac{df}{dt}$ to make it clear what we are differentiating with respect to, and similarly if some other variable is used instead of x or t.

You may sometimes see a *dependent* variable y used to represent a function of an *independent* variable x, and then the notation $\frac{dy}{dx}$, y', or $\dot{y}$ is sometimes used for the derivative.

Finally, you may even see the notation Df for the derivative. This is a convenient notation if we want to talk about the *one-sided* derivatives $D^+f(x) = \lim_{h\to 0^+} \frac{f(x+h)-f(x)}{h}$ and $D^-f(x) = \lim_{h\to 0^-} \frac{f(x+h)-f(x)}{h}$. By Theorem 8.3, we see that f is differentiable at x if and only if $D^+f(x)$ and $D^-f(x)$ both exist and take the same value. Example 12.3 shows that the absolute value function $f(x) = |x|$ has $D^-f(0) = -1$ and $D^+f(0) = 1$.

13.4. Higher derivatives

If f is differentiable on (a, b), then f' is a function on (a, b) in its own right, and it is reasonable to ask whether this function is itself differentiable. If it is, then we refer to its derivative

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

as the *second derivative* of f, and denote it by $f''(x)$. This last notation is often pronounced "f double prime".

Another common notation for the second derivative of f with respect to x is

$$f''(x) = \frac{d^2}{dx^2} f(x) = \frac{d^2 f}{dx^2}(x).$$

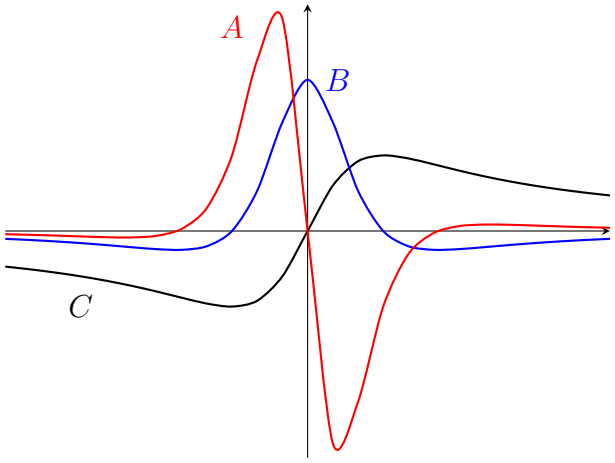
It should be noted that the appearance of a superscript "2" here is *not* used to indicated that any quantity is squared, or multiplied by itself; rather, it indicates that the operation of differentiation is done twice. One might also see the notation $\ddot{f}$, especially when f is a function of time, although we will generally stick to the notation f''.

Example 13.8. If $f(t)$ represents the height of an object at time t, then $f'(t)$ represents its velocity at time t, and $f''(t)$ represents the rate at which its velocity is changing; in other words, its *acceleration*.

One can go further and define the third derivative $f'''(x)$ to be the derivative of $f''(x)$, and so on. In general we use the notation $f^{(n)}(x)$ or $\frac{d^n}{dx^n} f(x)$ of $\frac{d^n f}{dx^n}$ for the nth derivative of f at x, when it exists; this can be defined iteratively by

$$f^{(n)}(x) = \frac{d^n}{dx^n} f(x) := \frac{d}{dx} f^{(n-1)}(x).$$

We will be most interested in first- and second-order derivatives f' and f'', but higher-order derivatives will play a role next semester when we discuss *Taylor polynomials and series*.



Example 13.9. The picture above shows the graphs of f, f', and f'' for some function f. Which curve represents which function?

To answer this question, look at the intervals on which the curves labeled A, B, C are positive, and see if any of the other curves are increasing on these intervals. We see that the interval on which B is positive is the same as the interval on which C is increasing, so B should be the graph of the derivative of the function whose graph is C. Similarly, A is positive on the intervals where B is increasing, and negative on the intervals where B is decreasing. So we conclude that C is the graph of f, B is the graph of f', and A is the graph of f''.

Something to think about: is there a connection between the sign of f'' (as shown in the curve A) and the shape of the graph of f (as shown in the curve C)? We will return to this later when we discuss *convexity*.

Lecture 14 Derivatives of polynomials and exponentials

This lecture corresponds to §3.1 in Stewart and Chapter 10 in Spivak

14.1. Power rule and polynomial functions

Now we start developing systematic rules to evaluate derivatives of broad classes of functions. When we did a similar procedure for limits, we started with polynomial functions, and we will do the same thing here.

Start by recalling that Exercises 12.4, 12.5, and 12.6 gave us

$$\frac{d}{dx}c = 0, \qquad \frac{d}{dx}x = 1, \qquad \frac{d}{dx}x^2 = 2x.$$

What about the derivative of x^n for other values of n? Start with $n = 3$:

$$\begin{aligned}\frac{d}{dx}x^3 &= \lim_{h\to 0}\frac{(x+h)^3 - x^3}{h} = \lim_{h\to 0}\frac{x^3 + 3x^2h + 3xh^2 + h^3 - x^3}{h}\\ &= \lim_{h\to 0}\frac{3x^2h + 3xh^2 + h^3}{h} = \lim_{h\to 0}(3x^2 + 3xh + h^2) = 3x^2.\end{aligned}$$

A pattern is beginning to emerge, which is confirmed by the following theorem.

Theorem 14.1 (Power rule)**.** *If $f(x) = x^n$ for some $n \in \mathbb{N}$, then $f'(x) = nx^{n-1}$.*

First proof of the power rule. Start by recalling the *Binomial Theorem*, which says that given $x, h \in \mathbb{R}$ and $n \in \mathbb{N}$, we have

$$(x+h)^n = x^n + nx^{n-1}h + \frac{n(n-1)}{2}x^{n-2}h^2 + \frac{n(n-1)(n-2)}{3\cdot 2}x^{n-3}h^3 + \cdots + nh^{n-1} + h^n,$$

where the coefficient on the term $x^{n-k}h^k$ is given by

$$\binom{n}{k} := \frac{n(n-1)(n-2)\cdots(n-k+1)}{k(k-1)\cdots 2} = \frac{n!}{k!(n-k)!}$$

We can rewrite the expansion in the Binomial Theorem as

$$(x+h)^n = \sum_{k=0}^{n}\binom{n}{k}x^{n-k}h^k = x^n + nx^{n-1}h + \sum_{k=2}^{n}\binom{n}{k}x^{n-k}h^k, \tag{14.1}$$

where in the last expression we have separated out the first two terms for reasons that will become clear momentarily.

Exercise 14.2. Prove the Binomial Theorem (72.7) using induction on n.

Using (72.7) we can compute $f'(x)$ for $f(x) = x^n$ as follows:

$$\begin{aligned}f'(x) &= \lim_{h\to 0}\frac{1}{h}((x+h)^n - x^n) = \lim_{h\to 0}\frac{1}{h}\Big(x^n + nx^{n-1}h + \Big(\sum_{k=2}^{n}\binom{n}{k}x^{n-k}h^k\Big) - x^n\Big)\\ &= \lim_{h\to 0}\frac{1}{h}\Big(nx^{n-1}h + \Big(\sum_{k=2}^{n}\binom{n}{k}x^{n-k}h^k\Big)\Big) = nx^{n-1} + \sum_{k=2}^{n}\lim_{h\to 0}\binom{n}{k}x^{n-k}h^{k-1},\end{aligned}$$

where the last equality uses Limit Law 1 for addition. For every $k = 2, 3, \ldots, n$, we have $\lim_{h\to 0}\binom{n}{k}x^{n-k}h^{k-1} = 0$ since $k - 1 \geq 1$, and thus we conclude that $f'(x) = nx^{n-1}$. □

An alternate proof of the power rule uses the following exercise, which generalizes the formula $x^2 - y^2 = (x-y)(x+y)$ for a difference of squares.

Exercise 14.3. Prove that for every $x, y \in \mathbb{R}$ and $n \in \mathbb{N}$, we have

$$y^n - x^n = (y-x)(y^{n-1} + y^{n-2}x + y^{n-3}x^2 + \cdots + yx^{n-2} + x^{n-1}) = (y-x)\sum_{k=1}^{n} y^{n-k}x^{k-1}. \tag{14.2}$$

Second proof of the power rule. Using (14.2), we have

$$\begin{aligned}\frac{d}{dx}x^n &= \lim_{y\to x}\frac{y^n - x^n}{y-x} = \lim_{y\to x}\frac{(y-x)(y^{n-1} + y^{n-2}x + y^{n-3}x^2 + \cdots + yx^{n-2} + x^{n-1})}{y-x} \\ &= \lim_{y\to x}\sum_{k=1}^{n} y^{n-k}x^{k-1} = \sum_{k=1}^{n}\lim_{y\to x} y^{n-k}x^{k-1} = \sum_{k=1}^{n} x^{n-1} = nx^{n-1}. \qquad \square\end{aligned}$$

By a short application of the limit laws, we can find the derivatives of cf and $f \pm g$ if f', g' are known and $c \in \mathbb{R}$.

Theorem 14.4. *If $c \in \mathbb{R}$ and f is differentiable at a, then the function cf is differentiable at a and $(cf)'(a) = c \cdot f'(a)$.*

Proof. Writing $g(x) = cf(x)$, Limit Law 3 gives

$$g'(x) = \lim_{h\to 0}\frac{cf(x+h) - cf(x)}{h} = c\lim_{h\to 0}\frac{f(x+h) - f(x)}{h} = cf'(x). \qquad \square$$

Theorem 14.5. *If f, g are differentiable at a, then so are the functions $f \pm g$, and $(f \pm g)'(a) = f'(a) \pm g'(a)$.*

Proof. Writing $F(x) = f(x) + g(x)$, Limit Law 4 gives

$$\begin{aligned}F'(x) &= \lim_{h\to 0}\frac{f(x+h) + g(x+h) - f(x) - g(x)}{h} \\ &= \lim_{h\to 0}\frac{f(x+h) - f(x)}{h} + \lim_{h\to 0}\frac{g(x+h) - g(x)}{h} = f'(x) + g'(x). \qquad \square\end{aligned}$$

Now we can differentiate any polynomial function.

Example 14.6. If $f(x) = 2 + 3x^2 + 7x^5$, then

$$f'(x) = \frac{d}{dx}2 + \frac{d}{dx}(3x^2) + \frac{d}{dx}(7x^5) = 0 + 3\frac{d}{dx}(x^2) + 7\frac{d}{dx}(x^5) = 6x + 35x^4,$$

where the first equality uses Theorem 14.4, the second uses Theorem 14.5, and the third uses the power rule for differentiation.

In fact the power rule holds more generally. We saw in Example 49.3 that $\frac{d}{dx}\frac{1}{x} = -\frac{1}{x^2}$, and in Remark 13.1 that $\frac{d}{dx}\sqrt{x} = \frac{1}{2\sqrt{x}}$; these can be written as

$$\frac{d}{dx}x^{-1} = -x^{-2} \quad \text{and} \quad \frac{d}{dx}x^{\frac{1}{2}} = -\frac{1}{2}x^{-\frac{1}{2}},$$

which both have the same form as the power rule.

Theorem 14.7. *For every $\beta \in \mathbb{R}$, the function $f(x) = x^\beta$ has $f'(x) = \beta x^{\beta-1}$.*

Proof. Deferred; first we need to study derivatives of exponential functions, logarithmic functions, and the chain rule. □

Using this, we can also differentiate functions that are not polynomials but can be written as sums of power functions.

Example 14.8. If $g(t) = \sqrt{t}(t-1)$, then we can write

$$g'(t) = \frac{d}{dt}\left(t^{3/2} - t^{1/2}\right) = \frac{d}{dt}t^{3/2} - \frac{d}{dt}t^{1/2} = \frac{3}{2}t^{1/2} - \frac{1}{2}t^{-1/2}.$$

14.2. Exponential functions

Fix $a > 0$ and let $f(x) = a^x$. To find the derivative of f at x (if it exists) we write

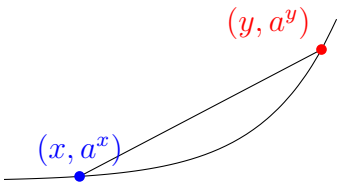
$$f'(x) = \lim_{h\to 0}\frac{a^{x+h} - a^x}{h} = \lim_{h\to 0} a^x \cdot \frac{a^h - 1}{h} = a^x \lim_{h\to 0}\frac{a^h - 1}{h}. \tag{14.3}$$

If this last limit exists, it is $f'(0)$. But why should it exist? Recall that a two-sided limit exists if and only if the one-sided limits both exist and agree, so we might make progress by studying the one-sided limits

$$D^+f(0) = \lim_{h\to 0^+}\frac{a^h - 1}{h} \quad \text{and} \quad D^-f(0) = \lim_{h\to 0^-}\frac{a^h - 1}{h}.$$

Do they exist? Do they agree?

If we accept that the function f has a graph that "bends upwards" as shown in the picture, then the following property seems like it should be true.



Conjecture 14.9. *The slope of the secant line from (x, a^x) to (y, a^y) increases if we move either of x or y to the right, and decreases if we move either of them to the left.*

The picture shows the case $a > 1$; for $0 < a < 1$ we get a decreasing function and the secant lines have negative slopes, but it still looks as though the conjecture should be true. We will prove Conjecture 14.9 later, but the proof drags us into some technicalities that would distract from the main story here, so for the time being let us simply accept the conjecture as true and observe the following consequences.

- For every $h < 0$, the slope of the secant line from (h, a^h) to $(0, 1)$ (which is $\frac{a^h-1}{h}$) is smaller than the slope of the secant line from $(0, 1)$ to $(1, a)$ (which is $a - 1$).
- As h increases to 0 from the left, the quantity $\frac{a^h-1}{h}$ increases.
- Since $\frac{a^h-1}{h}$ increases with h and is bounded above by $a - 1$, the Monotone Convergence Theorem (which we formulated for sequences, but which also holds for functions) implies that $D^-f(0)$ exists, and $D^-f(0)$.
- A similar argument (where now we use the secant line from $(-1, \frac{1}{a})$ to $(0, 1)$ in the first step) shows that $D^+f(0)$ exists and $D^+f(0)$.

The same computation as in (14.3) shows that

$$D^+f(x) = a^x D^+f(0) \quad \text{and} \quad D^-f(x) = a^x D^-f(0) \tag{14.4}$$

for all x; in particular, the left and right derivatives exist at every point. The following two exercises can be completed using Conjecture 14.9.

Exercise 14.10. Prove that $1 - \frac{1}{a} \le D^- f(0) \le D^+ f(0) \le a - 1$.

Exercise 14.11. Prove that given any $x < y$, we have $D^+ f(x) \le \frac{a^y - a^x}{y - x} \le D^- f(y)$.

Now we can prove that $D^- f(0) = D^+ f(0)$ and conclude that f is differentiable at 0 (and hence everywhere).[21] Let $c := D^+ f(0)/D^- f(0)$. We have $c \ge 1$ by Exercise 14.10, and by (14.4) we have $c = D^+ f(x)/D^- f(x)$ for all x. Now given any $n \in \mathbb{N}$, we have

$$\begin{aligned}\frac{D^- f(1)}{D^- f(0)} &= \Big(\frac{D^+ f(0)}{D^- f(0)} \frac{D^- f(\frac{1}{n})}{D^+ f(0)}\Big)\Big(\frac{D^+ f(\frac{1}{n})}{D^- f(\frac{1}{n})} \frac{D^- f(\frac{2}{n})}{D^+ f(\frac{1}{n})}\Big) \cdots \Big(\frac{D^+ f(\frac{n-1}{n})}{D^- f(\frac{n-1}{n})} \frac{D^- f(1)}{D^+ f(\frac{n-1}{n})}\Big) \\ &\ge \frac{D^+ f(0)}{D^- f(0)} \frac{D^+ f(\frac{1}{n})}{D^- f(\frac{1}{n})} \cdots \frac{D^+ f(\frac{n-1}{n})}{D^- f(\frac{n-1}{n})} = c^n,\end{aligned}$$

where the inequality uses Exercise 14.11. It follows that $1 \le c \le (D^- f(1)/D^- f(0))^{1/n}$ for all $n \in \mathbb{N}$, and thus $c = 1$. We conclude that $f'(0)$ exists, and thus we have proved the following (contingent on eventually proving Conjecture 14.9).

Theorem 14.12. *For any $a > 0$, the exponential function $f(x) = a^x$ is differentiable on $\mathbb{R}$ and has derivative given by*

$$f'(x) = a^x \lim_{h \to 0} \frac{a^h - 1}{h} = f'(0) a^x. \tag{14.5}$$

Theorem 14.12 is not entirely satisfying because we still do not have a clear understanding of what the limit $\lim_{h \to 0} \frac{a^h - 1}{h}$ is. How does it depend on a? Can we choose a such that this limit takes a simple value, like 1?

For convenience let us denote this limit by

$$g(a) := \lim_{h \to 0} \frac{a^h - 1}{h}, \tag{14.6}$$

and study the function $g \colon (0, \infty) \to \mathbb{R}$. It follows from Exercise 14.10 that $1 - \frac{1}{a} \le g(a) \le a - 1$. In particular, for $a > 1$ we have $g(a) \ge 1 - \frac{1}{a} > 0$, and for $a < 1$ we have $g(a) \le a - 1 < 0$, so $g(a) \ne 0$ for all $a \ne 1$.

To proceed further we need to think for a moment about how the exponential function $x \mapsto a^x$ changes when we vary the base a. Given $a, b > 0$, consider the functions $f_1(x) = a^x$ and $f_2(x) = b^x$. We can use the relationship $a = b^{\log_b(a)}$ to write

$$a^x = b^{\log_b(a) x} \quad \Rightarrow \quad f_1(x) = f_2(cx) \quad \text{for } c = \log_b(a). \tag{14.7}$$

Geometrically, the graph of f_1 is the result of scaling the graph of f_2 horizontally by a factor of $\frac{1}{c}$, which we expect to result in all lines (including tangent lines) having their slopes scaled by a factor of c. More formally, we can write

$$\begin{aligned} f_1'(x) &= \lim_{h \to 0} \frac{f_1(x + h) - f_1(x)}{h} = \lim_{h \to 0} \frac{f_2(cx + ch) - f_2(cx)}{h} \\ &= \lim_{h \to 0} c \cdot \frac{f_2(cx + ch) - f_2(cx)}{ch} = c f_2'(cx). \end{aligned} \tag{14.8}$$

For the functions in question we have $f_1'(x) = g(a) a^x$ and $f_2'(y) = g(b) b^y$, so this gives

$$g(a) a^x = f_1'(x) = c f_2'(cx) = c g(b) b^{cx} = (\log_b(a)) g(b) a^x,$$

[21] I got this argument from a post by Todd Trimble at nLab.

where the last equality uses the fact that $b^c = a$. Since $a^x > 0$ we conclude that

$$g(a) = (\log_b(a))g(b) \text{ for all } a, b > 0. \tag{14.9}$$

Fixing $b = 2$ (indeed we could use any $b \neq 1$) we have $g(2) \neq 0$, and recall that $\log_2 \colon (0,\infty) \to \mathbb{R}$ is a continuous bijection (since it is the inverse of the continuous bijection $x \mapsto 2^x$), so $g = g(2)\log_2$ is also a continuous bijection from $(0,\infty) \to \mathbb{R}$. This proves the following.

Theorem 14.13. *There is a unique real number $e > 0$ such that $g(e) = \lim_{h\to 0} \frac{e^h - 1}{h} = 1$. Taking $b = e$ in* (14.9) *gives $g(a) = \log_e(a)$ for all $a \in \mathbb{R}$. Writing $\ln(a) = \log_e(a)$ for this* natural logarithm, (14.6) *gives*

$$\ln(a) = \lim_{h\to 0} \frac{a^h - 1}{h}, \tag{14.10}$$

and the formula (14.5) *for the exponential derivative becomes*

$$\frac{d}{dx}a^x = (\ln a)a^x; \quad \textit{in particular, } \frac{d}{dx}e^x = e^x. \tag{14.11}$$

Exercise 14.14. Prove the fact that $\ln(ab) = \ln(a) + \ln(b)$ in two different ways: as a direct consequence of (14.10), and as a consequence of the fact that $e^{x+y} = e^x e^y$.

Exercise 14.15. Use Conjecture 14.9 and the secant lines at $\pm\frac{1}{n}$ (following the same idea as in Exercise 14.10) to prove that

$$n - na^{-\frac{1}{n}} \leq \ln a \leq na^{\frac{1}{n}} - n \text{ for all } n, \tag{14.12}$$

and that both the lower and upper bounds improve when n increases:

$$1 - a^{-1} \leq 2 - 2a^{-\frac{1}{2}} \leq 3 - 3a^{-\frac{1}{3}} \leq \cdots \leq \ln a \leq \cdots \leq 3a^{\frac{1}{3}} - 3 \leq 2a^{\frac{1}{2}} - 2 \leq a - 1.$$

By finding the values of a at which the lower and upper bounds in (14.12) are equal to 1, prove that

$$\left(1 + \frac{1}{n}\right)^n = \left(\frac{n+1}{n}\right)^n \leq e \leq \left(\frac{n}{n-1}\right)^n = \left(1 - \frac{1}{n}\right)^{-n}, \tag{14.13}$$

and that once again the lower and upper bounds improve when n increases:

$$2 \leq \left(\frac{3}{2}\right)^2 \leq \left(\frac{4}{3}\right)^3 \leq \left(\frac{5}{4}\right)^4 \leq \cdots \leq e \leq \cdots \leq \left(\frac{4}{3}\right)^4 \leq \left(\frac{3}{2}\right)^3 \leq 2^2.$$

Prove that the ratio between the upper and lower bounds in (14.13) goes to 1 as $n \to \infty$, and use this fact to deduce that both the lower and upper bounds converge to e. In particular, we have

$$e = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n. \tag{14.14}$$

Using a calculator or computer, use (14.13) to approximate e to within one digit after the decimal.

Remark 14.16. The relationship in (14.8) can be rewritten as

$$\frac{d}{dx}f(cx) = cf'(cx) \tag{14.15}$$

and is true for any differentiable function f and any $c \in \mathbb{R}$. This is the first instance of a more general result, the *chain rule*, that we will come to soon. It is worth noting that while $\frac{d}{dx}f(x)$ and $f'(x)$ mean the same thing, (14.15) shows that $\frac{d}{dx}f(cx)$ and $f'(cx)$ mean *different* things. To make sense of this, observe that $x \mapsto f(cx)$ is the composition of two distinct functions; first we multiply by c, then we apply f. This can be represented by the following diagram:

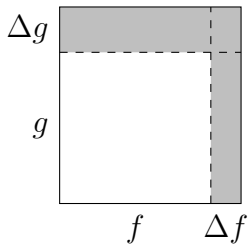
$$x \xmapsto{\cdot c} cx \xmapsto{f} f(cx).$$

In $\frac{d}{dx}f(cx)$, we are measuring the response of the final output to a variation in the **initial input** x. In $f'(cx)$, on the other hand, we are measuring the response of the final output to a variation in the **input of the function** f, which is cx. The two do not give the same result in general, and the relationship between them is given by (14.15).

Lecture 15 Product and quotient rules

This lecture corresponds to §3.2 in Stewart and Chapter 10 in Spivak.

Suppose f, g are differentiable at x. What can we say about $(fg)'(x)$? Does it exist? What is its value?



The first thing to observe is that $(fg)'(x)$ is **not** given by the naive formula $f(x)g(x)$; indeed, if $f(x) = g(x) = x$, then $f'(x) = g'(x) = 1$ and $(fg)(x) = x^2$, so $(fg)'(x) = 2x \neq 1 = f'(x)g'(x)$.

To find the correct formula, we start by imagining a rectangle whose width and height vary with time. Let $f(t)$ be the width and $g(t)$ the height at time t. Then the area at time t is $A(t) = f(t)g(t)$. If we go from time t to time $t + \Delta t$, then the rectangle at the new time has width $f + \Delta f$ and height $g + \Delta g$, as shown in the picture, and the change in the area is

$$\Delta(fg) = \Delta A = (f + \Delta f)(g + \Delta g) - fg = (\Delta f)g + f(\Delta g) + \Delta f \Delta g.$$

Observe that the three terms in this sum correspond to the three shaded rectangles in the picture. We see that

$$\frac{\Delta(fg)}{\Delta t} = \frac{\Delta f}{\Delta t}g(t) + f(t)\frac{\Delta g}{\Delta t} + \frac{\Delta f \Delta g}{\Delta t} \approx f'(t)g(t) + f(t)g'(t) + f'(t)\Delta g$$

taking Δt tending to 0 suggests that $(fg)'(t) = f'(t)g(t) + f(t)g'(t)$. Now we make this formal.

Theorem 15.1 (Product Rule)**.** *If f and g are differentiable at x, then fg is also differentiable at x, and $(fg)'(x) = f'(x)g(x) + f(x)g'(x)$.*

Proof. The derivative of fg at x, if it exists, is given by the limit

$$\begin{aligned}(fg)'(x) &= \lim_{y \to x} \frac{f(y)g(y) - f(x)g(x)}{y - x} \\ &= \lim_{y \to x} \frac{f(y)g(y) - f(x)g(y) + f(x)g(y) - f(x)g(x)}{y - x}\end{aligned}$$

$$= \underbrace{\lim_{y\to x} \frac{f(y)-f(x)}{y-x}}_{\text{I}} \underbrace{\lim_{y\to x} g(y)}_{\text{II}} + f(x) \underbrace{\lim_{y\to x} \frac{g(y)-g(x)}{y-x}}_{\text{III}},$$

where the first line is the definition of derivative, the second line comes by adding and subtracting $f(x)g(y)$, and the third line uses the limit laws for addition and multiplication. In order for this to be valid, we need to verify that the limits in the last line exist. Limits I and III exist and are equal to $f'(x)$ and $g'(x)$, respectively, because we assumed that f and g are differentiable at x. Because g is differentiable at x, it is also continuous at x by Theorem 13.3; thus limit II exists and is equal to $g(x)$. This completes the proof of the theorem. □

Example 15.2. In Example 14.8 we evaluted the derivative of $f(x) = \sqrt{x}(x-1)$ by expanding it as $x^{3/2} - x^{1/2}$ and using the power rule. We can also evaluate it without expanding by using the product rule (and then the power rule):

$$f'(x) = \Big(\frac{d}{dx}\sqrt{x}\Big)(x-1) + \sqrt{x}\frac{d}{dx}(x-1) = \frac{1}{2\sqrt{x}}(x-1) + \sqrt{x}$$
$$= \frac{\sqrt{x}}{2} - \frac{1}{2\sqrt{x}} + \sqrt{x} = \frac{3}{2}\sqrt{x} - \frac{1}{2\sqrt{x}},$$

which agrees with our earlier answer.

The product rule lets us give a third proof of the power rule $\frac{d}{dx}x^n = nx^{n-1}$, by induction this time: the power rule for $n = 1$ is just the observation that $\frac{d}{dx}x = 1$, and if the power rule is true for a given value of n, then the product rule gives

$$\frac{d}{dx}x^{n+1} = \frac{d}{dx}(x^n x) = \Big(\frac{d}{dx}x^n\Big)x + x^n\frac{d}{dx}x = nx^{n-1}\cdot x + x^n\cdot 1 = (n+1)x^n,$$

so the power rule is true for $n+1$. By induction, the power rule holds for all $n \in \mathbb{N}$.

To compute the derivative of a quotient function $f(x)/g(x)$, we start by considering the case when $f(x) = 1$. Suppose that g is differentiable at x and that $g(x) \neq 0$; suppose moreover that $h(x) := 1/g(x)$ is differentiable at x; then we have $(gh)(x) = g(x)h(x) = 1$, and differentiating both sides gives

$$0 = \frac{d}{dx}1 = \frac{d}{dx}(gh)(x) = g'(x)h(x) + g(x)h'(x).$$

Solving for h' gives

$$\frac{d}{dx}\frac{1}{g(x)} = h'(x) = -\frac{g'(x)h(x)}{g(x)} = -\frac{g'(x)}{g(x)^2},$$

where the last equality uses the definition of $h(x)$. This suggests what the formula for the derivative of a reciprocal should be. However, it does **not** prove quite the result that we want; recall Theorem 15.1, where we only needed to assume that f and g were differentiable at x, and then were able to conclude that fg was also differentiable at x. Here we are forced to assume differentiability of $1/g$, when really this ought to be one of the conclusions of the theorem. In order to do this we need to go back to the definition of derivative rather than relying on the product rule.

Theorem 15.3 (Reciprocal rule)**.** *If g is differentiable at x and $g(x) \neq 0$, then $1/g$ is also differentiable at x and $(1/g)'(x) = -g'(x)/g(x)^2$.*

Proof. Use the definition of derivative together with the limit laws:

$$\begin{aligned}\Big(\frac{1}{g}\Big)'(x) &= \lim_{y\to x} \frac{\frac{1}{g(y)} - \frac{1}{g(x)}}{y-x} = \lim_{y\to x} \frac{1}{y-x}\frac{g(x)-g(y)}{g(y)g(x)} \\ &= -\frac{1}{g(x)} \underbrace{\lim_{y\to x} \frac{1}{g(y)}}_{\text{I}} \underbrace{\lim_{y\to x} \frac{g(y)-g(x)}{y-x}}_{\text{II}},\end{aligned}$$

provided limits I and II exist. Since g is differentiable at x, II exists and is equal to $g'(x)$; moreover, Theorem 13.3 implies that g is continuous at x, so I exists and is equal to $1/g(x)$, which proves the theorem. □

The reciprocal rule lets us prove the power rule for negative integers:

$$\frac{d}{dx}(x^{-n}) = \frac{d}{dx}\frac{1}{x^n} = -\frac{\frac{d}{dx}x^n}{(x^n)^2} = -\frac{nx^{n-1}}{x^{2n}} = -nx^{-n-1}.$$

However, we are still awaiting a proof for the case when n is not an integer.

Putting the product rule and reciprocal rule together gives the quotient rule.

Theorem 15.4 (Quotient rule)**.** *If f, g are differentiable at x and $g(x) \neq 0$, then f/g is also differentiable at x and*

$$\Big(\frac{f}{g}\Big)'(x) = \frac{g(x)f'(x) - f(x)g'(x)}{g(x)^2}. \tag{15.1}$$

Proof. The reciprocal rule implies that $1/g$ is differentiable; then the product rule implies that $f/g = f \cdot \frac{1}{g}$ is differentiable. Its derivative can be computed by combining the two formulas given above:

$$\begin{aligned}\Big(\frac{f}{g}\Big)'(x) = \Big(f\cdot\frac{1}{g}\Big)'(x) &= f'(x)\cdot\frac{1}{g(x)} + f(x)\Big(\frac{1}{g}\Big)'(x) \\ &= \frac{f'(x)}{g(x)} - \frac{f(x)g'(x)}{g(x)^2} = \frac{g(x)f'(x)}{g(x)^2} - \frac{f(x)g'(x)}{g(x)^2}.\end{aligned}$$

□

Formula (15.1) is mildly clunky but is very important and should be memorized. I find the incantation "bottom times derivative of top, minus top times derivative of bottom, over bottom squared" helpful. It may also be helpful to observe the similarity between the numerator in (15.1) and the product rule; the only difference is the negative sign, and to remember where this goes you can remind yourself of the principle that differentiating something in the denominator tends to produce a negative sign, as in the reciprocal rule or as in $\frac{d}{dx}(\frac{1}{x}) = -\frac{1}{x^2}$.

Example 15.5. Suppose that x and y are related by the formula $y = \frac{x^2+x+1}{x^2-1}$. To find $\frac{dy}{dx}$, we can use the quotient rule and get

$$\frac{dy}{dx} = \frac{(x^2-1)\frac{d}{dx}(x^2+x+1) - (x^2+x+1)\frac{d}{dx}(x^2-1)}{(x^2-1)^2}$$

$$= \frac{(x^2-1)(2x+1)-(x^2+x+1)(2x)}{(x^2-1)^2} = \frac{2x^3+x^2-2x-1-(2x^3+2x^2+2x)}{(x^2-1)^2}$$
$$= \frac{-x^2-4x-1}{(x^2-1)^2}.$$

Example 15.6. The quotient rule leads to complicated enough computations that it is often better to use a different rule if we have the option. For example, we could differentiate the function $f(x) = (4x + 3\sqrt[3]{x})/\sqrt{x}$ using the quotient rule, but it is easier to write

$$f(x) = 4\sqrt{x} + 3x^{\frac{1}{3}-\frac{1}{2}} = 4x^{\frac{1}{2}} + 3x^{-\frac{1}{6}}$$

and then apply the power rule to each term directly.

Lecture 16 Trigonometric functions

This lecture corresponds to §3.3 in Stewart and Chapter 15 in Spivak, although Spivak's approach to the trigonometric functions is a little different and involves integration, which we did not discuss yet.

In Example 12.8 we proved that the sine function is differentiable at 0, with derivative 1; in other words,

$$\lim_{x\to 0} \frac{\sin x}{x} = 1. \tag{16.1}$$

We can use this, together with some trigonometric identities, to find the derivative of sin everywhere, similarly to the way that we used the property $e^{x+y} = e^x e^y$ of the exponential function to differentiate e^x everywhere once we understood its derivative at 0. For sin, given any x, h, we have

$$\sin(x+h) = \sin x \cos h + \cos x \sin h,$$

and thus

$$\frac{d}{dx}\sin x = \lim_{h\to 0}\frac{\sin(x+h)-\sin(x)}{h} = \lim_{h\to 0}\frac{\sin x\cos h + \cos x \sin h - \sin x}{h}$$
$$= \lim_{h\to 0} \sin x \frac{\cos h - 1}{h} + \cos x \frac{\sin h}{h} = \sin x \underbrace{\lim_{h\to 0}\frac{\cos h - 1}{h}}_{\text{I}} + \cos x \underbrace{\lim_{h\to 0}\frac{\sin h}{h}}_{\text{II}}.$$

The limit in II exists and is equal to 1 by (16.1). For the limit in I, we observe that

$$\lim_{h\to 0}\frac{\cos h - 1}{h} = \lim_{h\to 0}\frac{\cos h - 1}{h}\cdot\frac{\cos h + 1}{\cos h + 1} = \lim_{h\to 0}\frac{\cos^2 h - 1}{h(\cos h + 1)}$$
$$= \lim_{h\to 0}\frac{-\sin^2 h}{h(\cos h + 1)} = \Big(\lim_{h\to 0}\frac{\sin h}{h}\Big)\Big(\lim_{h\to 0}\frac{-\sin h}{\cos h + 1}\Big) = 1\cdot\frac{-0}{2} = 0,$$

and we conclude that

$$\frac{d}{dx}\sin x = \cos x. \tag{16.2}$$

Exercise 16.1. Use the formula $\cos(x+h) = \cos x \cos h - \sin x \sin h$ to prove that $\frac{d}{dx}\cos x = -\sin x$.

Remark 16.2. *Euler's formula* states that $e^{ix} = \cos x + i \sin x$. Differentiating this using what we just proved gives

$$\frac{d}{dx}e^{ix} = \frac{d}{dx}\cos x + i\frac{d}{dx}\sin x = -\sin x + i\cos x. \tag{16.3}$$

Recall that when we studied derivatives of exponential functions, we proved that $\frac{d}{dx}e^{cx} = ce^{cx}$. If this holds true for complex values of c as well, then we could also differentiate e^{ix} as

$$\frac{d}{dx}e^{ix} = ie^{ix} = i(\cos x + i\sin x) = i\cos x - \sin x. \tag{16.4}$$

Observe that (16.3) and (16.4) agree. However, one should observe that there's something a little bit suspect going on here; we never defined the value of the exponential function for arguments that are not real numbers, so we never actually defined e^{ix}, let alone prove that Euler's formula holds for it. In fact, one option is to use Euler's formula as the definition of e^{ix}; then the above computations verify that differentiation still behaves as we expect, and the formula $e^{i(x+y)} = e^{ix}e^{iy}$ continues to be true as a consequence of the formulas for $\cos(x+y)$ and $\sin(x+y)$, though we omit the computation.

Now the derivatives of the other trigonometric functions can be computed by using the product and quotient rules, for example

$$\frac{d}{dx}\tan x = \frac{d}{dx}\frac{\sin x}{\cos x} = \frac{\cos x \frac{d}{dx}\sin x - \sin x \frac{d}{dx}\cos x}{\cos^2 x} = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = \frac{1}{\cos^2 x} = \sec^2 x.$$

Exercise 16.3. Prove that

$$\frac{d}{dx}\cot x = -\csc^2 x, \qquad \frac{d}{dx}\sec x = \sec x \tan x, \qquad \frac{d}{dx}\csc x = -\csc x \cot x.$$

Example 16.4. The derivative of $f(x) = \frac{\cos x}{1+\cot x}$ can be found using the quotient rule and the formulas $\frac{d}{dx}\cos x = -\sin x$, $\frac{d}{dx}\cot x = -\csc^2 x$:

$$\begin{aligned} f'(x) &= \frac{(1+\cot x)(-\sin x) - \cos x(-\csc^2 x)}{(1+\cot x)^2} = \frac{-\sin x - \frac{\cos x}{\sin x}\sin x + \cos x \frac{1}{\sin^2 x}}{(1+\cot x)^2} \\ &= \frac{-\sin x + \frac{\cos x}{\sin^2 x}(1-\sin^2 x)}{(1+\frac{\cos x}{\sin x})^2} = \frac{-\sin x + \frac{\cos^3 x}{\sin^2 x}}{(\sin x + \cos x)^2/\sin^2 x} = \frac{\cos^3 x - \sin^3 x}{(\cos x + \sin x)^2}. \end{aligned}$$

Remark 16.5. Since there are many trigonometric identities that allow us to convert write formulas of one trig function in terms of other trig functions, any example such as the previous one will often have multiple ways of writing the final answer. As a general rule, you should simplify as much as possible but be aware that the expression you write down might look different from what somebody else gets, even if both are correct. For example, in the previous computation we could just as easily have written the final answer as $\frac{-\sin x + \cos x \cot^2 x}{(1+\cot x)^2}$.

Higher derivatives of sine and cosine can be easily computed since

$$\begin{aligned}\frac{d}{dx}\sin x = \cos x \quad &\Rightarrow \quad \frac{d^2}{dx^2}\sin x = \frac{d}{dx}\cos x = -\sin x \\ &\Rightarrow \quad \frac{d^3}{dx^3}\sin x = \frac{d}{dx}(-\sin x) = -\frac{d}{dx}\sin x = -\cos x \\ &\Rightarrow \quad \frac{d^4}{dx^4}\sin x = \frac{d}{dx}(-\cos x) = -\frac{d}{dx}\cos x = \sin x,\end{aligned}$$

and after this the pattern repeats with every fourth derivative. Keeping Remark 16.2 in mind, it is worth comparing this to the observation that

$$\begin{aligned}\frac{d}{dx}e^{ix} = ie^{ix} \quad &\Rightarrow \quad \frac{d^2}{dx^2}e^{ix} = \frac{d}{dx}ie^{ix} = i^2e^{ix} = -e^{ix} \\ &\Rightarrow \quad \frac{d^3}{dx^3}e^{ix} = \frac{d}{dx}(-e^{ix}) = -ie^{ix} \\ &\Rightarrow \quad \frac{d^4}{dx^4}e^{ix} = \frac{d}{dx}(-ie^{ix}) = (-i)ie^{ix} = e^{ix},\end{aligned}$$

which admits a geometric interpretation: multiplying a complex number by i corresponds to rotating it by 90 degrees ($\pi/2$ radians) around the origin; doing this twice corresponds to a 180 degree rotation, which is the same as multiplication by -1, and doing it four times returns us to where we started. For a similar geometric interpretation of the derivatives of sine, we can invoke some trigonometric identies and observe that

$$\sin\left(x+\frac{\pi}{2}\right) = \cos\frac{\pi}{2}\sin x + \sin\frac{\pi}{2}\cos x = 0\cdot\sin x + 1\cdot\cos x = \cos x = \frac{d}{dx}\sin x$$

$$\sin(x+\pi) = \cos\pi\sin x + \sin\pi\cos x = -1\cdot\sin x + 0\cdot\cos x = -\sin x = \frac{d^2}{dx^2}\sin x,$$

$$\sin\left(x+\frac{3\pi}{2}\right) = \cos\frac{3\pi}{2}\sin x + \sin\frac{3\pi}{2}\cos x = 0\cdot\sin x + -1\cdot\cos x = -\cos x = \frac{d^3}{dx^3}\sin x.$$

In other words, differentiating the sine function has the effect of translating the argument by $\pi/2$; doing this twice has the effect of changing the sign, since $\sin(x+\pi) = -\sin x$, and doing it four times has the same effect as doing nothing, since the function is periodic with period 2π.

Example 16.6. We can compute $\frac{d^{53}}{dx^{53}}\sin x$ by observing that $53 = 4\cdot 13+1$, so the first $52 = 4\cdot 13$ derivatives get us back to $\sin x$, and we thus have $\frac{d^{53}}{dx^{53}}\sin x = \frac{d}{dx}\sin x = \cos x$.

Of particular importance is the fact that if $f(x) = \sin x$ or $f(x) = \cos x$ (or more generally if $f(x) = a\sin x + b\cos x$ for some constants $a, b \in \mathbb{R}$), then

$$f''(x) = -f(x) \text{ for all } x \in \mathbb{R}. \tag{16.5}$$

This is a tremendously important example of a *differential equation*, which arises naturally in many different areas of science.

Example 16.7. Consider an object on a spring that at time t is displaced by a distance r from its rest position; then the force acting on the object has magnitude kr, where $k > 0$ is the *spring constant* that measures the strength of the spring, and this force

is directed in the opposite direction of the displacement. Thus if m is the mass of the object, then Newton's second law gives

$$\frac{d^2r}{dt^2} = \text{acceleration} = \frac{\text{force}}{\text{mass}} = -\frac{kr}{m},$$

and if we define a function f by $f(x) = r(x/\omega)$ where $\omega := \sqrt{k/m}$, we can use (14.15) to get

$$f'(x) = \frac{d}{dt}r\Big(\frac{x}{\omega}\Big) = \frac{1}{\omega}r'\Big(\frac{x}{\omega}\Big) \quad \Rightarrow \quad \begin{cases} f''(x) = \dfrac{d}{dx}\Big(\dfrac{1}{\omega}r'\Big(\dfrac{x}{\omega}\Big)\Big) = \dfrac{1}{\omega^2}r''\Big(\dfrac{x}{\omega}\Big) \\ \qquad = -\dfrac{1}{\omega^2}\dfrac{k}{m}r\Big(\dfrac{x}{\omega}\Big) = -r\Big(\dfrac{x}{\omega}\Big) = f(x), \end{cases}$$

which means that f must be a function that satisfies the differential equation in (16.5). One can prove (though we won't do it yet) that *every* function satisfying this differential equation can be written as $f(x) = a\sin x + b\cos x$, and we conclude that the displacement of the object is given by

$$r(t) = f(\omega t) = a\sin(\omega t) + b\cos(\omega t). \tag{16.6}$$

Exercise 16.8. Prove that if $r(t)$ is given by (16.6), then there are $A > 0$ and $\phi \in \mathbb{R}$ such that $r(t) = A\sin(\omega t + \phi)$.

The exercise demonstrates that the function r describing the position of an object on a spring can be written in terms of sine (or cosine). This system is an example of a *simple harmonic oscillator*.

Lecture 17 Convexity of the exponential function

This lecture fills in some details that do not appear in either Stewart or Spivak, although convexity is discussed in the Appendix to Chapter 11 of Spivak, and in §4.3 of Stewart (from a different point of view).

Fix $a > 0$ and let $f(x) = a^x$. In Lecture 14.2 we computed $f'(x)$ by writing

$$f'(x) = \lim_{h\to 0}\frac{a^{x+h} - a^x}{h} = \lim_{h\to 0} a^x \cdot \frac{a^h - 1}{h} = a^x \lim_{h\to 0}\frac{a^h - 1}{h} = (\ln a)a^x, \tag{17.1}$$

where our argument that $\lim_{h\to 0}\frac{a^h-1}{h} = \ln a$ relied on Conjecture 14.9, which stated that the slope of the secant line from (x, a^x) to (y, a^y) increases if we move either of x or y to the right, and decreases if we move either of them to the left. This looked plausible based on the shape of the graph of the exponential function, and in particular the fact that it "bends upwards", but now we need to provide an actual proof in order to prove that our computation of $\frac{d}{dx}a^x$ was actually valid.

Let us start by recalling our notion of *increasing* and *decreasing* functions, which are illustrated in Figure 3. There are multiple ways to describe the property of being increasing that is exhibited in the left-hand picture.

- The slope of every tangent line is positive.
- The derivative f' is positive everywhere.

- The slope of every secant line is positive.
- As x increases, the value of $f(x)$ increases.

The first two of these only make sense if f is differentiable, while the last two make sense for every f. (Observe that the last two are equivalent to each other.) Thus we can take either of these last two as the definition of increasing.

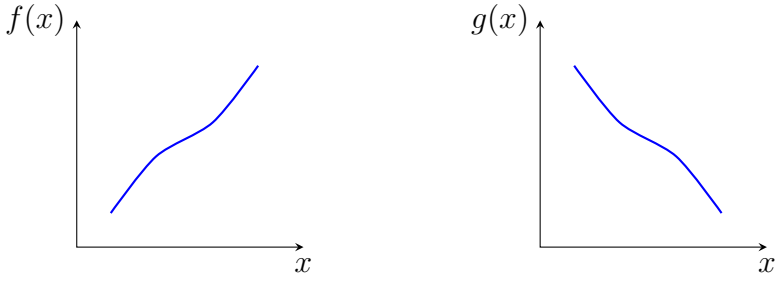


FIGURE 3. An increasing function f and a decreasing function g.

Now consider the two functions illustrated in Figure 4. The function f is an example of a *convex* function, and the function g is *concave*. But what do these words mean? How would you describe the difference between f and g to someone who could not see the pictures of their graphs?

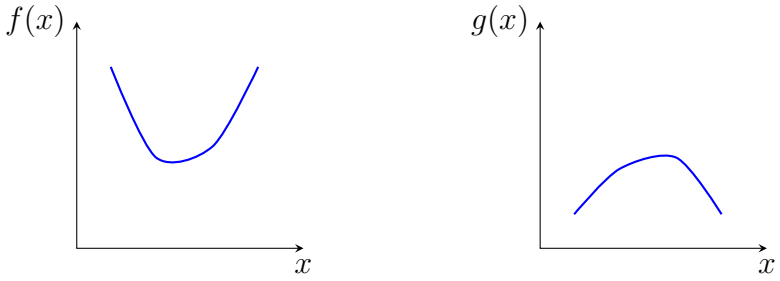


FIGURE 4. A convex function f and a concave function g.

A first attempt might be to say that the graph of f "bends upwards", or "opens upwards", while the graph of g "bends downwards". One way to make this more precise would be to phrase it in terms of the tangent lines: the function f has the following two properties.

- The graph of f lies above all its tangent lines.
- As x moves to the right, the slope of the tangent line at x increases.

The second of these can also be written in terms of the derivative, since $f'(x)$ is the slope of the tangent line at x.

- As x increases, the value of $f'(x)$ increases.

If f is twice differentiable, then the derivative f' being increasing corresponds to the second derivative f'' being positive. So we might define convexity as

- the second derivative satisfies $f''(x) > 0$ everywhere.

On the other hand, we might want to consider functions that are *not* differentiable. What if we put a 'corner' in the graph of f, but kept the same basic shape? None of the previous descriptions would make sense, because f' would not be defined everywhere. However, if we replace 'tangent line' with 'secant line', then we get a description that works even without differentiability: f has the property stated in Conjecture 14.9:

- As either x or y moves to the right (with the other one staying in the same place), the slope of the secant line from $(x, f(x))$ to $(y, f(y))$ increases.

Suppose $x < y$. In order for the slope of the secant line to increase as y increases, we need the graph of f to lie below the secant line just to the left of y, and above the secant line just to the right of y. Similarly, the graph should lie below the secant line just to the right of x, and above it just to the left of x. So we give the following description.

- Between x and y, the secant line through those points lies above the graph.

Exercise 17.1. Using proof by contradiction, show that if the graph of f has the last property above, then at any z that is *not* between x and y, the secant line lies above the graph. Use this to prove the property about the slope of the secant line increasing.

Exercise 17.2. Determine the relationship between the six conditions listed above. Are they all equivalent when f is differentiable, or twice differentiable? Or are there some functions which satisfy certain conditions but not others?

All of the descriptions above capture various aspects of the 'shape' of the graph of f. For now we will take the last one as our definition of 'convex'. To formulate it precisely, we observe that given any z between x and y, we can write $z = tx + (1-t)y$ for some $t \in [0,1]$ (the choice $t = \frac{z-y}{x-y}$ works), and the vertical coordinate of the secant line at this point is $tf(x) + (1-t)f(y)$. With this in mind, we make the following definition.

Definition 17.3. Let $I \subset \mathbb{R}$ be an interval. A function $f\colon I \to \mathbb{R}$ is *convex* if for every $x, y \in I$ and every $t \in [0,1]$, we have

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y). \tag{17.2}$$

We call f *strictly convex* if $\leq$ can be replaced by $<$ in (17.2) whenever $x \neq y$ and $t \in (0,1)$.

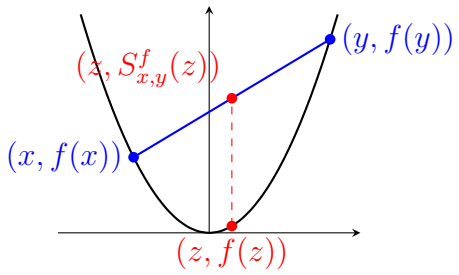
Example 17.4. The function $f(x) = x^2$ is strictly convex. To see this, we write the difference between the right- and left-hand sides of (17.2) as

$$\begin{aligned} &tf(x) + (1-t)f(y) - f(tx + (1-t)y) \\ &\quad = tx^2 + (1-t)y^2 - (t^2x^2 + 2t(1-t)xy + (1-t)^2y^2) \\ &\quad = (t - t^2)x^2 - 2(t-t^2)xy + (1 - t - (1 - 2t + t^2))y^2 \\ &\quad = (t-t^2)(x^2 - 2xy + y^2) = t(1-t)(x-y)^2. \end{aligned}$$

This last quantity is > 0 for every $x \neq y$ and $t \in (0,1)$, which proves (17.2) with a strict inequality. Thus f is strictly convex.

Remark 17.5. Later, in Lecture 27.2, we will see an easier way to establish convexity of $f(x) = x^2$ using second derivatives.

As described above, the definition of convexity says that on the interval between x, y, the graph of the function f lies below the secant line through the points $(x, f(x))$ and $(y, f(y))$. The picture at right illustrates this for $f(x) = x^2$, using the notation $z = tx + (1-t)y$ and writing $S^f_{x,y}$ for the function whose graph is the secant line through $(x, f(x))$ and $(y, f(y))$.[22]



To find a formula for $S^f_{x,y}$, start by observing that the secant line through $(x, f(x))$ and $(y, f(y))$ is the set of all points (z, w) such that[23]

$$\frac{w - f(x)}{z - x} = \frac{f(y) - f(x)}{y - x}.$$

This is equivalent to $w - f(x) = \frac{f(y)-f(x)}{y-x}(z - x)$, and thus we see that the secant line is the graph of the function

$$S^f_{x,y}(z) := \frac{f(y) - f(x)}{y - x}(z - x) + f(x). \tag{17.3}$$

For later use we rewrite (17.3) as

$$S^f_{x,y}(z) = f(y)\frac{z - x}{y - x} + f(x)\Big(1 - \frac{z - x}{y - x}\Big) = f(y)\frac{z - x}{y - x} + f(x)\frac{y - z}{y - x}. \tag{17.4}$$

Exercise 17.6. Prove that if $z = tx + (1-t)y$ for some $t \in [0, 1]$, then $S^f_{x,y}(z) = tf(x) + (1-t)f(y)$. Use this to deduce that a function $f\colon I \to \mathbb{R}$ is convex if and only if for every $x, y \in I$ and every z between x and y, we have $f(z) \le S^f_{x,y}(z)$, and strictly convex if this inequality is strict whenever $z \neq x, y$.

Now fix $a > 0$ and consider the function $f(x) = a^x$. Our goal is to prove that this function is convex[24] on $\mathbb{R}$, which amounts to showing that

$$(a^x)^t(a^y)^{1-t} \le ta^x + (1-t)a^y \tag{17.5}$$

for all $x \neq y \in \mathbb{R}$ and $t \in (0, 1)$. It is not immediately clear how to prove this for an arbitrary t; however, in the specific case $t = 1/2$, our desired inequality reduces to

$$\sqrt{a^x a^y} \le \frac{a^x + a^y}{2}, \tag{17.6}$$

which seems more manageable. Can we prove that (17.6) holds for all $a > 0$ and $x \neq y$?

Our immediate response to something like (17.6) is to square both sides (perhaps after multiplying both sides by 2); since both sides are positive, we see that (17.6) is true if and only if

$$4a^x a^y \le (a^x + a^y)^2 = a^{2x} + 2a^x a^y + a^{2y}.$$

[22] The subscripts x, y and superscript f are just there to remind us which points and which function we used in defining $S^f_{x,y}$.

[23] Of course we would naturally write (x, y) for a point in $\mathbb{R}^2$, but we already used these symbols, so we must pick different ones; and after all, what's in a name?

[24] In fact it turns out to be *strictly* convex, but we will settle for convex for now.

But this is equivalent to $0 \le a^{2x} - 2a^x a^y + a^{2y}$, which we know is always true since the right-hand side can be written as $(a^x - a^y)^2$. Thus we have proved that (17.5) is true for all $x, y \in \mathbb{R}$ and $t = 1/2$; in other words, while we don't yet know that the graph of f lies below the secant line for x, y *everywhere* on the interval between x and y, it at least does so at the midpoint of this interval. We formulate this fact as a lemma, whose proof is given by the above discussion.

Lemma 17.7. *Given any $a > 0$, the exponential function $f(x) = a^x$ has the following property: for all $x, y \in \mathbb{R}$ with $x \neq y$, the point $c = \frac{x+y}{2}$ satisfies $f(c) < \frac{1}{2}(f(x)+f(y)) = S^f_{x,y}(c)$.*

In our proof of convexity, we will find it useful to have a tool that lets us translate inequalities such as the one in Lemma 17.7 into statements comparing secant lines.

Lemma 17.8. *Let I be an interval, $f\colon I \to \mathbb{R}$ any function, and $x, y \in I$ two numbers with $x < y$. If $c \in (x, y)$ has the property that $f(c) \le S^f_{x,y}(c)$, then we have*

$$\operatorname{slope}(S^f_{x,c}) \le \operatorname{slope}(S^f_{x,y}) \le \operatorname{slope}(S^f_{c,y}). \tag{17.7}$$

In particular, we have

$$\begin{aligned} &S^f_{c,y}(z) \le S^f_{x,y}(z) \text{ for all } z \text{ between } c \text{ and } y, \\ &S^f_{x,c}(z) \le S^f_{x,y}(z) \text{ for all } z \text{ between } x \text{ and } c. \end{aligned} \tag{17.8}$$

Proof. The inequalities in (17.8) follow directly from (17.7) by noting where the secant lines intersect each other. To prove (17.7), observe that

$$\operatorname{slope}(S^f_{x,c}) = \frac{f(c) - f(x)}{c - x} \le \frac{S^f_{x,y}(c) - f(x)}{c - x} = \operatorname{slope}(S^f_{x,y}),$$

and similarly,

$$\operatorname{slope}(S^f_{c,y}) = \frac{f(y) - f(c)}{y - c} \ge \frac{f(y) - S^f_{x,y}(c)}{y - c} = \operatorname{slope}(S^f_{x,y}). \qquad \square$$

Now the way to prove convexity of the exponential function is to iterate Lemma 17.7 using bisection sequences as in the proof of the Intermediate Value Theorem, as suggested by Figure 5. In fact we will prove a more general result, valid for any continuous function satisfying a convexity-type inequality at midpoints.

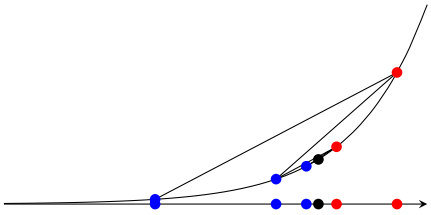


FIGURE 5. Midpoint convexity and continuity imply convexity.

Theorem 17.9. *Let $f\colon I \to \mathbb{R}$ be a continuous function with the property that $f(\frac{x+y}{2}) \le \frac{1}{2}(f(x) + f(y))$ for all $x, y \in I$. Then f is convex.*

Proof. We will use the characterization of convexity from Exercise 17.6. Fix $x, y \in \mathbb{R}$; without loss of generality assume that $x < y$. Given a point $z \in (x, y)$, define a pair of bisection sequences $b_n, r_n \to z$ recursively as follows (see Figure 5):

- Start with $b_1 = x$ and $r_1 = y$.
- For every $n \geq 1$, once b_n and r_n have been determined, let $m_n = \frac{1}{2}(b_n + r_n)$ and proceed by cases:
  (1) if $b_n \leq z \leq m_n$, then let $b_{n+1} = b_n$ and $r_{n+1} = m_n$;
  (2) if $m_n < z \leq r_n$, then let $b_{n+1} = m_n$ and $r_{n+1} = r_n$.

This procedure should be familiar from our proof of the Intermediate Value Theorem. Note that we have $b_n \leq z \leq r_n$ for every n, and since both sequences are monotonic, their limits exist. Since $r_n - b_n \to 0$, the limits are the same, and equal to z. Finally, continuity of f and of $S^f_{x,y}$ shows that

$$f(z) = \lim_{n\to\infty} f(b_n), \quad S^f_{x,y}(z) = \lim_{n\to\infty} S^f_{x,y}(b_n). \tag{17.9}$$

(The same would be true if we replaced b_n with r_n.) At each step the number m_n is the midpoint of b_n, r_n, and by the definition of b_{n+1}, r_{n+1}, it follows from (17.8) that

$$S^f_{b_{n+1},r_{n+1}}(w) \leq S^f_{b_n,r_n}(w) \text{ for all } w \in [b_n, r_n].$$

Since $b_n \in [b_k, r_k]$ for every $k \leq n$, it follows by induction that

$$f(b_n) = S^f_{b_n,r_n}(b_n) \leq S^f_{x,y}(b_n).$$

Taking a limit as $n \to \infty$ and using (17.9), this implies that $f(z) \leq S^f_{x,y}(z)$ by the theorem on monotonicity of limits. Since this holds for any $z \in (x, y)$, we conclude that f is convex. □

Corollary 17.10. *For every $a > 0$, the exponential function $f(x) = a^x$ is convex.*

Remark 17.11. The proof of Theorem 17.9 can be modified to show that the exponential function is in fact *strictly* convex, using the fact that the inequality in Lemma 17.7 is strict.

By Exercise 17.1, convexity of the exponential function implies the conclusion of Conjecture 14.9, and thus this completes the proof of the results claimed in Lecture 14.2; the exponential function is differentiable everywhere, and $\frac{d}{dx}a^x = (\ln a)a^x$.

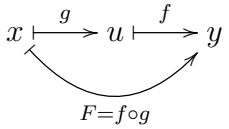
Lecture 18 — Chain rule

This lecture corresponds to §3.4 in Stewart and Chapter 10 in Spivak.

In (14.15), we saw that $\frac{d}{dx}f(cx) = cf'(cx)$. The geometric interpretation of this is that $x \mapsto f(cx)$ has a graph given by taking the graph of $f(x)$ and "squashing" it in the x-direction by a factor of $\frac{1}{c}$, which has the effect of multiply slopes by c.

Now we prove a more general version of this result. Suppose that f and g are two functions for which $f(g(a))$ is well-defined (that is, a is in the domain of g and $g(a)$ is in

the domain of f). Given $x \approx a$, let $u = g(x)$ and $y = f(g(x))$, so that x, u, y are related as follows:



Writing $F = f \circ g$, we want to find a formula for F' in terms of f, g, and their derivatives. Intuitively, the idea is that $F' = \frac{dy}{dx}$ measures the ratio $\frac{\Delta y}{\Delta x}$, where Δy is the amount by which $y = F(x) = f(g(x))$ changes in response to a small change Δx in x. Using our (by now standard) trick of multiplying and dividing by the same thing, we can write

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta u}\frac{\Delta u}{\Delta x} = \lim_{\Delta u \to 0} \frac{\Delta y}{\Delta u} \cdot \lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} = \frac{dy}{du}\frac{du}{dx}. \tag{18.1}$$

Since $\frac{dy}{dx}|_{x=a} = (f \circ g)'(a)$, $\frac{dy}{du}|_{u=g(a)} = f'(g(a))$, and $\frac{du}{dx}|_{x=a} = g'(a)$, we can write (18.1) as

$$(f \circ g)'(a) = f'(g(a))g'(a), \tag{18.2}$$

an identity which is called the *chain rule*. There is one small problem, though; what if $\Delta u = 0$ in (18.1)? Then the computation is not valid, because we divided by 0. Thus (18.1) is not quite a proof. The trick to making it a rigorous proof is to think about what quantity should replace $\frac{\Delta y}{\Delta u}$ when $\Delta u = 0$; presumably in this case we should use the derivative $\frac{dy}{du}$ itself. This reasoning leads us to the following argument.

Theorem 18.1 (Chain rule)**.** *Let f and g be functions and $a \in \mathbb{R}$ a point such that g is differentiable at a, and f is differentiable at $g(a)$.*[25] *Then $f \circ g$ is differentiable at a, and the identity* (18.2) *holds.*

Proof. In terms of our usual notation for computing derivatives, we have $\Delta x = h$, $\Delta u = g(a + h) - g(a)$, and $\Delta y = f(g(a + h)) - f(g(a))$. We would like to write the second step in (18.1) as

$$\frac{f(g(a+h)) - f(g(a))}{h} = \frac{f(g(a+h)) - f(g(a))}{g(a+h) - g(a)} \cdot \frac{g(a+h) - g(a)}{h},$$

but the first fraction on the right-hand side may be undefined if $g(a + h) - g(a) = 0$. However, this represents a *removable* discontinuity, because f is differentiable at $g(a)$.

Lemma 18.2. *The function ϕ defined by*

$$\phi(h) = \begin{cases} \frac{f(g(a+h)) - f(g(a))}{g(a+h) - g(a)} & \text{if } g(a+h) - g(a) \neq 0, \\ f'(g(a)) & \text{if } g(a+h) - g(a) = 0, \end{cases}$$

is continuous at 0*.*

Proof. Because f is differentiable at $g(a)$, we have

$$\lim_{t \to 0} \frac{f(g(a) + t) - f(g(a))}{t} = f'(g(a)).$$

[25] In particular, this requires that a is in the domain of g, and $g(a)$ is in the domain of $f(a)$.

Thus for every $\epsilon > 0$ there exists $\delta' > 0$ such that

$$\text{if } 0 < |t| < \delta', \text{ then } \left| \frac{f(g(a)+t) - f(g(a))}{t} - f'(g(a)) \right| < \epsilon. \tag{18.3}$$

Moreover, since g is differentiable at a, it is continuous there by Theorem 13.3. Thus there is $\delta > 0$ such that

$$\text{if } |h| < \delta, \text{ then } |g(a+h) - g(a)| < \delta'. \tag{18.4}$$

Now given h with $|h| < \delta$, we can write $t = g(a+h) - g(a)$ so that $g(a+h) = g(a) + t$.

- Case I. If $t \neq 0$, then $0 < |t| < \delta'$ by (18.4), and so
$$\phi(h) = \frac{f(g(a+h)) - f(g(a))}{g(a+h) - g(a)} = \frac{f(g(a)+t) - f(g(a))}{t}$$
satisfies $|\phi(h) - \phi(0)| = |\phi(h) - f'(g(a))| < \epsilon$ by (18.3).
- Case II. If $t = 0$, then $\phi(h) = f'(g(a)) = \phi(0)$ by definition.

In both cases, we conclude that $|\phi(h) - \phi(0)| < \epsilon$, which proves continuity at 0. □

Returning to the proof of the chain rule, we observe that for every h, we have

$$\frac{f(g(a+h)) - f(g(a))}{h} = \phi(h) \cdot \frac{g(a+h) - g(a)}{h},$$

where when $g(a+h) \neq g(a)$ this follows from the definition of ϕ, and when $g(a+h) = g(a)$ it follows since both sides are equal to 0. Thus we have

$$\begin{aligned}(f \circ g)'(a) &= \lim_{h\to 0} \frac{f(g(a+h)) - f(g(a))}{h} = \lim_{h\to 0} \phi(h) \cdot \lim_{h\to 0} \frac{g(a+h) - g(a)}{h} \\ &= \phi(0) \cdot g'(a) = f'(g(a))g'(a),\end{aligned}$$

where the first equality on the second line uses Lemma 18.2, and the second uses the definition of $\phi(0)$. □

Example 18.3. Consider the function $F(x) = \sqrt{x^2 - 1}$. We can write this as the composition of the functions $g(x) = x^2 - 1$ and $f(u) = \sqrt{u}$, whose derivatives we know to be

$$f'(u) = \frac{1}{2\sqrt{u}} \quad \text{and} \quad g'(x) = 2x.$$

By the chain rule, we have

$$F'(x) = (f \circ g)'(x) = f'(g(x))g'(x) = \frac{1}{2\sqrt{g(x)}} \cdot 2x = \frac{x}{\sqrt{x^2 - 1}}.$$

Example 18.4. The function $f(x) = \sin \frac{1}{x}$ can be written as $f = g \circ h$ where $g(u) = \sin u$ and $h(x) = \frac{1}{x}$. The chain rule gives

$$f'(x) = g'(h(x))h'(x) = \cos(h(x)) \cdot \left(-\frac{1}{x^2}\right) = -\frac{\cos \frac{1}{x}}{x^2}.$$

Example 18.5. For any differentiable function g, we can compute $\frac{d}{dx}(g(x))^n$ by recalling that $f(u) = u^n$ has $f'(u) = nu^{n-1}$ and using the chain rule to get

$$\frac{d}{dx}(g(x))^n = f'(g(x))g'(x) = n(g(x))^{n-1}g'(x).$$

Example 18.6. In (14.8) we saw that $\frac{d}{dx}a^x = \frac{d}{dx}b^{\log_b(a)x} = \log_b(a)\frac{d}{dx}b^x$ for any $a, b > 0$ with $b \neq 1$; we used this with $b = e$ to get the formula $\frac{d}{dx}a^x = (\ln a)a^x$ in (14.11). In (14.15) we pointed out that (14.8) is a special case of the more general formula $\frac{d}{dx}f(cx) = cf'(cx)$. Now we see that this in turn is a special case of the chain rule, where we put $g(x) = cx$ and get

$$\frac{d}{dx}f(g(x)) = g'(x)f'(g(x)) = cf'(cx).$$

Remark 18.7. There is a subtlety here that is worth emphasizing. We are used to writing $f'(x)$ and $\frac{d}{dx}f(x)$ to mean the same thing – the derivative of f with respect to x. However, $f'(cx)$ and $\frac{d}{dx}f(cx)$ do *not* mean the same thing. The latter – $\frac{d}{dx}f(cx)$ – means the derivative of $x \mapsto f(cx)$ with respect to x, which gives the sensitivity of $f(cx)$ to a small change in x. The former – $f'(cx)$ – means the derivative with respect to the input of f, which in this case is cx, and so it gives the sensitity of $f(cx)$ to a small change in cx. You should keep this in mind when using and reading the notations $\frac{d}{dx}$ and f': the first of these always means a rate of change with respect to x, while the second means a rate of change with respect to the input of f, whatever that input is called.

Example 18.8. Using the chain rule and the power rule, we get

$$\begin{aligned}\frac{d}{dx}\frac{1}{\sqrt{x^2+1}} &= \frac{d}{dx}(x^2+1)^{-1/2} \\ &= -\frac{1}{2}(x^2+1)^{-3/2}\frac{d}{dx}(x^2+1) = \frac{-2x}{2(x^2+1)^{3/2}} = \frac{-x}{(x^2+1)^{3/2}}.\end{aligned}$$

Note that although we did not write the functions out explicitly, we took $g(x) = x^2 + 1$ and $f(u) = u^{-1/2}$, so that $f \circ g(x) = f(x^2+1) = (x^2+1)^{-1/2}$; then the chain rule is used to get the first equality on the second line.

Example 18.9. The rational function $F(t) = (\frac{t+1}{t-1})^2$ can be differentiated in several ways. We could write it as $\frac{(t+1)^2}{(t-1)^2}$ and use the quotient rule directly. Or we can use the chain rule with $g(t) = \frac{t+1}{t-1}$ and $f(u) = u^2$, then apply the quotient rule to g (which is simpler than F) and obtain

$$\begin{aligned}F'(t) = \frac{d}{dt}\left(\left(\frac{t+1}{t-1}\right)^2\right) &= 2\cdot\frac{t+1}{t-1}\cdot\frac{d}{dt}\left(\frac{t+1}{t-1}\right) \\ &= 2\cdot\frac{t+1}{t-1}\left(\frac{t-1-(t+1)}{(t-1)^2}\right) = \frac{-4(t+1)}{(t-1)^3}.\end{aligned}$$

Yet another way to obtain $F'(t)$ is to rewrite the quotient as $\frac{t+1}{t-1} = \frac{t-1+2}{t-1} = 1 + \frac{2}{t-1}$, so that

$$F(t) = \left(\frac{t+1}{t-1}\right)^2 = \left(1 + \frac{2}{t-1}\right)^2 = 1 + \frac{4}{t-1} + \frac{4}{(t-1)^2}.$$

The power rule and chain rule together imply that $\frac{d}{dt}(t-1)^n = n(t-1)^{n-1}$ for all n, so we get

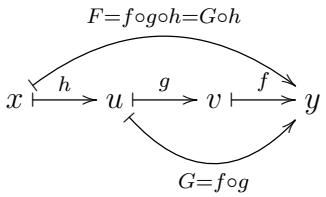
$$F'(t) = \frac{-4}{(t-1)^2} - 2\cdot\frac{4}{(t-1)^3}.$$

A little more algebra shows that this is equal to the expression we got earlier. This example illustrates that there can be more than one correct way to compute a given derivative, and that the appearance of the answer may depend on the method chosen, even though all answers will be equivalent.

Example 18.10.

$$\frac{d}{dx}e^{\cos x} = e^{\cos x}\frac{d}{dx}\cos x = -\sin x e^{\cos x}.$$

The chain rule can be used to differentiate compositions of more than two functions; indeed, this motivates its name, since we can think of this as a 'chain' of functions composed with each other. For example, given three functions $f, g, h\colon \mathbb{R} \to \mathbb{R}$, we can visualize $F = f \circ g \circ h$ as follows:



Writing $G = f \circ g$, we can use the chain rule to write G' in terms of f' and g', then again to write F' in terms of G' and h', obtaining

$$F'(x) = (G \circ h)'(x) = G'(h(x))h'(x) = (f \circ g)'(h(x))h'(x) = f'(g(h(x))) \cdot g'(h(x)) \cdot h'(x).$$

It is perhaps easiest to view this using the alternate notation

$$\frac{dy}{dx} = \frac{dy}{dv}\frac{dv}{du}\frac{du}{dx}, \tag{18.5}$$

where $u = h(x)$, $v = g(u) = g(h(x))$, and $y = f(v) = f(g(h(x))$. We reiterate that although this notation makes it very tempting to view the chain rule as just a matter of 'canceling the dv and the du from the numerator and denominator', this is not an actual proof of anything, because the quantities $\frac{dy}{dv}$, $\frac{dv}{du}$, and $\frac{du}{dx}$ are derivatives, not fractions, and the symbols du and dv have no independent meaning.

Example 18.11. The function $F(x) = \sin(e^{3x})$ can be differentiated by writing $u = 3x$, $v = e^u = e^{3x}$, and $y = \sin v = \sin e^{3x}$. We get

$$F'(x) = \frac{dy}{dx} = \frac{dy}{dv}\frac{dv}{du}\frac{du}{dx} = (\cos v)(e^u)(3) = 3e^{3x}\cos e^{3x}.$$

Note that although applying (18.5) gave us an expression in terms of u, v, x, this was not yet our final answer; it was necessary to go one step further and write u and v in terms of the original variable x. You will encounter this phenomenon in other places as well.

Lecture 19 — Implicit differentiation

This lecture corresponds to §3.5 in Stewart and Chapter 12 in Spivak.

19.1. Implicitly defined functions

Suppose (x, y) is a point on the unit circle $x^2 + y^2 = 1$. What is the slope of the tangent line to the circle at (x, y)?

One approach is to write y as a function of x, and then differentiate. If $y > 0$, then this means that $y = \sqrt{1 - x^2}$, and we can use the chain rule with $g(x) = 1 - x^2$ and $f(z) = z^{1/2}$ to deduce that

$$\text{slope} = \frac{dy}{dx} = \frac{d}{dx}\underbrace{(1-x^2)^{1/2}}_{f\circ g(x)} = \underbrace{\frac{1}{2}(1-x^2)^{-1/2}}_{f'(g(x))}\underbrace{(-2x)}_{g'(x)} = \frac{-2x}{2\sqrt{1-x^2}} = -\frac{x}{y}.$$

A similar computation gives the same result if $y < 0$. Note that if $y = 0$ then the tangent line is vertical so the slope is undefined.

This approach works fine in this case, although the computation with the chain rule is a little messy. But what if we want to find the tangent line to a more complicated curve, such as the curve defined by the equation $x^3 + y^3 = xy$, which is shown in Figure 6? (Note that this curve was drawn by appealing to a computer program; we do not yet have the tools to plot such curves by hand.) In this case it is not so easy to solve the equation and write y as a function of x. Fortunately, there is another approach available to us.

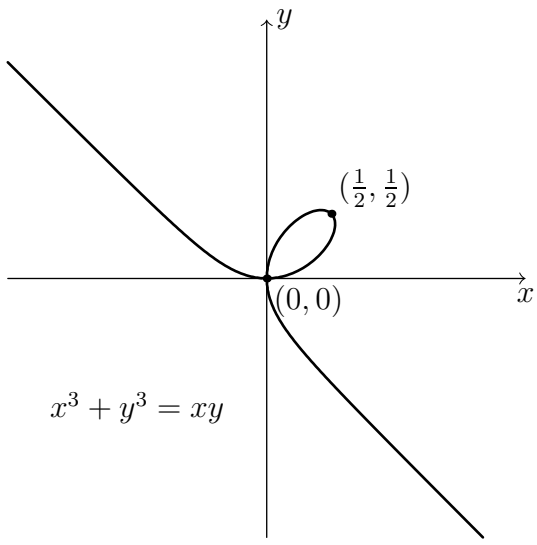


FIGURE 6. An implicitly defined function.

First let us return to the circle. The idea is that if $y = y(x)$ is a function that gives the dependence of y on x, then this function must satisfy the relationship $x^2 + (y(x))^2 = 1$. Both sides of this equation are functions of x, which we can differentiate using the rules we have discovered so far:

$$\frac{d}{dx}(x^2 + (y(x))^2) = 2x + 2y(x)\frac{dy}{dx},$$

$$\frac{d}{dx}1 = 0.$$

In the first equation we use the chain rule to differentiate $y(x)^2$. Because the two functions $x^2 + y(x)^2$ and 1 are equal, their derivatives are as well, and we conclude that

$$2x + 2y\frac{dy}{dx} = 0.$$

Solving for $\frac{dy}{dx}$ gives $\frac{dy}{dx} = -\frac{x}{y}$, just as before.

Remark 19.1. Note that this approach only works if we know beforehand that the point (x, y) is actually on the circle! If you use the formula $\frac{dy}{dx} = -\frac{x}{y}$ for a point that is not on the circle, you will get a meaningless answer.

The benefit of this second approach – *implicit differentiation* – is seen by considering the second example, $x^3 + y^3 = xy$. Here we differentiate both sides and get

$$3x^2 + 3y^2\frac{dy}{dx} = x\frac{dy}{dx} + y \quad \Rightarrow \quad \frac{dy}{dx} = \frac{y - 3x^2}{3y^2 - x}. \tag{19.1}$$

Note that in both examples, we obtain an expression for $\frac{dy}{dx}$ that involves both x and y; this is typical of solutions obtained using implicit differentiation.

Example 19.2. Let us find the points at which the curve $x^3 + y^3 = xy$ has a horizontal tangent line. From the computation in (19.1) we have $\frac{dy}{dx} = 0$ if and only if $y = 3x^2$ and $3y^2 \neq x$. If the point (a, b) lies on the curve and satisfies $b = 3a^2$, then we have

$$0 = a^3 + b^3 - ab = a^3 + (3a^2)^3 - a(3a^2) = 27a^6 - 2a^3 = a^3(27a^3 - 2),$$

so $a = 0$ or $a = \sqrt[3]{2/27} = \frac{1}{3}\sqrt[3]{2}$. In the first case we have $b = 3a^2 = 0$, in the second we have $b = 3a^2 = 3 \cdot \frac{1}{9} \cdot 2^{2/3} = \frac{1}{3}2^{2/3}$. At $(a, b) = (0, 0)$ the denominator in the formula for $\frac{dy}{dx}$ vanishes, while at $(a, b) = (\frac{1}{3}2^{1/3}, \frac{1}{3}2^{2/3})$ we have $3b^2 - a = \frac{3}{9}2^{4/3} - \frac{1}{9}2^{1/3} = \frac{5}{9}2^{1/3} \neq 0$, so at this point we have $\frac{dy}{dx} = 0$ and the curve has a horizontal tangent line.

Remark 19.3. The problem with the point $(0, 0)$ in the previous example has to do with the fact that the curve crosses itself at this point, as shown in Figure 6, so it does not have a uniquely defined tangent line.

Example 19.4. We can find the points with a vertical tangent line by reversing the roles of x and y, so that x is a function of y. Mimicking (19.1), we can differentiate both sides of $x^3 + y^3 = xy$ with respect to y and get

$$3x^2\frac{dx}{dy} + 3y^2 = x + y\frac{dx}{dy} \quad \Rightarrow \quad \frac{dx}{dy} = \frac{x - 3y^2}{3x^2 - y}.$$

A similar computation to the one in the previous example shows that there is a vertical tangent line at $(a, b) = (\frac{1}{3}2^{2/3}, \frac{1}{3}2^{1/3})$.

Example 19.5. We can iterate this process to find higher derivatives. Consider the circle $x^2 + y^2 = 1$. Differentiating this gave $2x + 2yy' = 0$, which we solved to get $y' = -\frac{x}{y}$. We can differentiate this to get

$$y'' = -\frac{y \cdot 1 - xy'}{y^2} = -\frac{1}{y} + \frac{x}{y^2}\Big(-\frac{x}{y}\Big) = -\frac{1}{y} - \frac{x^2}{y^3}.$$

Observe that such computations will usually yield an expression involving y', which needs to be substituted in order to simplify and obtain the final answer. We could also have proceeding by differentiating both sides of $x + yy' = 0$ to get

$$1 + (y')^2 + yy'' = 0 \quad \Rightarrow \quad y'' = \frac{-1 - (y')^2}{y} = -\frac{1}{y} - \frac{x^2}{y^3}.$$

Our discussion so far has sidestepped an important issue. The graphs associated to the equations $x^2 + y^2 = 1$ and $x^3 + y^3 = xy$ do not satisfy the vertical line test, so they are not graphs of functions; so what do we mean when we write $y = y(x)$?

The resolution of this is that if (a, b) is a point on the graph of one of these equations, then in most cases there is a small piece of the graph near (a, b) that *is* the graph of a function. For the circle, if $b > 0$ then for $x \approx a$ we can write $y = \sqrt{1 - x^2}$, while if $b < 0$ we can write $y = -\sqrt{1 - x^2}$. The exception comes at the points $(\pm 1, 0)$, for which $b = 0$; there is no neighborhood of these points on which the circle is the graph of a function $y = y(x)$. Whenever we use implicit differentiation, we must restrict ourselves to points near which y can be written as a function of x; we also need to know that $y(x)$ is differentiable at these points. The theoretical tool for determining which points satisfy these criteria is the following theorem; this theorem really belongs to multivariable calculus, so our description here is just to provide some intuition rather than to give a complete justification. In particular, this theorem highlights one of the recurring themes of calculus, which is the power of linear approximations.

Theorem 19.6 (Implicit Function Theorem)**.** *Let $F(x, y)$ be a continuously differentiable*[26] *real-valued function of* two *variables, and suppose that $a, b, c \in \mathbb{R}$ are such that $F(a, b) = c$. Suppose that the function $g(y) := F(a, y)$ has the property that $g'(b) \neq 0$. Then there exists $\epsilon > 0$ and a differentiable function $f\colon (a - \epsilon, a + \epsilon) \to \mathbb{R}$ such that given $x \in (a - \epsilon, a + \epsilon)$, $y = f(x)$ is the only solution of $F(x, y) = c$ that is near b.*

Idea of proof. The idea is to use the fact that near (a, b), the continuously differentiable function F has the linear approximation $F(x, y) \approx F(a, b) + \ell(x - a) + m(y - b)$, where $\ell, m \in \mathbb{R}$ are the *partial derivatives* that represent how sensitive $F(x, y)$ is to changes in x and y, respectively. In particular, $m = g'(b) \neq 0$, so the equation $F(x, y) = c$ is very close to $F(a, b) + \ell(x - a) + m(y - b) = c$. Since $F(a, b) = c$, the solutions of the latter equation form a line $y - b = -\frac{\ell}{m}(x - a)$; this is where we use the assumption that $m \neq 0$. Then one chooses $\epsilon > 0$ small enough that the linear approximation is accurate enough to give a function $f(x) \approx b - \frac{\ell}{m}(x - a)$ with the desired property. □

Example 19.7. With $F(x, y) = x^2 + y^2$ and $c = 1$, at a point (a, b) on the unit circle we have $g(y) = a^2 + y^2$, so $g'(y) = 2y$. Thus we have $g'(b) = 0$ if and only if $b = 0$. The points on the circle with $b = 0$ are $(\pm 1, 0)$, which are exactly the points near which the circle *cannot* be represented as the graph of a function $y = f(x)$. Near every other point on the circle, the theorem applies.

[26] A precise definition of "continuously differentiable" for functions of two variables involves more technicalities than are appropriate for this course. Basically it means that near (a, b), there is a good linear approximation $F(x, y) \approx F(a, b) + \ell(x - a) + m(y - b)$ for some $\ell, m \in \mathbb{R}$. Most 'nice' functions you encounter will have this property.

Example 19.8. With $F(x,y) = x^3 + y^3 - xy$, for a given a we have $g(y) = a^3 + y^3 - ay$, so $g'(y) = 3y^2 - a$.

Lecture 20 Inverse functions

This lecture corresponds to §§3.5–3.6 in Stewart and Chapter 12 in Spivak.

We have encountered several functions (e^x, $\sin x$, $\cos x$) whose inverses ($\ln x$, $\arcsin x$, $\arccos x$) are important functions in their own right. In particular, it would be helpful to be able to compute derivatives of the inverse functions in terms of the original functions. Implicit differentiation suggests a solution: if f^{-1} is differentiable at $b = f(a)$ and f is differentiable at a, then differentiating both sides of $y = f(f^{-1}(y))$ with respect to y (using the chain rule) and evaluating at $y = b$ gives

$$1 = (f^{-1})'(b)f'(a) \quad \Rightarrow \quad (f^{-1})'(b) = \frac{1}{f'(a)},$$

which can be rewritten as

$$(f^{-1})'(y) = \frac{1}{f'(f^{-1}(y))}. \tag{20.1}$$

There is one caveat. As with our previous applications of implicit differentiation, in order for this to be valid we need to know that f^{-1} is differentiable at $b = f(a)$. This can be guaranteed by the Implicit Function Theorem: we want $x = f^{-1}(y)$ to satisfy $f(x) - y = 0$, so we take $F(x,y) = f(x) - y$. Now the Implicit Function Theorem says that x is a differentiable function of y near the values $y = b$, $x = a$ as long as the derivative $\frac{d}{dx}(f(x) - y)|_{x=a}$ is nonzero.[27] This derivative is $f'(a)$, and so we see that f^{-1} is differentiable at $b = f(a)$ whenever $f'(a) \neq 0$, and in this case $(f^{-1})'(b) = 1/f'(a)$.

This formula is easy to remember if we write it in the form

$$\frac{dx}{dy} = \frac{1}{\frac{dy}{dx}}.$$

Once again, though, we emphasize that $\frac{dy}{dx}$ and $\frac{dx}{dy}$ are not fractions, so this relationship is a theorem that makes our notation reasonable, rather than a simple consequence of how fractions work.

20.1. Inverse trigonometric functions

Since we proved that $\frac{d}{dx}\sin x = \cos x$, it follows from the formula for derivatives of inverse functions that

$$\frac{d}{dx}\sin^{-1} x = \frac{1}{\cos(\sin^{-1} x)}.$$

[27] The hypothesis in the Implicit Function Theorem placed a requirement on the derivative with respect to y. But that was because the theorem was set up to find y as a function of x. Here we are doing the reverse, and the general rule is that we need to check that the derivative *with respect to the desired dependent variable* is nonzero.

This can be simplified further by observing that if $t = \sin^{-1} x$, then $-\frac{\pi}{2} \le t \le \frac{\pi}{2}$ and $\sin t = x$, so $\cos t = \sqrt{1 - \sin^2 t} = \sqrt{1 - x^2}$, and we deduce that

$$\frac{d}{dx} \sin^{-1} x = \frac{1}{\sqrt{1 - x^2}}. \tag{20.2}$$

Exercise 20.1. Mimic the above computation to prove that $\frac{d}{dx} \cos^{-1} x = -1/\sqrt{1 - x^2}$.

We can also use implicit differentiation directly: for example, if $y = \tan^{-1} x$, then $x = \tan y(x)$, and differentiating both sides with respect to x gives

$$1 = (\sec^2 y)\frac{dy}{dx} \quad \Rightarrow \quad \frac{dy}{dx} = \frac{1}{\sec^2 y} = \frac{1}{1 + \tan^2 y} = \frac{1}{1 + x^2}.$$

20.2. Logarithms

Given $a > 0$, the function $y = \log_a x$ is the inverse of $x = a^y$, and differentiating gives

$$1 = (\ln a)a^y \frac{dy}{dx} = x \ln a \frac{dy}{dx},$$

so we conclude that

$$\frac{d}{dx} \log_a x = \frac{1}{x \ln a}, \tag{20.3}$$

and in particular,

$$\frac{d}{dx} \ln x = \frac{1}{x}. \tag{20.4}$$

This has many important consequences. For example, we can now prove the power rule stated in Theorem 14.7: given any $\beta \in \mathbb{R}$, the function $f\colon (0, \infty) \to \mathbb{R}$ defined by $f(x) = x^\beta$ can be rewritten as $f(x) = e^{\beta \ln x}$, and then the chain rule together with the rules for differentiating exponentials and logarithms gives

$$f'(x) = \Big(\frac{d}{dx}(\beta \ln x)\Big) e^{\beta \ln x} = \frac{\beta}{x} e^{\beta \ln x} = \frac{\beta}{x} x^\beta = \beta x^{\beta - 1}.$$

Example 20.2. If g is a differentiable positive function, then the chain rule gives

$$\frac{d}{dx} \ln g(x) = \frac{g'(x)}{g(x)}. \tag{20.5}$$

The expression on the right-hand side of (20.5) is called the *logarithmic derivative* of g.

Example 20.3. If $y = \ln(x^2 + 3x + 4)$, then

$$\frac{dy}{dx} = \frac{\frac{d}{dx}(x^2 + 3x + 4)}{x^2 + 3x + 4} = \frac{2x + 3}{x^2 + 3x + 4}.$$

Example 20.4.

$$\frac{d}{dx} \ln(\cos x) = \frac{-\sin x}{\cos x} = -\tan x.$$

Example 20.5. The function $g\colon (-\infty,0) \to (0,\infty)$ defined by $g(x) = -x = |x|$ has $g'(x) = -1$, so

$$\frac{d}{dx}\ln g(x) = \frac{g'(x)}{g(x)} = \frac{-1}{-x} = \frac{1}{x},$$

which is the same formula as for $\frac{d}{dx}\ln x$ when $x > 0$, and we conclude that

$$\frac{d}{dx}\ln|x| = \frac{1}{x} \text{ for all } x \neq 0. \tag{20.6}$$

As with previous methods we introduced, we now have multiple options for how to evaluate certain derivatives.

Example 20.6. Logarithmic differentiation and the quotient rule give

$$\frac{d}{dx}\ln\Big(\frac{x+1}{x-1}\Big) = \frac{1}{\frac{x+1}{x-1}}\cdot\Big(\frac{(x-1)-(x+1)}{(x-1)^2}\Big) = \frac{-2}{(x+1)(x-1)} = \frac{2}{1-x^2}.$$

We could also compute this by using properties of the logarithm before taking the derivative:

$$\frac{d}{dx}\ln\Big(\frac{x+1}{x-1}\Big) = \frac{d}{dx}\big(\ln(x+1)-\ln(x-1)\big) = \frac{1}{x+1}-\frac{1}{x-1} = \frac{(x-1)-(x+1)}{(x+1)(x-1)} = \frac{2}{1-x^2}.$$

Even when the original problem does not include a logarithm, it is sometimes easier to evaluate the derivative by using logarithms instead of the potentially messier product and quotient rules.

Example 20.7. If $y = x^{2/3}\sqrt{x^2+2}/(2x+1)^3$, then we can evaluate $\frac{dy}{dx}$ by observing that

$$\ln y = \frac{2}{3}\ln x + \frac{1}{2}\ln(x^2+1) - 3\ln(2x+1),$$

and differentiating gives

$$\frac{1}{y}\frac{dy}{dx} = \frac{2}{3x} + \frac{x}{x^2+1} - \frac{6}{2x+1},$$

which we solve to get

$$\frac{dy}{dx} = \frac{x^{2/3}\sqrt{x^2+2}}{(2x+1)^3}\Big(\frac{2}{3x} + \frac{x}{x^2+1} - \frac{6}{2x+1}\Big).$$

We can also use logarithmic differentiation to address expressions such as $F(x) = f(x)^{g(x)}$, where both the base and the power are functions of x, so neither the power rule nor exponential differentiation are sufficient on their own. By taking logarithms we get

$$\ln F(x) = g(x)\ln f(x) \quad \Rightarrow \quad \frac{F'(x)}{F(x)} = g'(x)\ln f(x) + g(x)\frac{f'(x)}{f(x)},$$

and thus

$$F'(x) = g'(x)f(x)^{g(x)}\ln f(x) + g(x)f'(x)f(x)^{g(x)-1}. \tag{20.7}$$

Note that the first term behaves like an exponential derivative (the original expression is still there, multiplied by the rate of change in the exponent and by the natural logarithm of the base), while the second behaves like the power rule (the power gets decreased by 1

and we multiply by the original power, as well as by a factor that comes from the chain rule). Indeed, if $g(x) = \beta$ in (20.7) is constant, then the equation reduces to the rule

$$\frac{d}{dx} f(x)^\beta = \beta f'(x) f(x)^{\beta-1},$$

and if $f(x) = a$ is constant, then we get

$$\frac{d}{dx} a^{g(x)} = (\ln a) g'(x) a^{g(x)}.$$

We can also use the formula for $\frac{d}{dx} \ln x$ to recover an alternate expression for the natural logarithmic base e. On the one hand, $f(x) = \ln x$ has $f'(1) = \frac{1}{1} = 1$. On the other hand,

$$f'(1) = \lim_{x\to 0} \frac{f(1+x) - f(1)}{x} = \lim_{x\to 0} \frac{\ln(1+x)}{x} = \lim_{x\to 0} \ln\left((1+x)^{1/x}\right).$$

Since the exponential function is continuous, we get

$$e = e^1 = e^{f'(1)} = \lim_{x\to 0} e^{\ln((1+x)^{1/x})} = \lim_{x\to 0} (1+x)^{1/x}.$$

In particular, since $\frac{1}{n} \to 0$ as $n \to \infty$, this proves that

$$e = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n. \tag{20.8}$$

Lecture 21 Rates of change in sciences

This lecture corresponds to §3.7 in Stewart

In this lecture we briefly discuss various places in physics, chemistry, and biology where rates of change, and hence derivatives, naturally appear. Given more time, one could extend this discussion further to include other sciences and many more applications.

21.1. Physics

21.1.1. *Velocity and acceleration*

Consider an object moving along a straight line. If $s = s(t)$ represents the *position* of the object at time t, then the rate of change of the object's position is its *velocity* $v = \frac{ds}{dt} = s'$. This is also sometimes denoted $\dot{s}$. As we discussed when we first defined the derivative, the average velocity from time t_1 to time t_2 is $\frac{s(t_2)-s(t_1)}{t_2-t_1}$, and the (instantaneous) velocity v is the limit of this quantity as $t_2 \to t_1$.

Differentiating again gives the rate at which velocity is changing, which is is the *acceleration* $a = \frac{dv}{dt} = \frac{d^2s}{dt^2} = v' = s'' = \dot{v} = \ddot{s}$.

21.1.2. *Density*

Consider a straight rod whose density may vary from point to point. Given a point on the rod, let x denote the distance from the beginning of the rod to this point, and let $m(x)$ denote the mass of that part of the rod. Then the mass of the part of the rod lying

between points x_1 and x_2 is $m(x_2)-m(x_1)$, and the *average density* of this part of the rod is $\frac{m(x_2)-m(x_1)}{x_2-x_1}$. Taking a limit as $x_2 \to x_1$ gives the *linear density* $\rho = \rho(x) := \frac{dm}{dx} = m'(x)$.

21.1.3. *Electricity*

Consider an electric current passing through a wire. Fix a cross-section of the wire, and let $Q = Q(t)$ be the total amount of electrical charge that has passed through that cross-section by time t. Then $\frac{\Delta Q}{\Delta t} = \frac{Q(t_2)-Q(t_1)}{t_2-t_1}$ gives the *average current* between time t_1 and time t_2, and $I = I(t) = \frac{dQ}{dt}$ gives the *current* at time t.

21.2. Chemistry

21.2.1. *Rate of reaction*

Suppose we have a chemical reaction in which two substances A and B combine to form a third substance C. Let $[A]$, $[B]$, $[C]$ denote the *concentrations* of these substances at a given time t; this is measured in moles (a unit for number of molecules) per liter. The *average rate of reaction of* C is $\frac{\Delta[C]}{\Delta t} = \frac{[C](t_2)-[C](t_1)}{t_2-t_1}$. Taking the limit as $\Delta t \to 0$ gives the *instantaneous rate of reaction of* C, $\frac{d[C]}{dt}$; this is the rate at which C is being produced. The rates of reaction for A and B are defined similarly.

Suppose that 2 molecules of A and 1 molecules of B are required to produce each molecules of C. Then we write the reaction as $2A + B \to C$, and we have $\frac{d[A]}{dt} = -2\frac{d[C]}{dt}$ and $\frac{d[B]}{dt} = -\frac{d[C]}{dt}$.

21.2.2. *Compressibility*

Let V denote the volume of a certain quantity of gas, and P the pressure at which the gas is held. Then V is determined by P, and in particular we can consider the rate of change of V with respect to P, which is given by the derivative $\frac{dV}{dP}$. It is reasonable to expect that increasing the pressure will cause the volume to decrease by an amount proportional to V; if a given pressure increase causes 10 liters of gas to compress by 1 liter, then we expect that 20 liters would compress by 2 liters under the same pressure increase. Thus the relevant quantity is the rate of change of V per unit V, and with this in mind we define the *isothermal*[28] *compressibility* of the gas to be $\beta := -\frac{1}{V}\frac{dV}{dP}$. Note that this is equal to $-\frac{d}{dP}\log V$. The negative sign appears because we measure 'compressibility', which is the rate at which the volume compresses (decreases) as pressure increases.

21.3. Biology

21.3.1. *Blood flow*

Under certain assumptions,[29] a fluid flowing through a cylindrical tube, such as blood flowing through a vein, flows according to *Poiseuille's law of laminar flow*, which says that if R is the radius of the tube, η the viscosity of the fluid, P the difference in pressure

[28] 'Isothermal' refers to the fact that we are considering a gas held at a constant temperature.

[29] The fluid should be incompressible and Newtonian, the flow should be laminar rather than turbulent, and the tube should be substantially longer than it is wide.

between the two ends of the tube, and ℓ the length of the tube, then a particle of fluid at a distance r from the center of the tube flows with velocity

$$v = \frac{P}{4\eta\ell}(R^2 - r^2). \tag{21.1}$$

To use this equation to determine the overall rate at which fluid is flowing through a given cross-section, we would need to use the theory of *integration*, which we will develop near the end of the course. For the time being we point out that the sensitivity of velocity to changes in the distance from the center is given by the *velocity gradient* $\frac{dv}{dr} = -\frac{Pr}{2\eta\ell}$.

21.3.2. *Population growth*

Consider a population that changes over time, with $n = n(t)$ giving the size of the population at time t. Technically n should only take integer values, since the number of individuals in a population is always a whole number; however, when the population is large enough, it is useful to approximate the true population with a differentiable function of time, to which the tools of calculus can be applied. The average rate of growth $\frac{\Delta n}{\Delta t}$ makes sense in either case, but if we allow n to take non-integer values then it is reasonable to also consider the instantaneous rate of growth $\frac{dn}{dt}$.

As an important example, suppose we have a colony of bacteria that reproduces at such a rate that its population doubles every hour. Writing n_0 for the initial population at time 0, and $n(t)$ for the population at time t, we see that

$$n(1) = 2n_0, \quad n(2) = 2^2 n_0, \quad n(3) = 2^3 n_0, \quad \dots \quad n(k) = 2^k n_0.$$

If we use the same formula for times t that are not whole multiples of an hour, we see that the population formula $n(t) = 2^t n_0$ satisfies the given growth condition. With this formula, the rate of growth of the population is

$$\frac{dn}{dt} = (\ln 2)2^t n_0 = (\ln 2)n(t).$$

This situation, where the rate of growth of a quantity is proportional to the quantity itself, leads to exponential growth or decay, and we will examine this phenomenon in more detail shortly. For the moment we observe that if this were to really happen, and if we were to begin with just a single bacteria with mass $\approx 10^{-12}$ g, then after 50 hours (a little over 2 days) the total population would be $2^{50} \approx (10^3)^5 = 10^{15}$ (here we use the approximation $2^{10} = 1024 \approx 10^3$) and thus the total mass would be $10^{15} \cdot 10^{-12}$ g $= 10^3$ g $= 1$ kg. After 80 more hours (130 hours total), the mass would be $2^{80} = (2^{10})^8 \approx (10^3)^8 = 10^{24}$ kg. The mass of the earth is $\approx 6 \times 10^{24}$ kg, so after 133 hours of exponential growth the total mass of the bacteria ($\approx 8 \times 10^{24}$ kg) would exceed that of the earth. Clearly this does not happen in practice, which illustrates the limitations of the exponential growth model; it is valid only as long as there are sufficient resources to permit unconstrained growth. Next semester when we study differential equations we will examine some more realistic models.

Lecture 22 Exponential growth and decay

This lecture corresponds to §3.8 in Stewart (related rates are in §3.9)

22.1. Solving a differential equation

As we saw in the previous lecture, the simplest model for population growth leads to the *differential equation*

$$\frac{dy}{dt} = ky, \tag{22.1}$$

where $y = y(t)$ is the population at time t, and $k > 0$ is a constant parameter that determines the rate of growth. We know that $y = e^{kt}$ is a solution, and so is $y = Ce^{kt}$ for every $C \geq 0$, since

$$\frac{d}{dt}(Ce^{kt}) = C\frac{d}{dt}e^{kt} = Cke^{kt}.$$

Note that $y(0) = Ce^{k\cdot 0} = Ce^0 = C$, so, we can write this as

$$y(t) = y(0)e^{kt}. \tag{22.2}$$

There are two natural questions to ask.

(1) Suppose we didn't know in advance that (22.2) is a solution of (22.1). How could we find the solution?
(2) Is this the only solution? Or could there be another function $y(t)$ that also satisfies (22.1) but is not given by (22.2)?

To address the first question, we can divide both sides of (22.1) by y and obtain $y'/y = k$; since the left-hand side is the logarithmic derivative of y, we get

$$\frac{d}{dt}\ln y = k.$$

This is easier to solve: indeed, we know that given any $\ell \in \mathbb{R}$, the function $f(t) = kt + \ell$ has the property that $f'(t) = k$ for every t. Thus $\ln y = kt + \ell$ gives a solution of (22.1), and exponentiating both sides gives $y(t) = e^{kt}e^{\ell}$. But is it the only solution? Note that if $f'(t) = k$ for every t, then $g(t) := f(t) - kt$ has the property that $g'(t) = f'(t) - k = 0$ for every t. Our intuition strongly suggests that in order to have zero derivative at every point, g must be a constant function. And indeed, this is true.

Theorem 22.1. *If $g\colon (a,b) \to \mathbb{R}$ is differentiable and $g'(x) = 0$ for every $x \in (a,b)$, then g is constant on (a,b).*

It is easiest to prove Theorem 22.1 as a corollary of the Mean Value Theorem, which we will prove later on. It is possible to prove it now using bisection sequences, and we include this proof here for completeness.

Proof of Theorem 22.1. We start by proving that

$$\text{for every } \varepsilon > 0, \text{ we have } |g(x) - g(y)| \leq \varepsilon|x - y| \text{ for every } x, y \in (a,b). \tag{22.3}$$

This implies the conclusion of the theorem because if $x, y \in (a,b)$ were such that $g(x) \neq g(y)$, then taking $\varepsilon = \frac{1}{2}\frac{|g(x)-g(y)|}{|x-y|}$ would give $|g(x)-g(y)| = 2\varepsilon|x-y|$, contradicting (22.3).

To prove (22.3), fix $\varepsilon > 0$ and $a < x < y < b$. (The case $y > x$ is similar, and $y = x$ is automatic.) Aiming for a contradiction, suppose that $|g(x) - g(y)| > \varepsilon|x - y|$. Let m be the midpoint of x and y and observe that $|g(x) - g(y)| \le |g(x) - g(m)| + |g(m) - g(y)|$; moreover, $|x - m| = |m - y| = \frac{|x-y|}{2}$, so

$$\varepsilon < \frac{|g(x) - g(y)|}{|x - y|} \le \frac{1}{2}\Big(\frac{|g(x) - g(m)|}{|x - m|} + \frac{|g(m) - g(y)|}{|m - y|}\Big).$$

It follows that at least one of $\frac{|g(x)-g(m)|}{|x-m|}$ and $\frac{|g(m)-g(y)|}{|m-y|}$ must exceed ε. Replacing the interval (x, y) with whichever of the intervals (x, m) or (m, y) has this property, we can iterate this argument to produce two bisection sequences; more precisely, we write $b_0 = x$, $r_0 = y$, and observe that if $|g(r_n) - g(b_n)| > \varepsilon|r_n - b_n|$, then we can choose $b_{n+1}, r_{n+1} \in [b_n, r_n]$ such that

- one of b_{n+1}, r_{n+1} is the midpoint of $[b_n, r_n]$, and the other is at an endpoint of that interval;
- $|g(r_{n+1}) - g(b_{n+1})| > \varepsilon|r_{n+1} - b_{n+1}|$.

As in our previous proofs with bisection sequences, the sequences b_n and r_n converge to a common limit, which we denote by

$$c = \lim_{n\to\infty} b_n = \lim_{n\to\infty} r_n.$$

For every n, we have

$$|g(b_n) - g(r_n)| \le |g(b_n) - g(c)| + |g(c) - g(r_n)|$$

and

$$|b_n - r_n| \ge |b_n - c|, \quad |b_n - r_n| \ge |c - r_n|;$$

dividing gives

$$\varepsilon < \frac{|g(b_n) - g(r_n)|}{|b_n - r_n|} \le \frac{|g(b_n) - g(c)|}{|b_n - c|} + \frac{|g(c) - g(r_n)|}{|c - r_n|},$$

and thus there is $x_n \in \{b_n, r_n\}$ such that $|g(x_n) - g(c)|/|x_n - c| > \varepsilon/2$. Sending $n \to \infty$ we get

$$g'(c) = \lim_{n\to\infty} \frac{|g(x_n) - g(c)|}{|x_n - c|} \ge \frac{\varepsilon}{2} > 0,$$

contradicting the assumption that g' vanishes on the entire interval. This proves (22.3), and by the argument given at the start of the proof, this is enough to prove Theorem 22.1. □

Returning to the question of solutions of (22.1), we see that every solution $y = y(t)$ has the property that $g(t) = \ln(y(t)) - kt$ has zero derivative at all t. By Theorem 22.1, this implies that $\ln(y(t)) - kt = g(t) = g(0) = \ln(y(0))$ for all t, and exponentiating gives $y(t)e^{-kt} = y(0)$, demonstrating that (22.2) is the only solution of (22.1).

22.2. Carbon dating

The differential equation (22.1) and its solution (22.2) also arise in the process of *radioactive decay*. A carbon atom has 6 protons and 6 electrons, and can have either 6, 7, or 8 neutrons. Roughly 99% of carbon atoms have 6 neutrons, giving "carbon-12", and nearly all remaining carbon atoms have 7 neutrons, giving "carbon-13". Roughly one in every trillion carbon atoms has 8 neutrons, giving "carbon-14". The ratio is so low because this is an unstable isotope; given a sample of carbon with N_0 atoms of carbon-14, after ∼5,730 years about half of these atoms will have undergone radioactive decay and been transformed into nitrogen, so that the number of carbon-14 atoms remaining is $\frac{1}{2}N_0$. Extrapolating, we see that the number $N(t)$ of carbon-14 atoms after time t is

$$N(t) = \left(\frac{1}{2}\right)^{\frac{t}{5730}} N_0 = e^{(\frac{-\ln 2}{5730})t} N_0, \tag{22.4}$$

so that $N(t)$ obeys (22.1) and (22.2) with $k = -(\ln 2)/5730$. This is the calculation that leads to the process of "carbon dating": the proportion of carbon-14 in the atmosphere is relatively stable, because the exponential decay is balanced out by the production of carbon-14 in the upper atmosphere via cosmic rays, and thus as long as a plant or animal is alive, it exchanges carbon with its environment so that the proportion of carbon-14 in its body is stable. Once it dies, however, it no longer takes in new carbon-14 atoms, and the decay process begins. Thus if we wish to find the age of a particular sample of organic matter, we can measure its current carbon-14 content to determine $N(t)$, and then since N_0 is known (via the background carbon-14 level in the atmosphere), we can take logs in (22.4) and solve for t, obtaining.

$$\ln N(t) = kt + \ln N_0 \quad \Rightarrow \quad t = \frac{1}{k}(\ln N(t) - \ln N_0).$$

22.3. Newton's law of cooling

Suppose that an object has temperature $T(t)$ at time t, and that its surroundings have constant temperature T_s. Newton's law of cooling states that

$$\frac{dT}{dt} = k(T - T_s),$$

where k is a constant that depends on physical properties of the object and its surroundings. Note that $k < 0$ since we expect the object's temperature to decrease when $T > T_s$. Writing $y = T - T_s$, we get $y' = \frac{dT}{dt} = ky$, so y is a solution of (22.1), and thus (22.2) gives

$$T(t) = T_s + y(t) = T_s + y(0)e^{kt} = T_s + (T(0) - T_s)e^{kt}.$$

Note that this discussion makes some simplifying assumptions.

- We assume that the object is always at a single temperature throughout; of course in reality a spatially extended object can vary in temperature from point to point, and if this variation becomes significant then a more refined model is needed.
- We assume that the object's surroundings remain at a constant temperature even as they absorb the heat from the object (or give heat to the object). This is of course not true in the strictest sense, but as long as the surroundings are

sufficiently large relative to the object, it is a reasonable approximation to make. This assumption that the environment functions as a "heat bath" is a common one in the study of thermodynamics.

22.4. Compound interest

Suppose I have a bank account in the amount $A(t)$ at time t. Suppose also that I make no withdrawals or deposits, but that the money grows with annual interest rate r (here $r = 0.03$ would correspond to 3% interest). If interest is applied once per year, then after 1 year the amount is $A(1) = (1 + r)A(0)$, after 2 years it is $A(2) = (1 + r)^2 A(0)$, and in general after t years it is

$$A(t) = (1 + r)^t A(0).$$

If interest is applied twice per year, then every 6 months the amount of money in the account is multiplied by $(1 + \frac{r}{2})$, and we get

$$A(t) = \left(1 + \frac{r}{2}\right)^{2t} A(0).$$

In general, if interest is applied n times per year, then the amount is multiplied by $(1+\frac{r}{n})$ every time interest is applied, and after t years, interest has been applied nt times, so

$$A(t) = \left(1 + \frac{r}{n}\right)^{nt} A(0).$$

Recall from a homework assignment that $\lim_{n\to\infty}(1 + \frac{r}{n})^n = e^r$; thus in the limit as $n \to \infty$, we get

$$A(t) = e^{rt} A_0;$$

this is the case of *continuously compounded* interest.

Remark 22.2. The first three equations above have a small problem; they are only valid when nt is an integer, since otherwise the expression $(1+\frac{r}{n})^{nt}$ includes a pro-rated share of the next term's interest, which has not yet been applied.

Lecture 23 Related rates; linear approximation

This lecture corresponds to §§3.9–3.10 in Stewart

23.1. Related rates

It is often the case that we are interested in determining the rate of change of one quantity, while the information we are given is in terms of a different quantity; in this case we need to write down the function that relates the two, and apply the chain rule. For example, suppose we fill a balloon with air at the constant rate of 100 cm^3/s, and we want to determine the rate at which the radius is increasing when the diameter is 50 cm. Writing $V(t)$ for the volume at time t and $r(t)$ for the radius at time t, we have the following.

$$\text{Given quantity: } \frac{dV}{dt} \qquad \text{Desired quantity: } \frac{dr}{dt} \qquad \text{Relationship: } V = \frac{4}{3}\pi r^3$$

The last equation, which expresses volume as a function of radius, is valid as long as we assume that the balloon is spherical, and will be proved later on after we have studied integration. For now we simply take it as a given fact from geometry. Applying the chain rule and the power rule, we get

$$\frac{dV}{dt} = \frac{dV}{dr}\frac{dr}{dt} = (4\pi r^2)\frac{dr}{dt},$$

and solving for $\frac{dr}{dt}$ gives

$$\frac{dr}{dt} = \frac{1}{4\pi r^2}\frac{dV}{dt}.$$

When the diameter is 50 cm, the radius is 25 cm, so

$$\frac{dr}{dt}\Big|_{r=25} = \frac{1}{4\pi \cdot 625 \text{ cm}^2} \cdot 100 \text{ cm}^3/\text{s} = \frac{1}{25\pi} \text{ cm/s} \approx 0.0127 \text{ cm/s}.$$

Example 23.1. Suppose I have two rings that can slide along a straight track, which are connected by a string of length 20 m. I pull the middle of the string in a direction perpendicular to the track with a constant velocity v. How fast is the distance between the two rings decreasing when they are 10 m apart?

Solution: Let $x(t)$ denote the distance between the two rings, and let $y(t)$ denote the distance that I have pulled the middle of the string away from the track. Then we have

$$\text{Given quantity: } \frac{dy}{dt} \qquad \text{Desired quantity: } \frac{dx}{dt} \qquad \text{Relationship: } \left(\frac{x}{2}\right)^2 + y^2 = 10^2$$

where the last equation comes by looking at the right triangle with vertices at my hand, the midpoint of the rings, and one of the two rings. Using implicit differentiation and the chain rule we get

$$0 = \frac{d}{dt}10^2 = \frac{d}{dt}\Big(\frac{x^2}{4} + y^2\Big) = \frac{x}{2}\frac{dx}{dt} + 2y\frac{dy}{dt},$$

and solving for $\frac{dx}{dt}$ gives

$$\frac{dx}{dt} = -\frac{4y}{x}\frac{dy}{dt}.$$

When the rings are 10 m apart, we have $x = 10$ and $y = \sqrt{10^2 - 5^2} = \sqrt{75} = 5\sqrt{3}$, and we are given that $\frac{dy}{dt}$ takes the constant value v, so we conclude that at this moment we have

$$\frac{dx}{dt} = -\frac{4 \cdot 5\sqrt{3}}{10}v = -2\sqrt{3}v.$$

Thus the distance between the rings is decreasing at an instantaneous rate of $2\sqrt{3}v$ at the moment when they are 10 m apart.

Example 23.2. Consider a conical tank whose top is a circle of radius 2 m and whose height is 5 m. Suppose that water drains from the bottom of the tank at a rate of 10 L/min. How fast is the height of the water decreasing when the tank is half empty?

Step 1: Identify the quantities involved and assign notation. Let V be the volume of water remaining at time t, h the height of the top of the water (relative to the lowest point of the cone), and r the radius of the circle formed by the top of the water.

Step 2: Identify the given information and the desired information in terms of derivatives. We are given that the water is draining at a rate of 10 L/min, so $\frac{dV}{dt} = -10$ L/min.

The rate at which the height of the water is changing is $\frac{dh}{dt}$. We want the rate at which it is *decreasing*, so our desired quantity is $-\frac{dh}{dt}$. (If the derivative is negative, then the height is decreasing and this number will be positive.)

Step 3: Write down formulas relating the various quantities. The water is in the shape of a cone with height h whose base is a circle of radius r: recalling a formula from geometry,[30] the volume of this cone is $V = \frac{1}{3}\pi r^2 h$. We also need to relate r and h; their ratio is the same as the ratio of the radius of the tank to its height, so $\frac{r}{h} = \frac{2}{5}$. Writing this as $r = \frac{2}{5}h$, we can relate V and h by

$$V = \frac{4}{75}\pi h^3. \tag{23.1}$$

Step 4: Differentiate using the chain rule. Differentiating (23.1) gives

$$\frac{dV}{dt} = \frac{4}{25}\pi h^2 \frac{dh}{dt}.$$

Step 5: Substitute the given information to obtain the solution. We are given that $\frac{dV}{dt} = -10$ L/min, so

$$\frac{dh}{dt} = \frac{25}{4\pi h^2}\frac{dV}{dt} \text{ L/min} = -\frac{250}{4\pi h^2} \text{ L/min}. \tag{23.2}$$

To obtain the final solution we need to know the value of h at the desired instant. When half the water is gone, the volume of the water is half the volume of the tank, so

$$\underbrace{\frac{4}{75}\pi h^3}_{\text{volume of water}} = \frac{1}{2} \cdot \underbrace{\frac{4}{75}\pi 5^3}_{\text{volume of tank}} \quad \Rightarrow \quad h^3 = \frac{5^3}{2} \quad \Rightarrow \quad h = \frac{5}{\sqrt[3]{2}} \text{ m}.$$

Together with (23.2) this gives

$$\frac{dh}{dt} = -\frac{250}{4\pi} \cdot \frac{2^{2/3}}{5^2} \text{ L/min/m}^2 = -\frac{5}{\pi\sqrt[3]{2}} \text{ L/min/m}^2.$$

We are not quite done, because we need to put our answer in the correct units. The volume was given in liters, and we have 1 L = 1000 cm^3. Note that 1 m^3 = 100^3 cm^3 = 10^6 cm^3, so 1 L = 10^{-3} m^3. We conclude that at the moment the tank is half full, we have

$$\frac{dh}{dt} = -\frac{5}{\pi\sqrt[3]{2}} \cdot 10^{-3} \text{ m}^3/\text{min/m}^2 = -\frac{1}{200\pi\sqrt[3]{2}} \text{ m/min} = -\frac{1}{2\pi\sqrt[3]{2}} \text{ cm/min}.$$

Thus the water level is falling at a rate of $\frac{1}{2\pi\sqrt[3]{2}} \approx 0.126$ cm/min.

23.2. Linear approximation and differentiating inverse functions

We have already mentioned several times that the derivative of a function gives a linear approximation to that function via the tangent line. This is made precise by the following.

Proposition 23.3. *Consider a function $f\colon I \to \mathbb{R}$, where I is an interval. Given a point a in the interior of I, a real number $m \in \mathbb{R}$ is the derivative of f at a if and only if there is a function $r\colon I \to \mathbb{R}$ such that*

[30] When we study integrals later, we will see how to obtain this formula directly.

- $f(x) = f(a) + m(x-a) + r(x)$ *for all* $x \in I$, *and*
- $\lim\limits_{x\to a} \dfrac{r(x)}{x-a} = 0$.

Proof. If f is differentiable at a, then writing $m := f'(a)$, the function $r(x) := f(x) - (f(a) + f'(a)(x-a))$ satisfies the first condition by definition, and moreover

$$\lim_{x\to a} \frac{r(x)}{x-a} = \lim_{x\to a} \frac{f(x) - (f(a) + f'(a)(x-a))}{x-a} = \lim_{x\to a} \frac{f(x)-f(a)}{x-a} - f'(a) = 0$$

by the definition of derivative. Conversely, if there is a function r satisfying the two conditions given, then

$$\lim_{x\to a} \frac{f(x)-f(a)}{x-a} = \lim_{x\to a} \frac{m(x-a)+r(x)}{x-a} = \lim_{x\to a} m + \frac{r(x)}{x-a} = m,$$

and thus $m = f'(a)$. □

Remark 23.4. Later, in multivariable calculus, the characterization of derivative in Proposition 23.3 turns out to be the best way to define derivative for functions of multiple variables. For a function of n variables, $x - a$ is a vector with n components, not a real number, and the derivative is an $n \times n$ *matrix*.

Remark 23.5. The function $r(x)$ can be thought of as a 'remainder term' that captures the difference between $f(x)$ and the linear approximation given by the tangent line at a. We will encounter generalizations of this later on when we study Taylor series. A simple first step in this direction would be to ask for the *quadratic* function that gives the "best fit" to $f(x)$ at the point a. The best linear approximation was the function $L(x) = c_0 + c_1 x$ such that $L(a) = f(a)$ and $L'(a) = f'(a)$, so similarly the best quadratic approximation should be the function $Q(x) = c_0 + c_1 x + c_2 x^2$ such that $Q(a) = f(a)$, $Q'(a) = f'(a)$, and $Q''(a) = f''(a)$. Rewriting as $Q(x) = A + B(x-a) + C(x-a)^2$ we see that

$$Q(a) = A, \qquad Q'(x) = B + 2C(x-a) \Rightarrow Q'(a) = B, \qquad Q''(x) = 2C.$$

Thus $A = Q(a)$, $B = Q'(a)$, and $C = \frac{1}{2}Q''(a)$, so the best quadratic approximation to $f(x)$ near a is

$$Q(x) = f(a) + f'(a)(x-a) + \frac{1}{2} f''(a)(x-a)^2.$$

This is the *quadratic Taylor polynomial*.

We can use Proposition 23.3 to study derivatives of inverse functions; in particular, to show that they exist (and thus are given by (20.1)) *without* relying on the Implicit Function Theorem from Lecture 19 (which after all we did not really prove). Suppose that $f\colon I \to \mathbb{R}$ is injective, continuous, and that $f'(a)$ exists and is not equal to 0. By Theorem 11.6, f^{-1} is continuous. Given $y \approx b$, let $x = f^{-1}(y)$. By Proposition 23.3, we have

$$y = f(x) = f(a) + f'(a)(x-a) + r(x),$$

and solving for x gives

$$y - f(a) - r(x) = f'(a)(x-a) \quad \Rightarrow \quad f^{-1}(y) = x = a + \frac{y - f(a) - r(x)}{f'(a)}.$$

Since $a = f^{-1}(b)$ and $x = f^{-1}(y)$, we can rewrite this as

$$f^{-1}(y) = f^{-1}(b) + \frac{1}{f'(a)}(y-b) - \underbrace{\frac{r(f^{-1}(y))}{f'(a)}}_{R(y)}.$$

We want to apply Proposition 23.3 by proving that $R(y)/(y-b) \to 0$ as $y \to b$. Indeed, since f^{-1} is continuous at b, we have $x \to a$ as $y \to b$, and thus

$$\begin{aligned}\lim_{y\to b} \frac{R(y)}{y-b} &= \lim_{x\to a} \frac{r(x)}{f'(a)(f(x)-f(a))} = \lim_{x\to a} \frac{r(x)}{f'(a)\big(f'(a)(x-a)+r(x)\big)} \\ &= \lim_{x\to a} \frac{r(x)/(x-a)}{f'(a)^2 + f'(a)r(x)/(x-a)} = \frac{0}{f'(a)^2+0} = 0,\end{aligned}$$

where we have used the fact that $f'(a) \neq 0$. This proves that f^{-1} is differentiable at $b = f(a)$, and that

$$(f^{-1})'(b) = \frac{1}{f'(a)} = \frac{1}{f'(f^{-1}(b))}. \tag{23.3}$$

23.3. Using linear approximations

Linear approximations are useful when we only need a rough estimate for a function that is perhaps difficult to compute exactly. For example, if we want a reasonable estimate for $\sqrt{4.1}$, we might observe that $f(x) = \sqrt{x}$ has $f(4) = 2$ and $f'(x) = \frac{1}{2\sqrt{x}}$, so $f'(4) = \frac{1}{4}$; thus the linear approximation to f near 4 is

$$\sqrt{x} = f(x) \approx f(4) + f'(4)(x-4) = 2 + \frac{1}{4}(x-4).$$

With $x = 4.1$ we get

$$\sqrt{4.1} \approx 2 + \frac{1}{4}(.1) = 2.025.$$

In fact the first few digits of the true value are $2.02485\ldots$, so this is a reasonable approximation.

Note that we must be careful in how we use linear approximations. For $x = 4.1$ we got a reasonable approximation. But if we go too far, things break down. For example, at $x = 9$ we have $\sqrt{9} = 3$, but the linear approximation gives $2 + \frac{1}{4}(9-4) = 2 + \frac{5}{4} = 3.25$, which is not particularly close. It is a very important problem to understand how the error term between the linear approximation and the true value of the function can be controlled, and we will return to this near the end of next semester when we study Taylor polynomials.

Lecture 24 — Hyperbolic functions

This lecture corresponds to §3.11 in Stewart

The standard trigonometric functions are defined by considering the unit circle $x^2 + y^2 = 1$, and writing $\cos t$ and $\sin t$ for the x- and y-coordinates of the point on the circle

obtained by moving a distance t counterclockwise from $(1,0)$. Another useful way of describing this point is that if $\gamma(s)$ is a parametrization of the unit circle that moves counterclockwise and satisfies $\gamma(0) = 1$, and if $L(s)$ is the length of γ from 0 to s, then $(\cos t, \sin t) = \gamma(L^{-1}(t))$.

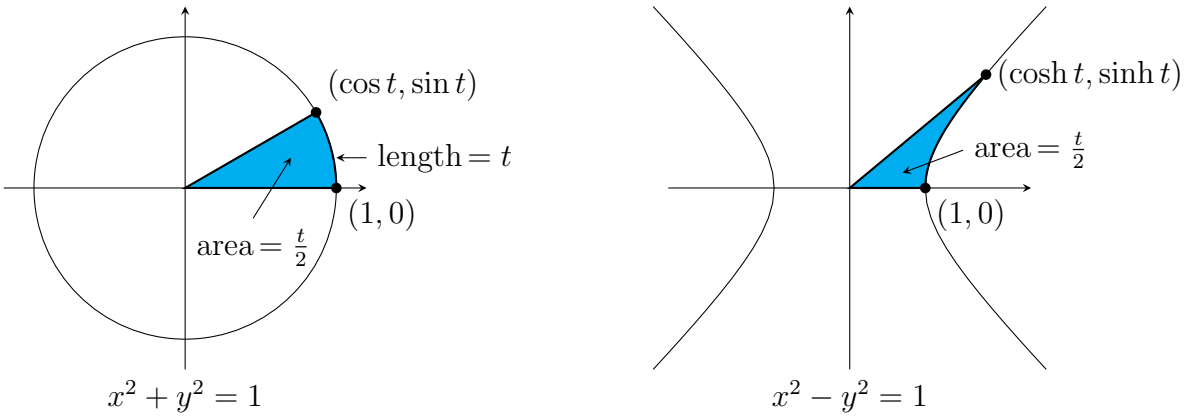


FIGURE 7. The circle and the hyperbola.

To define the hyperbolic functions we start by observing that $\cos t$ and $\sin t$ can also be characterized in terms of area. If we write $A(s)$ for the area swept out by the line from $(0,0)$ to $\gamma(s)$, as the parameter moves from 0 to s, then $(\cos t, \sin t) = \gamma(A^{-1}(t/2))$. In other words, $(\cos t, \sin t)$ is the point at which this line has swept out an area of $t/2$. Now replace the circle with the hyperbola $x^2 - y^2 = 1$ as shown in Figure 7. Let γ be a parametrization of this hyperbola such that $\gamma(0) = (1,0)$ and $\gamma(s)$ moves into the first quadrant as s increases. Again writing $A(s)$ for the area swept out by the line from $(0,0)$ to γ as the parameter moves from 0 to s, we define $(\cosh t, \sinh t) = \gamma(A^{-1}(t/2))$, so that $(\cosh t, \sinh t)$ is the point on the hyperbola at which this line has swept out an area of $t/2$.

When we first discussed arc length, we observed that it needs to be defined using some kind of limiting procedure. A similar statement is true for area; we know what is meant by "area of a rectangle" – length times height – but what is meant by "area of the region swept out by such-and-such a line"? To make this properly precise requires integration, and so we defer it until later; for the time being we merely observe that it is once again a limiting procedure, in which the region whose area we wish to determine is approximated by simpler regions (rectangles) whose area we know. Once we have developed the theory of integration, we will see that cosh and sinh admit the following formulas:

$$\cosh(t) = \frac{e^t + e^{-t}}{2}, \qquad \sinh(t) = \frac{e^t - e^{-t}}{2}. \tag{24.1}$$

Compare these to the formulas using $e^{\pm it}$ for cos and sin.

Because $(\cosh t, \sinh t)$ lies on the hyperbola $x^2 - y^2 = 1$, we obtain the fundamental identity

$$\cosh^2 t - \sinh^2 t = 1 \text{ for all } t \in \mathbb{R}, \tag{24.2}$$

which is analogous to the identity $\cos^2 t + \sin^2 t = 1$. Note that this identity can also be deduced directly from (24.1). We also see from (24.1), or from the geometric description, that sinh is an odd function, while cosh is even.

Example 24.1. Hyperbolic functions arise naturally in various applications. For example, if a cable is supported at its endpoints and hangs between them under its own weight (such as a power line, a suspension bridge, etc.), then it takes the shape of a *catenary*, which is a curve given as the graph of the function $y = c + a\cosh(x/a)$, where a, c are parameters determined by the physical characteristics of the situation. (We will prove this next semester.) Another example comes if we consider a wave of wavelength L propagating in water of depth d; the velocity of the wave is $\sqrt{\frac{gL}{2\pi}\tanh(\frac{2\pi d}{L})}$, where $g = 9.8\ m/s^2$ is the force of gravity, and $\tanh t = \sinh t/\cosh t$.

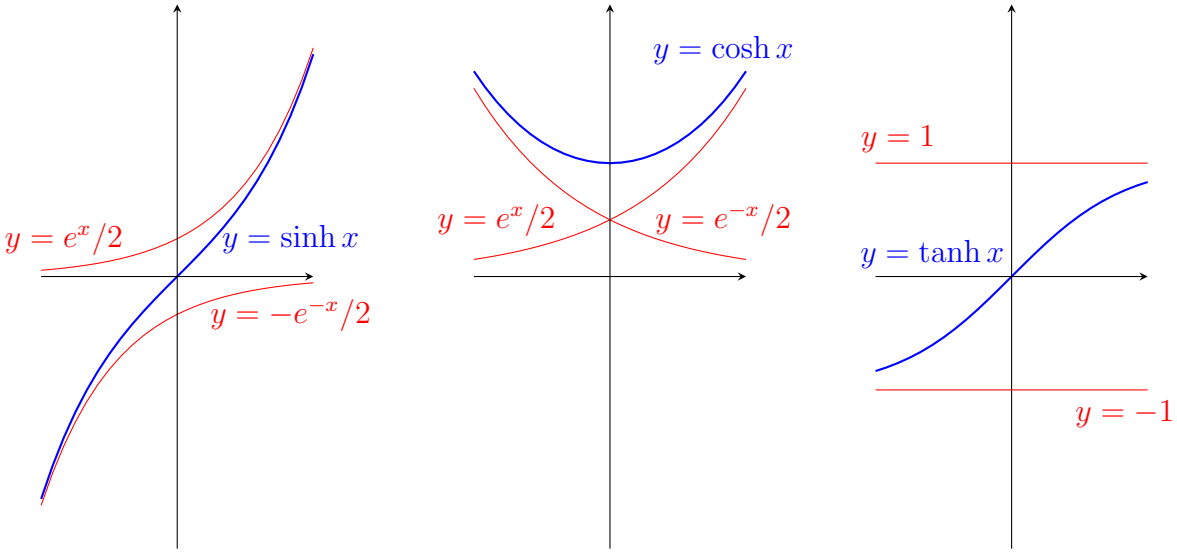


FIGURE 8. Hyperbolic functions.

The graphs of the functions cosh, sinh, and tanh are shown in Figure 8; the shapes of these graphs are relatively easy to deduce from (24.1). Note that the slope of sinh is positive at all points, and is equal to 1 as it goes through the origin; this is because

$$\frac{d}{dt}\sinh(t) = \frac{d}{dt}\frac{e^t - e^{-t}}{2} = \frac{e^t + e^{-t}}{2} = \cosh(t),$$

and $\cosh(0) = 1$. A similar observation applies to $\tanh(t)$, which we can justify either by differentiating $\tanh(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}}$, or more cleanly by observing that the linear approximations to sinh and cosh near $t = 0$ are given by

$$\sinh(t) \approx t, \qquad \cosh(t) \approx 1,$$

and thus $\tanh(t) \approx t$ for $t \approx 0$. Note also that

$$\frac{d}{dt}\cosh(t) = \frac{d}{dt}\frac{e^t + e^{-t}}{2} = \frac{e^t - e^{-t}}{2} = \sinh(t),$$

so that once again we recover equations very similar to the ones we are familiar with from trigonometric functions, but with some interchange of positive and negative signs.

One difference between trigonometric functions and hyperbolic functions is that the latter can be explicitly inverted. Indeed, if $y = \cosh^{-1}(x)$, where we use the branch $y \geq 0$ – note that cosh is invertible on $[0, \infty)$ and on $(-\infty, 0]$ – then we have

$$x = \cosh y = \frac{e^y + e^{-y}}{2} \Rightarrow e^y - 2x + e^{-y} = 0 \Rightarrow e^{2y} - 2xe^y + 1 = 0,$$

and the quadratic formula gives

$$e^y = x \pm \sqrt{x^2 - 1}.$$

For $y \geq 0$ we have $e^y \geq 1$, so we use the sum instead of the difference, and get

$$\cosh^{-1} x = y = \ln(x + \sqrt{x^2 - 1}). \tag{24.3}$$

Differentiating gives

$$\frac{d}{dx} \cosh^{-1} x = \frac{1}{x + \sqrt{x^2 - 1}} \Big(1 + \frac{2x}{2\sqrt{x^2 - 1}}\Big) = \frac{1}{x + \sqrt{x^2 - 1}} \frac{\sqrt{x^2 - 1} + x}{\sqrt{x^2 - 1}} = \frac{1}{\sqrt{x^2 - 1}}.$$

Alternately we could obtain this via implicit differentiation:

$$y = \cosh^{-1} x \Rightarrow \cosh y = x \Rightarrow (\sinh y) \frac{dy}{dx} = 1 \Rightarrow \frac{d}{dx} \cosh^{-1} x = \frac{1}{\sinh x} = \frac{1}{\sqrt{x^2 - 1}}.$$

Lecture 25 The Extreme Value Theorem

Stewart §4.1, Spivak Chapter 11

Now we start to look at applications of differentiation.

Definition 25.1. A function $f\colon D \to \mathbb{R}$ has an *absolute maximum* (or *global maximum*) at $c \in D$ if $f(c) \geq f(x)$ for all $x \in D$. The value $f(c)$ is called the *maximum value* of f.

Similarly, if $c \in D$ is such that $f(c) \leq f(x)$ for all $x \in D$, then f has an *absolute minimum* (or *global minimum*) at c, and $f(c)$ is called the *minimum value* of f.

The maximum and minimum values of a function are called its *extreme values*.

Many problems in science and engineering can be formulated as the search for the extreme points and/or values of some function; these are sometimes referred to as *optimization problems*.

First observe that global maxima and minima may or may not exist, and may or may not be unique.

Example 25.2. If $f(x) = \sin x$, then $-1 \leq f(x) \leq 1$ for all $x \in \mathbb{R}$. Moreover $f(x) = 1$ for all $x = \frac{\pi}{2} + 2n\pi$, $n \in \mathbb{Z}$, and so each of these points is a global maximum, and 1 is the maximum value. Similarly $f(x) = -1$ for all $x = -\frac{\pi}{2} + 2n\pi$, so each of these is a global minimum, and -1 is the minimum value.

Example 25.3. $f(x) = -x^2$ has a global maximum at $x = 0$ (the maximum value is 0), but no global minimum.

Example 25.4. $f(x) = x^3$ has no global maximum or minimum.

Example 25.5. $f(x) = \frac{1}{x}$ has no global maximum or minimum.

Example 25.6. $f(x) = \frac{1}{x^2+1}$ has a global maximum at 0 (the maximum value is 1), but no global minimum.

Note that the last two examples are *bounded below* in the sense that there is $M \in \mathbb{R}$ such that $f(x) \geq M$ for every x – such an M is called a *lower bound* for f – but because of their asymptotic behavior they never achieve a lower bound and so do not have global minima. This sort of behavior can be avoided by considering continuous functions on closed and bounded intervals.

Theorem 25.7 (Extreme Value Theorem)**.** *Let $f\colon [a,b] \to \mathbb{R}$ be a continuous function. Then f has a global maximum and minimum.*

Before proving the theorem we make a few observations.

- The maximum and minimum need not be unique; for example, if f is constant then every point is both a global maximum and a global minimum.
- The theorem doesn't give any information about how to actually find the maximum and minimum.
- The conclusion can fail if either hypothesis is violated; if $f(x) = x$ for $x \in (0,1)$, then f has no global maximum or minimum on $(0,1)$ – here the domain of f is not a closed bounded interval. If we define f like this on $(0,1)$ and then put $f(0) = f(1) = \frac{1}{2}$, we get a discontinuous function on $[0,1]$ that has no global maximum or minimum.

Proof of the EVT. It suffices to prove existence of a global maximum; then the result for a global minimum follows by finding a global maximum for $-f$. We break the proof into two halves, both of which we prove using bisection sequences.

(1) If $f\colon [a,b] \to \mathbb{R}$ is continuous, then it is bounded above.
(2) If $f\colon [a,b] \to \mathbb{R}$ is continuous and bounded above, then it has a global maximum.

For the first half, we argue by contradiction. Suppose that f is not bounded above on $[a,b]$, and construct a pair of bisection sequences using the question: "Is f bounded above on the interval $[a,m]$?" Here m is the midpoint of the current points in each sequence. Note that the answer is 'yes' for $m = a$ and 'no' for $m = b$. Thus we iteratively construct sequences

$$a = r_1 \leq r_2 \leq r_3 \leq \cdots \leq b_3 \leq b_2 \leq b_1 = b$$

with the following properties:

- f is bounded above on each interval $[a, r_n]$;
- f is not bounded above on $[a, b_n]$ for any n;
- the sequences r_n and b_n converge to a common limit $c \in [a,b]$.

By continuity there is $\delta > 0$ such that for every $x \in [a,b]$ with $|x - c| < \delta$, we have $|f(x) - f(c)| < 1$. In particular, for every $x \in (c-\delta, c+\delta)$, we have $f(x) \leq f(c) + 1$. Now we complete the proof of the first half as follows.

- Since $r_n \to c$, there is $n \in \mathbb{N}$ such that $r_n \in (c-\delta, c+\delta)$.

- By the definition of the sequences, f is bounded above on $[a, r_n]$, so there is $M \in \mathbb{R}$ such that $f(x) \leq M$ for all $x \in [a, r_n]$.
- For every $x \in [a, c+\delta)$, we either have $x \in [a, r_n]$, in which case $f(x) \leq M$, or $x \in (r_n, c+\delta) \subset (c-\delta, c+\delta)$, in which case $f(x) \leq f(c)+1$. Thus f is bounded above by $\max(M, f(c)+1)$ on $[a, c+\delta)$.
- Since $b_n \to c$, there is $n \in \mathbb{N}$ such that $b_n \in (c-\delta, c+\delta)$, and the previous step shows that f is bounded above on $[a, b_n]$.
- This contradicts the definition of the sequences, and we conclude that f is bounded above on $[a, b]$.

Now we carry out the second part of the proof. Let $M \in \mathbb{R}$ be an upper bound for f. Choose any $m \in \text{range}(f)$. Clearly $m \leq M$. If $m = M$ then there is $c \in [a, b]$ such that $f(c) = m$ (since m is part of the range), and moreover $f(x) \leq f(c)$ for all $x \in [a, b]$ (since $f(c) = M$ is an upper bound). So we consider the case $m < M$. Our goal is to find a value in $[m, M]$ that is simultaneously (1) in the range of f and (2) an upper bound for f.

To this end, construct a pair of bisection sequences

$$m = r_1 \leq r_2 \leq r_3 \leq \cdots \leq b_3 \leq b_2 \leq b_1 = M$$

via the rule "color the midpoint red if it is in the range of f, and blue otherwise", so that we have

- $r_n \in \text{range}(f)$ for every n;
- $b_n \notin \text{range}(f)$ for every n;
- the sequences r_n and b_n converge to a common limit $L \in [m, M]$.

We prove that L is an upper bound for f and that it is in the range of f.

Upper bound: Suppose L is not an upper bound for f. Then there is $x \in [a, b]$ such that $f(x) > L$. Since $b_n \searrow L$, there is n such that $b_n \in (L, f(x))$. But then $m < b_n < f(x)$, and since m is in the range of f, the Intermediate Value Theorem implies that b_n is also in the range of f, contradicting the definition of the sequences. Thus L is an upper bound for f.

In the range: Suppose L is not in the range of f. Then $g(x) := \frac{1}{L-f(x)}$ is a continuous function on $[a, b]$, so by the first half of the EVT it is bounded above by some $K \in \mathbb{R}$. But $g(x) \leq K$ is equivalent to $L - f(x) \geq \frac{1}{K}$, which is equivalent to $f(x) \leq L - \frac{1}{K}$. Since $r_n \nearrow L$, there is n such that $r_n \in (K - \frac{1}{L}, K)$, but since $r_n \in \text{range}(f)$ this means that there is $x \in [a, b]$ such that $f(x) = r_n > L - \frac{1}{K}$, contradicting boundedness of g. This contradition implies that L is in the range of f, so there is $c \in [a, b]$ such that $f(c) = L$. Since L is an upper bound for f, this means that c is a global maximum. $\square$

Lecture 26 Local extrema; Mean Value Theorem

Stewart §4.2, Spivak Chapter 11

26.1. Local extrema and Fermat's Theorem

Definition 26.1. A function $f\colon D \to \mathbb{R}$ has a *local maximum* (or *relative maximum*) at $c \in D$ if there exists $\delta > 0$ such that for all $x \in D \cap (c-\delta, c+\delta)$, we have $f(x) \leq f(c)$. Similarly, f has a *local maximum* (or *relative minimum*) at $c \in D$ if there exists $\delta > 0$ such that for all $x \in D \cap (c-\delta, c+\delta)$, we have $f(x) \geq f(c)$.

Remark 26.2. Stewart's definition of local extreme points differs slightly from the one above; he requires that a local maximum or minimum have the property that $(c-\delta, c+\delta) \subset D$ for some $\delta > 0$. In other words, he does not allow endpoints of intervals to be local maxima or minima. We follow Spivak's definitions instead.

Clearly a global maximum is a local maximum, and similarly for minima. The converse is not true.

Exercise 26.3. Prove that $f(x) = (x^2-1)^2$ has a local maximum at 0, but that this is not a global maximum.

Theorem 26.4 (Fermat's theorem)**.** *If $f\colon (a,b) \to \mathbb{R}$ has a local maximum or minimum at $c \in (a,b)$, and if the derivative $f'(c)$ exists, then $f'(c) = 0$.*

Proof. We give the proof for a local maximum. The other case is similar (or we can just consider $-f$). Since $f'(c)$ exists, the left and right derivatives of f at c exist and agree. Let $\delta > 0$ be such that every $x \in (c-\delta, c+\delta)$ satisfies $x \in (a,b)$ and $f(x) \leq f(c)$. Then for every $h \in (0, \delta)$ we have $f(c+h) \leq f(c)$, and thus $\frac{f(c+h)-f(c)}{h} \leq 0$. The monotonicity property of limits gives

$$f'(c) = D^+ f(c) = \lim_{h \to 0^+} \frac{f(c+h)-f(c)}{h} \leq 0.$$

Similarly, for every $h \in (-\delta, 0)$, we have $\frac{f(c+h)-f(c)}{h} \geq 0$ since the numerator is ≤ 0 and the denominator is < 0. Thus

$$f'(c) = D^- f(c) = \lim_{h \to 0^-} \frac{f(c+h)-f(c)}{h} \geq 0.$$

Since $f'(c)$ is simultaneously ≤ 0 and ≥ 0, we must have $f'(c) = 0$. □

Definition 26.5. We say that c is a *critical point* for f if $f'(c) = 0$ or if $f'(c)$ does not exist. If c is a critical point, then $f(c)$ is called a *critical value*.

Fermat's theorem can be reformulated in the following way.

Corollary 26.6. *If $f\colon [a,b] \to \mathbb{R}$ has a local maximum or minimum at $c \in [a,b]$, then either c is an endpoint of the interval, or c is a critical point for f.*

It is important to stress that the converse of this result is not true; critical points (and endpoints) need not be local maxima or minima. For example, $f(x) = x^3$ has a critical point at $x = 0$, but this is neither a local maximum nor a local minimum. Also, the case of nondifferentiability really does need to be included in the definition of critical point; $f(x) = |x|$ has $f'(x) = \pm 1$ wherever the derivative exists, and a local (and global) minimum at $x = 0$, where it is not differentiable.

Corollary 26.6 gives us a method for finding the global maxima and minima of a continuous function $f\colon [a,b] \to \mathbb{R}$:

(1) Find the critical points of f on (a, b).
(2) Compare the values of f at its critical points and at the endpoints a and b. This will often be a finite set of points, and the largest and smallest values from this list give the extreme values of f on $[a, b]$.

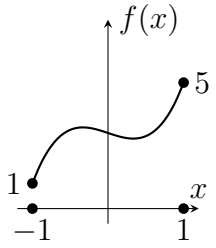
Example 26.7. Consider the function $f(x) = 3x^3 - x + 3$ on $[-1, 1]$. Because the interval is closed and the function is continuous, the method above will work to find the maxima and minima. To find the critical points, we compute

$$0 = f'(x) = 9x^2 - 1 \quad \Leftrightarrow \quad x^2 = \frac{1}{9} \quad \Leftrightarrow \quad x = \pm\frac{1}{3}.$$

Thus there are four points to check: the endpoints ± 1 and the critical points $\pm\frac{1}{3}$. We see that

- $f(-1) = 3(-1)^3 - (-1) + 3 = 1$;
- $f(1) = 3 - 1 + 3 = 5$;
- $f(-\frac{1}{3}) = 3(-\frac{1}{27}) - (-\frac{1}{3}) + 3 = -\frac{1}{9} + \frac{1}{3} + 3 = 3 + \frac{2}{9} = \frac{29}{9}$;
- $f(\frac{1}{3}) = 3(\frac{1}{27}) - \frac{1}{3} + 3 = 3 - \frac{2}{9} = \frac{25}{9}$.

Thus $f(-1) < f(\frac{1}{3}) < f(-\frac{1}{3}) < f(1)$, and we see that the global maximum is at $x = 1$ and the global minimum is at $x = -1$. The picture at right shows the shape of the graph; it appears that f has a local maximum at $-\frac{1}{3}$ and a local minimum at $\frac{1}{3}$. This can be verified by observing that on $[-1, 0]$, the only critical point is $-\frac{1}{3}$, and that $f(-\frac{1}{3}) > \max(f(-1), f(0))$; that $x = \frac{1}{3}$ is a local minimum can be proved similarly.



26.2. The Mean Value Theorem

Given a function $f\colon [a, b] \to \mathbb{R}$, the average rate of change of f over the entire interval is $\frac{f(b)-f(a)}{b-a}$. It is often useful to know that when f is differentiable, there is some point $c \in (a, b)$ at which the *instantaneous* rate of change $f'(c)$ is equal to the average rate of change; this is the content of the *Mean Value Theorem*. Before proving this general case of the theorem, we start with the case when $f(b) = f(a)$.

Theorem 26.8 (Rolle's Theorem)**.** *Let f be continuous on $[a, b]$ and differentiable on (a, b). If $f(a) = f(b)$, then there exists $c \in (a, b)$ such that $f'(c) = 0$.*

Proof. The following is almost a proof: "By the extreme value theorem, f has a global maximum at some point $x = c$. This global maximum is also a local maximum, so by Fermat's theorem it is a critical point. Since f is differentiable, we have $f'(c) = 0$."

The only thing wrong with this "proof" is that a global maximum might occur at an endpoint. But this can be dealt with as follows. If $c \in (a, b)$ then the above argument indeed gives $f'(c) = 0$. So suppose we have $c = a$ or $c = b$. We have $f(x) \le f(c)$ for all $x \in (a, b)$ by the definition of global maximum. If $f(x) = f(c)$ for all $x \in (a, b)$, then $f'(x) = 0$ for all $x \in (a, b)$ because constant functions have zero derivatives. On the other hand, if there is $x \in (a, b)$ such that $f(x) < f(c)$, then taking p to be the global minimum of f, we see that $p \in (a, b)$ since $f(a) = f(b) = f(c)$, and thus p is a critical point, with $f'(p) = 0$. □

Rolle's theorem can be used to prove that certain equations have a *unique* solution, even without finding an exact value.

Example 26.9. Consider the function $f(x) = x^5 + x^3 + x - 1$. We claim that there is exactly one real number r such that $f(r) = 0$. First note that $f(0) = -1$ and $f(1) = 2$, and that f is continuous, so by the Intermediate Value Theorem there exists $r \in [0, 1]$ such that $f(r) = 0$. Now suppose there exists $s \in \mathbb{R}$ such that $f(s) = 0$ and $s \neq r$. Then by Rolle's theorem, there exists c between r and s such that $f'(c) = 0$. However, we have $f'(x) = 5x^4 + 3x^2 + 1 \geq 1$ for all $x \in \mathbb{R}$, so no such s can exist.

Theorem 26.10 (Mean Value Theorem)**.** *Let f be continuous on $[a, b]$ and differentiable on (a, b). Then there exists $c \in (a, b)$ such that $f'(c) = \frac{f(b)-f(a)}{b-a}$, or equivalently, $f(b) - f(a) = f'(c)(b - a)$.*

Proof. The line between $(a, f(a))$ and $(b, f(b))$ has equation $y = f(a) + \frac{f(b)-f(a)}{b-a}(x - a)$. Let $h(x)$ denote the difference between $f(x)$ and this line, so

$$h(x) = f(x) - f(a) - \frac{f(b) - f(a)}{b - a}(x - a).$$

Observe that h is continuous on $[a, b]$ and differentiable on (a, b); moreover, $h(a) = h(b) = 0$, so Rolle's theorem gives $c \in (a, b)$ such that $h'(c) = 0$. Since

$$h'(x) = f'(x) - \frac{f(b) - f(a)}{b - a},$$

this implies that $f'(c) = \frac{f(b)-f(a)}{b-a}$, and completes the proof. □

If we are given bounds on f', we can use the Mean Value Theorem to bound f.

Example 26.11. If a differentiable function f satisfies $f(0) = 1$ and $f'(x) \leq 2$ for all x, then we can get an upper bound for $f(3)$ by applying the MVT to $[0, 3]$ to get some $c \in (0, 3)$ with

$$\frac{f(3) - f(0)}{3 - 0} = f'(c) \leq 2.$$

Multiplying by 3 gives $f(3) - f(0) \leq 6$, so $f(3) \leq f(0) + 6 = 7$.

We can use the MVT to give a short proof of Theorem 22.1, which said that if $f\colon (a, b) \to \mathbb{R}$ is differentiable and $f'(x) = 0$ for every $x \in (a, b)$, then f is constant on (a, b). Indeed, given any $x < y$ in (a, b), the MVT gives $c \in (x, y)$ such that

$$f(y) - f(x) = f'(c)(y - x) = 0(y - x) = 0 \quad \Rightarrow \quad f(y) = f(x).$$

This fact has the following important consequence.

Corollary 26.12. *If f and g are differentiable on (a, b) and have $f'(x) = g'(x)$ for every $x \in (a, b)$, then there exists $C \in \mathbb{R}$ such that $f(x) = g(x) + C$ for all $x \in (a, b)$.*

Proof. Let $F(x) = f(x) - g(x)$; then $F'(x) = f'(x) - g'(x) = 0$, so F is constant on (a, b) by Theorem 22.1. □

Example 26.13. It is crucial that the set of x on which we apply these results is a single open interval. The Heaviside function $H(x)$ defined by $H(x) = 0$ for $x < 0$ and $H(x) = 1$ for $x \geq 0$ has the property that $H'(x) = 0$ for all $x \neq 0$, but it is not constant on any interval that contains both positive and negative numbers.

Example 26.14. We saw in Lecture 22 that $\frac{dy}{dt} = ky$ has $y(t) = y(0)e^{kt}$ as a solution. In fact it is the *only* solution; any solution $y(t)$ must have

$$\frac{d}{dt}(\ln y(t) - kt) = \frac{y'}{y} - k = 0,$$

and thus $\ln y(t) - kt = \ln y(0)$ for all t, which gives the formula above.

Example 26.15. We can use Theorem 22.1 to prove that $\tan^{-1} x + \cot^{-1} x = \frac{\pi}{2}$ for all x; differentiating the left-hand side gives $\frac{1}{1+x^2} - \frac{1}{1+x^2} = 0$, and at $x = 1$ we have $\tan^{-1} 1 = \cot^{-1} 1 = \frac{\pi}{4}$. (One could also give a direct proof of this identity.)

Lecture 27 Shapes of graphs

Stewart §4.3, Spivak Chapter 11

27.1. Monotonicity and first derivatives

Back in Lecture 13.2, we observed that a function f with positive derivative f' should be increasing and a function with negative derivative should be decreasing. One can prove this directly from the definition of derivative, though we did not do so; instead we opted to wait until now, since with the MVT in hand we can give a quick proof.

Theorem 27.1. *Let f be continuous on $[a, b]$ and differentiable on (a, b). If $f'(x) > 0$ for every $x \in (a, b)$, then f is strictly increasing on $[a, b]$ (this means that $f(x_2) > f(x_1)$ whenever $x_2 > x_1$ for $x_1, x_2 \in [a, b]$). Similarly, if $f'(x) < 0$ for every $x \in (a, b)$, then f is strictly decreasing on $[a, b]$. If we replace $>$ and $<$ with $\geq$ and $\leq$, we get a similar result with "strictly increasing" replaced by "nondecreasing", and "strictly decreasing" replaced by "nonincreasing".*

Proof. Suppose $f'(x) > 0$ for every $x \in (a, b)$. Then given any $x_1 < x_2$ in $[a, b]$, the MVT gives $c \in (x_1, x_2)$ such that

$$f(x_2) - f(x_1) = f'(c)(x_2 - x_1) > 0.$$

Thus f is increasing. The proof for the case $f' < 0$ is the same, simply reverse the last inequality. □

Remark 27.2. This theorem is not a dichotomy; there are plenty of functions that do not satisfy either of the derivative conditions on a given interval, and that are neither increasing or decreasing. (Though they may be increasing or decreasing on a smaller interval.)

Remark 27.3. The converse of the theorem fails; a function can be strictly increasing even if its derivative vanishes somewhere. For example, $f(x) = x^3$ is strictly increasing but $f'(0) = 0$.

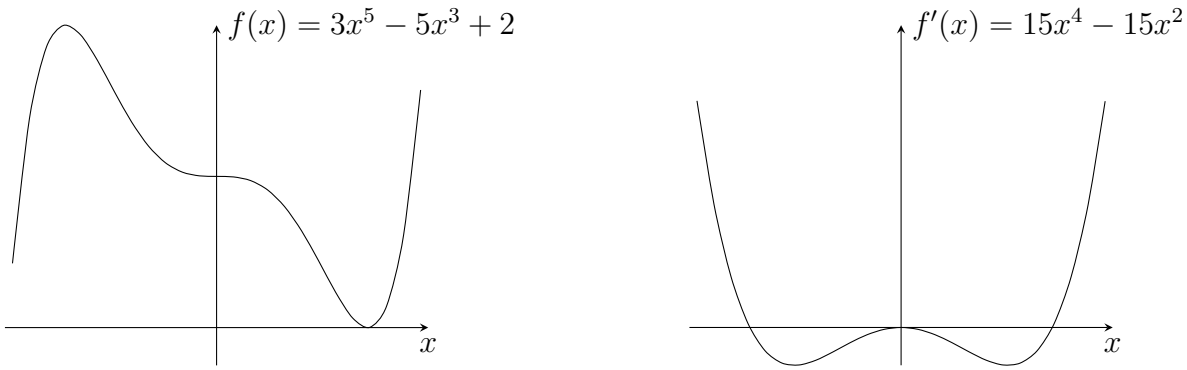


FIGURE 9. Connection between monotonicity of f and sign of f'.

Example 27.4. Consider the function $f(x) = 3x^5 - 5x^3 + 2$; the graphs of f and f' are shown in Figure 9. Observe that

$$f'(x) = 15x^4 - 15x^2 = 15x^2(x^2 - 1) = 15x^2(x - 1)(x + 1).$$

We identify the intervals on which f' is positive and negative by checking the sign of each of its factors between its zeros.

	$15x^2$	$x - 1$	$x + 1$	$f'(x)$
$x < -1$	$+$	$-$	$-$	$+$
$-1 < x < 0$	$+$	$-$	$+$	$-$
$0 < x < 1$	$+$	$-$	$+$	$-$
$x > 1$	$+$	$+$	$+$	$+$

Using this table and Theorem 27.1, we conclude that f is increasing on $(-\infty, 1]$ and $[1, \infty)$, while it is decreasing on $[-1, 0]$ and $[0, 1]$ (and thus in fact it is decreasing on $[-1, 1]$).

We see that -1 is a local maximum for f, because for $x < -1$ we have $f(x) < f(-1)$ since f is increasing on this interval, while for $x \in (-1, 0)$ we have $f(x) < f(-1)$ since f is decreasing on this interval. Similar reasoning shows that 1 is a local minimum. Observe that 0 is a critical point that is neither a local maximum nor a local minimum.

The reasoning in the last paragraph of the example is worth codifying.

Theorem 27.5 (First Derivative Test)**.** *Let f be differentiable on an interval (a, b) that contains a critical point c.*

(1) If f' changes from positive to negative at c ($f' > 0$ just to the left of c, and $f' < 0$ just to the right of c), then f has a local maximum at c.
(2) If f' changes from negative to positive at c ($f' < 0$ just to the left of c, and $f' > 0$ just to the right of c), then f has a local maximum at c.
(3) If f' does not change sign at c (either $f'(x) > 0$ for all $x \approx c$, or $f'(x) < 0$ for all $x \approx c$), then f has neither a local maximum nor a local minimum at c.

27.2. Convexity and second derivative

We also saw in Lecture 17 that the sign of f'' should be related to *convexity* properties of f. Now we make this precise. Recall from Definition 17.3 that a function f is *convex* on an interval I if for every $x, y \in I$ and every $t \in [0,1]$, we have

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

Given $t \in [0,1]$, let $z = tx + (1-t)y$ be the point at which f is evaluated to get the left-hand side. Note that when $t = 0$ we have $z = y$, and when $t = 1$ we have $z = x$, so z slides from y to x as t goes from 0 to 1. The right-hand side is equal to $f(y)$ when $t = 0$ and $f(x)$ when $t = 1$, and represents the point above z on the line joining $(x, f(x))$ to $(y, f(y))$. Thus a function is convex if its graph lies on or below this line between x and y, as illustrated in the left half of Figure 10.

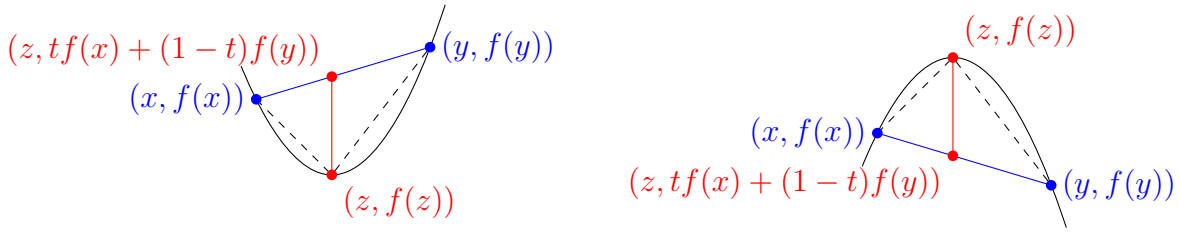


FIGURE 10. Convex (left) and concave (right).

Definition 27.6. A function f is *concave* on an interval I if for every $x, y \in I$ and every $t \in [0,1]$, we have

$$f(tx + (1-t)y) \geq tf(x) + (1-t)f(y).$$

A concave function is shown in the right half of Figure 10; it has the property that the graph of f lies above its secant line between x and y.

Remark 27.7. Some authors call convex functions *concave upward*, and concave functions *concave downward*.

The definition of convexity we gave above works whether or not f is differentiable. If f' exists then we can use it to give alternate characterizations of convexity.

Theorem 27.8. *Suppose f is differentiable on an interval I. Then the following are equivalent.*

(1) f is convex on I.
(2) $\frac{f(z)-f(x)}{z-x} \leq f'(z) \leq \frac{f(y)-f(z)}{y-z}$ for all $x < z < y$ in I.
(3) f' is a nondecreasing function on I.
(4) The graph of f lies on or above all its tangent lines on I.

The corresponding result for 'concave' holds, where we reverse all the inequalities, replace 'nondecreasing' by 'nonincreasing', and replace 'above' by 'below'.

Proof. We prove that (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (1), and that (2) $\Leftrightarrow$ (4).

(1) $\Rightarrow$ (2). Suppose that f is convex and consider $x < z < y$. Comparing the slopes of the two secant lines associated to x and z and to z and y (see the left half of Figure 10), we see that

$$\frac{f(z) - f(x)}{z - x} \leq \frac{f(y) - f(z)}{y - z}.$$

Taking a limit of the first quantity as $x \to z^-$ gives the second inequality in (2). Taking a limit of the second quantity as $y \to z^+$ gives the first inequality in (2).

(2) $\Rightarrow$ (3). Given any $a < b$ in I, the second inequality in (2) (with $z = a$ and $y = b$) gives $f'(a) \leq \frac{f(b)-f(a)}{b-a}$, while the first inequality in (2) (with $x = a$ and $z = b$) gives $\frac{f(b)-f(a)}{b-a} \leq f'(b)$. Combining these gives $f'(a) \leq f'(b)$, which proves (3).

(3) $\Rightarrow$ (1). Suppose that f is *not* convex. Then there exists $x < z < y$ such that $(z, f(z))$ lies above the line through $(x, f(x))$ and $(y, f(y))$ (see the right half of Figure 10), which in turn implies that

$$\frac{f(z) - f(x)}{z - x} > \frac{f(y) - f(z)}{y - z}.$$

By the MVT there are points $a \in (x, z)$ and $b \in (z, y)$ such that $f'(a) = \frac{f(z)-f(x)}{z-x}$ and $f'(b) = \frac{f(y)-f(z)}{y-z}$. But then we have $a < b$ and $f'(a) > f'(b)$, so that f' is *not* a nondecreasing function on I.

(2) $\Leftrightarrow$ (4). The tangent line at a has equation $g(x) = f(z) + f'(z)(x - z)$. For $x < a$, we see that the following are equivalent:

- $(x, f(x))$ lies above this tangent line;
- $f(x) \geq f(a) + f'(a)(x - a)$;
- $f(a) - f(x) \leq f'(a)(a - x)$;
- $\frac{f(a)-f(x)}{a-x} \leq f'(a)$.

For $y > a$, the following are equivalent.

- $(y, f(y))$ lies above the tangent line at a;
- $f(y) \geq f(a) + f'(a)(y - a)$;
- $f(y) - f(a) \geq f'(a)(y - a)$;
- $f'(a) \leq \frac{f(y)-f(a)}{y-a}$.

We conclude that the graph of f lies above its tangent lines if and only if for every $x < a < y$ we have $\frac{f(a)-f(x)}{a-x} \leq f'(a) \leq \frac{f(y)-f(a)}{y-a}$, which is (2). □

Combining Theorems 27.1 and 27.8, we can give a criterion for convexity using the second derivative.

Theorem 27.9. *Let f be twice differentiable on an interval I.*

(1) If $f''(x) > 0$ for all $x \in I$, then f is convex on I.
(2) If $f''(x) < 0$ for all $x \in I$, then f is concave on I.

Proof. If $f''(x) > 0$ for all $x \in I$, then Theorem 27.1 implies that f' is nondecreasing on I, and then Theorem 27.8 implies that f is convex on I. THe proof for the second half is similar. □

Definition 27.10. If f'' changes sign at x (goes from positive to negative or vice versa) then we say that x is an *inflection point* of f. Thus at an inflection point, f goes from being convex to being concave, or vice versa.

We can also use the second derivative to determine whether a critical point is a local maximum or minimum.

Theorem 27.11 (Second Derivative Test)**.** *Let f be twice differentiable on an interval (a, b) that contains a critical point c.*

(1) If $f''(c) < 0$, then f has a local maximum at c.
(2) If $f''(c) > 0$, then f has a local minimum at c.
(3) If $f''(c) = 0$, then c could be either a local maximum, a local minimum, or neither.

Proof. If $f''(c) > 0$ then we have

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

Thus there exists $\delta > 0$ such that $\frac{f'(c+h)}{h} > 0$ whenever $0 < |h| < \delta$. In particular, for $0 < h < \delta$ we have $f'(c+h) > 0$, and $f'(c-h) < 0$. By the first derivative test, this implies that c is a local minimum. The proof when $f''(c) < 0$ is similar. □

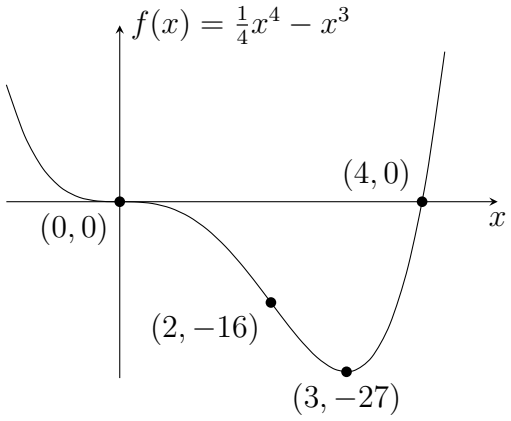


FIGURE 11. Roots, critical points, and inflection points.

Example 27.12. Consider the function $f(x) = \frac{1}{4}x^4 - x^3$, whose graph is shown in Figure 11. This function has roots at $x = 0$ and $x = 4$, is negative between the roots, and positive elsewhere. Differentiating gives

$$f'(x) = x^3 - 3x^2 = x^2(x-3) \quad \text{and} \quad f''(x) = 3x^2 - 6x = 3x(x-2).$$

Thus f has critical points at 0 and 3, and inflection points at 0 and 2. Because $f''(3) = 36 > 0$, the critical point at 3 is a local minimum by the Second Derivative Test. This text provides no information about the critical point at 0, and indeed checking the sign of f reveals that this is neither a local maximum nor a local minimum.

By considering the sign of f', we see that f is decreasing on $(-\infty, 3)$ and increasing on $(3, \infty)$. Considering the sign of f'', we see that f is convex on $(-\infty, 0)$ and $(2, \infty)$, and concave on $(0, 2)$.

Lecture 28 l'Hospital's rule

Stewart §4.4, Spivak Chapter 11

28.1. Indeterminate forms

The limit laws tell us how to compute $\lim \frac{f}{g}$ provided (1) $\lim f$ and $\lim g$ exist, and (2) $\lim g \neq 0$. But what if $\lim g = 0$? If $\lim f \neq 0$ then either $\lim \frac{f}{g} = \pm\infty$ (if g is consistently positive or consistently negative) or $\frac{f}{g}$ oscillates between $\pm\infty$ (if g oscillates between positive and negative values in the limit). If $\lim f = 0$, on the other hand, then the limit has the *indeterminate form* $\frac{0}{0}$, and many different outcomes are possible.

Example 28.1. We have seen the following examples in previous lectures.

(1) $\lim\limits_{x\to 1} \dfrac{x-1}{x^2-1} = \lim\limits_{x\to 1} \dfrac{1}{x+1} = \dfrac{1}{2}$ by algebraic simplifications (difference of squares).

(2) $\lim\limits_{x\to 1} \dfrac{\sqrt{x}-1}{x-1} = \lim\limits_{x\to 1} \dfrac{\sqrt{x}-1}{x-1} \cdot \dfrac{\sqrt{x}+1}{\sqrt{x}+1} = \lim\limits_{x\to 1} \dfrac{1}{\sqrt{x}+1} = \dfrac{1}{2}$ by multiplying by a conjugate and simplifying.

(3) $\lim\limits_{x\to 0} \dfrac{\sin x}{x} = 1$ by a geometric argument that we gave in Lecture 8.

(4) $\lim\limits_{x\to 0} \dfrac{\sin(2x)}{\sin x} = \lim\limits_{x\to 0} 2\dfrac{\sin(2x)}{2x}\dfrac{x}{\sin x} = 2$ using the previous example.

(5) $\lim\limits_{x\to\infty} \dfrac{x^2}{e^x} = 0$ by an argument from Lecture 24 using properties of exponentials.

(6) $\lim\limits_{x\to\infty} \dfrac{\ln x}{x} = 0$ by a similar argument, or as a consequence of the previous example.

The computation above for $\lim_{x\to 0} \frac{\sin 2x}{\sin x}$ suggests the following general approach:

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f(x)-f(a)}{x-a}\frac{x-a}{g(x)-g(a)} = \frac{f'(a)}{g'(a)},$$

where the first equality works provided $f(a) = g(a) = 0$, and the second works provided f, g are both differentiable at a. Another way of looking at this is as follows: if f, g are differentiable at a and vanish at a, then for $x \approx a$ we have the linear approximations $f(x) = f'(a)(x-a) + r(x)$ and $g(x) = g'(a)(x-a) + s(x)$, where r, s are error functions with the property that $\lim\limits_{x\to a} \dfrac{r(x)}{x-a} = \lim\limits_{x\to 0} \dfrac{s(x)}{x-a} = 0$. Thus

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(a)(x-a)+r(x)}{g'(a)(x-a)+s(x)} = \lim_{x\to a} \frac{f'(a)+\frac{r(x)}{x-a}}{g'(a)+\frac{s(x)}{x-a}} = \frac{f'(a)}{g'(a)}.$$

This gives the intuition behind l'Hospital's rule,[31] but there is a shortcoming inherent in this approach; these arguments only work if f, g are differentiable at a, and if $g'(a) \neq 0$.

[31] We follow Stewart's book in our spelling; Spivak's book uses the alternate spelling l'Hôpital.

Example 28.2. The limits $\lim\limits_{x\to 0} \dfrac{1-\cos x}{x^2}$ and $\lim\limits_{x\to 0^+} \dfrac{\sqrt{x}}{e^{-1/x}}$ have indeterminate form $\frac{0}{0}$, but *cannot* be computed via the above argument. In the first case, we have $f'(0) = g'(0) = 0$, so $f'(0)/g'(0)$ is undefined, and in the second case neither f nor g is differentiable at 0; indeed, g is not even defined at 0.

In light of Example 28.2, we might try to replace $f'(a)/g'(a)$ with $f'(x)/g'(x)$ for $x \approx a$, since this quantity *is* defined for both examples given there, and then take a limit as $x \to a$. This turns out to work, and we will soon prove the following result.

Theorem 28.3 (l'Hospital's rule, right-handed limits of the form 0/0)**.**
Suppose we are given functions f, g and $a < b$ such that the following are true:

(1) f, g are differentiable on (a, b) and $g'(x) \neq 0$ for all $x \in (a, b)$;
(2) $\lim\limits_{x\to a^+} f(x) = \lim\limits_{x\to a^+} g(x) = 0$;
(3) $\lim\limits_{x\to a^+} \dfrac{f'(x)}{g'(x)}$ exists.

Then $\lim\limits_{x\to a^+} \dfrac{f(x)}{g(x)}$ exists and is equal to $\lim\limits_{x\to a^+} \dfrac{f'(x)}{g'(x)}$.

28.2. Cauchy's MVT

As a first attempt to prove Theorem 28.3 and relate f'/g' to f/g, we could try to use the MVT. Indeed, if $x > a$ is such that f, g are continuous on $[a, x]$ and differentiable on (a, x), then the MVT gives $y, z \in (a, x)$ such that

$$f'(y) = \frac{f(x)-f(a)}{x-a} = \frac{f(x)}{x-a} \text{ and } g'(z) = \frac{g(x)-g(a)}{x-a} = \frac{g(x)}{x-a} \;\Rightarrow\; \frac{f'(y)}{g'(z)} = \frac{f(x)}{g(x)}.$$

Since $y, z \to a$ as $x \to a$, this *almost* does what we want. But not quite. We don't have any way to guarantee that y, z are the same point, and it is not clear why $f'(y)/g'(z)$ should tell us anything about $f'(y)/g'(y)$ or $f'(z)/g'(z)$.

The solution is to use a slightly stronger version of the Mean Value Theorem.

Theorem 28.4 (Cauchy's MVT)**.** *If f, g are continuous on $[a, b]$ and differentiable on (a, b), and if $g'(x) \neq 0$ for all $x \in (a, b)$, then there exists $c \in (a, b)$ such that*

$$\frac{f'(c)}{g'(c)} = \frac{f(b)-f(a)}{g(b)-g(a)}.$$

Note that if we choose $g(x) = x$, then this reduces to the standard MVT. And indeed, we can prove Cauchy's MVT from Rolle's Theorem by mimicking the proof of the MVT.

Proof of Cauchy's MVT. Consider the function

$$h(x) = f(x) - f(a) - \frac{f(b)-f(a)}{g(b)-g(a)}\big(g(x)-g(a)\big)$$

and observe that $h(a) = h(b) = 0$. Moreover, h is continuous on $[a, b]$ and differentiable on (a, b), so by Rolle's Theorem there is $c \in (a, b)$ such that $h'(c) = 0$. Note that

$$h'(x) = f'(x) - g'(x)\frac{f(b)-f(a)}{g(b)-g(a)},$$

so $h'(c) = 0$ implies $f'(c) = g'(c)\frac{f(b)-f(a)}{g(b)-g(a)}$. Since $g'(c) \neq 0$, this completes the proof. □

Exercise 28.5. Prove that under the conditions of Cauchy's MVT, we have $g(b) \neq g(a)$, so that the definition of $h(x)$ in the proof is valid.

28.3. l'Hospital's rule for limits of the form $0/0$

Now we can use Cauchy's MVT to prove l'Hospital's rule.

Proof of Theorem 28.3. Since none of the hypotheses or conclusions are affected by the value of f and g at a (or indeed by whether or not f and g are even defined at a), we may redefine f, g are this single point if necessary and assume that $f(a) = g(a) = 0$, so that f, g are continuous on $[a, b)$. Now let $L = \lim_{x\to a^+} \frac{f'(x)}{g'(x)}$. Then for every $\varepsilon > 0$, there is $\delta > 0$ such that

$$\left|\frac{f'(c)}{g'(c)} - L\right| < \varepsilon \text{ for all } c \in (a, a+\delta). \tag{28.1}$$

Given any $x \in (a, a+\delta)$ we can apply Cauchy's Mean Value Theorem to F, G on the interval $[a, x]$ and obtain $c \in (a, x)$ such that

$$\frac{f'(c)}{g'(c)} = \frac{f(x) - f(a)}{g(x) - g(a)} = \frac{f(x)}{g(x)}. \tag{28.2}$$

Combining (28.1) and (28.2) gives $\left|\frac{f(x)}{g(x)} - L\right| < \varepsilon$ for all $x \in (a, a+\delta)$. Since $\varepsilon > 0$ was arbitrary, this proves that $L = \lim_{x\to a^+} \frac{f(x)}{g(x)}$. □

Exercise 28.6. Prove the following versions of l'Hospital's rule for limits of the form $0/0$.

(1) *Left-handed limits*: Theorem 28.3 remains true if we replace $\lim\limits_{x\to a^+}$ by $\lim\limits_{x\to b^-}$.
(2) *Two-sided limits*: Theorem 28.3 remains true if we replace $\lim\limits_{x\to a^+}$ by $\lim\limits_{x\to a}$, provided there is an open interval I containing a such that f, g are differentiable on $I \setminus \{a\}$ and g' does not vanish on $I \setminus \{a\}$.

Exercise 28.7. Formulate and prove versions of l'Hospital's rule for right-handed, left-handed, and two-sided limits where we replace "$\lim f'/g'$ exists" with "$\lim f'/g' = \infty$" or "$\lim f'/g' = -\infty$.

Remark 28.8. We emphasize that l'Hospital's rule does not require f and g to be differentiable, continuous, or even defined at a itself.

Lecture 29 More on l'Hospital's rule

Stewart §4.4, Spivak Chapter 11

29.1. Examples, and limits at infinity

Now we can revisit some of the limits in Example 28.1 and compute them using l'Hospital's rule.

(1) $\lim_{x\to 1} \dfrac{x-1}{x^2-1} = \lim_{x\to 1} \dfrac{1}{2x} = \dfrac{1}{2}$.

(2) $\lim_{x\to 1} \dfrac{\sqrt{x}-1}{x-1} = \lim_{x\to 1} \dfrac{\frac{1}{2}x^{-1/2}}{1} = \dfrac{1}{2}$.

(3) $\lim_{x\to 0} \dfrac{\sin x}{x} = \lim_{x\to 0} \dfrac{\cos x}{1} = 1$. (Note, however, that the proof that $\frac{d}{dx}\sin x = \cos x$ required us to first compute this limit!)

(4) $\lim_{x\to 0} \dfrac{\sin(2x)}{\sin x} = \lim_{x\to 0} \dfrac{2\cos(2x)}{\cos x} = 2$.

Remark 29.1. Before moving on, a warning is in order. In order to apply l'Hospital's rule, it is crucial that the limit has indeterminate form. If we blindly differentiate top and bottom without first checking this condition, we can get ourselves into trouble. For example, with $f(x) = \sin x$ and $g(x) = \cos x$ we have

$$\lim_{x\to 0^+} \frac{\sin x}{\cos x} = 0 \quad \text{but} \quad \lim_{x\to 0^+} \frac{-\cos x}{\sin x} = -\infty.$$

This does not violate l'Hospital's rule because $\lim g \neq 0$, so the conditions of Theorem 28.3 are not met.

We cannot yet deal with the final two limits in Example 28.1, for two reasons: first, they are limits at ∞ instead of limits at $a \in \mathbb{R}$; second, they have the indeterminate form $\frac{\infty}{\infty}$, meaning that the numerator and denominator both go to ∞, instead of the indeterminate form $\frac{0}{0}$ covered by Theorem 28.3. Let us tackle these difficulties one at a time; start by considering the case where we have an indetermine form $\frac{0}{0}$ occuring in a limit at ∞. It turns out that we can deduce the result here as a consequence of Theorem 28.3 with just a bit more work.

Corollary 29.2 (l'Hospital's rule, limits at infinity of the form 0/0)**.**
Suppose we are given functions f, g and $r \in \mathbb{R}$ such that the following are true:

(1) f, g are differentiable on (r, ∞) and $g'(x) \neq 0$ for all $x > r$;

(2) $\lim_{x\to\infty} f(x) = \lim_{x\to\infty} g(x) = 0$;

(3) $\lim_{x\to\infty} \dfrac{f'(x)}{g'(x)}$ exists.

Then $\lim_{x\to\infty} \dfrac{f(x)}{g(x)}$ exists and is equal to $\lim_{x\to\infty} \dfrac{f'(x)}{g'(x)}$.

Proof. Define functions $F, G\colon (0, 1/r) \to \mathbb{R}$ by

$$F(x) = f(1/x) \text{ and } G(x) = g(1/x).$$

We claim that F, G satisfy the three conditions of Theorem 28.3. The chain rule gives

$$F'(x) = -x^{-2} f'(1/x) \text{ and } G'(x) = -x^{-2} g'(1/x),$$

so F, G are differentiable and G' is nonvanishing on $(0, 1/r)$, which verifies the first condition. The second condition follows from observing that $\lim_{x\to 0^+} F(x) = \lim_{x\to 0^+} f(1/x) = \lim_{t\to\infty} f(t) = 0$, and similarly for G. The third condition follows since

$$\lim_{x\to 0^+} \frac{F'(x)}{G'(x)} = \lim_{x\to 0^+} \frac{-x^{-2}f'(1/x)}{-x^{-2}g'(1/x)} = \lim_{x\to 0^+} \frac{f'(1/x)}{g'(1/x)} = \lim_{t\to\infty} \frac{f'(t)}{g'(t)}. \tag{29.1}$$

Thus Theorem 28.3 applies to F, G for the right-hand limit at 0, and we deduce that

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \lim_{x\to\infty} \frac{F(1/x)}{G(1/x)} = \lim_{t\to 0^+} \frac{F(t)}{G(t)} = \lim_{t\to 0^+} \frac{F'(t)}{G'(t)} = \lim_{x\to\infty} \frac{f'(x)}{g'(x)},$$

where the last equality uses (29.1). □

Example 29.3. Consider $\lim_{x\to\infty} x \ln(\frac{x+1}{x-1})$. Observe that $\lim_{x\to\infty} x = \infty$ and

$$\lim_{x\to\infty} \ln \frac{x+1}{x-1} = \ln\Big(\lim_{x\to\infty} \frac{x+1}{x-1}\Big) = \ln 1 = 0,$$

so this limit has indeterminate form $\infty \cdot 0$. In order to apply Corollary 29.2, we can rewrite it in the indeterminate form $\frac{0}{0}$ as

$$\begin{aligned}\lim_{x\to\infty} x \ln(\frac{x+1}{x-1}) &= \lim_{x\to\infty} \frac{\ln(\frac{x+1}{x-1})}{1/x} = \lim_{x\to\infty} \frac{\ln(x+1) - \ln(x-1)}{1/x} = \lim_{x\to\infty} \frac{\frac{1}{x+1} - \frac{1}{x-1}}{-1/x^2} \\ &= \lim_{x\to\infty} -x^2 \cdot \frac{(x-1)-(x+1)}{(x+1)(x-1)} = \lim_{x\to\infty} \frac{2x^2}{x^2-1} = 2.\end{aligned}$$

Here Corollary 29.2 is used to get the last equality on the first line.

29.2. Limits with indeterminate forms ∞/∞

Now we prove a version of l'Hospital's rule that can deal with the case when the numerator and denominator go to ∞.

Theorem 29.4 (l'Hospital's rule, limits at infinity of the form ∞/∞)**.**
Suppose we are given functions f, g and $r \in \mathbb{R}$ such that the following are true:

(1) f, g are differentiable on (r, ∞) and $g'(x) \neq 0$ for all $x > r$;
(2) $\lim_{x\to\infty} f(x) = \lim_{x\to\infty} g(x) = \infty$;
(3) $\lim_{x\to\infty} \frac{f'(x)}{g'(x)}$ exists.

Then $\lim_{x\to\infty} \frac{f(x)}{g(x)}$ exists and is equal to $\lim_{x\to\infty} \frac{f'(x)}{g'(x)}$.

Proof. Let r be as in the hypothesis, and let $L = \lim_{x\to\infty} f'(x)/g'(x)$. Given $\varepsilon > 0$, let $a \geq r$ be such that $\left|\frac{f'(x)}{g'(x)} - L\right| < \varepsilon/2$ for all $x > a$. Now given any $x > a$, first note that since $f, g \to \infty$ as $x \to \infty$, there is $b \in \mathbb{R}$ such that $f(x) > f(a)$, $g(x) > g(a)$, and $g(x) > 0$ for all $x > b$. For all such x, by Cauchy's MVT there is $c \in (a, x)$ such that

$$\frac{f'(c)}{g'(c)} = \frac{f(x) - f(a)}{g(x) - g(a)}. \tag{29.2}$$

Thus we have

$$\begin{aligned}\left|\frac{f(x)}{g(x)}-L\right| &\le \left|\frac{f(x)}{g(x)}-\frac{f'(c)}{g'(c)}\right|+\left|\frac{f'(c)}{g'(c)}-L\right| \le \left|\frac{f'(c)}{g'(c)}\right|\cdot\left|\frac{f(x)/g(x)}{f'(c)/g'(c)}-1\right|+\frac{\varepsilon}{2}\\ &\le (|L|+\varepsilon)\underbrace{\left|\frac{f(x)(g(x)-g(a))}{g(x)(f(x)-f(a))}-1\right|}_{\mathrm{I}}+\frac{\varepsilon}{2}.\end{aligned} \tag{29.3}$$

The limit laws give

$$\lim_{x\to\infty}\frac{f(x)}{g(x)}\frac{g(x)-g(a)}{f(x)-f(a)}=\lim_{x\to\infty}\frac{1-\frac{g(a)}{g(x)}}{1-\frac{f(a)}{f(x)}}=1$$

using the fact that $\lim f=\lim g=\infty$, so there exists $s\in\mathbb{R}$ such that for all $x>s$, we have $\mathrm{I}<\frac{\varepsilon}{2(|L|+\varepsilon)}$. For every such x, (29.3) gives $\left|\frac{f(x)}{g(x)}-L\right|<\varepsilon$, which completes the proof of Theorem 29.4. □

Exercise 29.5. Formulate and prove versions of l'Hospital's rule for one- and two-sided limits of the form ∞/∞, as well as a version that applies when $\lim f'/g'=\pm\infty$.

Now we can finally deal with the last two limits in Example 28.1

(5) $\lim\limits_{x\to\infty}\dfrac{x^2}{e^x}=\lim\limits_{x\to\infty}\dfrac{2x}{e^x}=\lim\limits_{x\to\infty}\dfrac{2}{e^x}=0$, where we have used l'Hospital's rule *twice*.

(6) $\lim\limits_{x\to\infty}\dfrac{\ln x}{x}=\lim\limits_{x\to\infty}\dfrac{1/x}{1}=0.$

29.3. Other indeterminate forms

Limits with other indeterminate forms, including $0\cdot\infty$, $\infty-\infty$, 0^0, ∞^0, and 1^∞, can often be evaluated using l'Hospital's rule by first relating them to a limit of the form $0/0$ or ∞/∞.

Example 29.6. The limit $\lim\limits_{x\to0^+}x\ln x$ has the indeterminate form $0\cdot\infty$ because $x\to0$ and $\ln x\to-\infty$. We can write it as a limit of form ∞/∞ by writing $x\ln x=\frac{\ln x}{1/x}$; then both numerator and denominator go to $\pm\infty$, and l'Hospital's rule gives

$$\lim_{x\to0^+}x\ln x=\lim_{x\to0^+}\frac{\ln x}{1/x}=\lim_{x\to0^+}\frac{1/x}{-1/x^2}=\lim_{x\to0^+}(-x)=0.$$

Note that if $f\to0$ and $g\to\infty$ then we can write $fg=\frac{f}{1/g}$ to get a limit of the form $0/0$, or $fg=\frac{g}{1/f}$ to get a limit of the form ∞/∞. We may need to choose wisely which we do: in the example above, making the other choice would give $x\ln x=\frac{x}{1/\ln x}$ and using l'Hospital's rule would require us to evaluate

$$\lim_{x\to0^+}\frac{1}{-\frac{1}{x}(\ln x)^2}=\lim_{x\to0^+}\frac{-x}{(\ln x)^2},$$

which is no easier to handle, and the situation would not improve upon further differentiations.

Example 29.7. $\lim_{x\to\frac{\pi}{2}^-}(\sec x - \tan x)$ has the indeterminate form $\infty - \infty$, but can be evaluated using l'Hospital's rule by first transforming it into a limit with the indeterminate form $0/0$:

$$\lim_{x\to\frac{\pi}{2}^-}(\sec x - \tan x) = \lim_{x\to\frac{\pi}{2}^-}\frac{1-\sin x}{\cos x} = \lim_{x\to\frac{\pi}{2}^-}\frac{-\cos x}{-\sin x} = 0.$$

Example 29.8. $\lim_{x\to 0^+} x^x$ has the indeterminate form 0^0, but can be evaluated using l'Hospital's rule by transforming it into the exponential of a limit that we already evaluated:

$$\lim_{x\to 0^+} x^x = \lim_{x\to 0^+} e^{x\ln x} = e^{\lim_{x\to 0^+} x\ln x} = e^0 = 1.$$

Here the second equality uses continuity of the exponential function.

Similarly, writing $f^g = e^{g\ln f}$ lets us evaluate limits of the forms 0^0, ∞^0, and 1^∞ in terms of limits of the form $0\cdot\infty$, which can then be dealt with as above.

Lecture 30 Curve sketching, optimization, Newton's method

Stewart §4.5, §4.7, and §4.8

30.1. Curve sketching

We can assemble the tools developed so far into a fairly robust procedure for qualitative curve sketching. If we are given a function f in terms of a formula $f(x)$, then to sketch its graph we should carry out the following steps, which we illustrate in Figure 12 for the example $f(x) = \frac{3x^2}{x^2-4}$.

(1) Determine the domain. If the domain is not explicitly stated then we take it to be the set of x for which the formula is well defined. In the example, f is a rational function so the domain is the set of x where the denominator is nonzero: $D = \mathbb{R}\setminus\{-2,2\} = (-\infty,-2)\cup(-2,2)\cup(2,\infty)$.

(2) Determine whether the function has any symmetry that should be taken into account: is it even, odd, or periodic? The example is an even function, so the graph on $(-\infty,0]$ will be the mirror image of the graph on $[0,\infty)$.

(3) Find the x- and y-intercepts and plot these points on the graph. In the example the only intercept occurs at the origin.

(4) Find the asymptotes and draw them as dashed lines; determine the corresponding limits and draw these "arms" of the graph. In the example there are vertical asymptotes at ± 2 since the denominator vanishes and the numerator does not, and we have

$$\lim_{x\to -2^+}\frac{3x^2}{x^2-4} = \lim_{x\to -2^-}\frac{3x^2}{x^2-4} = -\infty \quad\text{and}\quad \lim_{x\to -2^-}\frac{3x^2}{x^2-4} = \lim_{x\to 2^+}\frac{3x^2}{x^2-4} = +\infty,$$

as shown; note also that factoring the denominator reveals that $f(x)$ is positive on $(-\infty,-2)\cup(2,\infty)$, and negative on $(-2,2)$. There is a horizontal asymptote

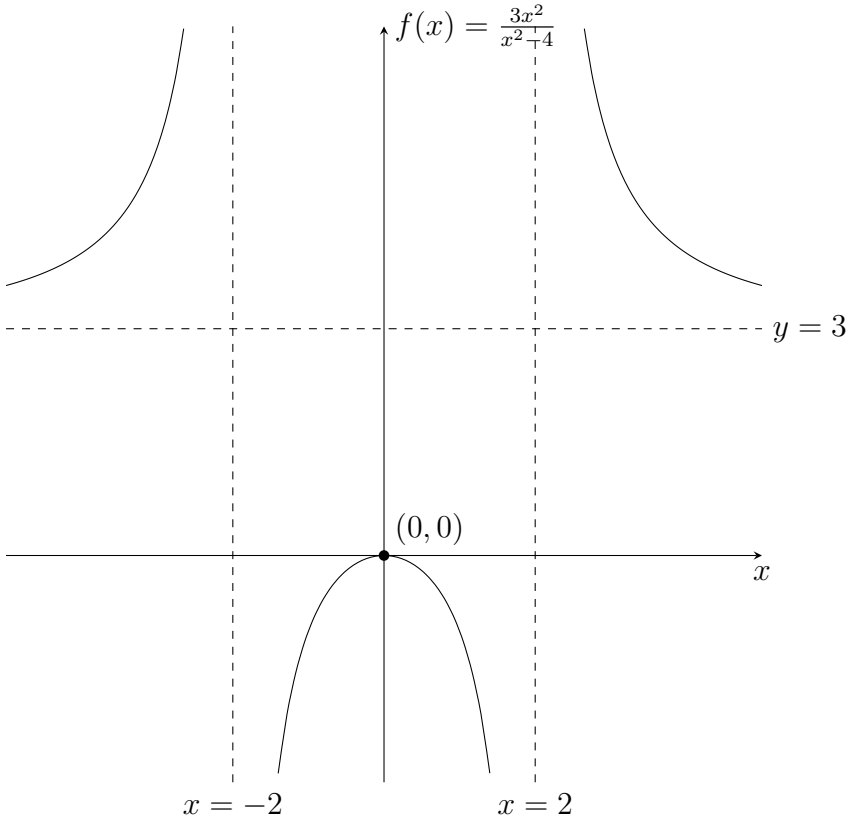


FIGURE 12. Roots, critical points, and inflection points.

at $y = 3$ because

$$\lim_{x\to\infty} \frac{3x^2}{x^2-4} = \lim_{x\to\infty} \frac{3}{1-\frac{4}{x^2}} = 3,$$

and similarly for $\lim_{x\to-\infty} f(x)$. Note that the function approaches the asymptote from above because $3x^2 > 3(x^2-4)$ and thus $\frac{3x^2}{x^2-4} > 3$ whenever $x^2-4 > 0$.

(5) Use f' to find the critical points of f and determine on which intervals f is increasing or decreasing. In the example, we have

$$f'(x) = \frac{6x(x^2-4) - 3x^2(2x)}{(x^2-4)^2} = \frac{-24x}{(x^2-4)^2},$$

so f is increasing on $(-\infty, -2)$ and $(-2, 0)$ – note that since -2 is not part of the domain, we cannot say "f is increasing on $(-\infty, 0)$" – and decreasing on $(0, 2)$ and $(2, \infty)$. The only critical point is $x = 0$.

(6) Compute f'' at the critical points (or use some other technique) to find the local maxima, local minima, and inflection points. In the example we have

$$f''(x) = \frac{(x^2-4)^2(-24) - (-24x)\cdot(2x)2(x^2-4)}{(x^2-4)^4} = 24\frac{4x^2-(x^2-4)}{(x^2-4)^3} = 24\frac{3x^2+4}{(x^2-4)^3}$$

and thus in particular $f''(0) = 24 \cdot 4/(-4)^3 = -\frac{3}{2} < 0$, so 0 is a local maximum. Alternately we could observe that f is increasing on $(-2, 0)$ and decreasing on $(0, 2)$, which is enough to conclude that 0 is a local maximum without computing f''.

(7) Use f'' to determine convexity and concavity. The above computation shows that f'' takes the same sign as $x^2 - 4$, and thus f is convex on $(-\infty, -2)$, concave on $(-2, 2)$, and convex on $(2, \infty)$.

(8) Use the information about local extrema, inflection points, monotonicity, and convexity to connect the dots and arms from the initial sketch.

We mention one more possibility that should be looked for when sketching the graph of a function. Consider the example $f(x) = \frac{x^2}{x+1}$. The domain is $(-\infty, -1) \cup (-1, \infty)$ and the function has no symmetry; the only intercept is at the origin, and there is a vertical asymptote at $x = -1$. There is no horizontal asymptote, but it turns out that there *is* a "slant asymptote", or "oblique asymptote". To explain this, we first observe that $f(x)$ is written as a sort of "improper fraction", where the numerator has larger degree than the denominator. We can use polynomial long division to rewrite $f(x)$ as a polynomial plus a rational function in "proper fraction" form, where the numerator has smaller degree than the denominator, as follows.

$$\begin{array}{r|l} & x - 1 \\ \hline x + 1 & x^2 \\ & -x^2 - x \\ \hline & \quad -x \\ & \quad\; x + 1 \\ \hline & \qquad 1 \end{array}$$

From this we conclude that $\frac{x^2}{x+1} = x - 1 + \frac{1}{x+1}$, and in particular

$$\lim_{x \to \pm\infty} \big(f(x) - (x - 1)\big) = \lim_{x \to \pm\infty} \frac{1}{x + 1} = 0.$$

Thus $f(x)$ is asymptotic to the line $y = x - 1$. More generally, we say that $y = mx + b$ is a slant asymptote, or oblique asymptote, for a function f if $f(x) - (mx + b)$ goes to 0 at ∞ or $-\infty$.

30.2. Optimization problems

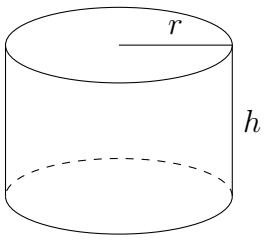
Fermat's theorem, which says that local maxima and minima of a function on a closed interval must occur at critical points or at endpoints, provides a good tool for finding extreme points. To apply this in examples, we must first translate the problem into a question of maximizing or minimizing a function on an interval.

Example 30.1. I want to make a five-gallon bucket that is cylindrical and has an open top. How much material do I need, assuming the thickness of the material is predetermined?

Step 1 is to understand the problem and identify the various quantities in play: which quantities are given, and which are unknown? In the example, the two most

important quantities are the volume of the bucket and the surface area of the bucket (which determines the amount of material I use).

Step 2 (which can be done in conjunction with Step 1) is to draw a diagram illustrating the situation. In the example, when we draw the cylinder, we see that both volume and surface area are determined by the height of the cylinder and the radius of the circle that forms its base, so these two quantities will also enter our analysis.



Step 3 is to set up notation for the various quantities in play, and to identify which quantity we need to maximize or minimize in order to solve the problem. In the example, we can write V for volume, A for surface area, h for height, and r for radius; then our goal is to minimize A, because we are trying to construct the bucket using the smallest amount of material possible.

Step 4 is to write the quantity to be optimized as a function of the other variables. In the example we see that $A = \pi r^2 + 2\pi rh$, where the first term represents the area of the base, and the second term is the area of the sides of the cylinder.

Step 5 is to use the relationship between the other variables to write the quantity to be optimized as a function of a *single* variable. This step is necessary because usually there are multiple other variables that are related by certain constraints that must be taken into account. In the example, the volume is fixed, so r and h are related by $\pi r^2 h = V$. Solving this for h gives $h = \frac{V}{\pi r^2}$, and thus

$$A = \pi r^2 + 2\pi r \frac{V}{\pi r^2} = \pi r^2 + \frac{2V}{r}.$$

Step 6 is to use our tools for finding extreme points to solve the problem. In the example we observe that

$$\frac{dA}{dr} = 2\pi r - \frac{2V}{r^2} = 0 \quad \Leftrightarrow \quad \pi r = \frac{V}{r^2} \quad \Leftrightarrow \quad r^3 = \frac{V}{\pi}$$

and thus there is a single critical point at $r = \sqrt[3]{V/\pi}$. Moreover, we have $\frac{dA}{dr} < 0$ to the left of this critical point, and $\frac{dA}{dr} > 0$ to its right, so this critical point is a global minimum. At this critical point we have

$$A = \pi r^2 + \frac{2V}{r} = \pi\Big(\frac{V}{\pi}\Big)^{2/3} + 2V\Big(\frac{\pi}{V}\Big)^{1/3} = \pi^{1/3}V^{2/3} + 2\pi^{1/3}V^{2/3} = 3\pi^{1/3}V^{2/3}.$$

Note that Step 5 required us to solve for h in terms of r. We could also have solved for r in terms of h, but this would have led to $r = \sqrt{\frac{V}{\pi h}}$ and thus

$$A = \pi\frac{V}{\pi h} + 2\pi\sqrt{\frac{V}{\pi h}}h = \frac{V}{h} + 2\sqrt{\pi V h},$$

which makes the following computations a little messier because of the square root (though we could certainly carry them out). An alternate approach is to use implicit differentiation: treating h as an (unknown) function of r and differentiating the formulas $A = \pi r^2 + 2\pi rh$ and $V = \pi r^2 h$ gives

$$\frac{dA}{dr} = 2\pi r + 2\pi h + 2\pi r\frac{dh}{dr} \quad \text{and} \quad 0 = \frac{dV}{dr} = 2\pi rh + \pi r^2\frac{dh}{dr}.$$

Solving the second of these gives $\frac{dh}{dr} = -2h/r$. Using this in the first equation gives

$$\frac{dA}{dr} = 2\pi\Big(r + h - r \cdot \frac{2h}{r}\Big) = 2\pi(r + h - 2h) = 2\pi(r - h).$$

Thus the critical point occurs when $r = h$, and some simple algebra gives the answer.

30.3. Newton's method

Consider the equation $x^5 + x + 1 = 0$. We do not have an analogue of the quadratic formula available to find explicitly the solution(s) of this equation. On the other hand, we can consider the function $f(x) = x^5 + x + 1$ and observe that f is continuous and $f(-1) = -1 < 0 < 1 = f(0)$, so by the IVT there is at least one solution of $f(x) = 0$ in the interval $[-1, 0]$. Moreover, $f'(x) = 5x^4 + 1 > 0$ for all $x \in \mathbb{R}$, so f is increasing and thus 1-1, which means that this is the *only* solution of $f(x) = 0$.

How can we approximate this solution numerically? One approach would be to mimic the proof of the IVT: build a pair of bisection sequences whose mutual limit is the unique root, and then take some element of one of those sequences, which gives a good approximation. Here, though, we outline another approach: *Newton's method*.

Start by making a guess at the solution, and call it x_0. For example, we might take $x_0 = -\frac{1}{2}$ since the desired value $f(x) = 0$ is midway between $f(-1) = -1$ and $f(0) = 1$. Then we consider the linear approximation to f at x_0, which is given by

$$g_0(x) = f(x_0) + f'(x_0)(x - x_0).$$

While solving the equation $f(x) = 0$ is hard, solving $g_0(x) = 0$ is quite easy, and if g_0 is a good enough approximation to f, we may hope that the solutions of the two equations are close together. Thus we let x_1 be defined by

$$0 = g_0(x_1) = f(x_0) + f'(x_0)(x_1 - x_0) \quad \Rightarrow \quad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

If we are lucky, then x_1 is a better approximation to the solution than x_0 is. If we want to keep improving, we can repeat the process: let $g_1(x) = f(x_1) + f'(x_1)(x - x_1)$ be the linear approximation to f at x_1, let x_2 be the solution to $g_1(x_2) = 0$, and so on. Thus in general, we define linear functions $g_n(x)$ and approximate solutions x_n by

$$g_n(x) = f(x_n) + f'(x_n)(x - x_n) \quad \text{and} \quad g_n(x_{n+1}) = 0.$$

More succinctly, we define a sequence x_n by

$$f(x_n) + f'(x_n)(x_{n+1} - x_n) = 0 \quad \Rightarrow \quad x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

If we carry this out for the example $f(x) = x^5 + x + 1$ with the starting guess $x_0 = -\frac{1}{2} = -0.5$, then using the fact that $f'(x) = 5x^4 + 1$, we get

$$\begin{aligned} x_1 &= x_0 - \frac{f(x_0)}{f'(x_0)} = x_0 - \frac{x_0^5 + x_0 + 1}{5x_0^4 + 1} = \frac{5x_0^5 + x_0 - (x_0^5 + x_0 + 1)}{5x_0^4 + 1} \\ &= \frac{4x_0^5 - 1}{5x_0^4 + 1} = \frac{4(-2)^{-5} - 1}{5(-2)^{-4} + 1} = \frac{-9/8}{21/16} = -\frac{6}{7} \approx -0.85714\ldots \\ x_2 &= \frac{4x_1^5 - 1}{5x_1^4 + 1} \approx \frac{4(-0.85714)^5 - 1}{5(-0.85714)^4 + 1} \approx \frac{-2.8506}{3.6989} \approx -0.77066\ldots \end{aligned}$$

$$x_3 = \frac{4x_2^5 - 1}{5x_2^4 + 1} \approx \frac{-2.0874}{2.7637} \approx -0.75529\ldots$$
$$x_4 = \frac{4x_3^5 - 1}{5x_3^4 + 1} \approx \frac{-1.9832}{2.6271} \approx -.75490\ldots$$
$$x_5 = \frac{4x_4^5 - 1}{5x_4^4 + 1} \approx \frac{-1.9806}{2.6238} \approx -.75486\ldots,$$

and so on, where the fact that successive iterates do not change by much suggests that we are approaching the true solution.

Part III. Integrals

Lecture 31 Antiderivatives

Stewart §4.9

Suppose we are interested in some function F, about which all we know is that its derivative is equal to some other function f. For example, this situation arises if we want to reconstruct a car's position over time based only on the information displayed on the speedometer; we want to know $s(t)$, the position at time t, and all we know at first is $v(t) = s'(t)$, the velocity at time t. This motivates the following definition.

Definition 31.1. A function F is an *antiderivative* of f on an interval $I \subset \mathbb{R}$ if $F'(x)$ exists and is equal to $f(x)$ for all $x \in I$.

Example 31.2. If $f(x) = 3x^2$, then $F(x) = x^3$ is an antiderivative of f on $\mathbb{R}$. So is $G(x) = x^3 + 10$. Indeed, for *any* $C \in \mathbb{R}$, the function $x \mapsto x^3 + C$ is an antiderivative of f.

This principle is quite general: if F is an antiderivative of f, then so is $x \mapsto F(x) + C$ for any constant C. In fact, this gives *all* the antiderivatives of f.

Theorem 31.3. *If F, G are antiderivatives of f on an interval I, then there is a constant C such that $G(x) = F(x) + C$ for every $x \in I$.*

Proof. By the definition of antiderivative, $(G-F)'(x) = G'(x) - F'(x) = f(x) - f(x) = 0$ for all $x \in I$, so by Theorem 22.1 (which follows quickly from the MVT), $G-F$ is constant on I. □

Geometrically, any two antiderivatives F and G have graphs with the property that one is a vertical translate of the other.

We stress that Theorem 31.3 only works on an interval. For example, $F(x) = \ln|x|$ is an antiderivative of $f(x)$, and so is

$$G(x) = \begin{cases} 1 + \ln x & x > 0, \\ \ln|x| & x < 0, \end{cases}$$

but $G - F$ is not a constant. The problem here is that f, F, G are not defined at 0, so we are not working on a single interval, but on the union of two disjoint intervals $(-\infty, 0)$ and $(0, \infty)$. On each of these intervals, any two antiderivatives must differ by a constant, but if we jump to a different interval then we need to allow for a new constant.

In order to recover a specific antiderivative F, we need to know enough information about F to determine the constant. For example, if F is an antiderivative of $f(x) = 3x^2$ with the property that $F(0) = 5$, then from Example 31.2 and Theorem 31.3 we see that $F(x) = 3x^2 + C$ for some constant C, and evaluating $F(0) = 3 \cdot 0^2 + C = C$ we see that $C = t$, so $F(x) = 3x^2 + 5$.

In some situations we may need to antidifferentiate more than once. For example, if we want to know the vertical position $s(t)$ of an object moving under the influence of gravity, then we are given not the velocity $v(t) = s'(t)$, but the acceleration $a(t) = v'(t) = s''(t)$.

With a constant gravitational field imparting a downward acceleration $g > 0$, we have $a(t) = -g$. Antidifferentiating once gives $v(t) = -gt + C$ for some constant C, and another antidifferentiation gives $s(t) = -\frac{g}{2}t^2 + Ct + D$ for some constant D. Thus in order to determine $s(t)$, we need to determine both C and D, which requires *two* pieces of information. For example, it would suffice to know the position at two different moments in time, or both the position and the velocity at a single moment.

It is natural at this point to ask whether every function has an antiderivative. We may reasonably start by listing the various formulas we have so far for derivatives of common functions, and obtain the following table; note that in each case we only write a single antiderivative, all the others can be obtained by adding a constant.

function $f(x)$	antiderivative $F(x)$
x^n	$\frac{1}{n+1}x^{n+1}$
$\frac{1}{x}$	$\ln \lvert x \rvert$
e^x	e^x
$\cos x$	$\sin x$
$\sin x$	$-\cos x$
$\sec^2 x$	$\tan x$
$\sec x \tan x$	$\sec x$
$\frac{1}{\sqrt{1-x^2}}$	$\arcsin x$
$\frac{1}{1+x^2}$	$\arctan x$

We could keep going and include the derivatives of the other trigonometric, inverse trigonometric, and hyperbolic functions, but it should quickly become apparent that this approach will only get us so far. For example, how are we to antidifferentiate something like $\ln x$? Does it even have an antiderivative?

First observe that since $(cF)' = c(F')$ for $c \in \mathbb{R}$, and $(F+G)' = F' + G'$, we can find an antiderivative of cf and $f+g$ provided we can antidifferentiate f and g individually.

Example 31.4. To find an antiderivative F of $f(x) = 2\sin x + \frac{2x^3 - \sqrt{x}}{x}$, we can rewrite the function as $f(x) = 2\sin x + 2x^2 - x^{-1/2}$ and find antiderivatives of each term individually, obtaining

$$F(x) = -2\cos x + \frac{2}{3}x^3 - 2\sqrt{x} + C.$$

This lets us assemble the functions from the table into a broader class of functions with antiderivatives, but still does not address all the possibilities. At this point it is useful to stop thinking about the *formulas* that define functions, and start thinking about other ways of representing functions. In particular, we consider the graph of f, which is a subset of $\mathbb{R}^2$. For concreteness, let $f(x) = x$, so that we know $F(x) = \frac{1}{2}x^2$ is an antiderivative. The graph of f is the line through the origin with slope 1, and we observe that for $x > 0$, the triangle with vertices at the origin, $(x, 0)$, and (x, x) has area $\frac{1}{2}x^2$. This triangle can be described in terms of the graph as the region that lies

underneath the graph of f, above the x-axis, to the right of the y-axis, and to the left of the vertical line through $(x, 0)$. So for this function at least, we have interpreted an antiderivative as the area of a specific region.

More generally, suppose that f is a positive function, and define a function F by letting $F(t)$ be the area of the region that is bounded above by the graph of f, to the left by the y-axis, below by the x-axis, and to the right by the line $x = t$. If we compare $F(t)$ and $F(t + h)$, we see that they are the areas of two regions which differ only in whether the strip from t to $t+h$ is included. This strip is close to being a rectangle with width h and height $f(t)$, so its area is approximately $h \cdot f(t)$. Thus we have the rough estimate

$$\frac{F(t+h) - F(t)}{h} = \frac{\text{area of the strip from } t \text{ to } t+h}{h} \approx \frac{h \cdot f(t)}{h} = f(t). \tag{31.1}$$

This suggests that as $h \to 0$, we may reasonably expect the difference quotient $\frac{F(t+h)-F(t)}{h}$ to converge to $f(t)$, which would show that $F'(t) = f(t)$.

There are a number of steps in this procedure that need to be made more precise.

(1) What exactly do we mean by "area"? How do we compute the area of a region that is irregular in shape?
(2) What does it mean to say that "this strip is close to being a rectangle"? Presumably we want some condition that guarantees that the approximation in (31.1) gets better and better as $h \to 0$; how do we guarantee this?

Obtaining precise and rigorous answers to these questions requires us to develop the theory of integration, which we do over the next few lectures. In a nutshell, the answers are as follows.

(1) We can introduce a precise definition of *definite integral* that plays the role of area for the regions we are concerned with, provided f is "nice enough". This leads to the notion of an *integrable function*.
(2) The approximation in (31.1) leads to correct conclusions whenever f is continuous. In particular, continuous functions are integrable, and every continuous function has an antiderivative that is produced via the procedure described above. This last fact is known as the *Fundamental Theorem of Calculus*.

Lecture 32 — Approximating areas by sums

Stewart §5.1, Spivak Chapter 13

In the next few lectures we develop the basic results of the theory of integration, including the definition of integrals and of integrable functions, the observation that continuous functions are integrable, and the Fundamental Theorem of Calculus that relates integration and differentiation. The treatment here is more theoretical and complete than that given in Stewart's book, but not as comprehensive as the one in Spivak's book. The order of the results, and the method of certain proofs, has been chosen to take the most efficient route that covers everything I want to say but does not bombard you with too much extra theory. In a number of places I have also borrowed ideas and

proofs from a book by Pete L. Clark,[32] which is closer to Spivak than to Stewart in its style but also has quite a few differences from Spivak.

32.1. Area and Riemann sums

Given a positive function f, last time we had the idea of producing an antiderivative F for f by letting $F(t)$ be the "area" of the region in $\mathbb{R}^2$ bounded by the x-axis, the graph of f, and the vertical lines $x = 0$ and $x = t$. But what do we mean by "area" of a region in $\mathbb{R}^2$? We start with three axioms that any reasonable notion of area should obey:

(1) The area of a rectangle is the product of its width and its height.
(2) If two regions X, Y do not overlap, or if their overlap has zero area, such as a straight line, then $\text{area}(X \cup Y) = \text{area}(X) + \text{area}(Y)$.
(3) If a region X is contained inside a region Y, then $\text{area}(X) \leq \text{area}(Y)$.

Let us write $\Gamma(f, a, b)$ for the region in $\mathbb{R}^2$ bounded by the x-axis, the graph of f, and the vertical lines $x = a$ and $x = b$. Let $\int_a^b f = \text{area}(\Gamma(f, a, b))$, for some reasonable notion of area. Then the three axioms above imply the following.

(1) If there is $C \in \mathbb{R}$ such that $f(x) = C$ for every $x \in [a, b]$, then $\int_a^b f = C(b - a)$.
(2) If $a < c < b$, then $\int_a^b f = \int_a^c f + \int_c^b f$.
(3) If $f(x) \leq g(x)$ for every $x \in [a, b]$, then $\int_a^b f \leq \int_a^b g$.

These axioms motivate the remainder of our discussion: we want to associate to every "reasonable" function $f\colon [a, b] \to \mathbb{R}$ a number $\int_a^b f$ in such a way that the properties above are satisfied. We will further clarify the meaning of "reasonable" later on.

Before making any formal definitions, we describe the general principle, which is to approximate $\Gamma(f, a, b)$ with a union of rectangles: for some large $n \in \mathbb{N}$ we choose points $a = x_0 < x_1 < \cdots < x_n = b$, and over each interval $[x_{i-1}, x_i]$ we draw a rectangle whose height is roughly equal to the value of the function on that interval. Adding up the areas of the rectangles gives us a number that we hope is a good approximation to the area of $\Gamma(f, a, b)$.

The set of points $\{x_i\}_{i=0}^n$ is said to *partition* the interval $[a, b]$ into smaller subintervals, and we often write $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$ to denote a specific partition. One way to choose the heights of the rectangles used in the approximation is to pick a point t_i inside each subinterval $[x_{i-1}, x_i]$, and use $f(t_i)$ as the height of the corresponding rectangle. We write τ to denote the set of points $(t_1, t_2, \ldots, t_n)$, and refer to the pair (P, τ) as a *tagged partition*. Each tagged partition corresponds to an approximation of $\Gamma(f, a, b)$ by rectangles, and the first two axioms tell us that the area of this approximation is

$$R(f, P, \tau) := \sum_{i=1}^{n} f(t_i)(x_i - x_{i-1}). \tag{32.1}$$

The notation $R(f, P, \tau)$ should be read as "the *Riemann sum* of f associated to the partition P with tags τ". The idea that drives the *Riemann integral* is that as we choose

[32] "Honors Calculus" by Pete L. Clark, Univ. of Georgia, `http://math.uga.edu/~pete/2400full.pdf`

tagged partitions for which the subinterval lengths $x_i - x_{i-1}$ get smaller and smaller, the corresponding Riemann sums should converge to a limit, which we will denote $\int_a^b f$. Making this precise, and giving conditions under which the limit actually exists, will require some nontrivial effort, and first we give an example that illustrates the idea.

Example 32.1. Consider $\int_0^b f$ when $f(x) = x$. Write P for the partition of $[0, b]$ into n subintervals of equal length $\frac{b}{n}$, and τ_{right} for tags chosen to be the right-hand endpoint of each subinterval. Thus $x_i = t_i = \frac{ib}{n}$, and the corresponding Riemann sum is

$$R(f, P, \tau_{\text{right}}) = \sum_{i=1}^{n} \frac{ib}{n} \cdot \frac{b}{n} = \frac{b^2}{n^2} \sum_{i=1}^{n} i = \frac{b^2}{n^2} \frac{n(n+1)}{2} = \frac{b^2}{2}\Big(1 + \frac{1}{n}\Big),$$

where we use the formula $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$. As $n \to \infty$, the length of the subintervals in the partition P goes to 0, and the Riemann sum converges to $\frac{b^2}{2}$, which we know from elementary geometry to be the area of the right triangle $\Gamma(f, 0, b)$. But what if we had chosen a different choice of tags? Would we still get the same conclusion? For example, suppose we let τ_{left} denote tags chosen at the left endpoint of each subinterval, so $t_i = x_{i-1} = \frac{(i-1)b}{n}$. Then we have

$$R(f, P, \tau_{\text{left}}) = \sum_{i=1}^{n} \frac{(i-1)b}{n} \cdot \frac{b}{n} = \frac{b^2}{n^2} \sum_{i=1}^{n} (i-1) = \frac{b^2}{n^2} \frac{(n-1)n}{2} = \frac{b^2}{2}\Big(1 - \frac{1}{n}\Big),$$

and once again we see that the Riemann sums converge to the appropriate limit. Of course, there are many other choices of tags we could make as well, and we could also have chosen different partitions, in which the subintervals did not have equal length. But it turns out that all of them lead to the same limit. Let us explain why the choice of tags does not affect the limit; we will postpone a discussion of unequal subintervals until later. Given the equal-length partition P into n subintervals, since f is increasing we have $f(x_{i-1}) \le f(x) \le f(x_i)$ for every $x \in [x_{i-1}, x_i]$. It follows that for *any* choice of tags τ, we will have

$$\frac{b^2}{2}\Big(1 - \frac{1}{n}\Big) = R(f, P, \tau_{\text{left}}) \le R(f, P, \tau) \le R(f, P, \tau_{\text{right}}) = \frac{b^2}{2}\Big(1 + \frac{1}{n}\Big). \tag{32.2}$$

By the squeeze theorem, we get the same limit no matter which tags we choose.

Lecture 33 Lower sums, upper sums, and integrals

Stewart §5.2, Spivak Chapter 13

33.1. Lower and upper sums

In general we need to deal with functions that are not necessarily increasing, but we can still obtain bounds similar to those in (32.2) as follows. Suppose that P is a partition of $[a, b]$ with endpoints x_i, and that $m_i, M_i \in \mathbb{R}$ have the property that $m_i \le f(x) \le M_i$

for all $x \in [x_{i-1}, x_i]$. Then for every choice of tags τ, we have $m_i \le f(t_i) \le M_i$, and thus

$$\sum_{i=1}^{n} m_i(x_i - x_{i-1}) \le \sum_{i=1}^{n} f(t_i)(x_i - x_{i-1}) = R(f, P, \tau) \le \sum_{i=1}^{n} M_i(x_i - x_{i-1}). \tag{33.1}$$

The first and last sums in (33.1) are not necessarily Riemann sums themselves, because there may be no choice of tags that gives $f(t_i) = m_i$ or $f(t_i) = M_i$. Nevertheless, these expressions turn out to be extremely useful for studying integrals and for developing the formal theory, because they allow us to invoke the third area axiom, which we have so far neglected, and observe that for each $1 \le i \le n$, we have

$$m_i(x_i - x_{i-1}) = \int_{x_{i-1}}^{x_i} m_i \le \int_{x_{i-1}}^{x_i} f \le \int_{x_{i-1}}^{x_i} M_i = M_i(x_i - x_{i-1}).$$

Here the two equalities use the first area axiom, and the two inequalities use the third axiom. Summing over i and using the second axiom gives

$$\sum_{i=1}^{n} m_i(x_i - x_{i-1}) \le \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f = \int_a^b f \le \sum_{i=1}^{n} M_i(x_i - x_{i-1}). \tag{33.2}$$

By comparing (33.1) and (33.2), we see that $\int_a^b f$ and $R(f, P, \tau)$ are both contained in the interval between the lower sum $\sum m_i(x_i - x_{i-1})$ and the upper sum $\sum M_i(x_i - x_{i-1})$. Intuitively, we would like to say that if we can make these two sums be arbitrarily close together by choosing x_i, m_i, M_i appropriately, then this gives a definition of the integral $\int_a^b f$ and shows that it is the limit of the Riemann sums.

33.2. The least upper bound property

To make this discussion precise, we need the following general notions.

Definition 33.1. An *upper bound* for a set $A \subset \mathbb{R}$ is a number $M \in \mathbb{R}$ such that $x \le M$ for every $x \in A$. We say that M is the *least upper bound* for A if every $M' < M$ is *not* an upper bound for A. In this case we also call M the *supremum* of A, and write $M = \sup A$. Equivalently, $M = \sup A$ if and only if the following are both true:

(1) $x \le M$ for every $x \in A$;
(2) for every $M' < M$, there is $x \in A$ such that $x > M'$.

If we reverse the direction of all inequalities in this definition, we get the definition of *lower bound* and *greatest lower bound*, also called *infimum* and denoted $\inf A$.

Example 33.2. $\sup[0, 1] = 1$ and $\inf[0, 1] = 0$, which suggests that sup and inf should be thought of as max and min, respectively. The difference is that the maximum of a set A must lie in A, while the supremum is not required to. Thus the open interval $(0, 1)$ has no max or min, while $\sup(0, 1) = 1$ and $\inf(0, 1) = 0$ still exist.

Recall that in an earlier lecture, we said that the defining characteristic of the real numbers (as compared to $\mathbb{Q}$) is the fact that every increasing sequence that is bounded above has a limit. This implies the following property.

Theorem 33.3 (Least Upper Bound property)**.** *If $A \subset \mathbb{R}$ is bounded above, then it admits a least upper bound in $\mathbb{R}$.*

Proof. Let b be an upper bound for A, and choose any $a \in A$. Define a pair of bisection sequences $a = a_1 \le a_2 \le a_3 \le \cdots \le b_3 \le b_2 \le b_1 = b$ by assigning each midpoint to b_n if it is an upper bound for A, and a_n if it is not. Thus every b_n is an upper bound for A, and no a_n is an upper bound for A. As usual, $c = \lim a_n = \lim b_n$ exists.

Given any $x \in A$, we have $x \le b_n$ for every n, and thus $c = \lim_{b_n} \ge x$. Thus c is an upper bound for A. Moreover, given any $c' < c$ there is $a_n \in (c', c)$, and since a_n is not an upper bound for A, neither is c'. Thus c is the least upper bound for A. □

This property fails in $\mathbb{Q}$: the set $A = \{p/q \in \mathbb{Q} : (p/q)^2 < 2\}$ is bounded above, but has no least upper bound in $\mathbb{Q}$. In $\mathbb{R}$, on the other hand, $\sqrt{2}$ is the least upper bound.

Exercise 33.4. Suppose we know that $\mathbb{R}$ has the least upper bound property, but do not yet know whether every bounded increasing sequence has a limit. Prove that if x_n is a bounded increasing sequence, then $\lim x_n$ exists and is equal to $\sup\{x_n : n \in \mathbb{N}\}$. This says that our description of $\mathbb{R}$ in terms of monotone convergence is equivalent to the description in terms of the least upper bound property, which is what Spivak uses.

Exercise 33.5. Use the least upper bound property to deduce that every set that is bounded below admits a greatest lower bound.

33.3. Lower and upper integrals

Returning to our discussion of $\int_a^b f$, which so far we have studied but not defined, we begin to make some formal definitions. In what follows, we require f to be bounded, since if a function can take arbitrarily large values then there is no reason to expect the associated area to be finite. However, we do not require f to be positive. If f is a negative function, then we should think of $\int_a^b f$ as a negative number whose absolute value is the area lying between the graph of f and the x-axis. If f takes both signs, then $\int_a^b f$ represents the difference between the area above the x-axis and the area below it.

Definition 33.6. As in the previous lecture, a *partition* of $[a, b]$ is a finite set $P \subset [a, b]$ that contains a and b. We will always write the elements of P in increasing order, so $P = \{a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b\}$.

Given a bounded function $f\colon [a, b] \to \mathbb{R}$ and a partition $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$ of $[a, b]$, consider for each $1 \le i \le n$ the quantities

$$m_i(f, P) = \inf\{f(x) : x \in [x_{i-1}, x_i]\} \quad \text{and} \quad M_i(f, P) = \sup\{f(x) : x \in [x_{i-1}, x_i]\}.$$

Note that the inf and sup exist because f is bounded above and below. (They may not be achieved, though; there may not be any $x \in [x_{i-1}, x_i]$ satisfying $f(x) = m_i(f, P)$ or $f(x) = M_i(f, P)$.) Define the *lower and upper sums* of f for P by

$$L(f, P) = \sum_{i=1}^{n} m_i(f, P)(x_i - x_{i-1}) \quad \text{and} \quad U(f, P) = \sum_{i=1}^{n} M_i(f, P)(x_i - x_{i-1}).$$

Recalling (33.2), we see that if we are ever to successfully define $\int_a^b f$ so that the three desired axioms are satisfied, then we must have

$$L(f, P) \le \int_a^b f \le U(f, P) \tag{33.3}$$

for every partition P of $[a,b]$. In other words $\int_a^b f$ should be simultaneously an upper bound for the set of all lower sums $L(f,P)$, and a lower bound for the set of all upper sums $U(f,P)$. This implies that

$$\underbrace{\sup\{L(f,P) : P \text{ is a partition of } [a,b]\}}_{\underline{\int}_a^b f} \le \int_a^b f \le \underbrace{\inf\{U(f,P) : P \text{ is a partition of } [a,b]\}}_{\overline{\int}_a^b f}.$$

We will refer to the quantities $\underline{\int}_a^b f$ and $\overline{\int}_a^b f$ as the *lower and upper integrals of* f, respectively. The lower integral can be thought of as the area of $\Gamma(f,a,b)$ that we can detect by approximating it from inside by rectangles; the upper integral is the area that we can detect by approximating it from outside by rectangles. Now we come to the crucial definition.

Definition 33.7. A function $f\colon [a,b] \to \mathbb{R}$ is *integrable* if it is bounded and if $\underline{\int}_a^b f = \overline{\int}_a^b f$. In this case the common value is called the *integral* of f on $[a,b]$ and is denoted $\int_a^b f$.

Remark 33.8. The quantity $\int_a^b f$ is often referred to as the *definite integral* to emphasize that it is a single number associated to a definite interval $[a,b]$. It is often denoted as $\int_a^b f(x)\,dx$, especially when we write the function f explicitly; for example, if $f(x) = x^2$, then we would write $\int_a^b f = \int_a^b x^2\,dx$. This has the same meaning as $\int_a^b t^2\,dt$, and as $\int_a^b y^2\,dy$, and so on. The symbol "$\int$" is called an *integral sign* and can be thought of as an elongated "S" (since integration is related to summation). The function f is called the *integrand*, and the values a, b are called the *limits of integration*.[33]

Example 33.9. With Example 32.1 in mind, let us consider $f(x) = x^2$ on $[0,b]$. Once again, we consider the partition $P_n = \{0 < \frac{b}{n} < \frac{2b}{n} < \cdots < \frac{(n-1)b}{n} < b\}$, and observe that since f is increasing, we have $m_i = f(x_{i-1}) = (\frac{(i-1)b}{n})^2$ and $M_i = f(x_i) = (\frac{ib}{n})^2$. Thus

$$L(f,P_n) = \sum_{i=1}^{n} \Big(\frac{(i-1)b}{n}\Big)^2 \frac{b}{n} = \frac{b^3}{n^3}(0^2 + 1^2 + 2^2 + 3^2 + \cdots + (n-1)^2),$$

$$U(f,P_n) = \sum_{i=1}^{n} \Big(\frac{ib}{n}\Big)^2 \frac{b}{n} = \frac{b^3}{n^3}(1^2 + 2^2 + 3^2 + 4^2 + \cdots + n^2).$$

It can be proved by induction that $1^2 + 2^2 + \cdots + n^2 = \frac{1}{6}n(n+1)(2n+1)$ – this is a good exercise to do if you haven't before – and thus

$$L(f,P_n) = \frac{b^3}{n^3} \cdot \frac{1}{6}(n-1)n(2n-1), \qquad U(f,P_n) = \frac{b^3}{n^3} \cdot \frac{1}{6}n(n+1)(2n+1).$$

In particular, as $n \to \infty$, both $L(f,P_n)$ and $U(f,P_n)$ converge to $\frac{b^3}{3}$. This strongly suggests that f is integrable, but is not quite a complete proof of this fact, since the definition of lower and upper integrals involves taking a supremum and infimum over *all* partitions, not just equally spaced ones. If we accept that f is integrable, the computations here show that $\int_0^b x^2\,dx = \frac{b^3}{3}$. In the next lecture, we will see an easier

[33] Note that this has nothing to do with our usual use of the word "limit".

way to both establish integrability rigorously and to compute the value of the integral without relying on summation formulas.

It is worth pointing out that not every bounded function is integrable.

Exercise 33.10. Define a function $f\colon [0,1] \to \mathbb{R}$ by $f(x) = 0$ if x is irrational, and $f(x) = 1$ if x is rational. Prove that $\underline{\int}_0^1 f = 0$ and $\overline{\int}_0^1 f = 1$, and thus f is not integrable.

We end this lecture by observing that lower and upper integrals both satisfy the three desired axioms. Two of these are easy and we leave them as exercises; the remaining axiom takes a short proof.

Exercise 33.11. Prove that $\underline{\int}_a^b C = \overline{\int}_a^b C = C(b-a)$ for every $a < b$ and $C \in \mathbb{R}$.

Exercise 33.12. Prove that if $f, g\colon [a,b] \to \mathbb{R}$ are bounded functions satisfying $f(x) \le g(x)$ for all $x \in [a,b]$, then $\underline{\int}_a^b f \le \underline{\int}_a^b g$ and $\overline{\int}_a^b f \le \overline{\int}_a^b g$.

Proposition 33.13. *If $f\colon [a,b] \to \mathbb{R}$ is bounded and $c \in [a,b]$, then $\underline{\int}_a^b f = \underline{\int}_a^c f + \underline{\int}_c^b f$ and $\overline{\int}_a^b f = \overline{\int}_a^c f + \overline{\int}_c^b f$.*

Proof. We give the proof for $\underline{\int}$ and leave the other half as an exercise. First observe that if Q and R are any partitions of $[a,c]$ and $[c,b]$, respectively, then $P := Q \cup R$ is a partition of $[a,b]$, and

$$L(f,P) = L(f,Q) + L(f,R).$$

Every partition P of $[a,b]$ that contains c arises in this way, and so taking a supremum over all Q and R gives

$$\underbrace{\sup\{L(f,P) : P \text{ is a partition of } [a,b] \text{ and } c \in P\}}_{\mathrm{I}} = \underline{\int}_a^c f + \underline{\int}_c^b f. \tag{33.4}$$

The quantity I is clearly $\le \underline{\int}_a^b f$. To conclude the proof we need to consider partitions of $[a,b]$ that do *not* contain c, and use the following lemma to see what happens when the point c is added to the partition.

Lemma 33.14. *If $f\colon [a,b] \to \mathbb{R}$ is bounded and P is a partition of $[a,b]$, then for every $c \in [a,b]$ we have $L(f,P) \le L(f, P \cup \{c\})$ and $U(f,P) \ge U(f, P \cup \{c\})$.*

Proof. We prove the inequality for L; the upper sums are left as an exercise. Let $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$ and choose $k \in \{1, \dots, n\}$ such that $c \in [x_{k-1}, x_k]$. Let

$$m_c^\ell := \inf\{f(x) : x \in [x_{k-1}, c]\} \quad \text{and} \quad m_c^r := \inf\{f(x) : x \in [c, x_k]\}.$$

Then we have

$$\begin{aligned} L(f, P \cup \{c\}) &= \left(\sum_{i \ne k} m_i(f,P)(x_i - x_{i-1}) \right) + m_c^r(x_k - c) + m_c^\ell(c - x_{k-1}) \\ &\ge \left(\sum_{i \ne k} m_i(f,P)(x_i - x_{i-1}) \right) + m_k(f,P)\big((x_k - c) + (c - x_{k-1})\big) \end{aligned}$$

$$= \sum_{i=1}^{n} m_i(f,P)(x_i - x_{i-1}) = L(f,P),$$

which proves the lemma. □

Returning to the proof of Proposition 33.13, we use Lemma 33.14 to deduce that

$$\underline{\int_a^b} f = \sup\{L(f,P) : P \text{ is a partition of } [a,b]\}$$
$$\leq \underbrace{\sup\{L(f, P \cup \{c\}) : P \text{ is a partition of } [a,b]\}}_{\mathrm{I}} \leq \underline{\int_a^b} f$$

and thus $\mathrm{I} = \underline{\int_a^b} f$; together with (33.4), this proves the proposition. □

Lecture 34 The Fundamental Theorem of Calculus

Stewart §5.3 and §5.4, Spivak Chapters 13 and 14

34.1. The Fundamental Theorem of Calculus

Now we address two important questions.

- Which functions are integrable?
- Does integration in fact give us an antiderivative, as we originally hoped?

We start with the following result: it is a little bit clunky because it works with $\underline{\int}$ and $\overline{\int}$, but it quickly implies three very important results that have cleaner formulations.

Theorem 34.1 (Fundamental Theorem of Calculus, Part 0)**.** *Let $f\colon [a,b] \to \mathbb{R}$ be a bounded function, and define $L, U\colon [a,b] \to \mathbb{R}$ by $L(x) := \underline{\int_a^x} f$ and $U(x) := \overline{\int_a^x} f$. If f is continuous at a point $c \in (a,b)$, then L and U are both differentiable at c, and satisfy $L'(c) = U'(c) = f(c)$.*

Proof. We prove the statement regarding $L(x)$; the proof for $U(x)$ is identical, since all we use are the properties of $\underline{\int}$ and $\overline{\int}$ from Proposition 33.13. Fix $c \in (a,b)$ such that f is continuous at c. Given $h > 0$, consider the quantities

$$m_h := \inf\{f(t) : c \leq t \leq c+h\} \quad \text{and} \quad M_h := \sup\{f(t) : c \leq t \leq c+h\}.$$

It follows from the first two statements in Proposition 33.13 that

$$m_h \cdot h = \underline{\int_c^{c+h}} m_h \leq \underline{\int_c^{c+h}} f \leq \underline{\int_c^{c+h}} M_h = M_h \cdot h.$$

The third statement in Proposition 33.13 gives

$$L(c+h) = \underline{\int_a^{c+h}} f = \underline{\int_a^{c}} f + \underline{\int_c^{c+h}} f = L(c) + \underline{\int_c^{c+h}} f.$$

Combining these, we get

$$m_h \le \frac{L(c+h) - L(c)}{h} \le M_h,$$

and a similar argument shows that it remains true when $h < 0$. Because f is continuous, we have $\lim_{h\to 0} m_h = \lim_{h\to 0} M_h = f(c)$, and so the squeeze theorem implies that $\lim_{h\to 0} \frac{1}{h}(L(c+h) - L(c))$ exists and is equal to $f(c)$, which proves the theorem. □

Remark 34.2. The only properties of $\underline{\int}$ and $\overline{\int}$ that we used in Theorem 34.1 were the three axioms; in particular, we did not need any knowledge of how these quantities are computed via partitions.

Theorem 34.1 has several tremendously important consequences.

Corollary 34.3. *If $f\colon [a,b] \to \mathbb{R}$ is continuous, then it is integrable.*[34]

Proof. By Theorem 34.1, $L(x) := \underline{\int}_a^x f$ and $U(x) := \overline{\int}_a^x f$ both define antiderivatives for f on $[a,b]$. By Theorem 22.1, this implies that $L - U$ is constant. But $L(a) = 0 = U(a)$, so we must have $L(x) = U(x)$ for all $x \in [a,b]$. In particular, $\underline{\int}_a^b f = L(b) = U(b) = \overline{\int}_a^b f$, so f is integrable. □

Corollary 34.4 (Fundamental Theorem of Calculus, Part 1)**.** *If $f\colon [a,b] \to \mathbb{R}$ is integrable and is continuous at a point $x \in (a,b)$, then $F(x) := \int_a^x f$ is differentiable at x and satisfies $F'(x) = f(x)$.*

In particular, to our original question about antiderivatives, we can now answer that every continuous function $f\colon [a,b] \to \mathbb{R}$ has an antiderivative, given by $F(x) = \int_a^x f$. Moreover, the correspondence goes both ways: if we happen to know an antiderivative of f, we can use it to compute the definite integral.

Example 34.5. We can use Corollary 34.4 together with the chain rule to differentiate functions where the upper limit of integration depends on x in a more complicated way:

$$\begin{aligned}\frac{d}{dx}\int_1^{x^2} \sqrt{1+t}\,dt &= \frac{d}{dx}\int_1^u \sqrt{1+t}\,dt \qquad\qquad (u = x^2)\\ &= \frac{du}{dx}\frac{d}{du}\int_1^u \sqrt{1+t}\,dt = (2x)\sqrt{1+u} = 2x\sqrt{1+x^2}.\end{aligned}$$

Corollary 34.6 (Fundamental Theorem of Calculus, Part 2)**.** *If $f\colon [a,b] \to \mathbb{R}$ is continuous and $F\colon [a,b] \to \mathbb{R}$ is an antiderivative of f, then $\int_a^b f = F(b) - F(a)$.*

Proof. By the previous corollary, $G(x) := \int_a^x f$ is an antiderivative of f, so $G - F$ is constant. Thus $G(b) - F(b) = G(a) - F(a) = -F(a)$, since $G(a) = \int_a^a f = 0$, which gives $G(b) = F(b) - F(a)$ and proves the corollary. □

[34]This theorem is proved in multiple ways in Spivak's book and in Clark's. To my mind, the proof here, which follows pages 292–293 in Spivak, is the simplest, but it does require setting up the FTC in a slightly non-standard way at first – usually the FTC is not stated for lower and upper integrals, but only for the integral itself – and there are certainly other ways of proving this result.

Example 34.7. Corollary 34.6 gives us a faster way to evaluate the integrals $\int_0^b x\,dx$ and $\int_0^b x^2\,dx$ from Examples 32.1 and 33.9:

$$\int_0^b x\,dx = \frac{1}{2}x^2\Big|_0^b = \frac{1}{2}b^2 - \frac{1}{2}0^2 = \frac{b^2}{2},$$
$$\int_0^b x^2\,dx = \frac{1}{3}x^3\Big|_0^b = \frac{1}{3}b^3 - \frac{1}{3}0^3 = \frac{b^3}{3}.$$

Here the notation $F(x)|_a^b$ is shorthand for $F(b) - F(a)$, and is especially convenient to use when the expression for $F(x)$ is quite complicated.

Remark 34.8. Corollary 34.6 actually remains true under the weaker assumption that f is integrable, but this requires a different proof using the MVT; see Theorem 14.2 in Spivak's book for details.

Remark 34.9. With a little more effort one can show that if $f\colon [a,b] \to \mathbb{R}$ is bounded and is continuous everywhere except at a finite set of points, then it is integrable; see §8.3.2 in Clark's book.

Remark 34.10. With a lot more effort one can give a complete characterization of which bounded functions are integrable: see §8.5 in Clark's book.

Example 34.11. Before applying Corollary 34.6, one must check that the conditions are satisfied. A too-naive application of the FTC gives

$$\int_{-2}^1 \frac{1}{x^2}\,dx = -\frac{1}{x}\Big|_{-2}^1 = -\frac{1}{1} - \left(\frac{-1}{-2}\right) = -1 - \frac{1}{2} = -\frac{3}{2},$$

despite the fact that $f \geq 0$ and so we expect to get a nonnegative integral. The issue is that we cannot apply Corollary 34.6 since $1/x^2$ is not continuous on the interval $[-2,1]$; in fact, it is not even bounded, so it is not integrable.

34.2. Basic properties of integrals

The following properties of definite integrals are immediate consequences of the corresponding properties in Exercise 33.11, Exercise 33.12, and Proposition 33.13.

Theorem 34.12. *Let $f, g\colon [a,b] \to \mathbb{R}$ be integrable functions. Then for every $c \in (a,b)$ and every $C \in \mathbb{R}$ we have*

(1) $\int_a^b C = C(b-a)$.
(2) $\int_a^b f = \int_a^c f + \int_c^b f$.
(3) If $f \leq g$ everywhere on $[a,b]$, then $\int_a^b f \leq \int_a^b g$.

Exercise 34.13. Prove that Theorem 34.12 continues to hold without any assumption on the ordering of a, b, c if we define $\int_b^a f = -\int_a^b f$ and $\int_a^a f = 0$.

Remark 34.14. The following two consequences of Theorem 34.12 come up so often that they are worth mentioning explicitly. Here $f\colon [a,b] \to \mathbb{R}$ is integrable and $m, M \in \mathbb{R}$.

$$f \geq 0 \quad \Rightarrow \quad \int_a^b f \geq 0, \tag{34.1}$$

$$m \le f \le M \quad \Rightarrow \quad m(b-a) \le \int_a^b f \le M(b-a). \tag{34.2}$$

Using Theorem 34.12, we can use the FTC to differentiate integrals where the lower limit of integration is variable: If f is integrable on $[a,b]$ then for $x \in (a,b)$ we have

$$\frac{d}{dx}\int_x^b f = \frac{d}{dx}\Big(\int_a^b f - \int_a^x f\Big) = -\frac{d}{dx}\int_a^x f = -f(x). \tag{34.3}$$

We can go one step further and differentiate integrals where both limits of integration are variable.

Example 34.15.

$$\frac{d}{dx}\int_{x^2}^{x^3} \sin t\,dt = \frac{d}{dx}\int_{x^2}^{a} \sin t\,dt + \frac{d}{dx}\int_a^{x^3} \sin t\,dt = (-2x)\sin(x^2) + (3x^2)\sin(x^3).$$

What happens to the integral $\int_a^b f$ if we multiply f by a scalar, or if we add two integrable functions together? First we consider this question in the case when f and g are continuous.

Theorem 34.16. *If $f, g\colon [a,b] \to \mathbb{R}$ are continuous and $c \in \mathbb{R}$, then*

$$\int_a^b (cf) = c\int_a^b f \quad \textit{and} \quad \int_a^b (f+g) = \int_a^b f + \int_a^b g.$$

Proof. By FTC1, f and g have antiderivatives $F, G\colon [a,b] \to \mathbb{R}$. By linearity of differentiation, we have $(cF)'(x) = c(F'(x)) = cf(x)$ and $(F+G)'(x) = F'(x) + G'(x) = f(x) + g(x) = (f+g)(x)$ for all $x \in (a,b)$, so cF and $F+G$ are antiderivatives of cf and $f+g$, respectively. By FTC2, this implies that

$$\int_a^b (cf) = (cF)(b) - (cF)(a) = c(F(b) - F(a)) = c\int_a^b f,$$
$$\int_a^b (f+g) = (F+G)(b) - (F+G)(a) = F(b) - F(a) + G(b) - G(a) = \int_a^b f + \int_a^b g,$$

which proves the theorem. □

Remark 34.17. In fact, a similar result holds for all integrable functions, not just continuous ones, but the proof is harder; see Theorem 35.16 below.

Lecture 35 More about integration

Spivak Chapter 13

35.1. Definitions via integrals

Using integrals, we can give alternate definitions of various numbers and functions that we have encountered so far. For example, a circle with radius 1 has area π, so half of the circle has area $\pi/2$. In particular, this is the area of the region between the curve $y = \sqrt{1-x^2}$ and the x-axis over the interval $[-1, 1]$, and we conclude that

$$\pi = 2\int_{-1}^{1} \sqrt{1-x^2}\,dx. \tag{35.1}$$

This gives one way of defining the number π.

Similarly, certain functions can be defined quite simply using integrals. Remember the amount of work we had to go through to define the exponential and logarithmic functions in earlier lectures. Since $\frac{d}{dx}\ln x = \frac{1}{x}$ and $\ln 1 = 0$, the FTC gives

$$\int_1^x \frac{1}{t}\,dt = \ln t\Big|_1^x = \ln x - \ln 1 = \ln x,$$

and thus we could just as well *define* the natural logarithm by $\ln x := \int_1^x \frac{1}{t}\,dt$. Indeed, this is exactly the approach taken in Spivak's book. Once ln has been defined, the function $x \mapsto e^x$ is then defined to be its inverse.

A similar approach works for trigonometric functions. Earlier we demonstrated that $\frac{d}{dx}\sin^{-1} x = \frac{1}{\sqrt{1-x^2}}$, so instead of the geometric definition, we could give an analytic definition of arcsine as

$$\arcsin x := \int_0^x \frac{1}{\sqrt{1-t^2}}\,dt.$$

Then sine can be defined as the inverse function of this, and extended periodically, and we recover all the trigonometric functions from the usual identities.

35.2. Indefinite integrals

Thanks to the Fundamental Theorem of Calculus, we have now seen that an antiderivative F of a continuous function f can be produced by defining $F(x) = \int_a^x f$, and the previous section showed that some important functions can be defined this way. Since such functions are obtained by integration, it is common to use the following terminology.

Definition 35.1. An antiderivative F of a continuous function f is also called an *indefinite integral* of f, and denoted $F(x) = \int f(x)\,dx$.

Remark 35.2. The word "integral" is used to refer both to definite and indefinite integrals. Recall that given a function f,

- the *definite integral* of f is a *number* associated to a specific interval $[a, b]$, so we must always specify the limits of integration and write $\int_a^b f(x)\,dx$; while
- an *indefinite integral* of f is a *function* that is not associated to any specific interval, and thus the notation $\int f(x)\,dx$ does not include any limits of integration.

We also stress that because a function has many antiderivatives on any given interval, indefinite integrals are only determined up to a constant, and indeed it is most appropriate to think of the indefinite integral of f as being a whole *family* of functions.

Example 35.3. The indefinite integral of x^2 is any of the family of functions $\frac{1}{3}x^3 + C$, where C is a constant, and we write

$$\int x^2\,dx = \frac{1}{3}x^3 + C.$$

The constant C is often called a *constant of integration*. We can look up many indefinite integrals directly using our table of antiderivatives from an earlier lecture. We can also use linearity to compute many indefinite integrals:

$$\int (5x^3 + 2\sec^2 x)\,dx = \frac{5}{4}x^4 + 2\tan x + C.$$

Note that even though we combine two functions here, only a single constant of integration is needed.

Sometimes a little bit of manipulation is needed in order to put a function in a form from which we can easily evaluate the indefinite integral:

$$\int \frac{\sin x}{\cos^2 x}\,dx = \int \frac{1}{\cos x}\frac{\sin x}{\cos x}\,dx = \int \sec x \tan x\,dx = \sec x + C.$$

Using the Fundamental Theorem of Calculus, we can write definite integrals in terms of indefinite integrals:

$$\int_0^1 \frac{1}{x^2+1}\,dx = \Big[\int \frac{1}{x^2+1}\,dx\Big]_0^1 = \Big[\arctan x\Big]_0^1 = \arctan(1) - \arctan(0) = \frac{\pi}{4}.$$

Technically the indefinite integral above should be "$\arctan x + C$" to include the constant of integration. However, because it does not matter *which* antiderivative we use in the FTC, the constant of integration can be omitted when computing a definite integral using indefinite integrals.

The FTC2 can be reformulated as follows.

Theorem 35.4 (Net Change Theorem)**.** *If $F\colon [a,b] \to \mathbb{R}$ is differentiable on (a,b), then $\int_a^b F'(x)\,dx = F(b) - F(a)$.*

The Net Change Theorem says that the net change in a quantity F is the integral of its rate of change. In applications of this theorem, we often (but not always) interpret F as a function of time t.

Example 35.5. If $V(t)$ represents the volume of water in a container at time t, then the Net Change Theorem says that $\int_{t_1}^{t_2} V'(t)\,dt = V(t_2) - V(t_1)$: the integral of the rate at which water enters or leaves the container is equal to the net change in the volume of water in the container.

Example 35.6. If $s(t)$ denotes the position of an object at time t, then $v(t) = s'(t)$ is its velocity, and the net displacement from time t_1 to time t_2 is $s(t_2) - s(t_1) = \int_{t_1}^{t_2} v(t)\,dt$. For example, if v is positive from t_1 to t', negative from t' to t'', and then positive from t'' to t_2, and if A_1, A_2, A_3 denote the areas of the regions between the curve and the x-axis in these three intervals, then the net displacement is $s(t_2) - s(t_1) = A_1 - A_2 + A_3$. The negative sign on A_2 comes because between times t' and t'' the object is moving in the negative direction.

If we want to compute the total distance traveled, then we need to integrate not v, but $|v|$: the total distance traveled is $\int_{t_1}^{t_2} |v(t)|\,dt$, which in the situation above is equal to $A_1 + A_2 + A_3$.

Example 35.7. To make the previous example concrete, suppose that the velocity at time t is $v(t) = t^2 - t - 6$ and that t ranges from 1 to 4. Then the net displacement is

$$\begin{aligned}\int_1^4 v(t)\,dt &= \int_1^4 (t^2 - t - 6)\,dt = \Big[\frac{1}{3}t^3 - \frac{1}{2}t^2 - 6\Big]_1^4 \\ &= \Big(\frac{64}{3} - 8 - 24\Big) - \Big(\frac{1}{3} - \frac{1}{2} - 6\Big) = -\frac{9}{2}.\end{aligned}$$

Note that $v(t) = (t-3)(t+2)$ is negative on $[1,3)$ and positive on $(3,4]$, so the total distance traveled is

$$\int_1^4 |v(t)|\,dt = \int_1^3 -v(t)\,dt + \int_3^4 v(t)\,dt - \Big[\frac{1}{3}t^3 - \frac{1}{2}t^2 - 6\Big]_1^3 + \Big[\frac{1}{3}t^3 - \frac{1}{2}t^2 - 6\Big]_3^4$$

which is equal to $\frac{61}{6}$ (after a little computation).

35.3. Approximations by lower and upper sums

Lemma 33.14 said that if we refine a partition by adding another point (thus dividing one of the subintervals into two pieces), then the corresponding lower sum does not get smaller, and the corresponding upper sum does not get bigger. Iterating this procedure gives the following result.

Lemma 35.8. *If $f\colon [a,b] \to \mathbb{R}$ is bounded and P, Q are partitions of $[a,b]$ with $P \subset Q$, then $L(f,P) \le L(f,Q)$ and $U(f,P) \ge U(f,Q)$.*

Proof. Choose partitions $P = P_0 \subset P_1 \subset P_2 \subset \cdots \subset P_m = Q$ such that each P_i is obtained from P_{i-1} by adding a single point. Then by Lemma 33.14,

$$L(f,P) = L(f,P_0) \le L(f,P_1) \le L(f,P_2) \le \cdots \le L(f,P_m) = L(f,Q).$$

The proof for U follows the same argument. □

Remark 35.9. The definition of lower and upper sums immediately gives $L(f,P) \le U(f,P)$ for every partition P, but did not immediately tell us that $L(f,P) \le U(f,Q)$ when P, Q are two different partitions. In particular, so far we did not even prove that $\underline{\int}_a^b f \le \overline{\int}_a^b f$ in general! Now we can do this, using Lemma 35.8. Indeed, given two partitions P, Q of $[a,b]$, Lemma 35.8 gives

$$L(f,P) \le L(f,P \cup Q) \le U(f,P \cup Q) \le U(f,Q).$$

Fixing Q and taking a supremum over all P gives $\underline{\int}_a^b f \le U(f,Q)$, and then taking an infimum over all Q gives $\underline{\int}_a^b f \le \overline{\int}_a^b f$. Putting it all together, we have shown that for any partitions P, Q and any bounded f, we have

$$L(f,P) \le L(f,P\cup Q) \le \underline{\int_a^b} f \le \overline{\int_a^b} f \le U(f,P\cup Q) \le U(f,Q). \tag{35.2}$$

Lemma 35.10. *Let $f\colon [a,b] \to \mathbb{R}$ be bounded. Then the following are equivalent:*

(1) f is integrable;
(2) for every $\varepsilon > 0$, there exists a partition P of $[a,b]$ such that $U(f,P) - L(f,P) < \varepsilon$.

Proof. If f is integrable, then by the definition of $\underline{\int}$ and $\overline{\int}$ there are partitions Q, R such that

$$\Big(\underline{\int_a^b} f\Big) - \frac{\varepsilon}{2} < L(f,Q) \le \underline{\int_a^b} f \le \overline{\int_a^b} f \le U(f,R) < \Big(\overline{\int_a^b} f\Big) + \frac{\varepsilon}{2},$$

and by (35.2) the partition $P = Q \cup R$ satisfies

$$U(f,P) - L(f,P) \le U(f,R) - L(f,Q) < \varepsilon.$$

Conversely, for every partition P we have

$$\overline{\int_a^b} f - \underline{\int_a^b} f \le U(f,P) - L(f,P),$$

so if the RHS can be made arbitrarily small by choosing P appropriately, then the LHS, which does not depend on P, must be equal to 0, so f is integrable. □

35.4. Riemann sums

In Examples 32.1 and 33.9, we saw that in order to compute the integrals $\int_0^b x\,dx$ and $\int_0^b x^2\,dx$, we did not really need to consider all partitions, or consider the lower and upper sums. It was enough to consider a single sequence of partitions for which the lengths of the subintervals became arbitrarily small, and to compute the Riemann sums associated to some convenient choice of tags. In fact, this always works, provided $f\colon [a,b] \to \mathbb{R}$ is integrable.

First we observe that as shown in (33.1), for every partition P and every choice of tags τ, we have

$$L(f,P) \le R(f,P,\tau) \le U(f,P),$$

where L, U are the lower and upper sums used in the definition of the (Darboux) integral and R is the Riemann sum associated to the tagged partition (P,τ). Moreover, by Lemma 35.10, if f is integrable then for every $\varepsilon > 0$ there is a partition P such that $U(f,P) - L(f,P) < \varepsilon$. In this case, every choice of tags τ will give a Riemann sum with the property that

$$\Big|R(f,P,\tau) - \int_a^b f\Big| < \varepsilon,$$

Thus the real question is to determine how we can find a partition with this 'good approximation' property. Intuitively we might expect that making the partition's subintervals small enough guarantees that the approximation is good, and indeed this turns out to be the case. To give a precise statement, we make the following definition

Definition 35.11. The *mesh* of a partition $P\{a = x_0 < x_1 < \cdots < x_n = b\}$ is

$$\operatorname{mesh}(P) := \max\{x_i - x_{i-1} : 1 \le i \le n\}.$$

Theorem 35.12. *Let $f\colon [a,b] \to \mathbb{R}$ be integrable. For every $\varepsilon > 0$, there exists $\delta > 0$ such that if P is any partition with $\operatorname{mesh}(P) < \delta$, then $U(f,P) - L(f,P) < \varepsilon$. In particular, for every tagged partition (P,τ) satisfying $\operatorname{mesh}(P) < \delta$, we have $|R(f,P,\tau) - \int_a^b f| < \varepsilon$.*

Proof. By Lemma 35.10, since f is integrable, there is a partition Q such that $U(f,Q) - L(f,Q) < \frac{\varepsilon}{2}$. Let N be the number of subintervals in Q, so $Q = \{a = y_0 < y_1 < \cdots < y_N = b\}$. Since f is bounded, there is $K > 0$ such that $|f(x)| \le K$ for all $x \in [a,b]$. Choose $\delta > 0$ small enough that $2KN\delta < \frac{\varepsilon}{4}$.

Now given any partition P with $\operatorname{mesh}(P) < \delta$, it follows from (35.2) that the partition $R = P \cup Q$ has $U(f,R) - L(f,R) < \frac{\varepsilon}{2}$. We claim that

$$U(f,P) < U(f,R) + \tfrac{\varepsilon}{4} \quad \text{and} \quad L(f,P) > L(f,R) - \tfrac{\varepsilon}{4}, \tag{35.3}$$

which will imply that $U(f,P) - L(f,P) < U(f,R) - L(f,R) + \frac{\varepsilon}{2} < \varepsilon$, and complete the proof. So our goal is to prove (35.3).

We prove the second half of (35.3); the first half is similar. Writing $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$, we have

$$L(f,P) = \sum_{i=1}^{n} m_i(f,P)(x_i - x_{i-1}).$$

For each $i \in \{1,\dots,n\}$, there are two possibilities.

- Case 1: there are no elements of R between x_{i-1} and x_i. In this case $[x_{i-1}, x_i]$ is also a subinterval in the partition R, and makes exactly the same contribution to the sum that defines $L(f,R)$.
- Case 2: there is at least one element of R between x_{i-1} and x_i. In this case, $R \cap [x_{i-1}, x_i]$ gives a partition of $[x_{i-1}, x_i]$ that we can write as $\{x_{i-1} = z_0 < z_1 < \cdots < z_\ell = x_i\}$, and the corresponding lower sum is

$$\begin{aligned}\sum_{j=1}^{\ell} \inf\{f(x) : z_{j-1} \le x \le z_j\}(z_j - z_{j-1}) &\le \sum_{j=1}^{\ell} (m_i(f,P) + 2K)(z_j - z_{j-1}) \\ &= (m_i(f,P) + 2K)(x_i - x_{i-1}) < m_i(f,P)(x_i - x_{i-1}) + 2K \operatorname{mesh}(P).\end{aligned}$$

Now we sum over all n values of i to get an estimate on $L(f,R)$. Because R contains at most N elements that are not one of the x_i's, Case 2 can happen at most N times, and we conclude that

$$L(f,R) \le \Big(\sum_{i=1}^{n} m_i(f,P)(x_i - x_{i-1}) \Big) + 2KN \operatorname{mesh}(P) = L(f,P) + 2KN \operatorname{mesh}(P).$$

When $\operatorname{mesh}(P) < \delta$, this gives

$$L(f,R) \le L(f,P) + 2KN\delta < L(f,P) + \tfrac{\varepsilon}{4},$$

which proves the second half of (35.3). The first half is similar, and as explained above, this completes the proof of Theorem 35.12. □

Remark 35.13. In particular, Theorem 35.12 shows that if (P_n, τ_n) is *any* sequence of tagged partitions of $[a,b]$ satisfying $\lim_{n\to\infty} \operatorname{mesh}(P_n) = 0$, then $\lim_{n\to\infty} R(f,P_n,\tau_n) = \int_a^b f$ for every integrable $f\colon [a,b] \to \mathbb{R}$. This includes the case when P_n is the partition into n subintervals of equal length, but is not restricted to that choice. Indeed, there are cases in which it is more convenient to use partitions into subintervals of unequal length, such as when f is known only approximately via data that has been collected for certain specific values of x that are not evenly spaced.

Remark 35.14. A function satisfying the conclusion of Theorem 35.12 is often called *Riemann integrable*, while a function satisfying $\underline{\int}_a^b f = \overline{\int}_a^b f$ is called *Darboux integrable*. Theorem 35.12 says that Darboux integrable functions are Riemann integrable, and the converse direction is straightforward (we leave it as Exercise 35.15 below), so in fact Riemann and Darboux integrability are equivalent, which justifies our cavalier use of the word "integrable" without specifying which notion is meant.

Exercise 35.15. Suppose that $f\colon [a,b] \to \mathbb{R}$ is a bounded function and that there is $I \in \mathbb{R}$ such that for every $\varepsilon > 0$, there exists $\delta > 0$ such that for every tagged partition (P,τ) satisfying $\operatorname{mesh}(P) < \delta$, we have $|R(f,P,\tau) - I| < \varepsilon$. Prove that f is (Darboux) integrable and that $\int_a^b f = I$.

35.5. Linearity of integrals

As mentioned in Remark 34.17, the linearity properties in Theorem 34.16 continue to hold for all integrable functions.

Theorem 35.16. *If $f, g\colon [a,b] \to \mathbb{R}$ are integrable and $c \in \mathbb{R}$, then*

(1) $cf\colon [a,b] \to \mathbb{R}$ is integrable, and $\int_a^b (cf) = c\int_a^b f$;

(2) $f+g\colon [a,b] \to \mathbb{R}$ is integrable, and $\int_a^b (f+g) = \int_a^b f + \int_a^b g$.

Before proving the theorem we point out the following example demonstrating the need for integrability.

Exercise 35.17. Let $f\colon [0,1] \to \mathbb{R}$ be the function from Exercise 33.10 and let $g = 1 - f$. Prove that $\underline{\int}_0^1 (-f) \neq -\underline{\int}_0^1 f$, and $\underline{\int}_0^1 (f+g) \neq \underline{\int}_0^1 f + \underline{\int}_0^1 g$ even though all these lower integrals are defined.

Proof of Theorem 35.16. We give the proof for $f+g$ and leave cf as an exercise. Fix $\varepsilon > 0$. By Theorem 35.12, there is $\delta > 0$ such that for every tagged partition (P,τ) of $[a,b]$ satisfying $\operatorname{mesh}(P) < \delta$, we have $|R(f,P,\tau) - \int_a^b f| < \frac{\varepsilon}{2}$ and $R(g,P,\tau) - \int_a^b g| < \frac{\varepsilon}{2}$. Moreover, observe that for every tagged point t_i, we have $(f+g)(t_i) = f(t_i) + g(t_i)$, and thus $R(f+g,P,\tau) = R(f,P,\tau) + R(g,P,\tau)$. It follows that

$$\left|R(f+g,P,\tau) - \Big(\int_a^b f + \int_a^b g\Big)\right| \leq \left|R(f,P,\tau) - \int_a^b f\right| + \left|R(g,P,\tau) - \int_a^b g\right| < \varepsilon.$$

By Exercise 35.15, this implies that $f+g$ is integrable and that $\int_a^b (f+g) = \int_a^b f + \int_a^b g$. □

Exercise 35.18. Adapt the above argument to prove the assertion in Theorem 35.16 regarding cf.

35.6. Alternate proof of linearity

We conclude by giving an alternate proof of Theorem 35.16 that works directly with the lower and upper sums and does not rely on Theorem 35.12. We start with the claim regarding cf, and consider the case $c \geq 0$. Given a partition $P = \{a = x_0 < x_1 < \cdots < x_n = b\}$, for each $1 \leq i \leq n$ we have

$$m_i(cf,P) = \inf\{cf(x) : x \in [x_{i-1},x_i]\} = c\inf\{f(x) : x \in [x_{i-1},x_i]\} = cm_i(f,P),$$

and thus

$$L(cf,P)=\sum_{i=1}^{n} m_i(cf,P)(x_i-x_{i-1})=\sum_{i=1}^{n} cm_i(f,P)(x_i-x_{i-1})=cL(f,P).$$

Taking a supremum over all partitions P gives $\underline{\int}_a^b cf = c\underline{\int}_a^b f$, and since f is integrable this gives $\int_a^b cf = c\int_a^b f$. When $c<0$, we have

$$m_i(cf,P)=\inf\{cf(x):x\in[x_{i-1},x_i]\}=c\sup\{f(x):x\in[x_{i-1},x_i]\}=cM_i(f,P),$$

and thus $L(cf,P)=cU(f,P)$. This time, taking a supremum over all partitions P gives

$$\underline{\int_a^b} cf=\sup_P L(cf,P)=\sup_P cU(f,P)=c\inf_P U(f,P)=c\overline{\int_a^b} f.$$

Since f is integrable this gives $\int_a^b cf = c\int_a^b f$.

The proof of the claim in Theorem 35.16 regarding $f+g$ requires a bit more work. Let $f,g\colon[a,b]\to\mathbb{R}$ be integrable, and consider $f+g$. We start by observing that for a given partition P, if we write $m_i^f=\inf\{f(x):x\in[x_{i-1},x_i]\}$ and define m_i^g, m_i^{f+g} similarly, then for every $x\in[x_{i-1},x_i]$ we have $f(x)\ge m_i^f$ and $g(x)\ge m_i^g$, so $(f+g)(x)\ge m_i^f+m_i^g$, and consequently

$$m_i^{f+g}\ge m_i^f+m_i^g.$$

Recalling the definition of the lower sums, we get

$$L(f+g,P)=\sum_{i=1}^{n} m_i^{f+g}(x_i-x_{i-1})\ge\sum_{i=1}^{n}(m_i^f+m_i^g)(x_i-x_{i-1})=L(f,P)+L(g,P).$$

A similar argument gives $M_i^{f+g}\le M_i^f+M_i^g$ and thus $U(f+g,P)\le U(f,P)+U(g,P)$. Putting it all together, we get

$$L(f,P)+L(g,P)\le L(f+g,P)\le U(f+g,P)\le U(f,P)+U(g,P) \tag{35.4}$$

for every partition P. Since f,g are integrable, for every $\varepsilon>0$, Lemma 35.10 lets us choose partitions Q,R such that

$$\begin{aligned}\Big(\int_a^b f\Big)-\frac{\varepsilon}{4}<L(f,Q)<\int_a^b f<U(f,Q)<\Big(\int_a^b f\Big)+\frac{\varepsilon}{4},\\ \Big(\int_a^b g\Big)-\frac{\varepsilon}{4}<L(g,R)<\int_a^b g<U(g,R)<\Big(\int_a^b g\Big)+\frac{\varepsilon}{4}.\end{aligned} \tag{35.5}$$

Now let $P=Q\cup R$ and use Lemma 33.14 together with (35.4) to get

$$\begin{aligned}\Big(\int_a^b f+\int_a^b g\Big)-\frac{\varepsilon}{2}&<L(f,Q)+L(g,R)\le L(f,P)+L(g,P)\\ &\le L(f+g,P)\le\underline{\int_a^b}(f+g)\le\overline{\int_a^b}(f+g)\le U(f+g,P)\\ &\le U(f,P)+U(g,P)\le U(f,Q)+U(g,R)<\Big(\int_a^b f+\int_a^b g\Big)+\frac{\varepsilon}{2}.\end{aligned}$$

This proves that $\overline{\int}_a^b (f+g) - \underline{\int}_a^b (f+g) < \varepsilon$, since both quantities are contained in the interval of length ε centered at $\int_a^b f + \int_a^b g$. Since this is true for any $\varepsilon > 0$, and the quantities $\overline{\int}_a^b f$, $\underline{\int}_a^b f$ do not depend on ε, it must be the case that $\overline{\int}_a^b (f+g) = \underline{\int}_a^b (f+g) = \int_a^b f + \int_a^b g$, which proves Theorem 35.16.

Lecture 36 Substitution rule

Stewart §5.5, Spivak Chapter 19

Now that we have a theory of integration, it is time to start developing the tools necessary to put it into practice. So far the only functions whose indefinite integrals we can find explicitly are the ones listed in the table in Lecture 31. The first step in extending our abilities is the *substitution rule*. Suppose we want to compute the indefinite integral

$$\int \frac{x}{\sqrt{1+x^2}}\,dx.$$

If we happen to guess that $\sqrt{1+x^2}$ might be an antiderivative, then it is easy to check that indeed it is, because we can write it as the composition of the functions $x \mapsto u = 1+x^2$ and $u \mapsto \sqrt{u}$, and then use the chain rule to obtain

$$\frac{d}{dx}\sqrt{1+x^2} = \frac{d}{dx}u^{1/2} = \frac{du}{dx}\frac{d}{du}u^{1/2} = 2x \cdot \frac{1}{2}u^{-1/2} = \frac{x}{\sqrt{u}} = \frac{x}{\sqrt{1+x^2}}.$$

But how would we come up with this guess in the first place? Again, the chain rule provides the clue. Looking at the integrand $\frac{x}{\sqrt{1+x^2}}$, we might decide that we would get a simpler expression if we made a change of variables and wrote $u = 1+x^2$. Then $\frac{du}{dx} = 2x$, and if we treat the notation as though it was genuinely a fraction (which it is not!) we might write $x\,dx = \frac{1}{2}du$ and obtain the formal, unjustified computation

$$\int \frac{x\,dx}{\sqrt{1+x^2}} = \int \frac{1}{2}u^{-1/2}du = u^{1/2} + C = \sqrt{1+x^2} + C.$$

The fact that this works provides some justification for why the notation is set up the way it is. This procedure is formalized by the following theorem.

Theorem 36.1 (Substitution rule)**.** *Let f be a continuous function on an interval I and g be a differentiable function whose range is contained in I. Write $u = g(x)$; then*

$$\int f(g(x))g'(x)\,dx = \int f(u)\,du.$$

Equivalently, if $F\colon I \to \mathbb{R}$ is an antiderivative of f, then $F \circ g$ is an antiderivative of $(f \circ g)\cdot(g')$.

Proof. This is a direct consequence of the chain rule:

$$(F \circ g)'(x) = F'(g(x))g'(x) = f(g(x))g'(x),$$

where the first equality is the chain rule and the second uses the fact that F is an antiderivative of f. □

Example 36.2. In the integral $\int x^2 \sin(x^3+1)\,dx$, we can write $u = x^3+1$ and obtain $du = 3x^2\,dx$, so

$$\int x^2 \sin(x^3+1)\,dx = \int \frac{1}{3}\sin u\,du = -\frac{1}{3}\cos u + C = -\frac{1}{3}\cos(x^3+1) + C.$$

The last step in the example is important; when we are computing an indefinite integral, we always need to find a final expression that is given in terms of the original variable, and not any intermediate variables such as u that we introduced along the way.

The hardest part of the procedure is choosing which function u to use. This is often a trial and error procedure, but there are some guidelines that are helpful to keep in mind.

- If some part of the integrand represents a function whose derivative also appears in the integrand, it may be worth setting u to be this part and seeing what happens.
- If there is some complicated expression that appears inside a square root, trigonometric function, logarithm, exponential, etc., then we might make progress by setting u to be this expression.

Example 36.3. In $\int \sqrt{3x+2}\,dx$, we can use the expression under the square root: $u = 3x+2$, so $du = 3\,dx$, and we get

$$\int \sqrt{3x+2}\,dx = \int \frac{1}{3}\sqrt{u}\,du = \frac{1}{3}\Big(\frac{1}{3/2}u^{3/2}\Big) + C = \frac{2}{9}u^{3/2} + C = \frac{2}{9}(3x+2)^{3/2} + C.$$

An alternate approach would be to take $u = \sqrt{3x+2}$ so that $u^2 = 3x+2$ and $2u\,du = 3\,dx$, giving

$$\int \sqrt{3x+2}\,dx = \int \frac{2}{3}u^2\,du = \frac{2}{9}u^3\,du + C = \frac{2}{9}(3x+2)^{3/2} + C.$$

Example 36.4. In $\int x^3\sqrt{1+x^2}\,dx$, we again use the expression under the square root: $u = 1+x^2$ gives $du = 2x\,dx$ and

$$\begin{aligned}\int x^3\sqrt{1+x^2}\,dx &= \int \big(x^2\sqrt{1+x^2}\big)x\,dx = \int (u-1)\sqrt{u}\frac{du}{2} = \frac{1}{2}\int (u^{3/2}-u^{1/2})\,du \\ &= \frac{1}{2}\Big(\frac{2}{5}u^{5/2} - \frac{2}{3}u^{3/2}\Big) + C = \frac{1}{5}u^{5/2} - \frac{1}{3}u^{3/2} + C \\ &= \frac{1}{5}(1+x^2)^{5/2} - \frac{1}{3}(1+x^2)^{3/2} + C.\end{aligned}$$

Example 36.5. To find the integral of $\tan x$, we can write it as $\frac{\sin x}{\cos x}$ and notice that the derivative of $\cos x$ appears in the numerator (up to a negative sign), so putting $u = \cos x$ gives $du = -\sin x\,dx$ and

$$\begin{aligned}\int \tan x\,dx = \int \frac{\sin x}{\cos x}\,dx = \int \frac{-du}{u} &= -\ln|u| + C = -\ln|\cos x| + C = \ln|1/\cos x| + C \\ &= \ln|\sec x| + C.\end{aligned}$$

This is an important enough example that it is worth remembering for future reference.

To compute *definite* integrals using the substitution rule, one option is to use the above procedure to evaluate the indefinite integral, and then apply the FTC. For example, we can use Example 36.3 to compute

$$\int_0^1 \sqrt{3x+2}\,dx = \frac{2}{9}(3x+2)^{3/2}\Big|_0^1 = \frac{2}{9}(5^{3/2}-2^{3/2}). \tag{36.1}$$

An alternate approach is to apply the substitution rule direction to the definite integral via the following result, which uses the above procedure and the FTC in its proof.

Theorem 36.6 (Substitution rule for definite integrals)**.** *Suppose $[a,b]$ and I are intervals in $\mathbb{R}$, and that we are given functions $g\colon [a,b]\to I$ and $f\colon I\to\mathbb{R}$ such that g' exists and is continuous on (a,b), and f is continuous on I. Then*

$$\int_a^b f(g(x))g'(x)\,dx = \int_{g(a)}^{g(b)} f(u)\,du.$$

Proof. Let $F\colon I\to\mathbb{R}$ be an antiderivative of f; then $\frac{d}{dx}F(g(x)) = F'(g(x))g(x)$ so the FTC2 gives

$$\int_a^b f(g(x))g'(x)\,dx = F(g(x))\Big|_a^b = F(g(b))-F(g(a)) = F(u)\Big|_{g(a)}^{g(b)} = \int_{g(a)}^{g(b)} f(u)\,du.$$

□

With this approach, we can evaluate the integral in (36.1) by using $u = g(x) = 3x+2$ to get $du = 3\,dx$ and $g(0)=2$, $g(1)=5$, so

$$\int_0^1 \sqrt{3x+2}\,dx = \int_2^5 \frac{1}{3}\sqrt{u}\,du = \frac{1}{3}\cdot\frac{2}{3}u^{3/2}\Big|_2^5 = \frac{2}{9}(5^{3/2}-2^{3/2}).$$

Observe that we need to change the limits of integration so that they are given in terms of the new variable. This example says that the area under the graph of $\sqrt{3x+2}$ between 0 and 1 is the same as the area under the graph of $\sqrt{u}/3$ between 2 and 5.

Example 36.7. To compute $\int_1^2 (1-2x)^{-2}\,dx$, we can write $u = 1-2x$ so that $du = -2\,dx$ and the new integral goes from $u=-1$ to $u=-3$:

$$\int_1^2 \frac{dx}{(1-2x)^2} = -\frac{1}{2}\int_{-1}^{-3} u^{-2}\,du = \frac{1}{2u}\Big|_{-1}^{-3} = \frac{1}{2(-3)} - \frac{1}{2(-1)} = -\frac{1}{6}+\frac{1}{2} = \frac{1}{3}.$$

The substitution rule lets us deduce certain properties of integrals of symmetric functions.

Theorem 36.8. *Let $f\colon [-a,a]\to\mathbb{R}$ be continuous.*

(1) If f is even, then $\int_{-a}^a f = 2\int_0^a f$.
(2) If f is odd, then $\int_{-a}^a f = 0$.

Proof. From properties of integrals, we have

$$\int_{-a}^a f = \int_{-a}^0 f + \int_0^a f = -\int_0^{-a} f + \int_0^a f. \tag{36.2}$$

The substitution $u = -x$ has $du = -dx$ and gives

$$-\int_0^{-a} f(x)\,dx = -\int_0^a f(-u)(-du) = \int_0^a f(-u)\,du.$$

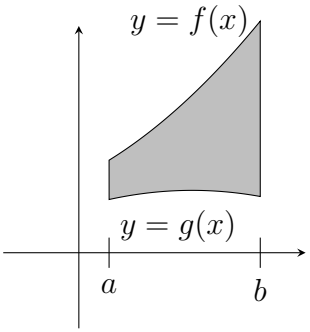
When f is even, we have $f(-u) = f(u)$, so this is equal to $\int_0^a f$. When f is odd, we have $f(-u) = -f(u)$, so this is equal to $-\int_0^a f$. Using these in (36.2) proves the theorem. □

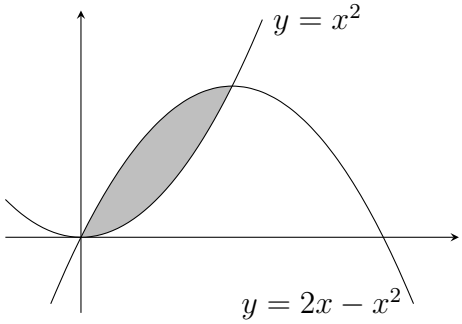
Lecture 37 Finding areas between curves

Stewart §6.1

Suppose we want to find the area of a region such as the one shown in the first picture below, which is bounded on the left and right by the vertical lines $x = a$ and $x = b$, below by the graph of $y = g(x)$, and above by the graph of $y = f(x)$, where $f(x) \geq g(x)$ for all $x \in [a, b]$. If we partition the interval $[a, b]$ into n subintervals $[x_{i-1}, x_i]$ for $i = 1, \ldots, n$, inside each of which we pick a 'tag' point t_i, then for each i we can consider the rectangle that ranges horizontally from x_{i-1} to x_i and vertically from $f(t_i)$ to $g(t_i)$, so its area is $(f(t_i) - g(t_i))(x_i - x_{i-1})$. The union of all of these rectangles has area $\sum_{i=1}^n (f(t_i) - g(t_i))(x_i - x_{i-1})$, which is a Riemann sum for the function $f - g$ on the interval $[a, b]$. Sending the mesh of the partition to 0, this sum converges to $\int_a^b (f - g)$, and we conclude that

$$\text{(37.1)} \quad \text{(area bounded by } f \text{ and } g \text{ between } x = a \text{ and } x = b) = \int_a^b \big(f(x) - g(x)\big)\,dx.$$





Example 37.1. To find the area enclosed by the two parabolas $y = x^2$ and $y = 2x - x^2$, we must first find the points where these curves intersect. If (x, y) lies on both curves, then we have $x^2 = y = 2x - x^2$, and so $2x^2 = 2x$, which implies $x = 0$ or $x = 1$. Thus the two intersection points are $(0, 0)$ and $(1, 1)$, and between these points we have $2x - x^2 \leq x^2$, so the area bounded by the curves is

$$\int_0^1 \big((2x - x^2) - x^2\big)\,dx = \int_0^1 (2x - 2x^2)\,dx = \Big[x^2 - \frac{2}{3}x^3\Big]_0^1 = 1 - \frac{2}{3} = \frac{1}{3}.$$

(This has the following interesting consequence: the unit square with vertices $(0,0)$, $(1,0)$, $(1,1)$, and $(0,1)$ is divided into three regions of equal area $\frac{1}{3}$ by the two parabolas in this example.)

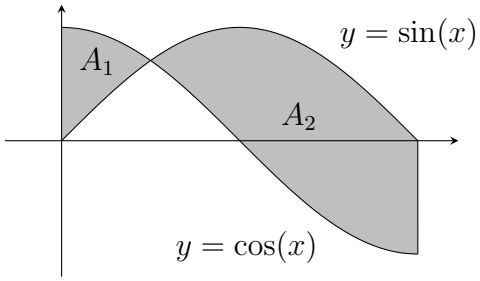
We may also be interested in situations where we want to include area on both sides of an intersection of two curves, as in the following example.

Example 37.2. Let us find the area between the curves $y = \sin x$ and $y = \cos x$ from $x = 0$ to $x = \pi$. The curves intersect exactly once in this interval, at $x = \frac{\pi}{4}$, as shown in the picture, and the area between them is the sum of the areas A_1 (to the left of $\frac{\pi}{4}$) and A_2 (to the right of $\frac{\pi}{4}$). On $[0, \frac{\pi}{4}]$ we have $\sin x \le \cos x$, while on $[\frac{\pi}{4}, \pi]$ we have $\cos x \le \sin x$. Thus we can compute A_1 and A_2 as follows:

$$\begin{aligned} A_1 &= \int_0^{\pi/4} (\cos x - \sin x)\,dx = \Big[\sin x + \cos x\Big]_0^{\pi/4} = \sin\frac{\pi}{4} + \cos\frac{\pi}{4} - \sin 0 - \cos 0 \\ &= \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2} - 0 - 1 = \sqrt{2} - 1, \\ A_2 &= \int_0^{\pi/4} (\sin x - \cos x)\,dx = \Big[-\cos x - \sin x\Big]_{\pi/4}^{\pi} = -\cos\pi - \sin\pi + \cos\frac{\pi}{4} + \sin\frac{\pi}{4} \\ &= -(-1) - 0 + \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2} = 1 + \sqrt{2}. \end{aligned}$$

Adding these gives

$$\text{total area} = \int_0^{\pi} |\sin x - \cos x|\,dx = A_1 + A_2 = (\sqrt{2} - 1) + (1 + \sqrt{2}) = 2\sqrt{2}.$$



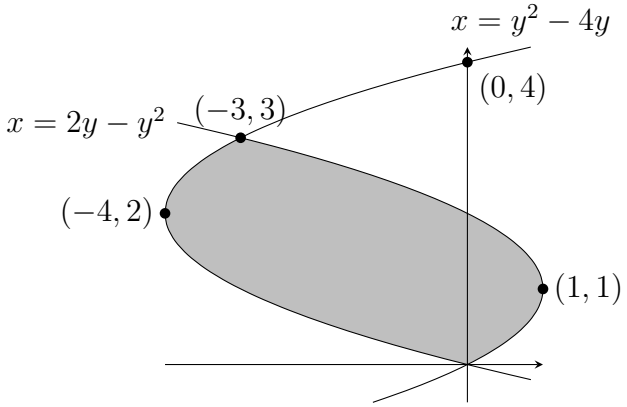
Remark 37.3. We previously encountered the need to integrate an absolute value in Lecture 35 when we studied the Net Change Theorem; we saw in Example 35.6 that if $v(t)$ is the velocity of an object moving along a line, then $\int_0^T |v(t)|\,dt$ is the total distance that object travels between times 0 and T, while $\int_0^T v(t)\,dt$ is its net displacement. If there are two objects, moving with velocities $v_1(t)$ and $v_2(t)$, that start in the same position at time 0, then $v_2(t) - v_1(t)$ gives the speed with which the second object is moving away from the first, and $\int_0^T (v_2(t) - v_1(t))\,dt$ gives the distance by which the second is ahead at time T.

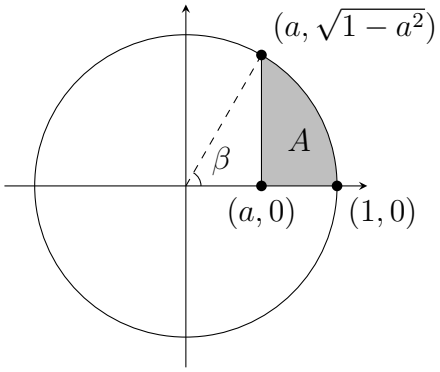
As the following example shows, sometimes it is easier to compute a certain area if we treat x as a function of y.

Example 37.4. Consider a parabola opening left with vertex at $(1,1)$ and y-intercepts at 0 and 2, and a parabola opening right with vertex at $(-4,2)$ and y-intercepts at 0 and 4. To find the area enclosed by these two curves, as shown in the first picture below, we need to first write equations for the parabolas, then find the points where they intersect, and then integrate. The first parabola has equation $x = 2y - y^2$, and the second has equation $x = y^2 - 4y$. These intersect when $2y - y^2 = x = y^2 - 4y$, that is, when $2y^2 - 6y = 0$, which occurs when $y = 0$ and when $y = 3$. (The former gives $x = 0$ and the latter gives $x = -3$.) The region enclosed by the parabolas corresponds to $0 \le y \le 3$, and in this range we have $y^2 - 4y \le 2y - y^2$, so the area is

$$\int_0^3 \big((2y - y^2) - (y^2 - 4y)\big)\, dy = \int_0^3 (6y - 2y^2)\, dy = \Big[3y^2 - \frac{2}{3}y^3\Big]_0^3 = 3 \cdot 3^2 - \frac{2}{3} \cdot 3^3$$

which works out to $27 - 18 = 9$.





Moving in the other direction, we can use geometric considerations to compute integrals that may be tricky to evaluate otherwise. The second picture in the figure shows a region whose area is $A = \int_a^1 \sqrt{1 - x^2}\, dx$. Since this region, together with the triangle to the left of it, makes up the circular sector ("pizza slice") with angle $\beta = \cos^{-1}(a)$, we see that

$$A + \text{area(triangle)} = \text{area(sector)} \quad \Rightarrow \quad A + \frac{1}{2}a\sqrt{1 - a^2} = \frac{1}{2}\beta = \frac{1}{2}\cos^{-1}(a),$$

and thus $\int_a^1 \sqrt{1 - x^2}\, dx = A = \frac{1}{2}(\cos^{-1}(a) - a\sqrt{1 - a^2})$. Could we compute this integral directly using the Fundamental Theorem of Calculus, without appealing to this geometric reasoning?

Since $\sqrt{1 - x^2}$ does not appear on our list of functions whose antiderivatives we know, and cannot be broken up as a sum of such functions, the only real tool we have us the substitution technique from the previous lecture. As a first attempt, we might try the substitution $u = 1 - x^2$, which gives $du = -2x\, dx$ and $x = \sqrt{1 - u}$, leading to

$$\int \sqrt{1 - x^2}\, dx = \int \sqrt{u}\Big(\frac{-du}{2x}\Big) = -\frac{1}{2}\int \sqrt{\frac{u}{1 - u}}\, du,$$

but then we get stuck; this is no easier to evaluate. So we might try a different substitution, say $y = \sqrt{1-x^2}$, which gives $y^2 = 1 - x^2$ and $y\,dy = -x\,dx$, so that

$$\int \sqrt{1-x^2}\,dx = \int y\Big(\frac{y\,dy}{-x}\Big) = -\int \frac{y^2}{\sqrt{1-y^2}}\,dy,$$

but again we are no better off; where do we go from here?

Our earlier geometric reasoning relied on knowing the area of the sector in terms of the angle β; inspired by this, we might associate to the point $(x, y = \sqrt{1-x^2})$ on the unit circle the angle θ for which $x = \cos\theta$ and $y = \sin\theta$. Treating $x = \cos\theta$ (or if you prefer, $\theta = \cos^{-1}(x)$) as a substitution with which to rewrite the integral, we get $dx = -\sin\theta\,d\theta$, and $\sqrt{1-x^2} = \sqrt{1-\cos^2\theta} = \sin\theta$, so the definite integral we are interested in can be rewritten as

$$A = \int_a^1 \sqrt{1-x^2}\,dx = \int_\beta^0 \sin\theta(-\sin\theta\,d\theta) = \int_0^\beta \sin^2\theta\,d\theta.$$

Although $\sin^2\theta$ does not appear on our list of functions whose antiderivatives we know, we can transform it using trigonometric identities into something more tractable: recall that $\cos(2\theta) = \cos^2\theta - \sin^2\theta = 1 - 2\sin^2\theta$, and so $\sin^2\theta = \frac{1}{2} - \frac{1}{2}\cos(2\theta)$, which lets us write

$$\begin{aligned} A &= \int_0^\beta \Big(\frac{1}{2} - \frac{1}{2}\cos(2\theta)\Big)\,d\theta = \Big[\frac{1}{2}\theta - \frac{1}{4}\sin(2\theta)\Big]_0^\beta = \frac{1}{2}\beta - \frac{1}{4}\sin(2\beta) \\ &= \frac{1}{2}\beta - \frac{1}{2}\sin\beta\cos\beta = \frac{1}{2}\big(\cos^{-1}(a) - a\sqrt{1-a^2}\big), \end{aligned}$$

in agreement with our earlier answer. This substitution technique is called *trigonometric substitution*, and we will study it at greater length next semester.

Lecture 38 Volumes

Stewart §§6.2–6.3

38.1. Volumes of "cylinders"

Now we go up a dimension and consider the problem of finding the volume of a region in $\mathbb{R}^3$. For area, the simplest region in the plane was a rectangle, as in the first picture below, where area is given by multiplying the two side lengths:

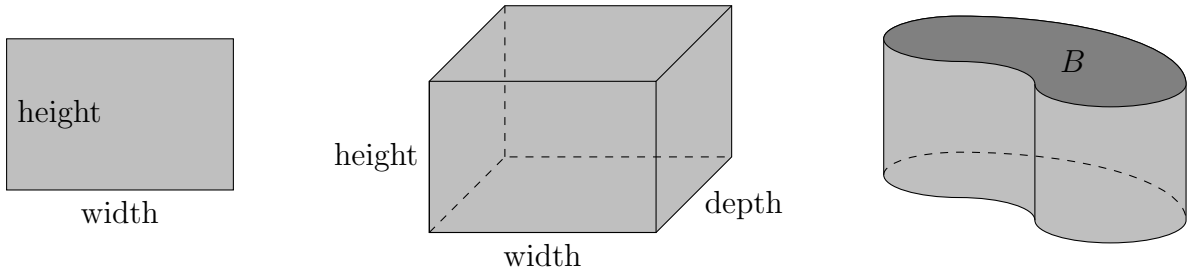
$$\text{area} = \text{width} \times \text{height}. \tag{38.1}$$

For volume, the simplest region is a "box" with all right angles, as in the second picture, where we multiply the three side lengths to get the volume:

$$\text{volume} = \text{width} \times \text{depth} \times \text{height}. \tag{38.2}$$

Observe that the rectangle forming the "base" of the box has area given by "width × depth", and thus (38.2) can be rewritten as

$$\text{volume} = (\text{area of base}) \times \text{height}. \tag{38.3}$$



This formula is a natural generalization of (38.1), where "width of base" is replaced by "area of base", and also has the benefit of extending to more irregularly shaped regions such as the third one shown in the figure. To describe this region a little more precisely, we make the following definition.

Definition 38.1. Let P_1 and P_2 be parallel planes in $\mathbb{R}^3$, and consider a region $B \subset P_1$. The *cylinder* above B between P_1 and P_2 is the solid region given as the union of all line segments starting in B, running perpendicular to P_1, and ending in P_2. The *height* of this cylinder is the length of each of these line segments (the distance between P_1 and P_2), and the *volume* of this cylinder is the area of B times the height.

Observe that when B is a disc (the region enclosed by a circle), Definition 38.1 gives the familiar shape that we usually associate with the word "cylinder". When B is a more irregular region, we obtain a more general notion of "cylinder", as shown in the picture above.

38.2. Volumes by cross-sections

Our motivation for developing the theory of integration was to determine the area of a region R in the plane; to do this, we approximated R by a union of rectangles, whose total area is given by a Riemann sum.

38.3. Cylindrical shells

Lec 38
§6.2
(11/28)

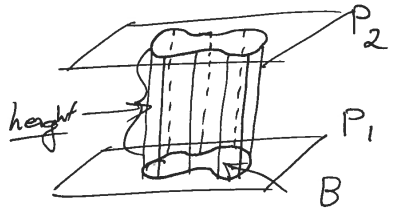
Volumes

(89)

Def Given a region B in a plane P_1 & a parallel plane P_2, the cylinder above B between P_1 & P_2 is the set of all pts in line segments starting in B, running $\perp$ to P_1, & ending in P_2.

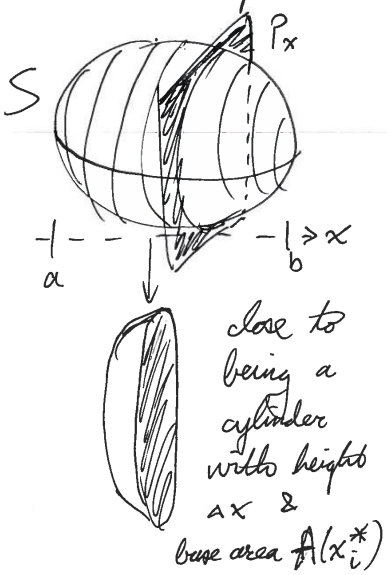
This solid region has

volume = area(B) · height



What about volumes of other regions?

- approx by unions of cylinders
- same procedure as for rectangles in $\mathbb{R}^2$.



$$V = \lim_{n\to\infty} \sum_{i=1}^{n} A(x_i^*)\Delta x = \int_a^b A(x)\,dx$$

for a solid between $x=a$ & $x=b$ s.t. cross-sectional area in P_x is $A(x)$.

(NB) If S is a cylinder then $A(x)$ is constant

$\therefore V = \int_a^b A\,dx = A(b-a)$

(90)

eg S = sphere w/ radius r centred at 0.

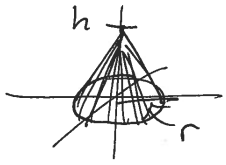
P_x intersects S in disc radius $\sqrt{r^2-x^2}$

$\therefore A(x) = \pi(r^2-x^2)$

$\therefore V = \int_{-r}^{r} \pi(r^2-x^2)\,dx = 2\pi\int_0^r (r^2-x^2)\,dx$

$= 2\pi\left[r^2x - \frac{1}{3}x^3\right]_0^r = 2\pi\left(r^3 - \frac{1}{3}r^3\right) = \frac{4\pi r^3}{3}$

eg Cone with height h & radius r at base.



Slice horizontally:

$V = \int_0^h A(z)\,dz$

$S \cap P_z$ is disc w/ radius $\frac{(h-z)r}{h}$

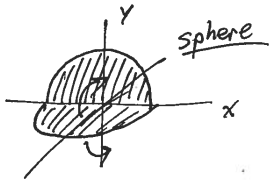
$\therefore A(z) = \pi\frac{(h-z)^2}{h^2}r^2$

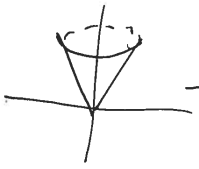
$\therefore V = \int_0^h \pi\frac{(h-z)^2}{h^2}r^2\,dz = \frac{\pi r^2}{h^2}\int_0^h (h-z)^2\,dz$

$= \frac{\pi r^2}{h^2}\left[-\frac{1}{3}(h-z)^3\right]_0^h = \frac{1}{3}\frac{\pi r^2}{h^2}h^3 = \frac{1}{3}\pi r^2 h$

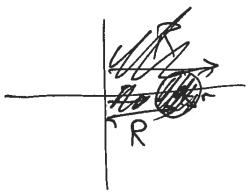
→ could make integral easier by flipping over, so $A = \pi\left(\frac{zr}{h}\right)^2$

These were both solids of revolution:

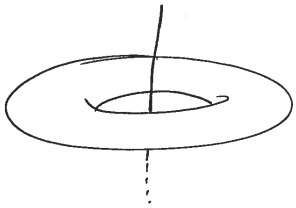




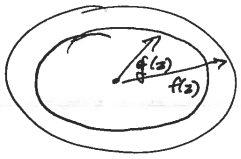
(9)

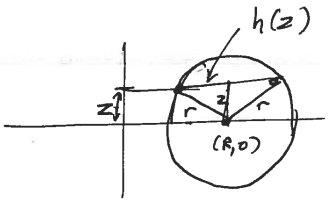


gives torus



w/ cross-section





$$A(z) = \pi(f(z)^2 - g(z)^2)$$

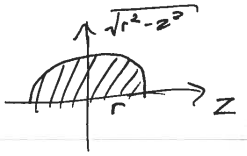
$$h(z)^2 = r^2 - z^2$$

$$f(z) = R + h(z) \qquad g(z) = R - h(z)$$

$$A(z) = \pi(f+g)(f-g) = \pi \cdot 2R \cdot 2h(z) = 4\pi R\, h(z)$$

$$= 4\pi R\sqrt{r^2 - z^2}$$

$$\therefore V = \int_{-r}^{r} 4\pi R\sqrt{r^2 - z^2}\, dz = 4\pi R \int_{-r}^{r} \sqrt{r^2 - z^2}\, dz$$



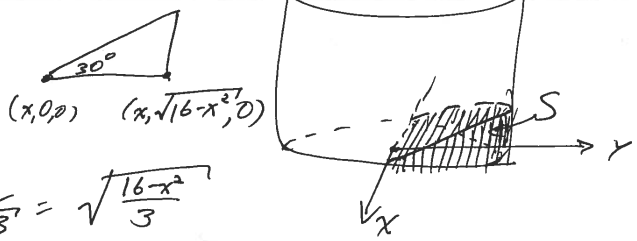
$$\int_{-r}^{r} \sqrt{r^2 - z^2}\, dz = \frac{1}{2}\pi r^2$$

$$\therefore \text{volume of torus} = 4\pi R\left(\tfrac{1}{2}\pi r^2\right) = 2\pi^2 R r^2$$

Lec 39 (11/30) @ Circular cylinder radius 4 – cut out wedge by 2 planes: one ⊥ to axis of cylinder, ~~the~~ second intersecting first at 30° along diam of cylinder.

Volume(wedge) = ?

$S \cap P_x$ = ~~triangle~~



height $= y \tan 30°$

$= y \frac{1/2}{\sqrt{3}/2} = \frac{y}{\sqrt{3}} = \sqrt{\frac{16-x^2}{3}}$

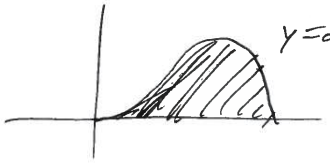
$A(x) = \frac{1}{2}\sqrt{16-x^2} \cdot \frac{1}{\sqrt{3}}\sqrt{16-x^2} = \frac{16-x^2}{2\sqrt{3}}$

$V = \int_{-4}^{4} A(x)\,dx = \frac{1}{\sqrt{3}} \int_0^4 16-x^2\,dx = \frac{1}{\sqrt{3}}\left[16x - \frac{1}{3}x^3\right]_0^4$

$= \frac{1}{\sqrt{3}}\left[64 - \frac{1}{3}64\right] = \frac{128}{3\sqrt{3}}$

§6.3 Volumes by cylindrical shells

Consider solid of revolution around y-axis

from 

$S \cap P_y$ annulus where inner & outer radii are

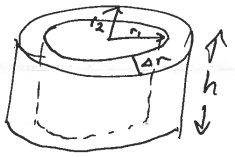
sides of $y = 2x^2 - x^3$ .... no nice formula

(93)

So instead, integrate out from y-axis:

each rectangle rotates into a cylindrical shell =

cylinder over an annulus



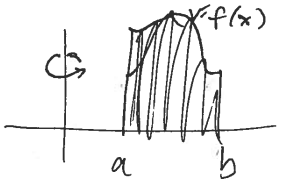
$r = \frac{1}{2}(r_1 + r_2)$

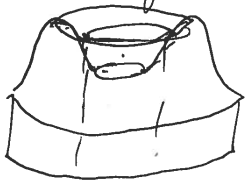
$A(\text{annulus}) = \pi r_2^2 - \pi r_1^2 = \pi(r_2^2 - r_1^2) = \pi(r_2 - r_1)(r_2 + r_1)$

$= 2\pi r \Delta r$

$V(\text{cylindrical shell}) = 2\pi r h \Delta r$

= circumference · height · thickness





Divide $[a, b]$ into subintervals $\Delta x = \frac{b-a}{n}$ width,

$\bar{x}_i$ = midpt of i^{th}, then rectangles rotate to shells,

& volume (i^{th} shell) $= 2\pi \bar{x}_i f(\bar{x}_i) \Delta x$

$\therefore$ Volume $(S) = \lim_{n\to\infty} \sum_{i=1}^{n} 2\pi \bar{x}_i f(\bar{x}_i) \Delta x = \int_a^b 2\pi x f(x) dx$

Eg $y = 2x^2 - x^3$ on $[0, 2]$:

$V = \int_0^2 (2\pi x)(2x^2 - x^3) dx = 2\pi \int_0^2 (2x^3 - x^4) dx$

$= 2\pi \left[\frac{1}{2}x^4 - \frac{1}{5}x^5 \right]_0^2 = 2\pi \left(8 - \frac{32}{5}\right) = \frac{16}{5}\pi$

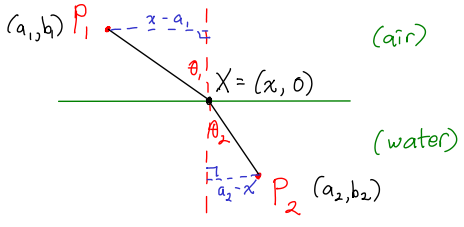
Lecture 39 Rainbows

In this lecture we will deviate from our work developing the theory of calculus, and take a sightseeing tour through an application to optics, which uses some of the tools we have developed so far to explain why rainbows form. This lecture draws from Chapter 9 of "Mathematical Understanding of Nature" by V.I. Arnold.

39.1. Snell's law

Our starting point is *Fermat's principle*, which states that light follows the path of least time: that is, given two points P_1 and P_2, a ray of light that travels from P_1 to P_2 will do so along the path that minimizes the amount of time required. If this takes place in a medium where the speed of light does not depend on direction (isotropic) or position (homogeneous), then travel time is directly proportional to length of the path followed, and since the path that minimizes length is a straight line, we conclude that light travels along straight lines. However, when light passes between two media in which it travels with different speeds, the time-minimizing path is no longer a straight line, and the phenomenon of *refraction* occurs. A quantitative description is given by *Snell's law*, which we derive next.

Remark 39.1. Fermat's principle appears mysterious at first. Why, and how, should light minimize time? Taken carelessly, the principle might seem to ascribe intent and purpose to the light, as if it is making a conscious decision to follow one path and not another with the goal of reaching its destination as quickly as possible. Historically, Fermat's principle was developed as a way of describing the phenomenon observed in Snell's law, rather than the other way round (as we do here), and was not initially given more of a justification than that. In fact, it turns out to be one specific manifestation of the *principle of least action* and more general *variational principles* that play a crucial role in physics; the full story involves Euler, Maupertuis, Lagrange, Hamilton, Dirac, Feynman (from which you may guess that quantum mechanics plays a role) and many others, and goes far beyond the scope of this calculus course. We will content ourselves with the statement that Fermat's principle can be deduced from the wave theory of light, and that its proper statement actually replaces the requirement that the path minimize time with the requirement that the path be a *critical point* for the function that assigns lengths to paths. Since the space of paths is infinite-dimensional, this notion of critical point is for an infinite-dimensional version of the derivative we have studied in this course, which appears in the *calculus of variations*.



Consider a beam of light traveling through the air and then passing into water. The light follows a straight line while it is in the air, and also while it is in the water, but at the interface between the two it is free to change direction, and will do so according to Fermat's principle. To describe this change in direction, let us use a coordinate system in which the interface occurs along the x-axis, with air above the axis and water below it. Suppose the light starts at point P_1 with coordinates (a_1, b_1), and travels to P_2 with coordinates

(a_2, b_2), and passes through a point X on the interface with coordinates $(x, 0)$. We must find the value of x that minimizes the total travel time; to this end, let $T(x)$ denote the time it takes light to travel along the path via $(x, 0)$. Let v_1 be the speed of light in air, and v_2 the speed of light in water. Then

$$T(x) = \frac{|P_1X|}{v_1} + \frac{|XP_2|}{v_2}.$$

It is standard to replace the velocity v_i with the *index of refraction* $n_i = c/v_i$, where c is the speed of light in a vacuum, and we get

$$\begin{aligned} cT(x) &= n_1|P_1X| + n_2|XP_2| \\ &= n_1\sqrt{(x-a_1)^2 + b_1^2} + n_2\sqrt{(x-a_2)^2 + b_2^2} \end{aligned}$$

To find the critical point, we differentiate using the chain rule and get

$$\begin{aligned} cT'(x) &= \frac{n_1(x-a_1)}{\sqrt{(x-a_1)^2 + b_1^2}} + \frac{n_2(x-a_2)}{\sqrt{(x-a_2)^2 + b_2^2}} \\ &= n_1\frac{(x-a_1)}{|P_1X|} - n_2\frac{(a_2-x)}{|XP_2|} = n_1\sin\theta_1 - n_2\sin\theta_2, \end{aligned}$$

where θ_1 and θ_2 are the angles that the ray makes with the vertical line through $(x, 0)$ on either side of the interface. Thus we see that $T'(x) = 0$ if and only if

$$n_1\sin\theta_1 = n_2\sin\theta_2, \tag{39.1}$$

which is exactly Snell's law. In general, θ_1 and θ_2 represent the angles between the ray of light and the normal direction perpendicular to the interface.

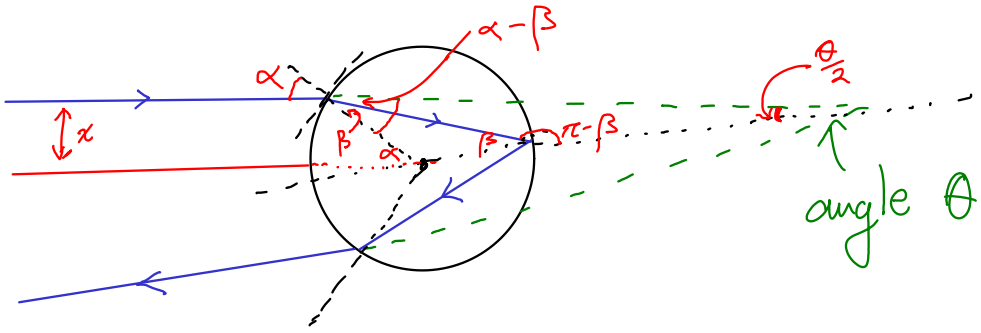
This reasoning works for any two media, not just air and water. For air we have $n_1 \approx 1.0003$, and for water we have $n_2 \approx 1.333$. In the remainder of the lecture we will need to use the ratio of these: $n = \frac{n_2}{n_1} \approx \frac{4}{3}$.

39.2. Refraction in a water drop

Now consider a beam of light that refracts upon entering a water droplet, then passes through the interior and reflects off the surface of the droplet, passing through the interior once more and then finally refracting a second time as it leaves the droplet.[35]

Suppose the droplet is a sphere; to make computations easier, let us work in units in which the radius of the sphere is 1. As shown in the picture, let x denote the distance between the incoming beam and a parallel straight line passing through the center of the sphere. The blue lines in the picture represent the path followed by the light, and the dotted green lines are the extensions of the incoming and outgoing lines; they meet at an angle θ, which is the angle by which the direction of the light is changed as a result of its interaction with the droplet. Our goal is to understand how θ depends on x; in particular, for reasons that will become clear in the next section, we want to find its critical value(s), that is, the value(s) taken by θ when $\frac{d\theta}{dx} = 0$.

[35]You may rightly ask why it should reflect exactly once, instead of just refracting twice with no reflections, or reflecting more than once before it refracts out. Indeed there are also light beams that follow these patterns, and it is a good exercise to understand what role, if any, they play in rainbows. Think about it after you finish this lecture.



Let α be the angle that the incoming beam makes with the normal line when it hits the sphere, and β the angle that it makes after refracting. Thanks to Snell's law (39.1), we have

$$\sin\alpha = n\sin\beta. \tag{39.2}$$

Consider the triangle in the picture whose edges are the top green line, the top blue line, and the dotted black line. Adding the angles of this triangle gives

$$\frac{\theta}{2} + (\alpha-\beta) + (\pi-\beta) = \pi \qquad \Rightarrow \qquad \frac{\theta}{2} + \alpha - 2\beta = 0,$$

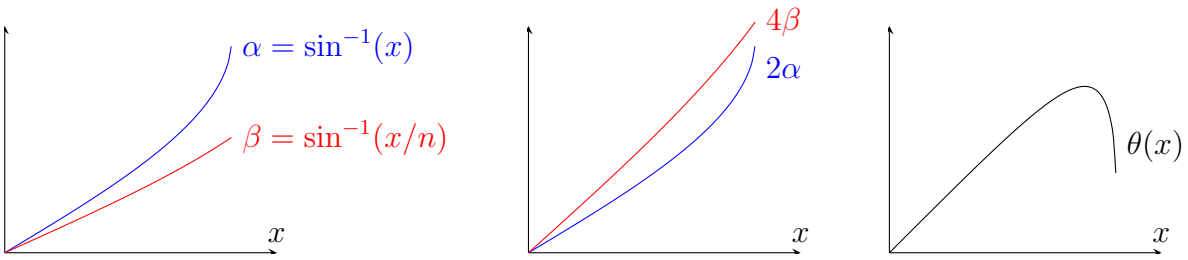
from which we deduce that

$$\theta = 4\beta - 2\alpha. \tag{39.3}$$

Since we work in units where the drop has radius 1, we have $x = \sin\alpha$, so $\alpha = \sin^{-1}(x)$. Using (39.2), we have $\sin\beta = \frac{\sin\alpha}{n} = \frac{x}{n}$, so $\beta = \sin^{-1}(\frac{x}{n})$. Putting these together with (39.3) gives

$$\theta = 4\sin^{-1}(x/n) - 2\sin^{-1}(x). \tag{39.4}$$

This is probably not a function that you have seen before, so it may not be immediately clear what the graph looks like. It is helpful to first sketch the graph of $\alpha = \sin^{-1}(x)$, which is familiar, and then of $\beta = \sin^{-1}(x/n)$, which is obtained from it by a horizontal scaling. Scaling these graphs vertically gives the second picture shown (the three pictures all have different vertical scales), and $\theta(x)$ is the difference between the red curve and the blue curve in that picture, so it has the shape shown in the third picture.



39.3. Most common outgoing angles

Differentiating gives

$$\frac{d\theta}{dx} = \frac{4/n}{\sqrt{1-(x/n)^2}} - \frac{2}{\sqrt{1-x^2}} = \frac{4}{\sqrt{n^2-x^2}} - \frac{2}{\sqrt{1-x^2}},$$

which vanishes when

$$2\sqrt{n^2-x^2} = 4\sqrt{1-x^2}$$

Dividing both sides by 2 and then squaring gives

$$n^2 - x^2 = 4(1-x^2) = 4 - 4x^2.$$

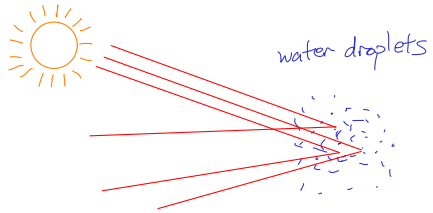
Solving for x^2 gives

$$3x^2 = 4 - n^2 \approx 4 - \left(\frac{4}{3}\right)^2 = 4 - \frac{16}{9} = \frac{20}{9} \quad \Rightarrow \quad x^2 = \frac{4-n^2}{3} \approx \frac{20}{27}.$$

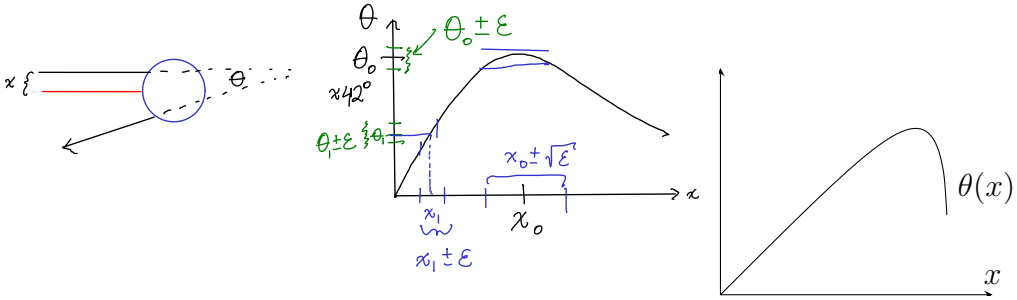
This allows us to compute

$$\alpha \approx 59.4^\circ, \qquad \beta \approx 40.2^\circ, \qquad \theta = 4\beta - 2\alpha \approx 42^\circ.$$

Call this θ_0. Two explanations for why this angle has the highest intensity of light.



Both rely on fact that incoming beams hit droplets at random positions, so x is random. One explanation observes that near θ_0, the function looks quadratic, and elsewhere it looks linear, so in general $\theta \pm \epsilon$ corresponds to range of x-values with size proportional to ϵ, but $\theta_0 \pm \epsilon$ corresponds to range of x-values with size proportional to $\sqrt{\epsilon}$, which is much larger. For example, $\epsilon = 10^{-6}$ has $\sqrt{\epsilon} = 10^{-3}$.



Another way of looking at it involves probabilities via integrals.
Assume $\mathbf{P}(a \le x \le b) = b - a$ (uniform distribution).
If the function $\theta = f(x)$ were invertible, with inverse $g(\theta) = x$, then we would have

$$\mathbf{P}(a \le \theta \le b) = \mathbf{P}\big(g(a) \le x \le g(b)\big) = g(b) - g(a) = \int_a^b g'(\theta)\, d\theta$$

where the last equality uses the Net Change Theorem. In fact many values of θ correspond to two distinct values of x, so there are two functions $g_1(\theta)$ and $g_2(\theta)$ with

$f(g_1(\theta)) = f(g_2(\theta)) = \theta$, and similar reasoning to the above gives

$$\mathbf{P}(a \le \theta \le b) = \int_a^b h(\theta)\, d\theta, \quad \text{where } h(\theta) = |g_1'(\theta)| + |g_2'(\theta)|.$$

Since $g_i'(\theta) = 1/f'(g_i(\theta))$ by the formula for derivatives of inverse functions, we see that $h(\theta)$ has a vertical asymptote near θ_0. This 'spike' indicates the increased intensity of light at that angle.

Finally, the explanation for the colors of the rainbow is that θ_0 depends on n, and n actually varies a little bit with the wavelength of light, so different wavelengths have a slightly different value of θ_0.

Final final remark: two reflections inside drop leads to "double rainbow" with a different angle (and fainter).

Part IV. Integration

Lecture 40 Review of integration and the substitution rule

Stewart §5.5, Spivak Ch. 19

40.1. Definite and indefinite integrals

Last semester, we motivated the introduction of integrals by considering the question of how to determine areas. This led us to two definitions:

(1) the definite integral $\int_a^b f(x)\,dx$ is a *number* obtained as a limit of Riemann sums, which depends on the interval $[a,b]$ and can be interpreted as an area;
(2) the indefinite integral $\int f(x)\,dx$ is a *function* whose derivative is $f(x)$.

The two are related by the Fundamental Theorem of Calculus, which has two halves.

The first half says that definite integrals can be used to find indefinite integrals (antiderivatives), since $\frac{d}{dx}\int_a^x f(t)\,dt = f(x)$.

The second half goes in the opposite direction, and says that indefinite integrals can be used to find definite integrals: if $F(x) = \int f(x)\,dx$ is an indefinite integral of f, so that $F'(x) = f(x)$ at every x, then $\int_a^b f(x)\,dx = F(b) - F(a)$.

Although the first half guarantees that every continuous function has an indefinite integral, it does not give a general procedure for writing down an elementary formula for $\int f(x)\,dx$. Our emphasis for the next little while will be on this process, which is essential if we are to use the second half of the FTC effectively.

By "elementary formula", we mean a formula that can be written down in terms of constants, polynomials, rational functions, exponentials, trigonometric functions, and logarithms using addition, subtraction, multiplication, and division. For example, $F(x) = \tan^{-1}(x)$ is an elementary formula, but $F(x) = \int_0^x \frac{1}{1+t^2}\,dt$ is not elementary because it involves an integral, even though it represents the same function.

Given an integral $\int f(x)\,dx$, then, our goal will be to find an elementary formula for it. Bear the following warning in mind, though: not every integral admits an elementary formula. For example, it is possible to show[36] that $\int \sin(x^2)\,dx$ does not have an elementary formula, and in fact there is a sense in which *most* indefinite integrals do not have elementary formulas. Nevertheless, a great many of them do, including some of the most important ones, and so we will turn our attention now to finding them.

40.2. Substitution rule

The first method of integration is by direct inspection: we have a list of functions $F(x)$ whose derivatives $f(x) = F'(x)$ are known, and if f happens to appear on the corresponding list of derivatives, then we can simply read off the indefinite integral $\int f(x)\,dx = F(x) + C$.

The second method, which we encountered briefly last semester, is the substitution rule. This is a consequence of the chain rule for differentiation, which says that if F, g are

[36] The proof involves tools that go beyond the scope of this course, and we will not discuss it.

differentiable functions, then $F \circ g$ is differentiable and has $(F \circ g)'(x) = F'(g(x))g'(x)$. In particular, if $F'(x) = f(x)$ so that F gives the indefinite integral of f, then we have $(F \circ g)' = (f \circ g) \cdot (g')$; this can be written in the form

$$\int f(g(x))g'(x)\,dx = F(g(x)).$$

It is usually easier to remember and apply this rule if we introduce a new variable $u = g(x)$, and observe that $\frac{d}{du}F(u) = f(u)$, so that the above formula becomes

$$\int f(g(x))g'(x)\,dx = \int f(u)\,du. \tag{40.1}$$

It is common to rewrite the formula $g'(x) = \frac{du}{dx}$ as $du = g'(x)\,dx$, in which case (40.1) appears to become almost trivial:

$$\int f(\underbrace{g(x)}_{u})\,\underbrace{g'(x)\,dx}_{du} = \int f(u)\,du.$$

We emphasize, though, that the formula $du = g'(x)\,dx$ is purely a bookkeeping device rather than a valid part of a proof, because we have not yet given du and dx any independent meaning of their own. We will continue to use it because it simplifies the appearance of various computation, but please remember the logical order of things: (40.1) justifies this formula, rather than the other way round.

Example 40.1. We can compute $\int x\sqrt{1+x^2}\,dx$ by putting $u = 1 + x^2$ so that $du = 2x\,dx$, and we obtain

$$\int x\sqrt{1+x^2}\,dx = \int \underbrace{\sqrt{1+x^2}}_{\sqrt{u}} \cdot \underbrace{x\,dx}_{\frac{1}{2}du} = \int \frac{1}{2}u^{1/2}\,du = \frac{1}{2}\cdot\frac{2}{3}u^{3/2} + C = \frac{1}{3}(1+x^2)^{3/2} + C.$$

Example 40.2. To find $\int \tan x\,dx$, we can write $\tan x = \frac{\sin x}{\cos x}$ and notice that the derivative of $\cos x$ appears in the numerator (up to a negative sign), so putting $u = \cos x$ gives $du = -\sin x\,dx$ and

$$\begin{aligned}\int \tan x\,dx = \int \frac{\sin x}{\cos x}\,dx = \int \frac{-du}{u} &= -\ln|u| + C = -\ln|\cos x| + C = \ln|1/\cos x| + C \\ &= \ln|\sec x| + C.\end{aligned}$$

There is no universal procedure telling us how to make the change of variables $u = g(x)$, but these examples illustrate some guidelines that are helpful to keep in mind: it is reasonable to try setting u as the input of some function in the integrand (the square root function in Example 40.1), or as an expression whose derivative also appears in the integrand (the cosine function in Example 40.2). Sometimes it even works to let u be the entire integrand: for example, in $\int \sqrt{2x+1}\,dx$ we can take $u = \sqrt{2x+1}$ so that $u^2 = 2x + 1$ and $2u\,du = 2\,dx$, and we get

$$\int \underbrace{\sqrt{2x+1}}_{u}\,\underbrace{dx}_{u\,du} = \int u \cdot u\,du = \frac{1}{3}u^3 + C = \frac{1}{3}(2x+1)^{3/2} + C.$$

Note that the substitution $u = 2x + 1$ would also work here; there is often more than one route to the correct answer!

When computing an indefinite integral via the substitution rule, it is important to remember that the final answer must always be written in terms of the *original* variable, not the substituted one. Thus the last step in each of the above examples was to convert an expression involving u into an expression involving x.

The substitution rule can also be used for definite integrals, either by first computing the indefinite integral and then applying the FTC, or by applying the change of variables $u = g(x)$ to the limits of integration as well.

Example 40.3. To compute $\int_1^2 (1-2x)^{-2}\,dx$, we can write $u = 1-2x$ so that $du = -2\,dx$ and the new integral goes from $u = -1$ to $u = -3$:

$$\int_1^2 \frac{dx}{(1-2x)^2} = -\frac{1}{2}\int_{-1}^{-3} u^{-2}\,du = \frac{1}{2u}\Big|_{-1}^{-3} = \frac{1}{2(-3)} - \frac{1}{2(-1)} = -\frac{1}{6} + \frac{1}{2} = \frac{1}{3}.$$

Lecture 41 Integration by parts

Stewart §7.1, Spivak Ch. 19

41.1. A consequence of the product rule

We found the substitution rule for integrals by looking at the chain rule for derivatives, and exploring its consequences for integrals. We can also do this with the product rule, which says that if f, g are differentiable functions, then

$$\frac{d}{dx}\big(f(x)g(x)\big) = f(x)g'(x) + g(x)f'(x).$$

This can be rewritten as

$$f(x)g'(x) = \frac{d}{dx}\big(f(x)g(x)\big) - g(x)f'(x) = \frac{d}{dx}\Big(f(x)g(x) - \int g(x)f'(x)\,dx\Big),$$

and we conclude that

$$\int f(x)g'(x)\,dx = f(x)g(x) - \int g(x)f'(x)\,dx. \tag{41.1}$$

We do not write a constant of integration because the right-hand side still contains an indefinite integral. The relationship (41.1) is called *integration by parts* and is a powerful tool for evaluating many integrals, especially when f, g can be chosen so that gf' is easier to integrate than fg'.

Example 41.1. Suppose we want to evaluate $\int x\cos x\,dx$. Then we might try $f(x) = x$ and $g'(x) = \cos x$; to get this, we should put $g(x) = \sin x$, and then (41.1) gives

$$\int x\cos x\,dx = x\sin x - \int \underbrace{(\sin x)}_{g(x)} \cdot \underbrace{1}_{f'(x)}\,dx = x\sin x - (-\cos x) + C = x\sin x + \cos x + C.$$

And indeed, we can verify this by differentiating and using the product rule:

$$\frac{d}{dx}(x\sin x + \cos x) = (\sin x + x\cos x) - \sin x = x\cos x.$$

Remark 41.2. Since antiderivatives are only determined up to a constant, the fact that $g'(x) = \cos x$ actually only tells us that $g(x) = \sin x + C$ for some C. You can check that using this $g(x)$ still gives us the same answer. Because we can choose $g(x)$ to be *any* antiderivative of $g'(x)$, we may as well choose it to be the antiderivative that is the simplest to write down, which usually happens when we put $C = 0$.

Remark 41.3. As with the substitution rule, not all choices are helpful! For example, if we put $f(x) = \cos x$ and $g'(x) = x$ in the example above, we would get $g(x) = \frac{1}{2}x^2$ and $f'(x) = -\sin x$, so

$$\int x \cos x \, dx = \frac{1}{2}x^2 \cos x - \int \frac{1}{2}x^2(-\sin x)\, dx = \frac{1}{2}\Big(x^2 \cos x + \int x^2 \sin x \, dx\Big).$$

We have done nothing wrong – the equation we derived is true – but we have not done anything helpful, either, since we do not know how to evaluate $\int x^2 \sin x \, dx$.

We pause a moment to recall that we can write the chain rule in an alternate form by writing $u = g(x)$ and $y = f(u) = f(g(x))$, so that we have the following diagram:

$$x \overset{g}{\longmapsto} u = g(x) \overset{f}{\longmapsto} y = f(u) = f(g(x)) \qquad (x \overset{f\circ g}{\longmapsto} y)$$

Then the chain rule becomes the very sensible-looking equation $\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$.

Remark 41.4. It looks like we are simply cancelling the two appearances of the term du, but this is not quite right; we have not given any independent meaning to the symbols dy, du, and dx outside of a derivative like $\frac{dy}{dx}$, or an integral like $\int f(x)\, dx$. Thus this should be regarded as a bookkeeping tool more than anything else; however, it is in some ways easier to remember, and the fact that it conforms to our expectation of how fractions should behave suggests that the notation $\frac{dy}{dx}$ is appropriate to use.

A similar bookkeeping tool is useful for integration by parts. Using the notation $u = f(x)$ and $v = g(x)$, we write $du = f'(x)\, dx$ and $dv = g'(x)\, dx$ (despite the fact that dx, du, and dv have no independent meaning in their own right!) and rewrite (41.1) in the following form, which is easier to remember:

$$\int u \, dv = uv - \int v \, du. \tag{41.2}$$

In Example 41.1 we would put $u = x$, $du = dx$, $dv = \cos x \, dx$, and $v = \sin x$, obtaining the same result as before.

Example 41.5. To evaluate $\int \ln x \, dx$, put $u = \ln x$, $dv = dx$, $du = \frac{1}{x}\, dx$, and $v = x$:

$$\int \underbrace{\ln x}_{u}\underbrace{dx}_{dv} = \underbrace{x}_{v}\underbrace{\ln x}_{u} - \int \underbrace{x}_{v}\underbrace{\frac{1}{x}\, dx}_{du} = x \ln x - \int 1 \, dx = x \ln x - x + C.$$

41.2. Iterated integration by parts

Example 41.6. To evaluate $\int t^2 e^t\,dt$, we can put $u = t^2$ and $dv = e^t\,dt$, so $du = 2t\,dt$ and $v = e^t$, giving

$$\int t^2 e^t\,dt = \underbrace{t^2 e^t}_{uv} - \underbrace{\int 2te^t\,dt}_{\int v\,du}. \tag{41.3}$$

To evaluate the last integral we use integration by parts a second time; bring out the factor of 2 and compute $\int te^t\,dt$ by putting $u = t$, $dv = e^t\,dt$, $du = dt$, $v = e^t$, giving

$$\int te^t\,dt = te^t - \int e^t\,dt = te^t - e^t.$$

Using this in (41.3) gives

$$\int t^2 e^t\,dt = t^2 e^t - 2\int te^t\,dt = t^2 e^t - 2(te^t - e^t) + C = t^2 e^t - 2te^t + 2e^t + C.$$

Exercise 41.7. Follow this same approach to show that if $f(t)$ is a polynomial of degree n, then using integration by parts n times gives

$$\int f(t)e^t\,dt = \big(f(t) - f'(t) + f''(t) - \cdots + (-1)^n f^{(n)}(t)\big)e^t + C.$$

Sometimes by using integration by parts multiple times, we end up with an expression that does not yield the integral directly, but which gives an equation that can be solved for it. This is best illustrated with an example.

Example 41.8. To evaluate $\int e^x \sin x\,dx$, we integrate by parts twice:

$$\begin{aligned}\int e^x \sin x\,dx &= -e^x\cos x + \int e^x\cos x\,dx && (u = e^x \text{ and } dv = \sin x\,dx)\\ &= -e^x\cos x + \Big(e^x\sin x - \int e^x\sin x\,dx\Big) && (u = e^x \text{ and } dv = \cos x\,dx).\end{aligned}$$

Since this last expression contains the original integral, one might at first think that we have gotten nowhere. But in fact, we are nearly done! Adding $\int e^x\sin x\,dx$ to both sides of the equation gives

$$2\int e^x\sin x\,dx = e^x(\sin x - \cos x) \qquad \Rightarrow \qquad \int e^x\sin x\,dx = \frac{1}{2}e^x(\sin x - \cos x) + C,$$

where we add a constant of integration to get the most general antiderivative.

41.3. Definite integrals

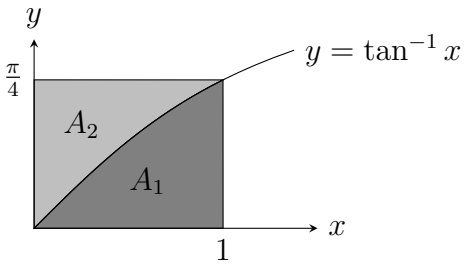
By the FTC, (41.1) has a counterpart for definite integrals:

$$\begin{aligned}\int_a^b f(x)g'(x)\,dx &= \big[f(x)g(x)\big]_a^b - \int_a^b g(x)f'(x)\,dx\\ &= f(b)g(b) - f(a)g(a) - \int_a^b g(x)f'(x)\,dx.\end{aligned} \tag{41.4}$$

Example 41.9. To evaluate $\int_0^1 \tan^{-1} x\,dx$, we put $f(x) = \tan^{-1} x$ and $g'(x) = 1$, so $g(x) = x$ and $f'(x) = \frac{1}{1+x^2}$, giving

$$\begin{aligned}\int_0^1 \tan^{-1} x\,dx &= \left[x\tan^{-1} x\right]_0^1 - \int_0^1 \frac{x}{1+x^2}\,dx && \text{(then substitute } u = 1+x^2\text{)}\\ &= \frac{\pi}{4} - \frac{1}{2}\int_1^2 \frac{1}{u}\,du && \text{(using } du = 2x\,dx\text{)}\\ &= \frac{\pi}{4} - \frac{\ln 2}{2}.\end{aligned}$$

Remark 41.10. The integral $\int_0^1 \tan^{-1} x\,dx$ represents the area A_1 in the diagram below. The area A_2 can be computed by observing that $A_1 + A_2 = \frac{\pi}{4}$, the area of the rectangle, or via the integral $A_2 = \int_0^{\pi/4} \tan y\,dy$, since it is the area to the left of the curve $x = \tan y$. Thus we conclude that $\int_0^{\pi/4} \tan y\,dy = \frac{\pi}{4} - A_1 = \frac{\ln 2}{2}$. This is consistent with the fact that (as we computed last semester using the substitution rule) $\int \tan y\,dy = \ln|\sec y|$ and thus $\int_0^{\pi/4} \tan y\,dy = \ln\sqrt{2}$.



Lecture 42 Trigonometric integrals

Stewart §7.2, Spivak Ch. 19

42.1. Powers of sine

As Remark 41.10 showed, there may be more than one way to correctly calculate a given integral. For another example of this, consider $\int \sin^2 x\,dx$. One approach is to use the identity

$$\cos(2x) = \cos^2 x - \sin^2 x = 1 - 2\sin^2 x \quad\Rightarrow\quad \sin^2 x = \frac{1}{2}(1 - \cos(2x)) \tag{42.1}$$

together with the substitution $u = 2x$, $du = 2\,dx$ to get

$$\begin{aligned}\int \sin^2 x\,dx &= \frac{1}{2}\int (1 - \cos(2x))\,dx = \frac{x}{2} - \frac{1}{4}\int \cos u\,du\\ &= \frac{x}{2} - \frac{1}{4}\sin u + C = \frac{x}{2} - \frac{1}{4}\sin(2x) + C.\end{aligned} \tag{42.2}$$

A second, equally good, approach is to use integration by parts to get

$$\begin{aligned}\int \sin^2 x\,dx &= \int \underbrace{(\sin x)}_{u}\underbrace{(\sin x)\,dx}_{dv} = \underbrace{(\sin x)}_{u}\underbrace{(-\cos x)}_{v} - \int \underbrace{(-\cos x)}_{v}\underbrace{(\cos x)\,dx}_{du} \\ &= -\sin x\cos x + \int \cos^2 x\,dx = -\sin x\cos x + \int (1-\sin^2 x)\,dx \\ &= -\sin x\cos x + x - \int \sin^2\,dx;\end{aligned}$$

then we can add $\int \sin^2 x\,dx$ to both sides and divide by 2, obtaining

$$\int \sin^2 x\,dx = \frac{1}{2}(x - \sin x\cos x) + C. \tag{42.3}$$

This agrees with (42.2) because $\frac{1}{4}\sin(2x) = \frac{1}{4}\cdot 2\sin x\cos x = \frac{1}{2}\sin x\cos x$.

So why bother with two different approaches? One reason is that they generalize to solve different classes of problems, as we will soon see: for some integrals, the first approach via trigonometric identities and substitution is better, while for others, the second approach via iterated integration by parts has advantages.

For now, let us return to the first way of computing $\int \sin^2 x\,dx$, where we used trigonometric identities and substitutions. Can we use this to compute $\int \sin^n x\,dx$ for other values of n?

For $n = 3$ we quickly see that the half-angle formula (42.1) does not seem to help:

$$\int \sin^3 x\,dx = \int \sin x\cdot\frac{1}{2}(1-\cos 2x)\,dx = \ldots?$$

For $n = 4$, on the other hand, we have

$$\begin{aligned}\int \sin^4 x\,dx &= \int \frac{1}{4}(1-\cos 2x)^2\,dx = \frac{1}{4}\int (1 - 2\cos 2x + \cos^2 2x)\,dx \\ &= \frac{1}{4}x - \frac{1}{4}\sin 2x + \frac{1}{4}\int \cos^2 2x\,dx;\end{aligned}$$

to compute this last integral, observe that (42.1) gives $\cos^2 y = \frac{1}{2}(1+\cos 2y)$, so

$$\int \cos^2 2x\,dx = \frac{1}{2}\int (1+\cos 4x)\,dx = \frac{1}{2}x + \frac{1}{8}\sin 4x + C,$$

and we conclude that

$$\int \sin^4 x\,dx = \frac{1}{4}x - \frac{1}{4}\sin 2x + \frac{1}{4}\Big(\frac{1}{2}x + \frac{1}{8}\sin 4x\Big) + C = \frac{3}{8}x - \frac{1}{4}\sin 2x + \frac{1}{32}\sin 4x + C.$$

A similar approach works for any even power of $\sin x$, but not for odd powers, as the case $n = 3$ illustrates. For odd powers, though, we can use the substitution rule without using a half-angle identity: writing $u = -\cos x$, so that $du = \sin x\,dx$, we get

$$\begin{aligned}\int \sin^3 x\,dx &= \int (\sin^2 x)(\sin x)\,dx = \int (1-\cos^2 x)\sin x\,dx = \int (1-u^2)\,du \\ &= u - \frac{1}{3}u^3 + C = -\cos x + \frac{1}{3}\cos^3 x + C.\end{aligned}$$

The same substitution will work for any odd power, although the computation will become longer. We could similarly compute $\int \cos^n x\,dx$ whenever n is odd by using $u = \sin x$, $du = \cos x\,dx$.

42.2. Products of sines and cosines

Now suppose that we want to evaluate an integral that involves powers of both sine and cosine, such as $\int \sin^2 x \cos^5 x\,dx$. By making the substitution $u = \sin x$, $du = \cos x\,dx$, we get

$$\begin{aligned}\int \sin^2 x \cos^5 x\,dx &= \int (\sin^2 x)(\cos^2 x)^2 \cos x\,dx = \int u^2(1-u^2)^2\,du \\ &= \int u^2(1-2u^2+u^4)\,du = \int u^2 - 2u^4 + u^6\,du \\ &= \frac{1}{3}u^3 - \frac{2}{5}u^5 + \frac{1}{7}u^7 + C = \frac{1}{3}\sin^3 x - \frac{2}{5}\sin^5 x + \frac{1}{7}\sin^7 x + C.\end{aligned}$$

Indeed, this substitution works to compute $\int \sin^m x \cos^n x\,dx$ whenever $m, n \geq 0$ and n is odd: if $n = 2k+1$, then $u = \sin x$, $du = \cos x\,dx$ gives

$$\int \sin^m x \cos^n x\,dx = \int \sin^m x(1-\sin^2 x)^k \cos x\,dx = \int u^m(1-u^2)^k\,du. \tag{42.4}$$

The last integral can be computed by expanding $(1-u^2)^k$ and using $\int u^\ell\,du = \frac{u^{\ell+1}}{\ell+1}$. In the case when m is odd, the substitution $u = \cos x$, $du = -\sin x\,dx$ lets us do a similar computation. Now we can summarize the overall strategy.

Technique 42.1. To compute $\int \sin^m x \cos^n x\,dx$ when $m, n \geq 0$, do the following:

(1) if n is odd, use the substitution $u = \sin x$, $du = \cos x$ as in (42.4);
(2) if m is odd, use the substitution $u = \cos x$, $du = -\sin x\,dx$;
(3) if n, m are both even, use the trigonometric identities $\sin^2 x = \frac{1}{2}(1-\cos 2x)$ and $\cos^2 x = \frac{1}{2}(1+\cos 2x)$ to rewrite the integral.

Example 42.2. With $m = 4$ and $n = 2$, we use the half-angle formulas to get

$$\begin{aligned}I = \int \sin^4 x \cos^2 x\,dx &= \int \frac{1}{4}(1-\cos 2x)^2 \frac{1}{2}(1+\cos 2x)\,dx \\ &= \frac{1}{8}\int (1 - 2\cos 2x + \cos^2 2x)(1+\cos 2x)\,dx \\ &= \frac{1}{8}\int (1 - \cos 2x - \cos^2 2x + \cos^3 2x)\,dx.\end{aligned}$$

The first two terms are easy to integrate. For the third we use the half-angle formula again to get

$$\int \cos^2 2x\,dx = \int \frac{1}{2}(1+\cos 4x)\,dx = \frac{1}{2}x + \frac{1}{8}\sin 4x + C.$$

For the fourth, we use $u = \sin 2x$ and $du = 2\cos 2x\,dx$ to get

$$\int \cos^3 2x\,dx = \int (1-\sin^2 2x)\cos 2x\,dx = \int (1-u^2)\frac{du}{2}$$

$$= \frac{1}{2}u - \frac{1}{6}u^3 + C = \frac{1}{2}\sin 2x - \frac{1}{6}\sin^3 2x + C.$$

Putting it all together gives

$$\begin{aligned} I &= \frac{1}{8}x - \frac{1}{16}\sin 2x - \frac{1}{8}\Big(\frac{1}{2}x + \frac{1}{8}\sin 4x\Big) + \frac{1}{8}\Big(\frac{1}{2}\sin 2x - \frac{1}{6}\sin^3 2x\Big) + C \\ &= \frac{1}{16}x - \frac{1}{64}\sin 4x - \frac{1}{48}\sin^3 2x + C. \end{aligned}$$

42.3. Products of tangents and secants

The technique above works well enough when $m, n \geq 0$. But what if one or both of them is negative? For the moment we consider the case when cos appears in the denominator, and see that converting the expression to tangents and secants is useful. (When sin is in the denominator, one should use cot and csc instead, and the story is similar. When both sin and cos are in the denominator, things become more difficult, and we will not consider this case.)

Example 42.3. $\displaystyle\int \frac{\sin x}{\cos^2 x}\,dx = \int \tan x \sec x\,dx = \sec x + C.$

More generally, whenever $n \geq m \geq 0$ we can write

$$\int \frac{\sin^m x}{\cos^n x}\,dx = \int \tan^m x \sec^{n-m} x\,dx,$$

so now we will study integrals of the form $\int \tan^m x \sec^k x\,dx$.

Exercise 42.4. If $m > n \geq 0$, show that we can always use the identity $\sin^2 x = 1 - \cos^2 x$ to write $\int \frac{\sin^m x}{\cos^n x}\,dx$ in terms of integrals of products of tangents and secants, as in the following:

$$\int \frac{\sin^3 x}{\cos^2 x} = \int \Big(\frac{\sin x}{\cos^2 x} - \frac{\sin x \cos^2 x}{\cos^2 x}\Big)\,dx = \int (\sec x \tan x - \sin x)\,dx = \sec x + \cos x + C.$$

Since the substitutions $u = \sin x$ and $u = \cos x$ worked in the previous section, it is natural to try the substitutions $u = \tan x$ and $u = \sec x$ to evaluate $\int \tan^m x \sec^k x\,dx$.

- $u = \tan x$ gives $du = \sec^2 x\,dx$, so for this to be effective we need to peel off a factor of $\sec^2 x$ and then be able to use the identity $\sec^2 x = \tan^2 x + 1$:

  $$\int \tan^m x \sec^k x\,dx = \int (\tan^m x \sec^{k-2} x)(\sec^2 x)\,dx = \int u^m (1+u^2)^{\frac{k-2}{2}}\,du.$$

  As long as k is even, this will lead to a polynomial that we can integrate.
- $u = \sec x$ gives $du = \sec x \tan x\,dx$, which helps if we can to remove a factor of $\sec x \tan x$ and then use the identity $\tan^2 x = \sec^2 x - 1 = u^2 - 1$:

  $$\int \tan^m x \sec^k x\,dx = \int (\tan^{m-1} x \sec^{k-1} x)(\sec x \tan x)\,dx = \int (u^2-1)^{\frac{m-1}{2}} u^{k-1}\,du$$

  This leads to a polynomial if m is odd.

Thus for $\int \tan^m x \sec^k x\,dx$, we have the following analogue of Technique 42.1: use the substitution $u = \tan x$ if $k \geq 2$ is even, and $u = \sec x$ if $m \geq 1$ is odd (as long as $k \geq 1$). If k is even and m is odd, then either substitution can be used.

Example 42.5. When $m = 2$ and $k = 4$ we use $u = \tan x$, $du = \sec^2 x\,dx$ to get

$$\int \tan^2 x \sec^4 x\,dx = \int \tan^2 x(1+\tan^2 x)\sec^2 x\,dx = \int u^2(1+u^2)\,du$$
$$= \int (u^2+u^4)\,du = \frac{1}{3}u^3 + \frac{1}{5}u^5 + C = \frac{1}{3}\tan^3 x + \frac{1}{5}\tan^5 x + C.$$

Example 42.6. When $m = k = 5$ we use $u = \sec x$, $du = \sec x \tan x\,dx$ together with $\tan^2 x = \sec^2 x - 1 = u^2 - 1$ to get

$$\int \tan^5 x \sec^5 x\,dx = \int (\tan^2 x)^2 \sec^4 x(\sec x \tan x)\,dx = \int (u^2-1)^2 u^4\,du$$
$$= \int (u^8 - 2u^6 + u^4)\,du = \frac{1}{9}\sec^9 x - \frac{2}{7}\sec^7 x + \frac{1}{5}\sec^5 x + C.$$

So far we have seen that $\int \tan^m x \sec^k x\,dx$ can be computed by the substitution $u = \tan x$ if $k \geq 2$ is even, and by $u = \sec x$ if $m \geq 1$ is odd (unless $k = 0$). The remaining cases not covered by this approach are the following:

(1) $k = 0$, so there are no powers of $\sec x$ to remove.
(2) The power on $\tan x$ is even, and the power on $\sec x$ is odd.

In the first case, we have $\int \tan^m x\,dx$. When $m = 1$ we recall that this can be computed by the substitution $u = \cos x$:

$$\int \tan x\,dx = \int \frac{\sin x}{\cos x}\,dx = -\int \frac{1}{u}\,du = -\ln|\cos x| + C = \ln|\sec x| + C.$$

For $m = 2$ we can use the identity $\tan^2 x = \sec^2 x - 1$ to get

$$\int \tan^2 x\,dx = \int (\sec^2 x - 1)\,dx = \tan x - x + C.$$

For $m \geq 3$, we can use the same identity and the substitution $u = \tan x$ to write

$$\int \tan^m x\,dx = \int \tan^{m-2} x \sec^2 x\,dx - \int \tan^{m-2} x\,dx = \int u^{m-2}\,du - \int \tan^{m-2} x\,dx$$
$$= \frac{1}{m-1}\tan^{m-1} x - \int \tan^{m-2} x\,dx.$$

Iterating this, we eventually reach either $\int \tan x\,dx$ or $\int \tan^2 x\,dx$.

Example 42.7. $\displaystyle\int \tan^3 x\,dx = \frac{1}{2}\tan^2 x - \int \tan x\,dx = \frac{1}{2}\tan^2 x - \ln|\sec x| + C.$

What about the second case above, $\int \tan^{2m} x \sec^{2k+1} x\,dx$? In this case we can still use the identity $\tan^{2m} x = (\tan^2 x)^m = (\sec^2 x + 1)^m$ to write the integral in terms of integrals of the form $\int \sec^{2\ell+1} x\,dx$. But how do we evaluate such integrals?

Lecture 43 More trigonometric integrals

Stewart §7.2, Spivak Ch. 19

43.1. The integral of secant

The last lecture left open the problem of how to evaluate $\int \sec^{2\ell+1} x\,dx$, where $\ell \geq 0$. Let us focus on the case $\ell = 0$ and compute $\int \sec x\,dx$. In the absence of any better ideas, we might observe that $\sec x$ is an odd power of cosine (albeit a negative power), and so perhaps the substitution $u = \sin x$ will benefit us here too. With this substitution we have $du = \cos x\,dx$, and we get

$$\int \sec x\,dx = \int \frac{1}{\cos x}\frac{du}{\cos x} = \int \frac{1}{1-\sin^2 x}\,du = \int \frac{1}{1-u^2}\,du.$$

At this point it is not so clear what to do. The way forward turns out to be the observation that

$$\frac{1}{1-u^2} = \frac{1}{(1+u)(1-u)} = \frac{1}{2}\Big(\frac{1}{1+u} + \frac{1}{1-u}\Big). \tag{43.1}$$

While you can verify easily enough that this is true, it is probably not the first thing that would have popped into your head, and you may reasonably wonder how one might think of doing this step! It will seem more natural once we have discussed *partial fractions*. At any rate, having reached this point we are in fact nearly done. Indeed, since $\int \frac{1}{1+u}\,du = \ln|1+u|$ and $\int \frac{1}{1-u}\,du = -\ln|1-u|$, we have

$$\int \sec x\,dx = \frac{1}{2}\big(\ln|1+u| - \ln|1-u|\big) = \frac{1}{2}\ln\Big|\frac{1+u}{1-u}\Big|,$$

omitting the constant of integration for the time being. Recalling that $u = \sin x$, we multiply top and bottom by $(1+u)$ to obtain $1-u^2 = 1-\sin^2 x = \cos^2 x$ in the denominator, and get

$$\int \sec x\,dx = \frac{1}{2}\ln\Big|\frac{(1+u)^2}{1-u^2}\Big| = \frac{1}{2}\ln\Big|\frac{(1+\sin x)^2}{\cos^2 x}\Big| = \ln\Big|\frac{1+\sin x}{\cos x}\Big| = \ln|\sec x + \tan x|.$$

In order to get the most general antiderivative we add the constant of integration:

$$\int \sec x\,dx = \ln|\sec x + \tan x| + C. \tag{43.2}$$

Although this argument does not use any rules that you have not learned yet, it is certainly not one that I would expect you to come up with on your own! It does, however, illustrate a little bit of the nature of computing indefinite integrals; there are many different steps that one might take next at any given stage, and it is a little bit of an art form to decide which one is most likely to be useful. Certain tricks appear over and over again – factor a difference of squares, add and subtract the same thing, multiply and divide by the same thing, look for any useful trigonometric identities that may be relevant – but in the end there is no substitute for just working through lots of problems and gaining practice and experience in integrating.

43.2. More trigonometric identities

One more set of trigonometric identities is worth mentioning at this point.

Exercise 43.1. Use the formulas for $\cos(A \pm B)$ and $\sin(A \pm B)$ to prove that

$$\sin A \cos B = \frac{1}{2}\big[\sin(A+B) + \sin(A-B)\big],$$
$$\sin A \sin B = \frac{1}{2}\big[\cos(A-B) - \cos(A+B)\big],$$
$$\cos A \cos B = \frac{1}{2}\big[\cos(A-B) + \cos(A+B)\big].$$

These identities can be used to evaluate integrals involving products of $\sin(mx)$ and $\cos(nx)$.

Example 43.2. The first identity above gives

$$\begin{aligned}\int \sin 2x \cos 7x\, dx &= \frac{1}{2}\int \big[\sin(2x+7x) + \sin(2x-7x)\big]\, dx \\ &= \frac{1}{2}\int \sin 9x\, dx - \frac{1}{2}\int \sin 5x\, dx = -\frac{1}{18}\cos 9x + \frac{1}{10}\cos 5x + C.\end{aligned}$$

43.3. *A reduction formula for powers of sine

Now let us return to the second approach given in §42.1 to compute $\int \sin^2 x\, dx$, using integration by parts, and see what happens if we try to use this approach to compute $\int \sin^n x\, dx$. Mimicking the integration by parts from that section, we can write

$$\begin{aligned}\int \sin^n x\, dx &= \int \underbrace{(\sin x)^{n-1}}_{u}\underbrace{(\sin x)\, dx}_{dv} \\ &= \underbrace{(\sin x)^{n-1}}_{u}\underbrace{(-\cos x)}_{v} - \int \underbrace{(-\cos x)}_{v}\underbrace{(n-1)(\sin x)^{n-2}(\cos x)\, dx}_{du} \\ &= -\cos x \sin^{n-1} x + (n-1)\int \sin^{n-2} x \cos^2 x\, dx.\end{aligned}$$

Then using $\sin^{n-2} x \cos^2 x = \sin^{n-2} x(1 - \sin^2 x) = \sin^{n-2} x - \sin^n x$, we get

$$\int \sin^n x\, dx = -\cos x \sin^{n-1} x + (n-1)\int \sin^{n-2} x\, dx - (n-1)\int \sin^n x\, dx.$$

Adding $(n-1)\int \sin^n x\, dx$ to both sides gives

$$n\int \sin^n x\, dx = -\cos x \sin^{n-1} x + (n-1)\int \sin^{n-2} x\, dx,$$

and dividing by n we obtain

$$\int \sin^n x\, dx = -\frac{1}{n}\cos x \sin^{n-1} x + \frac{n-1}{n}\int \sin^{n-2} x\, dx. \tag{43.3}$$

This is not a complete answer in and of itself, but it lets us reduce the problem to a similar question for a smaller value of n, and by iterating the procedure we will eventually reach $\int \sin x\, dx$ or $\int \sin^2 x\, dx$, both of which we know how to compute.

Example 43.3. Using $n = 3$ in (43.3) gives

$$\int \sin^3 x\,dx = -\frac{1}{3}\cos x \sin^2 x + \frac{2}{3}\int \sin x\,dx = -\frac{1}{3}\cos x \sin^2 x - \frac{2}{3}\cos x + C.$$

Exercise 43.4. Use integration by parts to prove a similar reduction formula for $\int \sec^2 x\,dx$; together with §43.1 this lets you compute $\int \sec^n x\,dx$ for all $n \in \mathbb{N}$.

43.4. *The Wallis product

Now suppose we compute the definite integrals associated to these examples over the interval $[0, \pi]$. That is, we consider for each $n = 0, 1, 2, 3, \ldots$ the real number

$$I_n = \int_0^\pi \sin^n x\,dx.$$

Before doing any computations, observe that the sequence I_n represents the areas of the regions shown here.

Remark 43.5. It appears that these regions are getting smaller and smaller, so that $\lim_{n\to\infty} I_n = 0$. This turns out to be true, but it takes a little bit of work to prove, and we will not do so here.

The first two terms are easy to compute:

$$I_0 = \int_0^\pi 1\,dx = \pi,$$
$$I_1 = \int_0^\pi \sin x\,dx = \Big[-\cos x\Big]_0^\pi = -\cos\pi + \cos 0 = 2.$$

For larger values of n, we use the reduction formula (43.3):

$$I_n = \int_0^\pi \sin^n x\,dx = \Big[-\frac{1}{n}\cos x \sin^{n-1} x\Big]_0^\pi + \frac{n-1}{n}\int_0^\pi \sin^{n-2} x\,dx.$$

Since $\sin 0 = \sin\pi = 0$, the first term on the RHS vanishes, and the last integral is just I_{n-2}, so we get

$$I_n = \frac{n-1}{n} I_{n-2}. \tag{43.4}$$

Thus the next few terms in the sequence are

$$I_2 = \frac{1}{2}I_0 = \pi\cdot\frac{1}{2}, \qquad I_3 = \frac{2}{3}I_1 = 2\cdot\frac{2}{3},$$
$$I_4 = \frac{3}{4}I_2 = \pi\cdot\frac{1}{2}\cdot\frac{3}{4}, \qquad I_5 = \frac{4}{5}I_3 = 2\cdot\frac{2}{3}\cdot\frac{4}{5},$$

and so on. The general formula is

$$I_{2n} = \pi\cdot\frac{1}{2}\cdot\frac{3}{4}\cdot\cdots\cdot\frac{2n-1}{2n}, \qquad I_{2n+1} = 2\cdot\frac{2}{3}\cdot\frac{4}{5}\cdot\cdots\cdot\frac{2n}{2n+1}.$$

So far this is kind of cute, but now something surprising happens. We see that there is one rule for the even terms in the sequence I_n, and another rule for the odd terms.

What happens if we compare two consecutive terms, one even and one odd? Are they close together, or far apart? Note that since $0 \le \sin x \le 1$ for all $x \in [0, \pi]$, we have $\sin^{n+1} x \le \sin^n x$ for all n, and thus

$$I_{n+1} = \int_0^\pi \sin^{n+1} x \, dx \le \int_0^\pi \sin^n x \, dx = I_n.$$

In particular, this gives $I_{2n} \ge I_{2n+1} \ge I_{2n+2}$, and dividing through by I_{2n} gives

$$1 = \frac{I_{2n}}{I_{2n}} \ge \frac{I_{2n+1}}{I_{2n}} \ge \frac{I_{2n+2}}{I_{2n}} = \frac{2n+1}{2n+2} \quad \text{using (43.4).}$$

Since $\lim_{n\to\infty} \frac{2n+1}{2n+2} = \lim_{n\to\infty} 1 = 1$, the Squeeze Theorem implies that

$$\lim_{n\to\infty} \frac{I_{2n+1}}{I_{2n}} = 1. \tag{43.5}$$

From the formulas for I_{2n} and I_{2n+1}, we see that

$$\frac{I_{2n+1}}{I_{2n}} = \frac{2 \cdot \frac{2}{3} \cdot \frac{4}{5} \cdot \cdots \cdot \frac{2n}{2n+1}}{\pi \cdot \frac{1}{2} \cdot \frac{3}{4} \cdot \cdots \cdot \frac{2n-1}{2n}} = \frac{2}{\pi} \cdot \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdots \frac{2n}{2n-1} \frac{2n}{2n+1}.$$

Together with (43.5), this implies that

$$1 = \lim_{n\to\infty} \Big(\frac{2}{\pi} \cdot \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdots \frac{2n}{2n-1} \frac{2n}{2n+1} \Big),$$

or if you prefer, after multiplying both sides by $\frac{\pi}{2}$,

$$\frac{\pi}{2} = \lim_{n\to\infty} \Big(\frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdots \frac{2n}{2n-1} \frac{2n}{2n+1} \Big). \tag{43.6}$$

This is the *Wallis product*, a formula for π that was discovered in 1655 by the English mathematician John Wallis. It is often written as an *infinite product*:

$$\frac{\pi}{2} = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdot \frac{8}{7} \cdot \frac{8}{9} \cdots .$$

We will have more to say about expressions like this when we study sequences and series at the end of the course; for the time being I merely urge extreme caution. Because this expression is infinite and not finite, it does not always behave in the way we might expect. For example, one might be tempted to say that because it does not matter in which order we multiply and divide things, we could just as well write the final expression as

$$\frac{2}{3} \cdot \frac{2}{3} \cdot \frac{4}{5} \cdot \frac{4}{5} \cdot \frac{6}{7} \cdot \frac{6}{7} \cdots \tag{43.7}$$

by moving all the denominators one spot to the left. But this turns out to be quite wrong! Indeed, if we write $x_n = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdots \frac{2n}{2n-1} \frac{2n}{2n+1}$ and $y_n = \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{4}{5} \cdot \frac{4}{5} \cdots \frac{2n}{2n+1} \frac{2n}{2n+1}$, then (43.6) says that $\lim_{n\to\infty} x_n = \frac{\pi}{2}$, and we see clearly that $y_n = \frac{x_n}{2n+1}$, so the product in (43.7) should be interpreted as

$$\lim_{n\to\infty} y_n = \lim_{n\to\infty} \frac{x_n}{2n+1} = 0.$$

But all this is really a discussion for another time, mentioned here merely to illustrate why we should exercise some care when treating infinite expressions.

Lecture 44 Trigonometric substitutions

Stewart §7.3, Spivak Ch. 19

44.1. Reversing the substitution rule

We know that the area of the unit circle is π, so the area under the curve $y = \sqrt{1-x^2}$ from 0 to 1 is $\frac{\pi}{4}$: in other words,

$$\int_0^1 \sqrt{1-x^2}\,dx = \frac{\pi}{4}.$$

Can we compute this integral using the fundamental theorem of calculus, by finding an antiderivative? The function $\sqrt{1-x^2}$ is not on our list of known derivatives, and integration by parts will not get us anywhere. Neither will the first substitutions we might try: $u = x^2$, $u = 1-x^2$, $u = \sqrt{1-x^2}$. On the other hand, there is a substitution we can make that helps, but it looks rather different: instead of replacing x with a new variable that is a function of x, we write x as a function of a new variable t by putting $x = \sin t$. Then we have $dx = \cos t\,dt$, and the usual trigonometric identities give

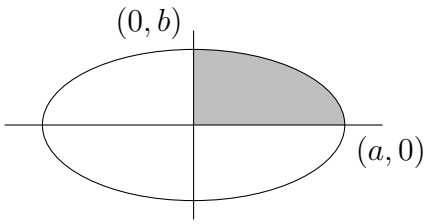
$$\begin{aligned}\int \sqrt{1-x^2}\,dx &= \int \left(\sqrt{1-\sin^2 t}\right)\cos t\,dt = \int \cos^2 t\,dt = \frac{1}{2}\int (1+\cos 2t)\,dt \\ &= \frac{1}{2}t + \frac{1}{4}\sin 2t + C = \frac{1}{2}t + \frac{1}{2}\sin t\cos t + C \\ &= \frac{1}{2}\sin^{-1} x + \frac{1}{2}x\sqrt{1-x^2} + C.\end{aligned}$$

In previous lectures we have made some effort to point out that dx and dt do not have an independent meaning in their own right, so the proper way to justify the above substitution is to define t by $t = \sin^{-1} x$, and then observe that writing $g(t) = \sin t$, the usual substitution rule gives

$$\int f(g(t))g'(t)\,dt = \int f(x)\,dx. \tag{44.1}$$

In this example, though we are going the reverse of the usual direction. In previous applications of the substitution rule, we wanted to find functions f and g so that the integral we were given could be rewritten as the LHS of (44.1), and then transformed into the RHS; in the present example, we start with the integral on the RHS and look for a function g such that transforming it into the LHS is productive.

Remark 44.1. Since sin is not 1-1 on its entire domain, any use of the inverse function $\sin^{-1}$ must always come with a choice of which branch we use. It is standard to choose $\sin^{-1} x \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, which is why we chose $\cos t = \sqrt{1-\sin^2 t}$ instead of $\cos t = -\sqrt{1-\sin^2 t}$ in the above computation; the latter choice would correspond to a different branch of the inverse function, although it would still lead to a valid antiderivative.



Example 44.2. To find the area enclosed by the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$, where $a, b > 0$ are the lengths of the two semi-axes of the ellipse, we can solve for y in the first quadrant and get

$$\frac{y^2}{b^2} = 1 - \frac{x^2}{a^2} = \frac{a^2 - x^2}{a^2} \quad \Rightarrow \quad y = \frac{b}{a}\sqrt{a^2 - x^2},$$

so that the total area of the ellipse is $A = 4\int_0^a \frac{b}{a}\sqrt{a^2 - x^2}\,dx$. Then we use the substitution $x = a\sin\theta$, $dx = a\cos\theta\,d\theta$, to get

$$\begin{aligned}\int_0^a \sqrt{a^2 - x^2}\,dx &= \int_0^{\pi/2} \sqrt{a^2 - a^2\sin^2\theta}\cdot a\cos\theta\,d\theta = a^2\int_0^{\pi/2}\cos^2\theta\,d\theta \\ &= \frac{a^2}{2}\int_0^{\pi/2}(1+\cos 2\theta)\,d\theta = \frac{a^2}{2}\Big[\theta + \frac{1}{2}\sin 2\theta\Big]_0^{\pi/2} = \frac{\pi a^2}{4},\end{aligned}$$

and we conclude that the area of the ellipse is

$$A = \frac{4b}{a}\int_0^a \sqrt{a^2 - x^2}\,dx = \frac{4b}{a}\cdot\frac{\pi a^2}{4} = \pi ab.$$

Notice that in the case $a = b$ this reduces to the familiar formula for the area of a circle.

Technique 44.3. Trigonometric substitutions such as the one above are useful for simplifying integrals involving quadratic polynomials inside square roots.

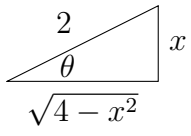
- If you see $\sqrt{a^2 - x^2}$, use $x = a\sin\theta$, $dx = a\cos\theta\,d\theta$, and $1 - \sin^2\theta = \cos^2\theta$.
- For $\sqrt{a^2 + x^2}$, use $x = a\tan\theta$, $dx = a\sec^2\theta\,d\theta$, and $1 + \tan^2\theta = \sec^2\theta$.
- For $\sqrt{x^2 - a^2}$, use $x = a\sec\theta$, $dx = a\sec\theta\tan\theta\,d\theta$, and $\sec^2\theta - 1 = \tan^2\theta$.

44.2. More examples

Example 44.4. To compute $\int \frac{\sqrt{4-x^2}}{x^2}\,dx$, put $x = 2\sin\theta$, $dx = 2\cos\theta\,d\theta$, and get

$$\begin{aligned}\int \frac{\sqrt{4-x^2}}{x^2}\,dx &= \int \frac{\sqrt{4-4\sin^2\theta}}{4\sin^2\theta}2\cos\theta\,d\theta = \int \frac{\cos^2\theta}{\sin^2\theta}\,d\theta = \int \cot^2\theta\,d\theta \\ &= \int(\csc^2\theta - 1)\,d\theta = -\cot\theta - \theta + C.\end{aligned}$$

To complete the solution we must write this in terms of x. It is useful to draw the triangle shown at right, where the edges are determined by the condition that $\sin\theta = x/2$ together with the Pythagorean theorem. We see that $\cot\theta = \frac{\sqrt{4-x^2}}{x}$, and obtain



$$\int \frac{\sqrt{4-x^2}}{x^2}\,dx = -\frac{\sqrt{4-x^2}}{x^2} - \sin^{-1}\frac{x}{2} + C.$$

Remark 44.5. We could also evaluate this integral using integration by parts, with $u = \sqrt{4-x^2}$ and $dv = x^{-2}\,dx$, so $v = -1/x$, and we would obtain the same result. But the approach using trigonometric substitution is a little more routine in that we do not have to guess at a choice of u and v that work.

Example 44.6. To compute $\int \frac{1}{x\sqrt{x^2+1}}\,dx$ we put $x = \tan\theta$, $dx = \sec^2\theta\,d\theta$, and get

$$\int \frac{1}{x\sqrt{x^2+1}}\,dx = \int \frac{\sec^2\theta}{\tan\theta\sqrt{\tan^2\theta+1}}\,d\theta = \int \frac{\sec^2\theta}{\tan\theta|\sec\theta|}\,dx.$$

To eliminate the absolute value signs we choose $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$ so that $\sec\theta > 0$, and we get

$$\int \frac{1}{x\sqrt{x^2+1}}\,dx = \int \frac{\sec\theta}{\tan\theta}\,d\theta = \int \frac{1/\cos\theta}{\sin\theta/\cos\theta}\,d\theta = \int \csc\theta\,d\theta.$$

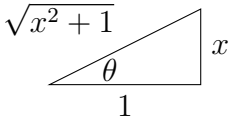
We have not evaluated this before, but we did compute $\int \sec\theta\,d\theta = \ln|\tan\theta + \sec\theta|$. Thus it is natural to expect that the integral of $\csc\theta$ is related to $\ln|\cot\theta + \csc\theta|$, and differentiating this expression gives

$$\frac{d}{d\theta}\ln|\cot\theta + \csc\theta| = \frac{-\csc^2\theta - \csc\theta\cot\theta}{\cot\theta + \csc\theta} = -\csc\theta.$$

We conclude that

$$\int \frac{1}{x\sqrt{x^2+1}}\,dx = \int \csc\theta\,d\theta = -\ln|\cot\theta + \csc\theta|.$$

To write this in terms of x, we use the triangle shown to get $\cot\theta = \frac{1}{x}$ and $\csc\theta = \frac{\sqrt{x^2+1}}{x}$, so that



$$\int \frac{1}{x\sqrt{x^2+1}}\,dx = -\ln\left|\frac{1}{x} + \frac{\sqrt{x^2+1}}{x}\right| = \ln\left|\frac{x}{1+\sqrt{x^2+1}}\right| = \ln|x| - \ln(1+\sqrt{x^2+1}) + C.$$

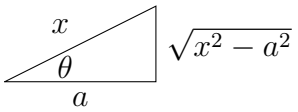
Example 44.7. For $\int \frac{x}{\sqrt{x^2+1}}\,dx$, we use $x = \tan\theta$ and $dx = \sec^2\theta\,d\theta$ with $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$ to get

$$\int \frac{x}{\sqrt{x^2+1}}\,dx = \int \frac{\tan\theta\sec^2\theta}{\sqrt{\tan^2\theta+1}}\,d\theta = \int \tan\theta\sec\theta\,d\theta = \sec\theta + C = \sqrt{x^2+1} + C.$$

Observe that we could also have used the substitution $u = x^2 + 1$.

Example 44.8. To compute $\int \frac{1}{\sqrt{x^2-a^2}}\,dx$, where $a > 0$, we put $x = a\sec\theta$ and $dx = a\sec\theta\tan\theta\,d\theta$, using the range $\theta \in [0, \frac{\pi}{2}) \cup [\pi, \frac{3\pi}{2})$ so that $\tan\theta > 0$, and get

$$\begin{aligned}\int \frac{1}{\sqrt{x^2-a^2}}\,dx &= \int \frac{a\sec\theta\tan\theta}{\sqrt{a^2\sec^2\theta - a^2}}\,d\theta = \int \frac{\sec\theta\tan\theta}{\tan\theta}\,d\theta \\ &= \int \sec\theta\,d\theta = \ln|\sec\theta + \tan\theta| + C \\ &= \ln\left|\frac{x}{a} + \frac{\sqrt{x^2-a^2}}{a}\right| + C.\end{aligned}$$



To obtain a marginally simpler expression we can absorb $-\ln a$ into the constant of integration and write $\int \frac{1}{\sqrt{x^2-a^2}}\,dx = \ln|x + \sqrt{x^2-a^2}| + C$.

Example 44.9. The previous example could also be computed by using the *hyperbolic* substitution $x = a\cosh t$, $dx = a\sinh t\,dt$, since then we have

$$\int \frac{1}{\sqrt{x^2-a^2}}\,dx = \int \frac{a\sinh t}{\sqrt{a^2\cosh^2 t - a^2}}\,dt = \int \frac{\sinh t}{\sinh t}\,dt = t + C = \cosh^{-1}\left(\frac{x}{a}\right) + C.$$

Recalling the definition $\cosh t = \frac{e^t+e^{-t}}{2}$, one can use the quadratic formula to verify that the two solutions agree with each other.

Lecture 45 Complicated quadratics; polynomial long division

Stewart §7.3–7.4, Spivak Ch. 19

45.1. Trigonometric substitutions for complicated quadratics

If an integral contains the square root of a quadratic polynomial in one of the three simple forms from Technique 44.3, then the corresponding trigonometric substitution is clear. For more complicated quadratic polynomials, a preliminary substitution can be used to bring the expression to one of the forms there.

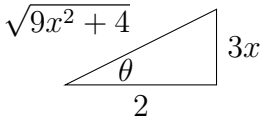
Example 45.1. If we are confronted with the integral $\int x^2(9x^2+4)^{-3/2}\,dx$, then we can first make the substitution $u = 3x$ to write $9x^2+4 = u^2+4$, and then make the trigonometric substitution $u = 2\tan\theta$. In terms of x, this is $3x = 2\tan\theta$, so $x = \frac{2}{3}\tan\theta$, $dx = \frac{2}{3}\sec^2\theta\,d\theta$, and we get

$$\begin{aligned}\int x^2(9x^2+4)^{-3/2}\,dx &= \int \frac{4}{9}\tan^2\theta(4\tan^2\theta+4)^{-3/2}\cdot\frac{2}{3}\sec^2\theta\,d\theta \\ &= \frac{8}{27}\int 4^{-3/2}\tan^2\theta(\sec^2\theta)^{-3/2}\sec^2\theta\,d\theta = \frac{1}{27}\int\frac{\tan^2\theta}{\sec\theta}\,d\theta,\end{aligned}$$

where we choose $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$ to guarantee that $\sec\theta > 0$. Writing everything in terms of sine and cosine gives

$$\begin{aligned}\int x^2(9x^2+4)^{-3/2}\,dx &= \frac{1}{27}\int\frac{\sin^2\theta}{\cos\theta}\,d\theta = \frac{1}{27}\int\left(\frac{1}{\cos\theta}-\cos\theta\right)d\theta \\ &= \frac{1}{27}\big(\ln|\sec\theta+\tan\theta| - \sin\theta\big) + C.\end{aligned}$$

Using the triangle at right, we have $\sec\theta = \frac{\sqrt{9x^2+4}}{2}$, $\tan\theta = \frac{3x}{2}$, and $\sin\theta = \frac{3x}{\sqrt{9x^2+4}}$, so



$$\int x^2(9x^2+4)^{-3/2}\,dx = \frac{1}{27}\left(\ln|\sqrt{9x^2+4}+3x| - \frac{3x}{\sqrt{9x^2+4}}\right) + C,$$

where as before we absorb a factor of $\frac{1}{27}\ln 2$ into the constant of integration.

If the quadratic contains a linear term, then we need to complete the square first.

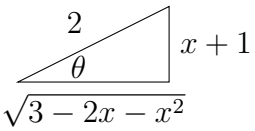
Example 45.2. To compute $I = \displaystyle\int \frac{x^2}{\sqrt{3-2x-x^2}}\,dx$, we complete the square as

$$3 - 2x - x^2 = -(x+1)^2 + 4,$$

and so the preliminary substitution to make is $u = x + 1$. Then the quantity inside the square root is $4 - u^2$, so we make the substitution $u = 2\sin\theta$. Doing both substitutions successively gives

$$\begin{aligned}\int \frac{x^2}{\sqrt{3-2x-x^2}}\,dx &= \int \frac{(u-1)^2}{\sqrt{4-u^2}}\,du = \int \frac{(2\sin\theta - 1)^2}{\sqrt{4-4\sin^2\theta}}\cdot 2\cos\theta\,d\theta \\ &= \int (4\sin^2\theta - 4\sin\theta + 1)\,d\theta = \int (2(1-\cos 2\theta) - 4\sin\theta + 1)\,d\theta \\ &= 2\theta - \sin 2\theta + 4\cos\theta + \theta + C.\end{aligned}$$

To transition back to x we use the triangle to get $\sin\theta = \frac{x+1}{2}$, $\cos\theta = \frac{1}{2}\sqrt{3-2x-x^2}$, and thus



$$\begin{aligned}I &= 3\sin^{-1}\left(\frac{x+1}{2}\right) - 2\cdot\frac{x+1}{2}\cdot\frac{\sqrt{3-2x-x^2}}{2} + 4\cdot\frac{\sqrt{3-2x-x^2}}{2} + C \\ &= 3\sin^{-1}\left(\frac{x+1}{2}\right) + (3-x)\frac{\sqrt{3-2x-x^2}}{2} + C.\end{aligned}$$

45.2. Polynomial long division

We know how to integrate polynomials using linearity and the power rule. But what about rational functions? The following example is instructive.

Example 45.3.

$$\int \frac{2x+3}{x+1}\,dx = \int \frac{2(x+1)+1}{x+1}\,dx = \int \left(2 + \frac{1}{x+1}\right)dx = 2x + \ln|x+1| + C.$$

Observe that while it was not clear how to integrate the original rational function directly, we were able to transform it into the sum of a polynomial (in this case the constant function 2) and a new rational function ($\frac{1}{x+1}$), where the new rational function has a simpler numerator and is easier to integrate.

We can carry this out more generally. Suppose we want to integrate a rational function $\frac{P(x)}{Q(x)}$, where P, Q are polynomials. The first step is to use *polynomial long division*.

Proposition 45.4. *Given any polynomials $P(x)$ and $Q(x)$, there are polynomials $S(x)$ and $R(x)$ such that*

$$P(x) = S(x)Q(x) + R(x) \quad \textit{and} \quad \deg R < \deg Q. \tag{45.1}$$

Proof. Let $P(x) = a_n x^n + f(x)$ and $Q(x) = b_k x^k + g(x)$, where $a_n, b_k \neq 0$ and f, g are polynomials with $\deg f < n$ and $\deg g < k$. Let $c_0 = a_n/b_k$ and observe that

$$c_0 x^{n-k} Q(x) = c_0 x^{n-k} b_k x^k + c_0 x^{n-k} g(x) = a_n x^n + c_0 x^{n-k} g(x),$$

where $\deg(c_0 x^{n-k} g(x)) = n - k + \deg g < n$. Thus we have

$$\begin{aligned} P(x) &= a_n x^n + f(x) = a_n x^n + c_0 x^{n-k} g(x) + \big(f(x) - c_0 x^{n-k} g(x)\big) \\ &= c_0 x^{n-k} Q(x) + \big(f(x) - c_0 x^{n-k} g(x)\big). \end{aligned}$$

Denote the quantity in brackets by $R_0(x)$. This is a polynomial with degree $< n = \deg(P)$. Repeating the above process with $R_0(x)$ in place of $P(x)$, we obtain $c_1 \in \mathbb{R}$ such that

$$R_0(x) = c_1 x^{n-k-1} Q(x) + R_1(x),$$

where R_1 is a polynomial with degree $< n - 1$. Combining this with the formula above for $P(x)$ gives

$$P(x) = (c_0 x^{n-k} + c_1 x^{n-k-1}) Q(x) + R_1(x),$$

and we can continue in this manner to obtain $c_i \in \mathbb{R}$ and $R_i(x)$ until the remainder term R_i has degree $< \deg Q$, at which point Proposition 45.4 is proved. □

Applying Proposition 45.4 to the rational function we wish to integrate, we get

$$\int \frac{P(x)}{Q(x)}\,dx = \int S(x)\,dx + \int \frac{R(x)}{Q(x)}\,dx.$$

The first integral on the RHS can be computed directly. For the second, we can compute it directly if Q is linear (so that $R(x)$ is constant), while for more general Q we will need the method of *partial fractions* that we introduce in the next section.

Example 45.5. Suppose we want to compute $\int \frac{x^3-x}{x+2}\,dx$. Then $P(x) = x^3 - x$ and $Q(x) = x + 2$, so polynomial long division gives $x^3 - x = (x^2 - 2x + 3)(x + 2) - 6$, via the following computation.

$$\begin{array}{r|l} & x^2 - 2x + 3 \\ \hline x+2 & x^3 \quad\quad - x \\ & \underline{-x^3 - 2x^2} \\ & \quad -2x^2 \;\; - x \\ & \quad \underline{2x^2 + 4x} \\ & \quad\quad\;\; 3x \\ & \quad\quad \underline{-3x - 6} \\ & \quad\quad\quad -6 \end{array}$$

Recall how the algorithm goes: since $\deg P = 3$ and $\deg Q = 1$, we must have $\deg S = 3 - 1 = 2$, so we start by determining the quadratic coefficient of S. This must be 1 in order for the leading terms of $P(x)$ and $S(x)Q(x)$ to agree, so we write x^2 in the top line and then subtract $x^2 Q(x)$ from $S(x)$, obtaining $-2x^2 - x$, and the relationship

$$x^3 - x = x^2(x + 2) - 2x^2 - x.$$

Since the remainder $-2x^2 - x$ still has degree $\geq \deg Q = 1$, we repeat the process with this polynomial in the role of P, obtaining the following at the next two steps:

$$\begin{aligned} x^3 - x &= (x^2 - 2x)(x + 2) + 3x, \\ x^3 - x &= (x^2 - 2x + 3)(x + 2) - 6. \end{aligned}$$

This last result has a remainder with degree smaller than $\deg Q = 1$, and this is what we want. Now using the result of the long division, we get

$$\int \frac{x^3 - x}{x+2}\,dx = \int \frac{(x^2-2x+3)(x+2)-6}{x+2} = \int \left(x^2 - 2x + 3 - \frac{6}{x+2}\right) dx$$
$$= \frac{1}{3}x^3 - x^2 + 3x - 6\ln|x+2| + C.$$

Lecture 46 Partial fraction decompositions

Stewart §7.4, Spivak Ch. 19

46.1. Distinct linear factors

The procedure in the previous lecture is not always enough; it may still not be immediately clear how to integrate the resulting rational function. For example, when we derived the formula for $\int \sec\theta\,d\theta$, an important step was to compute $\int \frac{1}{1-x^2}\,dx$, which is not as easy as integrating $\frac{1}{ax+b}$. The trick was to notice that

$$\frac{1}{1-x} + \frac{1}{1+x} = \frac{(1+x)+(1-x)}{(1-x)(1+x)} = \frac{2}{1-x^2},$$

which let us write

$$\int \frac{1}{1-x^2}\,dx = \frac{1}{2}\int \frac{1}{1-x} + \frac{1}{1+x}\,dx = \frac{1}{2}(\ln|1+x| - \ln|1-x|) + C.$$

This is an example of a *partial fraction decomposition*. But how did we come up with it? And how can we use a similar trick to help us compute other integrals?

Example 46.1. To compute $\int \frac{x+5}{x^2-2x-3}\,dx$, we can factor the denominator as $x^2-2x-3 = (x-3)(x+1)$ and conjecture that

$$\frac{x+5}{x^2-2x-3} = \frac{A}{x-3} + \frac{B}{x+1} \text{ for some choice of } A, B \in \mathbb{R}. \tag{46.1}$$

Observe that the RHS can be rewritten as

$$\frac{A(x+1)+B(x-3)}{(x-3)(x+1)} = \frac{(A+B)x + (A-3B)}{x^2-2x-3},$$

and so we want to choose A, B such that

$$(A+B)x + (A-3B) = x + 5 \text{ for every } x \in \mathbb{R}.$$

This happens if and only if $A+B=1$ and $A-3B=5$; this is a system of two equations in two variables, which we can easily solve to get $A=2$, $B=-1$, so that

$$\int \frac{x+5}{x^2-2x-3}\,dx = \int \frac{2}{x-3} - \frac{1}{x+1}\,dx = 2\ln|x-3| - \ln|x+1| + C.$$

Technique 46.2. A similar method works anytime the denominator can be factored into distinct linear polynomials: if $\deg P < \deg Q = n$ and $Q(x) = (x - r_1) \cdots (x - r_n)$, so that $r_1, \ldots, r_n$ are the roots of Q, then our goal is to find $A_1, \ldots, A_n \in \mathbb{R}$ such that

$$\frac{P(x)}{Q(x)} = \frac{A_1}{x - r_1} + \cdots + \frac{A_n}{x - r_n} \text{ for every } x \in \mathbb{R} \setminus \{r_1, \ldots, r_n\}. \tag{46.2}$$

Putting the RHS over a common denominator equal to $Q(x)$, one sees that (46.2) is true if and only if $A_1, \ldots, A_n$ satisfy a certain system of n linear equations in n variables, obtained by comparing the coefficients of the polynomial $P(x)$ (up to degree $n - 1$) to the coefficients of the numerator on the RHS. Once the values of $A_1, \ldots, A_n$ are found, the RHS can easily be integrated using $\int \frac{A}{x-r}\, dx = A \ln |x - r|$.

The procedure of rewriting the rational function $\frac{P(x)}{Q(x)}$ as the RHS of (46.2) is called a *partial fraction decomposition*.

Remark 46.3. We stress that (46.2) is *not* an equation to be solved for x; rather, it is a condition that is supposed to hold for *every* x, and this then determines the values of the numbers $A_1, \ldots, A_n$ using comparison of coefficients.

Technique 46.4. An alternate technique for finding the coefficients in (46.2) is to put the RHS over a common denominator and then instead of comparing coefficients, evaluate both $P(x)$ and the numerator of the RHS at n specific points. It makes sense to choose points where the RHS takes a simple form, and one can often achieve this by evaluating it at the points $r_1, \ldots, r_n$.

Example 46.5. In Example 46.1, we could rewrite (46.1) as

$$\frac{x + 5}{x^2 - 2x - 3} = \frac{A(x + 1) + B(x - 3)}{(x - 3)(x + 1)},$$

just as we did before, so that we want to find A, B such that

$$x + 5 = A(x + 1) + B(x - 3) \text{ for all } x;$$

then instead of comparing coefficients, we could evaluate this equation at $x = 3$ and $x = -1$, where it gives

$$3 + 5 = A(3 + 1) + B \cdot 0 \text{ and } -1 + 5 = A \cdot 0 + B(-1 - 3),$$

which are easily solved to give $A = 2$ and $B = -1$, just as before.

It is important to observe that both of these methods *fail* as currently formulated if the factors of $Q(x)$ are not distinct.

Example 46.6. If $P(x) = x$ and $Q(x) = (x + 1)^2$, then (46.2) becomes

$$\frac{x}{(x + 1)^2} = \frac{A_1}{(x + 1)^2} + \frac{A_2}{(x + 1)^2} = \frac{A_1 + A_2}{(x + 1)^2};$$

this can only be satisfied if $A_1 + A_2 = x$ for every x, which is impossible. (The corresponding system of linear equations is $A_1 + A_2 = 0$, $0 = 1$.)

Now there are three questions that need to be addressed.

(1) Does the system of equations coming from (46.2) always have a solution if the linear factors are all distinct?
(2) What do we do if the factors are not distinct; how do we deal with repeated roots of $Q(x)$?
(3) What do we do if $Q(x)$ does not factor into linear polynomials? For example, what if $Q(x) = x^2 + 1$?

We will defer the first question to the next lecture, and will spend the remainder of this lecture addressing the second and third questions.

46.2. Repeated factors

What do we do when the denominator $Q(x)$ has a repeated root, as in Example 46.6 so that $\frac{P(x)}{Q(x)}$ does not admit a partial fraction decomposition of the form (46.2)? Observe that in that example, we can still perform the following computation:

$$\int \frac{x}{(x+1)^2}\,dx = \int \frac{(x+1)-1}{(x+1)^2}\,dx = \int \frac{1}{x+1} - \frac{1}{(x+1)^2}\,dx = \ln|x+1| + \frac{1}{x+1} + C.$$

This suggests a general way of dealing with repeated factors.

Technique 46.7. To integrate $\frac{P(x)}{(x-r)^n}$, where P is a polynomial of degree $< n$, find real numbers $A_1, \ldots, A_n$ such that

$$P(x) = A_1(x-r)^{n-1} + A_2(x-r)^{n-2} + \cdots + A_{n-1}(x-r) + A_n.$$

Then we obtain the following result:

$$\begin{aligned}\int \frac{P(x)}{(x-r)^n} &= \int \frac{A_1}{x-r} + \frac{A_2}{(x-r)^2} + \cdots + \frac{A_n}{(x-r)^n}\,dx \\ &= A_1 \ln|x-r| - \frac{A_2}{x-r} - \frac{1}{2}\cdot\frac{A_3}{(x-r)^2} - \cdots - \frac{1}{n-1}\cdot\frac{A_n}{(x-r)^{n-1}} + C.\end{aligned}$$

Example 46.8. To integrate $\frac{1}{x(x+1)^3}$, we combine the ideas from (46.2) and Technique 46.7: we want to find real numbers A, B, C, D such that

$$\frac{1}{x(x+1)^3} = \frac{A}{x} + \frac{B}{x+1} + \frac{C}{(x+1)^2} + \frac{D}{(x+1)^3}.$$

Putting everything over a common denominator, we see that our goal is

$$\frac{1}{x(x+1)^3} = \frac{A(x+1)^3 + Bx(x+1)^2 + Cx(x+1) + Dx}{x(x+1)^3}.$$

This holds if and only if the numerators agree for all x, that is, if

$$1 = A(x+1)^3 + Bx(x+1)^2 + Cx(x+1) + Dx. \tag{46.3}$$

To find A, B, C, D, we can use either Technique 46.2 and compare coefficients, or Technique 46.4 and evaluate both sides at appropriate values of x.

Comparing coefficients: Expanding the RHS of (46.3) and comparing coefficients between the two sides, we obtain a system of four linear equations in four variables:

$$\begin{aligned}1 &= A(x^3 + 3x^2 + 3x + 1) + Bx(x^2 + 2x + 1) + C(x^2 + x) + Dx \\ &= (A+B)x^3 + (3A + 2B + C)x^2 + (3A + B + C + D)x + A,\end{aligned}$$

which yields

$$\begin{aligned} 0 &= A + B && \text{cubic coefficients} \\ 0 &= 3A + 2B + C && \text{quadratic coefficients} \\ 0 &= 3A + B + C + D && \text{linear coefficients} \\ 1 &= A && \text{constant coefficients.} \end{aligned}$$

The fourth equation gives $A = 1$, then the first gives $B = -A = -1$, then the second gives $C = -3A - 2B = -3 + 2 = -1$, then the third gives $D = -3A - B - C = -3 + 1 + 1 = -1$.

Evaluating at specific values: When $x = 0$, the RHS of (46.3) is equal to A, so we conclude that $A = 1$. Similarly, when $x = -1$, the RHS is equal to $-D$, and we conclude that $D = -1$. Because 0 and -1 are the only roots of $Q(x)$, it is not clear what two other values of x to use in order to find B and C. Choosing two values essentially at random would lead to a system of two equations in two variables, which could then be solved. Alternately, we can take one of the following two approaches.

Simplify and divide: Using $A = 1$ and $D = -1$, (46.3) can be rewritten as

$$1 = (x+1)^3 + Bx(x+1)^2 + Cx(x+1) - x; \tag{46.4}$$

adding $x - (x+1)^3$ to both sides gives

$$\begin{aligned} Bx(x+1)^2 + Cx(x+1) &= 1 + x - (x+1)^3 = (x+1)(1 - (x+1)^2) \\ &= (x+1)(1 - x^2 - 2x - 1) = -x(x+1)(x+2). \end{aligned}$$

Divide both sides by $x(x+1)$ to get

$$B(x+1) + C = -x - 2.$$

Again, recall that this is an equation that we want to be true for all x. Evaluating both sides at $x = -1$ gives $C = -(-1) - 2 = 1 - 2 = -1$, and thus we are left with

$$B(x+1) = -x - 2 - (-1) = -x - 1 = -(x+1),$$

so $B = -1$.

Differentiate: Another way to find B and C is to observe that the LHS and RHS of (46.4) are two different ways of writing the same function, so their derivatives must also be equal:

$$0 = 3(x+1)^2 + B(x+1)^2 + 2Bx(x+1) + C(x+1) + Cx - 1.$$

Evaluating this at $x = -1$ gives $C = -1$, and so the equation becomes

$$0 = 3(x+1)^2 + B(x+1)^2 + 2Bx(x+1) - 2(x+1).$$

Differentiating again gives

$$0 = 6(x+1) + 2B(x+1) + 2B(x+1) + 2Bx - 2,$$

and putting $x = -1$ gives $B = -1$.

Whichever of the above approaches we use, we conclude that $A = 1$ and $B = C = D = -1$, so

$$\int \frac{1}{x(x+1)^3}\,dx = \int \frac{1}{x} - \frac{1}{x+1} - \frac{1}{(x+1)^2} - \frac{1}{(x+1)^3}\,dx$$

$$= \ln|x| - \ln|x+1| + \frac{1}{x+1} + \frac{1}{2(x+1)^2} + C,$$

where this last C is a constant of integration (not the coefficient from the earlier computations).

At some level it is a matter of taste which approach we use to determine the coefficients in any given problem. However, it may be the case that one of the methods is easier than the others, and thus it is useful to be familiar with all of them. And indeed, there are other ways to proceed as well.

Example 46.9. Revisiting $\int \frac{1}{x(x+1)^3}\,dx$, suppose we start by only going partway in the partial fraction decomposition (here A, B, C, D are not the same as in the previous computations):

$$\frac{1}{x(x+1)^3} = \frac{A}{x} + \frac{Bx^2+Cx+D}{(x+1)^3}$$

Collecting the RHS over a common denominator we get

$$1 = A(x+1)^3 + Bx^3 + Cx^2 + Dx = (A+B)x^3 + (3A+C)x^2 + (3A+D)x + A,$$

so $A = 1$, $B = -1$, $C = -3$, and $D = -3$. Then we make the substitution $u = x+1$ and get

$$\begin{aligned}
\int \frac{1}{x(x+1)^3}\,dx &= \int \frac{1}{x} - \frac{x^2+3x+3}{(x+1)^3}\,dx = \ln|x| - \int \frac{(u-1)^2+3(u-1)+3}{u^3}\,du \\
&= \ln|x| - \int \frac{u^2+u+1}{u^3}\,du = \ln|x| - \int (u^{-1}+u^{-2}+u^{-3})\,du \\
&= \ln|x| - \ln|u| - u^{-1} - \frac{1}{2}u^{-2} + C \\
&= \ln|x| - \ln|x+1| - \frac{1}{x+1} - \frac{1}{2(x+1)^2} + C.
\end{aligned}$$

46.3. Irreducible quadratic factors

Now we have enough tools to integrate $\frac{P(x)}{Q(x)}$ whenever $Q(x)$ factors into linear terms. But what if it does not? The prototypical example of a polynomial that does not factor into linear terms is $Q(x) = x^2+1$, and we recall from our work on differentiating trigonometric functions that

$$\int \frac{1}{1+x^2}\,dx = \tan^{-1} x + C. \tag{46.5}$$

More generally, given $a > 0$ we can use the trigonometric substitution $x = a\tan\theta$, $dx = a\sec^2\theta\,d\theta$ to obtain

$$\int \frac{1}{x^2+a^2}\,dx = \int \frac{a\sec^2\theta}{a^2\tan^2\theta + a^2}\,d\theta = \int \frac{1}{a}\,d\theta = \frac{\theta}{a} + C = \frac{1}{a}\tan^{-1}\frac{x}{a} + C. \tag{46.6}$$

Technique 46.10. Any integral of the form $\int \frac{Ax+B}{x^2+bx+c}\,dx$, where x^2+bx+c is irreducible, can be evaluated by completing the square as $x^2+bx+c = (x-\frac{b}{2})^2 + (c-\frac{b^2}{4})$, making

the substitution $u = x - \frac{b}{2}$ so that the denominator becomes $u^2 + a^2$ for $a = \sqrt{c - \frac{b^2}{4}}$, and then using (46.6).

For integrals of the form $\int \frac{Ax+B}{(x^2+bx+c)^k}\,dx$, we can do the same substitution to obtain an integrand with denominator $(u^2 + a^2)^k$, and then use the substitution $u = a\tan\theta$ as above to obtain an integrand in terms of tangents and secants.

This turns out to be the final piece of the puzzle.

Technique 46.11 (Integrating rational functions by partial fractions)**.** Given *any* rational function $\frac{P(x)}{Q(x)}$, we can use the method of partial fractions to integrate it by going through the following steps.

(1) Use polynomial long division to reduce to the case when $\deg P < \deg Q$.
(2) Factor $Q(x)$ as a product of linear and quadratic terms, where the quadratic terms have no real roots.[37]
(3) Use any of the techniques described earlier to find coefficients that let us write $\frac{P(x)}{Q(x)}$ as a sum of expressions of one of the following forms:
$$\frac{A}{x-r}, \quad \frac{A}{(x-r)^k}, \quad \frac{Ax+B}{x^2+bx+c}, \quad \frac{Ax+B}{(x^2+bx+c)^k}.$$
(4) Integrate each of these individually:
   (a) $\displaystyle\int \frac{A}{x-r}\,dx = A\ln|x-r|$;
   (b) $\displaystyle\int \frac{A}{(x-r)^k} = \frac{A}{k-1}(x-r)^{-(k-1)}$ when $k \geq 2$;
   (c) For the last two types, use Technique 46.10: complete the square, make a u-substitution, and then either use (46.6) or make a further trigonometric substitution.

Example 46.12. To compute $\int \frac{1}{1+x^3}\,dx$, we factor the denominator as
$$1 + x^3 = (1 + x)(1 - x + x^2),$$
but we cannot go any further because $1 - x + x^2$ has no real roots (this follows from the quadratic formula since $(-1)^2 - 4(1)(1) = -3 < 0$). As in Example 46.9, though, we can write
$$\frac{1}{1+x^3} = \frac{A}{1+x} + \frac{Bx+C}{1-x+x^2},$$
and taking a common denominator we see that A, B, C must satisfy
$$1 = A(1 - x + x^2) + (Bx + C)(1 + x) \text{ for all } x.$$
As before, there are several ways to solve this. Putting $x = -1$ immediately gives $1 = A(1 - (-1) + 1) = 3A$, so $A = \frac{1}{3}$. Although $Q(x) = 1 + x^3$ has no other real roots, we can observe that the expressions obtained for $x = 0$ and $x = 1$ are not so complicated:
$$x = 0 \;\Rightarrow\; 1 = \frac{1}{3} + C \quad \text{and} \quad x = 1 \;\Rightarrow\; 1 = \frac{1}{3} + 2(B + C).$$

[37]The fact that this is always possible is called the *Fundamental Theorem of Algebra*; see Lecture 47.2.

Thus $C = \frac{2}{3}$ and $B + C = \frac{1}{3}$, so $B = -\frac{1}{3}$, and we have

$$\int \frac{1}{1+x^3}\,dx = \frac{1}{3}\int \frac{1}{1+x} + \frac{-x+2}{1-x+x^2}\,dx.$$

The first part integrates as $\frac{1}{3}\int \frac{1}{1+x}\,dx = \frac{1}{3}\ln|x+1|$, so it remains to integrate the last part. Completing the square gives $x^2 - x + 1 = (x - \frac{1}{2})^2 + \frac{3}{4} = u^2 + a^2$ for $u = x - \frac{1}{2}$ and $a = \sqrt{3}/2$, so we get

$$\begin{aligned}\int \frac{-x+2}{1-x+x^2}\,dx &= \int \frac{-(u+\frac{1}{2})+2}{u^2+a^2}\,du = -\frac{1}{2}\int \frac{2u}{u^2+a^2}\,du + \frac{3}{2}\int \frac{1}{u^2+a^2}\,du \\ &= -\frac{1}{2}\ln(u^2+a^2) + \frac{3}{2}\frac{1}{a}\tan^{-1}\frac{u}{a} \\ &= -\frac{1}{2}\ln(x^2-x+1) + \sqrt{3}\tan^{-1}\frac{x-\frac{1}{2}}{\sqrt{3}/2} + C.\end{aligned}$$

Putting it all together gives

$$\int \frac{1}{1+x^3}\,dx = \frac{1}{3}\ln|x+1| - \frac{1}{6}\ln(x^2-x+1) + \frac{\sqrt{3}}{3}\tan^{-1}\frac{2x-1}{\sqrt{3}} + C.$$

We give one more example to illustrate the procedure in the presence of a repeated quadratic factor.

Example 46.13. To compute $\int \frac{1-x}{x(x^2+1)^2}\,dx$, we first find $A, B, C, D, E \in \mathbb{R}$ such that

$$\frac{1-x}{x(x^2+1)^2} = \frac{A}{x} + \frac{Bx+C}{x^2+1} + \frac{Dx+E}{(x^2+1)^2}$$

for all x, which upon putting things over a common denominator is equivalent to

$$1 - x = A(x^2+1)^2 + (Bx+C)(x^2+1)x + (Dx+E)x.$$

Evaluating at $x = 0$ gives $A = 1$, so we must find B, C, D, E satisfying

$$(Bx+C)(x^2+1)x + (Dx+E)x = 1 - x - (1 + 2x^4 + x^4) = -x - 2x^2 - x^4.$$

Dividing both sides by x, we want

$$\begin{aligned}-1 - 2x - x^3 &= (Bx+C)(x^2+1) + (Dx+E) \\ &= Bx^3 + Cx^2 + (B+D)x + (C+E),\end{aligned}$$

and comparing coefficients gives

$$B = -1, \quad C = 0, \quad B + D = -2, \quad C + E = -1 \quad \Rightarrow \quad D = -1, \quad E = -1.$$

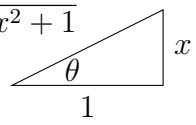
We conclude that

$$\begin{aligned}\int \frac{1-x}{x(x^2+1)^2}\,dx &= \int \frac{1}{x} - \frac{x}{x^2+1} - \frac{x+1}{(x^2+1)^2}\,dx \\ &= \ln|x| - \frac{1}{2}\ln|x^2+1| - \frac{1}{2}(x^2+1)^{-1} - \int \frac{1}{(x^2+1)^2}\,dx.\end{aligned}$$

To evaluate the final integral we use the substitution $x = \tan\theta$, $dx = \sec^2\theta\,d\theta$ and obtain

$$\int \frac{1}{(x^2+1)^2}\,dx = \int \frac{\sec^2\theta}{(\tan^2\theta+1)^2}\,d\theta = \int \frac{\sec^2\theta}{\sec^4\theta}\,d\theta = \int \cos^2\theta\,d\theta$$
$$= \frac{1}{2}\int (1+\cos 2\theta)\,d\theta = \frac{1}{2}\theta + \frac{1}{4}\sin 2\theta + C.$$

The triangle at right gives $\sin\theta = \frac{x}{\sqrt{x^2+1}}$ and $\cos\theta = \frac{1}{\sqrt{x^2+1}}$, so $\sin 2\theta = 2\sin\theta\cos\theta = \frac{2x}{x^2+1}$, and we get



$$\int \frac{1-x}{x(x^2+1)^2}\,dx = \ln|x| - \frac{1}{2}\ln(x^2+1) - \frac{1}{2}(x^2+1)^{-1} - \frac{1}{2}\tan^{-1}x + \frac{x}{2(x^2+1)} + C$$
$$= \ln|x| - \frac{1}{2}\ln(x^2+1) + \frac{x-1}{2(x^2+1)} - \frac{1}{2}\tan^{-1}x + C.$$

At this point we now have the ability to integrate *any* rational function, although the computations involved may be quite intimidating. For example, the partial fraction decomposition

$$\frac{x^3+x^2+1}{x(x-1)(x^2+x+1)(x^2+1)^2} = \frac{A}{x} + \frac{B}{x-1} + \frac{Cx+D}{x^2+x+1} + \frac{Ex+F}{x^2+1} + \frac{Gx+H}{(x^2+1)^2}$$

leads to a system of 8 linear equations in 8 variables, which would be tedious to solve by hand (though a computer would do it very quickly), and then we would be left with 8 separate integrals to compute, some immediate, others requiring appropriate substitutions.

Lecture 47 Why partial fraction decompositions work

Stewart §7.4, Spivak Ch. 19

Now we return to a question raised in the previous lecture: does the approach there always work? Certainly the techniques introduced there were enough to deal with all the examples we encountered so far, and it was strongly suggested that one can always find coefficients that make the partial fraction decompositions described work out the way they should; but why should this be true? How do we know that there is not some pathological rational function for which the partial fraction decomposition leads to a system of linear equations that has no solution?

47.1. Distinct linear factors

First let us consider the case of $\frac{P(x)}{Q(x)}$ where $Q(x)$ factors into linear terms. It is simplest to begin with the case $n = 2$, so $P(x) = ax+b$ and $Q(x) = (x-r_1)(x-r_2)$. Then (46.2) becomes

$$\frac{ax+b}{(x-r_1)(x-r_2)} = \frac{A_1}{x-r_1} + \frac{A_2}{x-r_2} = \frac{A_1(x-r_2)+A_2(x-r_1)}{(x-r_1)(x-r_2)}$$

and thus A_1, A_2 must satisfy

$$ax + b = A_1(x - r_2) + A_2(x - r_1) \text{ for every } x.$$

We could expand the RHS and compare coefficients, but it is easier to evaluate the above equation at $x = r_1$ and $x = r_2$, when it gives

$$ar_1 + b = A_1(r_1 - r_2) \quad \text{and} \quad ar_2 + b = A_2(r_2 - r_1).$$

If $r_1 \neq r_2$, then we can immediately solve for A_1 and A_2, and see that they are uniquely determined by a, b, r_1, r_2. On the other hand, we observe that $ar_i + b \neq 0$ for both $i = 1$ and $i = 2$ (since otherwise we could have simplified the expression $\frac{ax+b}{(x-r_1)(x-r_2)}$ by dividing top and bottom by $x - r_i$), and thus if $r_1 = r_2$ then there can be no solution A_1, A_2, since the right-hand sides vanish.

Proposition 47.1. *Suppose that $Q(x)$ factors as $Q(x) = (x - r_1)(x - r_2)\cdots(x - r_n)$, and $P(r_i) \neq 0$ for all i.*[38] *Then there are real numbers $A_1, \ldots, A_n$ satisfying (46.2) if and only if the roots $r_1, \ldots, r_n$ are all distinct. Morever, in this case there is exactly one solution: the values of $A_1, \ldots, A_n$ are uniquely determined by P, Q, and are all nonzero.*

Proof. Collecting the terms in the RHS of (46.2) over a common denominator, we get

$$P(x) = A_1(x - r_2)(x - r_3)\cdots(x - r_n) + A_2(x - r_1)(x - r_3)\cdots(x - r_n) + \cdots + A_n(x - r_1)\cdots(x - r_{n-1}),$$

where in each term we multiply A_j by the product of the factors $x - r_i$ taken over all $i \neq j$. In particular, the only term on the RHS that does *not* include a factor of $(x - r_1)$ is the first, and thus evaluating the above equation at $x = r_1$ gives

$$P(r_1) = A_1(r_1 - r_2)(r_1 - r_3)\cdots(r_1 - r_n).$$

Similarly, evaluating at $x = r_2$ gives

$$P(r_2) = A_2(r_2 - r_1)(r_2 - r_3)\cdots(r_2 - r_n).$$

Continuing in this way we get n equations, one for each A_i. If all the roots r_i are distinct, then each A_i is multiplied by a nonzero number to get $P(r_i)$, and we can solve for A_i to get the unique solution. On the other hand, if $r_i = r_j$ for some $i \neq j$, then we get $P(r_j) = A_j \cdot 0 = 0$, contradicting the assumption that $P(r_j) \neq 0$. Thus when the roots are not distinct, there is no solution. □

47.2. The Fundamental Theorem of Algebra

One step in the partial fraction decomposition described in Technique 46.11 was to factor the denominator as a product of linear and quadratic terms. But why is this always possible?

Theorem 47.2 (Fundamental Theorem of Algebra)**.** *Every nonconstant polynomial has a root in the complex numbers. That is, given $a_0, \ldots, a_n \in \mathbb{C}$ such that $a_n \neq 0$ and $n \geq 1$, and considering the polynomial $f(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n$, there exists $w \in \mathbb{C}$ such that $f(w) = 0$.*

[38] Again, the assumption on P is reasonable because otherwise $P(x)$ would have $(x - r_i)$ as a factor, and we could cancel this term from both P and Q.

Before outlining the proof, we observe two straightforward consequences that guarantee the factoring result we want.

Corollary 47.3. *Every nonconstant polynomial factors as a product of linear polynomials with complex coefficients. That is, given $f(z)$ as in Theorem 47.2, there exist $w_1, \dots, w_n \in \mathbb{C}$ such that $f(z) = a_n(z - w_1)(z - w_2)\cdots(z - w_n)$.*

Proof. Apply Theorem 47.2 to f to get w_1 such that $f(w_1) = 0$, and recall that this implies that $f(z) = (z - w_1)g(z)$ for some polynomial g with $\deg g = \deg f - 1$.[39] Then we apply Theorem 47.2 to g, and iterate the procedure until the desired factorization is reached. □

Corollary 47.4. *Let $f(x)$ be a nonconstant polynomial with real coefficients. Then f factors as a product of linear and quadratic polynomials with real coefficients.*

Proof. If $f(x)$ has real coefficients, then any $w \in \mathbb{C}$ with $f(w) = 0$ must have the property that $f(\overline{w}) = \overline{f(w)} = \overline{0} = 0$; its complex conjugate is also a root of f. Thus every factor in Corollary 47.3 is either a linear factor with real coefficients (if $w_i \in \mathbb{R}$), or has the property that it can be paired with another linear factor corresponding to its complex conjugate; multiplying the factors $(x - w)$ and $(x - \overline{w})$ together gives

$$(x - w)(x - \overline{w}) = x^2 - (w + \overline{w})x + w\overline{w} = x^2 - 2\,\mathrm{Re}(w)x + |w|^2,$$

which is a quadratic polynomial with real coefficients. □

Now we outline (one) proof of the Fundamental Theorem of Algebra; what follows is not a complete proof because it relies on certain concepts that take rather more work to make precise and rigorous, but it does at least convey the general idea.

Outline of proof of Theorem 47.2. Consider a polynomial $f(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n$, where $n \geq 1$ and $a_n \neq 0$. If $a_0 = 0$ then $f(0) = 0$ and we are done, so from now on we assume that $a_0 \neq 0$.

Given a real number $s > 0$, consider the following function:

$$\gamma_s \colon [0, 2\pi] \to \mathbb{C}, \qquad \gamma_s(t) = f(se^{it}).$$

This defines a continuous loop in $\mathbb{C}$, since $e^{2\pi i} = 1$ so $\gamma_s(2\pi) = f(s) = \gamma_s(0)$; this loop is the image under f of the circle $\{z \in \mathbb{C} : |z| = s\}$.

Figure 13 shows the loops γ_r and γ_R where $r > 0$ is small and R is large. (For illustration, we take $f(z) = 1 + i + z + z^3$.) Let us describe what "small" and "large" mean here.

For "r is small", first recall that we assumed $a_0 \neq 0$. Let k be the smallest natural number such that $a_k \neq 0$; then the small red dashed circle in the figure is the circle centred at a_0 with radius $a_k r^k$. One can take $r > 0$ small enough that this circle does not enclose the origin, and such that

$$|a_0 + a_k z^k| > |a_{k+1} z^{k+1} + \cdots + a_n z^n| \text{ for all } z \in \mathbb{C} \text{ with } |z| = r. \tag{47.1}$$

It follows that γ_r remains close enough to the small circle that it does not enclose the origin either.

[39] Indeed, polynomial long division gives $f(z) = (z - w_1)g(z) + h(z)$ for some polynomial h with $\deg h < \deg(z - w_1) = 1$, so $h(z)$ is constant, and putting $z = w_1$ we see that $h \equiv 0$.

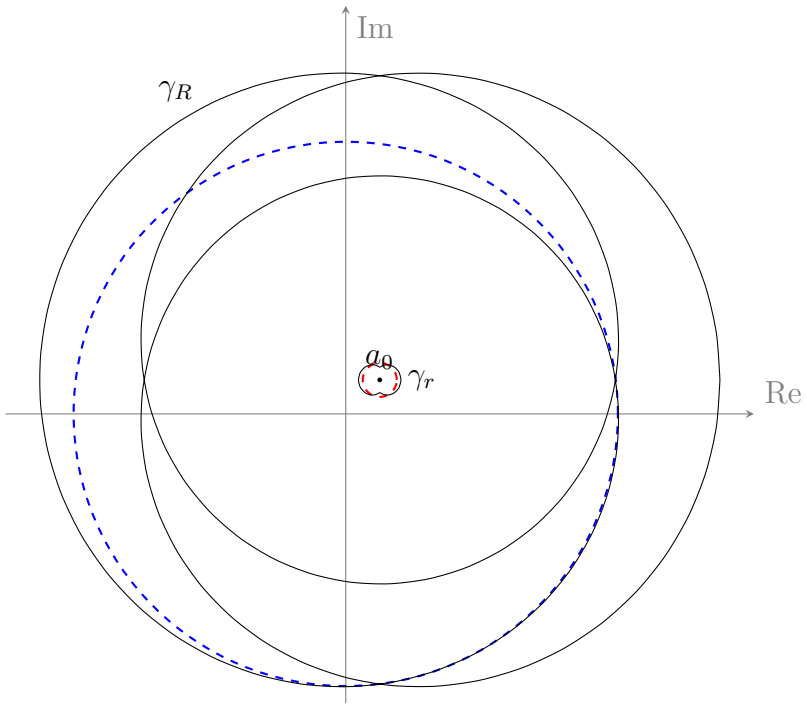


FIGURE 13. Proving the Fundamental Theorem of Algebra.

For "R is large", we consider the large blue dashed circle in Figure 13, which is the circle centered at the origin with radius $|a_n|R^n$. One can show that there is $R > 0$ large enough that

$$|a_n z^n| > |a_{n-1}z^n + \cdots a_k z^k + a_0| \text{ for all } z \in \mathbb{C} \text{ with } |z| = R. \tag{47.2}$$

Then as t goes from 0 to 2π, the curve $t \mapsto a_n(Re^{it})^n$ wraps around the blue circle n times, and thus (by the above inequality) the curve γ_R wraps around the origin n times.

Now consider the *family* of curves γ_s, where s varies from r to R. When s is close to r, the curve γ_s does not wrap around the origin, but when s is close to R, it does. Since the curve varies continuously in s, there must be some value of $s \in [r, R]$ for which γ_s passes *through* the origin (as it goes from not wrapping around at all, to wrapping around at least once). For this value of s there is $t \in [0, 2\pi]$ such that $f(se^{it}) = \gamma_s(t) = 0$, which gives the desired root. □

The outline just described gives the main ideas of the proof of the Fundamental Theorem of Algebra, but further work is required to make it a complete, precise, and rigorous proof. To accomplish this one must do the following.

- Give a careful proof of (47.1) and (47.2); this is something you should be able to do by now.
- Give a precise definition of what it means to say that "γ_r does not wrap around the origin" and "γ_R wraps around the origin n times"; this is the concept of *winding number* and requires some work to develop properly.
- Clarify the argument in the last paragraph for why the winding number cannot change unless the family of curves passes through the origin at some point.

47.3. The general case

Now we return to the question of proving that partial fractions can be used to integrate *any* rational function.

Definition 47.5. Given two polynomials $f(x)$ and $g(x)$, say that g is a *factor* of f if there exists a polynomial $h(x)$ such that $f(x) = g(x)h(x)$ for all x.

Definition 47.6. Say that two nonzero polynomials $f(x)$ and $g(x)$ are *coprime* if they have no common factors besides constant polynomials; that is, if any polynomial $h(x)$ that is a factor of both f and g has the property that h is constant.

Thanks to the Fundamental Theorem of Algebra (more precisely, Corollary 47.4), given any rational function $\frac{P(x)}{Q(x)}$, where P, Q are coprime, we can factor $Q(x)$ as[40]

$$Q(x) = F_1(x)^{k_1} \cdots F_n(x)^{k_n} \tag{47.3}$$

where each F_i is either a linear or irreducible quadratic polynomial in x, and all the F_i's are distinct (hence coprime). Our goal is to prove the following.

Theorem 47.7. *Given P, Q as above, the rational function $\frac{P(x)}{Q(x)}$ can be written as a sum of terms of the form*

$$\frac{A(x)}{F_i(x)^j} \text{ for some } 1 \le j \le k_i, \text{ with } \deg(A) < \deg(F_i). \tag{47.4}$$

Since we know how to integrate each expression in (47.4) individually, Theorem 47.7 gives a general method for integrating arbitrary rational functions. The theorem will be a consequence of the following two results.

Theorem 47.8. *Suppose that $P(x)$ and $Q(x)$ are nonzero coprime polynomials, and that $Q(x)$ factors as $Q(x) = R(x)S(x)$, where R, S are also coprime polynomials. Then there exist polynomials $A(x), B(x)$ such that $\deg(A) < \deg(R)$, $\deg(B) < \deg(S)$, and*

$$\frac{P(x)}{Q(x)} = \frac{P(x)}{R(x)S(x)} = \frac{A(x)}{R(x)} + \frac{B(x)}{S(x)} \text{ for all } x. \tag{47.5}$$

Theorem 47.9. *If $A(x)$ and $R(x)$ are coprime polynomials such that $\deg(A) < \deg(R)$ and $R(x) = T(x)^n$ for some polynomial $T(x)$ and natural number $n \in \mathbb{N}$, then there are polynomials $A_1(x), \ldots, A_n(x)$ such that $\deg A_i < \deg T$ and*

$$\frac{A(x)}{R(x)} = \frac{A(x)}{T(x)^n} = \frac{A_1(x)}{T(x)} + \frac{A_2(x)}{T(x)^2} + \cdots + \frac{A_n(x)}{T(x)^n} \text{ for all } x. \tag{47.6}$$

Before proving Theorems 47.8 and 47.9, we explain why they imply Theorem 47.7. Given P, Q as in that theorem, we can apply Theorem 47.8 with $R(x) = F_1(x)^{k_1}$ and $S(x) = F_2(x)^{k_2} \cdots F_n(x)^{k_n}$ to write

$$\frac{P(x)}{Q(x)} = \frac{A(x)}{F_1(x)^{k_1}} + \frac{B(x)}{F_2(x)^{k_2} \cdots F_n(x)^{k_n}}$$

[40]There is a subtlety here that we are glossing over. The FTA guarantees the *existence* of such a factorization, but does not provide a procedure for actually finding formulas for the factors in terms of the coefficients of $Q(x)$. Indeed, one can prove using Galois theory that no such formula exists in general, even though there are good numerical methods to compute the factors to arbitrary precision.

for some polynomials $A(x), B(x)$ such that in each expression, the numerator has smaller degree than the denominator. Then we can apply Theorem 47.9 to write $\frac{A(x)}{F_1(x)^{k_1}}$ as a sum of terms of the form (47.4). Iterating this process lets us peel off the coprime factors in (47.3) one at a time, until all have been dealt with and Theorem 47.7 is proved.

Now we prepare for the proofs of Theorems 47.8 and 47.9, both of which rely on polynomial long division. Before attacking Theorem 47.8 itself, it is perhaps useful to consider an analogous problem with "polynomial" replaced by "integer", so that we deal with rational numbers instead of rational functions. Suppose we want to write $\frac{61}{115}$ as a sum of fractions with prime denominators. Since $115 = 5 \cdot 23$ this means writing

$$\frac{61}{115} = \frac{A}{5} + \frac{B}{23} \text{ for some integers } A, B.$$

Putting the right-hand side over a common denominator, we see that this is equivalent to finding integers A, B that solve $23A + 5B = 61$. Ideally we would like $0 \le A < 5$ and $0 \le B < 23$. Since 5 and 23 are coprime, we can use the Euclidean algorithm to find integers m, n such that $23m + 5n = 1$: first perform iterated divisions to get

$$\begin{aligned} 23 &= 4 \cdot 5 + 3, & 3 &= 23 - 4 \cdot 5, \\ 5 &= 1 \cdot 3 + 2, & 2 &= 5 - 1 \cdot 3, \\ 3 &= 1 \cdot 2 + 1, & 1 &= 3 - 1 \cdot 2, \end{aligned}$$

then combine the equations at the end of each line to get

$$1 = 3 - 2 = 3 - (5 - 3) = 2 \cdot 3 - 5 = 2 \cdot (23 - 4 \cdot 5) - 5 = 2 \cdot 23 - 9 \cdot 5.$$

Multiplying through by 61 gives

$$61 = 122 \cdot 23 - 549 \cdot 5.$$

Of course 122 and 549 do not lie in the ranges we wanted A, B to lie in. But we can observe that $122 = 24 \cdot 5 + 2$, and $549 = 24 \cdot 23 - 3$, so that

$$\frac{61}{115} = \frac{122}{5} - \frac{549}{23} = \left(24 + \frac{2}{5}\right) - \left(24 - \frac{3}{23}\right) = \frac{2}{5} + \frac{3}{23}.$$

The proof of Theorem 47.8 will follow a similar procedure, but with polynomials in place of integers. We start with the following lemma.

Lemma 47.10 (Euclidean algorithm for polynomials)**.** *If $f(x), g(x)$ are nonzero coprime polynomials, then there are polynomials $U(x)$ and $V(x)$ such that $U(x)f(x)+V(x)g(x) = 1$ for all x.*

Proof. The proof will go by induction in $\deg(f) + \deg(g)$. For the base case of the induction, suppose that one of the polynomials f, g is constant. If there is a real number $c \neq 0$ such that $f(x) = c$ for all x, then we can take $U(x) = 1/c$ and $V(x) = 0$; the case when g is constant is similar.

Now suppose that both f, g are nonconstant – that is, both have degree ≥ 1. Let $n = \deg(f) + \deg(g)$, and suppose (for our inductive hypothesis) that the result has been proved for all pairs whose sum of degrees is $< n$. Without loss of generality assume that $\deg(f) \le \deg(g)$ (otherwise just reverse the roles and continue with the proof). Then polynomial long division gives polynomials $h(x)$ and $k(x)$ such that

$$g(x) = h(x)f(x) + k(x) \text{ for all } x, \text{ and } \deg(k) < \deg(f) \le \deg(g). \tag{47.7}$$

We will apply the inductive hypothesis to the pair of polynomials $f(x), k(x)$. To see that we can do this, observe first that $\deg(f) + \deg(k) < \deg(f) + \deg(g)$, so the sum of the degrees is smaller, and to apply the inductive hypothesis it suffices to show that $f(x), k(x)$ are nonzero and coprime.

To check that $k(x)$ is not the zero polynomial we observe that if $k \equiv 0$ then $g = hf$ so f is a factor of g, hence it is a common factor of both f and g, but it is not constant, contradicting the assumption that f, g are coprime. Thus k is nonzero.

To see that f, k are coprime, suppose $p(x)$ is a polynomial that is a factor of both f and k; then there are polynomials $q_1(x)$ and $q_2(x)$ such that $f(x) = p(x)q_1(x)$ and $k(x) = p(x)q_2(x)$, and (47.7) gives

$$g(x) = h(x)p(x)q_1(x) + p(x)q_2(x) = p(x)(h(x)q_1(x) + q_2(x)),$$

so $p(x)$ is a factor of both f and g. By the assumption that f and g are coprime, this implies that p is a constant polynomial. Since p was an arbitrary common factor of f, k, this implies that f and k are coprime.

We can now apply the inductive hypothesis to deduce that there are polynomials $U(x)$ and $V(x)$ such that $U(x)f(x) + V(x)k(x) = 1$. Using (47.7), this gives

$$1 = U(x)f(x) + V(x)\big(g(x) - h(x)f(x)\big) = \big(U(x) - V(x)h(x)\big)f(x) + V(x)g(x),$$

which proves the result for f and g. By induction, we are done. □

Proof of Theorem 47.8. Applying Lemma 47.10 to $R(x)$ and $S(x)$, there are polynomials $U(x), V(x)$ such that $1 = R(x)U(x) + S(x)V(x)$ for all x. Now let $f(x) = U(x)P(x)$ and $g(x) = V(x)P(x)$, so that

$$R(x)f(x) + S(x)g(x) = P(x) \quad \Rightarrow \quad \frac{P(x)}{Q(x)} = \frac{g(x)}{R(x)} + \frac{f(x)}{S(x)}. \tag{47.8}$$

We cannot immediately use f, g to produce a solution to (47.5), since they probably have large degree. However, we can use polynomial long division to divide $g(x)$ by $R(x)$ and obtain

$$g(x) = h(x)R(x) + A(x) \text{ for some polynomials } h(x), A(x) \text{ with } \deg(A) < \deg(R).$$

Together with the first part of (47.8), this gives

$$P = Rf + Sg = Rf + S(hR + A) = (f + Sh)R + AS,$$

where we omit the argument x to make the equation more readable. Subtracting AS from both sides gives

$$P - AS = (f + Sh)R.$$

Now we observe that $\deg(P) < \deg(RS)$ (by assumption on P, Q) and $\deg(AS) < \deg(RS)$ since $\deg(A) < \deg(R)$, so the left-hand side has degree $< \deg(RS)$. Thus the right-hand side does as well, and this implies that $\deg(f + Sh) < \deg(S)$. In particular, writing $B(x) = f(x) + S(x)h(x)$, we get $\deg(B) < \deg(S)$ and $P(x) = A(x)S(x) + B(x)R(x)$, which completes the proof of Theorem 47.8. □

To conclude our discussion of partial fractions, it remains only to prove Theorem 47.9.

Proof of Theorem 47.9. Putting the right-hand side of (47.6) over a common denominator, we see that we must find polynomials $A_1(x), \dots, A_n(x)$ such that $\deg A_i < \deg T$ and

$$A(x) = A_1(x)T(x)^{n-1} + A_2(x)T(x)^{n-2} + \cdots + A_{n-1}(x)T(x) + A_n(x). \tag{47.9}$$

Once again, we rely on polynomial long division. Dividing $A(x)$ by $T(x)^{n-1}$ gives polynomials $A_1(x), R_1(x)$ such that

$$A(x) = A_1(x)T(x)^{n-1} + R_1(x) \text{ for all } x, \text{ and } \deg(R_1) < \deg(T^{n-1}) = (n-1)\deg(T).$$

Observe that $\deg A_1 + (n-1)\deg T = \deg A < n \deg T$, so $\deg A_1 < \deg T$. Now we can iterate this procedure; divide $R_1(x)$ by $T(x)^{n-2}$ to get $A_2(x), R_2(x)$ such that

$$R_1(x) = A_2(x)T(x)^{n-2} + R_2(x), \quad \deg(R_2) < (n-2)\deg(T).$$

Note that combining these last two formulas gives

$$A(x) = A_1(x)T(x)^{n-1} + A_2(x)T(x)^{n-2} + R_2(x),$$

and that $\deg A_2 < \deg T$ for the same reason as with $\deg A_1$. Continuing in this manner we eventually get

$$A(x) = A_1(x)T(x)^{n-1} + A_2(x)T(x)^{n-2} + \cdots + A_{n-1}(x)T(x) + R_{n-1}(x),$$

with $\deg A_i < \deg T$ for all i, and $\deg R_{n-1} < \deg T$. Putting $A_n(x) = R_{n-1}(x)$ completes the proof. □

Lecture 48 Numerical integration

Stewart §7.7, Spivak Ch. 19

48.1. Endpoint and midpoint rules

When we are tasked with evaluating a definite integral $\int_a^b f(x)\,dx$, our usual approach is to find an antiderivative $F(x) = \int f(x)\,dx$ and then apply the FTC to get $\int_a^b f(x)\,dx = F(b) - F(a)$. However, as we have mentioned several times already, we may not be able to find a formula for an antiderivative, even if we know the formula for f. And it may be the case that we do not even know the formula for f, for example if the function is known only experimentally. In such cases we turn to a different approach to computing definite integrals, which goes back to their original definition via Riemann sums.

Recall that given $n \in \mathbb{N}$, we can partition the interval $[a, b]$ into n subintervals $[x_{i-1}, x_i]$ for $i = 1, \dots, n$, where $x_i = a + i\Delta x$, and $\Delta x = \frac{b-a}{n}$ is the width of each subinterval. Then upon selecting a point x_i^* inside each subinterval $[x_{i-1}, x_i]$, the corresponding Riemann sum is

$$\sum_{i=1}^{n} f(x_i^*)\Delta x.$$

There are three natural choices to make for x_i^*: we might choose the left endpoint, the right endpoint, or the midpoint of $[x_{i-1}, x_i]$. Using left endpoints gives the *left endpoint approximation*

$$L_n = \sum_{i=1}^{n} f(x_{i-1})\Delta x,$$

and similarly, the *right endpoint approximation* is

$$R_n = \sum_{i=1}^{n} f(x_i)\Delta x.$$

Choosing $x_i^* = \bar{x}_i := \frac{1}{2}(x_{i-1} + x_i)$ gives the *midpoint approximation*

$$M_n = \sum_{i=1}^{n} f(\bar{x}_i)\Delta x.$$

It follows from the general theory of integration that all three approximations converge to $\int_a^b f(x)\,dx$ as $n \to \infty$; however, we are also interested in the *speed* of approximation. Indeed, if you have an application that requires a numerical answer precise to within 10^{-4}, then it is necessary to know how large n must be in order to guarantee this degree of precision. We state the following theorem without proof.

Theorem 48.1. *If $f\colon [a,b] \to \mathbb{R}$ is twice differentiable and $K \in \mathbb{R}$ has the property that $|f''(x)| \le K$ for all $x \in [a,b]$, then the error term in the midpoint approximation can be bounded as follows:*

$$\left| M_n - \int_a^b f(x)\,dx \right| \le \frac{K(b-a)^3}{24n^2}.$$

For simplicity's sake, suppose that a, b, K have the property that $K(b-a)^3/24 = 1$. Then the error bound is n^{-2}, and so to guarantee precision of 10^{-4} using the midpoint rule, we would need to take $n = 10^2 = 100$. It turns out that the corresponding error estimate for the left and right endpoint approximations has a factor of n, not n^2, in the denominator, and to get $n^{-1} = 10^{-4}$ requires $n = 10^4$; this illustrates that to get an estimate with a very small error bound, it is useful to use the more efficient midpoint approximation.

48.2. Trapezoid rule

There are two more methods that are worth mentioning here; both involve replacing the rectangles used in Riemann sums with more general shapes.

For the *trapezoid rule*, instead of using a rectangle with height $f(x_i^*)$ for some $x_i^* \in [x_{i-1}, x_i]$, we use a trapezoid with two vertices on the x-axis, at $(x_{i-1}, 0)$ and $(x_i, 0)$, and the other two vertices on the graph of the function, at $(x_{i-1}, f(x_{i-1}))$ and $(x_i, f(x_i))$. The area of this trapezoid is

$$\text{average height} \times \text{base} = \frac{f(x_{i-1}) + f(x_i)}{2} \cdot \Delta x,$$

and adding up the areas of the n trapezoids over the intervals $[x_{i-1}, x_i]$ for $i = 1, \dots, n$, we get the following approximation for $\int_a^b f(x)\,dx$:

$$T_n := \sum_{i=1}^{n} \frac{f(x_{i-1}) + f(x_i)}{2} \cdot \Delta x, \qquad \Delta x := \frac{b-a}{n}, \quad x_i := a + i\Delta x. \tag{48.1}$$

We can rewrite this as

$$T_n = \frac{\Delta x}{2}\big(f(x_0) + 2f(x_1) + 2f(x_2) + \cdots + 2f(x_{n-1}) + f(x_n)\big).$$

Theorem 48.1 has an analogue for the trapezoid rule, except that the 24 in the denominator is replaced by 12; the trapezoid rule is actually not quite as good as the midpoint rule in general.

48.3. Simpson's rule

Another way to interpret the trapezoid rule is that on each interval $[x_{i-1}, x_i]$, we replaced the function f with a linear function $g \approx f$ that agrees with f at the endpoints, and then integrated g instead of f. (Of course, we use a different function g on each small interval $[x_{i-1}, x_i]$.) To get a better approximation, we might try using a quadratic function instead. Recall that a quadratic function is determined by its values at three points, so now we should ask for g to agree with f at the endpoints *and* at the midpoint. Notationally, it will be easier to assume that n is even and then approximate by quadratics on $[x_0, x_2]$, $[x_2, x_4]$, and so on. The key computation is contained in the following lemma.

Lemma 48.2. *Suppose we are given $h > 0$, three points $x_0 < x_1 < x_2$ related by $x_1 = x_0 + h$ and $x_2 = x_1 + h$, and three values $y_0, y_1, y_2 \in \mathbb{R}$. Let $g(x)$ be the unique quadratic polynomial $g(x)$ such that $g(x_i) = y_i$ for $i = 0, 1, 2$. Then*

$$\int_{x_0}^{x_2} g(x)\,dx = \frac{h}{3}(y_0 + 4y_1 + y_2). \tag{48.2}$$

Proof. Without loss of generality we can assume that $x_0 = -h$, $x_1 = 0$, and $x_2 = h$, since translating the graph horizontally does not change the area underneath it. Since g is a quadratic polynomial, we must have $A, B, C \in \mathbb{R}$ such that $g(x) = Ax^2 + Bx + C$. Evaluating this at $x = -h, 0, h$ and using the fact that $g(x_i) = y_i$ for $i = 0, 1, 2$, we get

$$\begin{aligned} y_0 &= Ah^2 - Bh + C, \\ y_1 &= C, \\ y_2 &= Ah^2 + Bh + C. \end{aligned}$$

Adding the first and the third equations gives $y_0 + y_2 = 2Ah^2 + 2C$, and the second equation gives $C = y_1$, so we can evaluate the integral as

$$\begin{aligned} \int_{-h}^{h} (Ax^2 + Bx + C)\,dx &= \Big[\frac{A}{3}x^3 + \frac{B}{2}x^2 + Cx\Big]_{-h}^{h} = \frac{2Ah^3}{3} + 2Ch = \frac{h}{3}(2Ah^2 + 6C) \\ &= \frac{h}{3}(y_0 + y_2 + 4C) = \frac{h}{3}(y_0 + y_2 + 4y_1), \end{aligned}$$

which proves the lemma. □

Now return to the question of approximating $\int_a^b f(x)\,dx$. Given $n \in \mathbb{N}$ even and $x_i = a + i\Delta x$, where $\Delta x = (b-a)/n$, let $y_i = f(x_i)$. When we add up the areas under the parabolas over $[x_0, x_2]$, $[x_2, x_4]$, and so on, we obtain the following approximation for $\int_a^b f(x)\,dx$, known as *Simpson's rule*:

$$\begin{aligned} S_n &= \frac{\Delta x}{3}(y_0 + 4y_1 + y_2) + \frac{\Delta x}{3}(y_2 + 4y_3 + y_4) + \cdots + \frac{\Delta x}{3}(y_{n-2} + 4y_{n-1} + y_n) \\ &= \frac{\Delta x}{3}(y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + \cdots + 2y_{n-2} + 4y_{n-1} + y_n). \end{aligned} \tag{48.3}$$

For Simpson's rule we have the following improvement on Theorem 48.1, which again we do not prove here.

Theorem 48.3. *If f is four times differentiable on $[a,b]$ and $K \in \mathbb{R}$ is such that $|f^{(4)}(x)| \le K$ for all $x \in [a,b]$, then the error term in Simpson's rule can be bounded as follows:*

$$\left| S_n - \int_a^b f(x)\,dx \right| \le \frac{K(b-a)^5}{180n^4}.$$

The factor of n^4 in the denominator means that we can obtain a very precise approximation with a relatively small value of n, which makes this approximation very useful and explains why we may consider it an improvement over the approximations described earlier.

Lecture 49 — Improper integrals

Stewart §7.8, Spivak exercises 14.25–30

49.1. Infinite width

We know that if $f\colon [a,b] \to (0,\infty)$ is a positive function, then $\int_a^b f(x)\,dx$ represents the area underneath the graph of $y = f(x)$ over the bounded interval $[a,b]$. But what if we consider the area under the graph over an *unbounded* interval? Can we still make sense of this notion?

Start with a concrete example: consider the region beneath the graph of $y = \frac{1}{x^2}$, above the x-axis, and to the right of the line $x = 1$. If we truncate this region by cutting off everything to the right of the line $x = t$ for some fixed $t > 1$, then the truncated region has area

$$A(t) = \int_1^t \frac{1}{x^2}\,dx = -\frac{1}{x}\Big|_1^t = 1 - \frac{1}{t}.$$

Viewing the truncated region as an approximation to the region we are interested in, we see that the approximation gets better the larger t gets, and that $\lim_{t\to\infty} A(t) = 1$, so it seems reasonable to say that the region originally described has area 1, and to write $\int_1^\infty \frac{1}{x^2}\,dx = 1$. This serves as a template for a more general definition.

Definition 49.1. Let $f\colon [a,\infty) \to \mathbb{R}$ be such that

(1) f is integrable on $[a,t]$ for every $t \geq a$, and
(2) $\lim_{t\to\infty} \int_a^t f(x)\,dx$ exists and is finite.

Then we write $\int_a^\infty f(x)\,dx := \lim_{t\to\infty} \int_a^t f(x)\,dx$, and call this an *improper integral (of type 1)*; in this case we say that the improper integral is *convergent*. If the limit does not exist, we say that it is *divergent*.

The improper integral $\int_{-\infty}^b f(x)\,dx$ is defined similarly when $f\colon (-\infty, b] \to \mathbb{R}$ is integrable on every $[t,b]$, provided the limit $\lim_{t\to-\infty} \int_t^b f(x)\,dx$ exists and is finite.

Finally, if both $\int_{-\infty}^a f(x)\,dx$ and $\int_a^\infty f(x)\,dx$ are convergent, then we write $\int_{-\infty}^\infty f(x)\,dx = \int_{-\infty}^a f(x)\,dx + \int_a^\infty f(x)\,dx$.

Exercise 49.2. Prove that in the last part of Definition 49.1, it does not matter what value of a we choose: if the two improper integrals are convergent for some value of a, then they are convergent for any other value of a, and their sum has the same value.

In the case when $f \geq 0$, we can interpret an improper integral as an area, just as with more familiar definite integrals.

Example 49.3. $\lim\limits_{t\to\infty} \int_1^t \dfrac{1}{x}\,dx = \lim\limits_{t\to\infty} \big[\ln x\big]_1^t = \lim\limits_{t\to\infty} \ln t = \infty$, so the improper integral $\int_1^\infty \frac{1}{x}\,dx$ is divergent.

Example 49.4.

$$\begin{aligned}\int_{-\infty}^0 xe^x\,dx &= \lim_{t\to-\infty} \int_t^0 \underbrace{x}_{u}\underbrace{e^x\,dx}_{dv} = \lim_{t\to-\infty} xe^x\Big|_t^0 - \int_t^0 e^x\,dx = \lim_{t\to-\infty} \Big[xe^x - e^x\Big]_t^0 \\ &= \lim_{t\to-\infty} 0e^0 - e^0 - te^t + e^t = -1,\end{aligned}$$

so the improper integral is convergent.

Example 49.5. To evaluate $\int_{-\infty}^\infty \frac{1}{1+x^2}\,dx$, we first compute

$$\int_0^\infty \frac{1}{1+x^2}\,dx = \lim_{t\to\infty} \int_0^t \frac{1}{1+x^2}\,dx = \lim_{t\to\infty} \Big[\tan^{-1} x\Big]_0^t = \lim_{t\to\infty} \tan^{-1} t = \frac{\pi}{2},$$

and a similar computation gives $\int_{-\infty}^0 \frac{1}{1+x^2}\,dx = \frac{\pi}{2}$, so

$$\int_{-\infty}^\infty \frac{1}{1+x^2}\,dx = \int_{-\infty}^0 \frac{1}{1+x^2}\,dx + \int_0^\infty \frac{1}{1+x^2}\,dx = \frac{\pi}{2} + \frac{\pi}{2} = \pi.$$

Example 49.6. Suppose we fix a positive real number $p > 0$ and consider the improper integral $\int_1^\infty \frac{1}{x^p}\,dx$. For which values of p is this integral convergent? Note that Example 49.3 showed that it is divergent when $p = 1$. For $p \neq 1$, we have

$$\lim_{t\to\infty} \int_1^t \frac{1}{x^p}\,dx = \lim_{t\to\infty} \Big[\frac{x^{1-p}}{1-p}\Big]_1^t = \lim_{t\to\infty} \frac{t^{1-p}-1}{1-p} = \begin{cases}\infty & \text{if } p < 1, \\ \frac{1}{p-1} & \text{if } p > 1.\end{cases}$$

Thus $\int_1^\infty \frac{1}{x^p}\,dx$ is convergent if $p > 1$, and divergent if $p \leq 1$.

49.2. Infinite height

Another type of improper integral arises from vertical asymptotes. Recall that our original definition of the definite integral $\int_a^b f(x)\,dx$ required the function f to be bounded on $[a,b]$ (as well as some other requirements). If f has a vertical asymptote at one of the endpoints, then we can define $\int_a^b f(x)\,dx$ as an improper integral by using a limiting procedure similar to the one in the previous section.

Definition 49.7. Let $f\colon [a,b) \to \mathbb{R}$ be continuous and suppose that

(1) $\lim_{x\to b^-} f(x)$ does not exist, and
(2) $\lim_{t\to b^-} \int_a^t f(x)\,dx$ exists and is finite.

Then we write $\int_a^b f(x)\,dx := \lim_{t\to b^-} \int_a^t f(x)\,dx$ for the corresponding *improper integral (of type 2)*, which we call *convergent*. If the limit does not exist, we say that the improper integral is *divergent*.

Similarly, if f is continuous everywhere on $[a,b]$ except for the left endpoint a, then we write $\int_a^b f(x)\,dx = \lim_{t\to a^-} \int_t^b f(x)\,dx$ provided the limit exists.

Finally, if f is continuous everywhere on $[a,b]$ except for some point $c \in (a,b)$, then we write $\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$ *provided both improper integrals converge.*

Exercise 49.8. Prove that if f is continuous on all of $[a,b]$, then all of the definitions above agree with the usual definition of $\int_a^b f(x)\,dx$.

Example 49.9. $f(x) = \frac{1}{\sqrt{x-2}}$ has a vertical asymptote at $x = 2$, so

$$\int_2^5 \frac{1}{\sqrt{x-2}}\,dx = \lim_{t\to 2^+} \int_t^5 \frac{dx}{\sqrt{x-2}} = \lim_{t\to 2^+} \Big[2\sqrt{x-2}\Big]_t^5 = \lim_{t\to 2^+} 2\sqrt{3} - 2\sqrt{t-2} = 2\sqrt{3},$$

and the improper integral converges.

Example 49.10. $\sec x$ has a vertical asymptote at $x = \pi/2$, so

$$\int_0^{\frac{\pi}{2}} \sec x\,dx = \lim_{t\to \frac{\pi}{2}^-} \int_0^t \sec x\,dx = \lim_{t\to \frac{\pi}{2}^-} \Big[\ln|\sec x + \tan x|\Big]_0^t = \lim_{t\to \frac{\pi}{2}^-} \ln|\sec t + \tan t| = \infty,$$

and the improper integral is divergent.

Example 49.11. To evaluate $\int_0^3 \frac{dx}{x-1}$, observe that $\frac{1}{x-1}$ has a vertical asymptote at $x = 1$, so we need to independently evaluate $\int_0^1 \frac{dx}{x-1}$ and $\int_1^3 \frac{dx}{x-1}$. The first of these is

$$\int_0^1 \frac{dx}{x-1} = \lim_{t\to 1^-} \int_0^t \frac{dx}{x-1} = \lim_{t\to 1^-} \Big[\ln|x-1|\Big]_0^t = \lim_{t\to 1^-} \ln|t-1| = -\infty,$$

and thus $\int_0^3 \frac{dx}{x-1}$ is divergent.

Remark 49.12. The previous example shows the need for caution when applying the FTC. It would be all too easy to unthinkingly push symbols around and write $\int_0^3 \frac{dx}{x-1} = [\ln|x-1|]_0^3 = \ln 2$, but this is wrong. Observe that the FTC does not apply here, because it requires the function to be integrable (and in particular, bounded) on the entire interval.

Example 49.13.

$$\int_0^1 \ln x\,dx = \lim_{t\to 0^+} [x\ln x - x]_t^1 = \lim_{t\to 0^+} (-1 - t\ln t + t) = -1,$$

where we use the fact that $\lim_{t\to 0^+} t\ln t = 0$. Thus the improper integral is convergent.

49.3. Comparison theorems

Sometimes determining whether or not an improper integral is convergent is significantly easier than establishing its numerical value.

Theorem 49.14. *Suppose $f, g\colon [a,\infty) \to \mathbb{R}$ are continuous and satisfy $f(x) \geq g(x) \geq 0$ for all $x \geq a$. Then the following are true.*

(1) If $\int_a^\infty f(x)\,dx$ is convergent, then $\int_a^\infty g(x)\,dx$ is convergent.
(2) If $\int_a^\infty g(x)\,dx$ is divergent, then $\int_a^\infty f(x)\,dx$ is divergent.

Proof. We start by proving the first claim. Suppose that $\int_a^\infty f(x)\,dx$ is convergent, and let $G(t) := \int_a^t g(x)\,dx$. For every $t > a$, we have

$$G(t) = \int_a^t g(x)\,dx \leq \int_a^t f(x)\,dx \leq \int_a^\infty f(x)\,dx,$$

where the first inequality uses $g \leq f$ and properties of integrals, and the second inequality uses the fact that $f \geq 0$. Thus the function G is bounded above. Moreover, for every $t_1 \leq t_2 > a$ we have

$$G(t_2) = \int_a^{t_2} g(x)\,dx = \int_a^{t_1} g(x)\,dx + \int_{t_1}^{t_2} g(x)\,dx = G(t_1) + \int_{t_1}^{t_2} g(x)\,dx \geq G(t_1),$$

where the last inequality uses the fact that $g \geq 0$. Thus G is a nondecreasing function. By the monotone convergence theorem, $\lim_{t\to\infty} G(t)$ exists (and is equal to $\sup\{G(t) : t > a\}$). This means that $\int_a^\infty g(x)\,dx$ is convergent.

The second claim in the theorem is equivalent to the first one, so this proves the theorem. □

Example 49.15. To determine convergence of $\int_0^\infty e^{-x^2}\,dx$, we observe that $x^2 \geq x$ for all $x \geq 1$, and thus $e^{-x^2} \leq e^{-x}$ for all $x \geq 1$. Since $\int_1^\infty e^{-x}\,dx$ is convergent, Theorem 49.14 implies that $\int_1^\infty e^{-x^2}\,dx$ is convergent as well. This in turn implies that $\int_0^\infty e^{-x^2}\,dx$ is convergent, because

$$\begin{aligned}\int_0^\infty e^{-x^2}\,dx = \lim_{t\to\infty}\int_0^t e^{-x^2}\,dx &= \lim_{t\to\infty}\int_0^1 e^{-x^2}\,dx + \int_1^t e^{-x^2}\,dx \\ &= \int_0^1 e^{-x^2}\,dx + \lim_{t\to\infty}\int_1^t e^{-x^2}\,dx = \int_0^1 e^{-x^2}\,dx + \int_1^\infty e^{-x^2}\,dx.\end{aligned}$$

Remark 49.16. In fact, using more sophisticated techniques it is possible to show that $\int_0^\infty e^{-x^2}\,dx = \sqrt{\pi}/2$, but this requires tools that we have not yet developed.

49.4. *Cauchy Principal Value integral

Let us return to the world of Example 49.11 for a moment.[41] We declared the integral $\int_0^3 \frac{1}{x-1}\,dx$ divergent because $\int_0^1 \frac{1}{x-1}\,dx = -\infty$. But someone looking at the picture might argue that the negative part of the graph here should exactly cancel with the positive part of the graph from $x = 1$ to $x = 2$, leaving us with a finite integral on the interval $[0, 3]$.

Consider the similar example $\int_{-1}^1 \frac{1}{x}\,dx$; the graph is symmetric around the origin, and so one may argue that the integral should be 0, despite the fact that $\int_{-1}^0 \frac{1}{x}\,dx$ and $\int_0^1 \frac{1}{x}\,dx$ both diverge. One way of making this precise is to use something called the *Cauchy Principal Value integral*, which says that if $f\colon [a, b] \to \mathbb{R}$ is continuous everywhere except for one point $c \in (a, b)$, then we put

$$\mathrm{PV}\!\int_a^b f(x)\,dx := \lim_{t\to 0^+} \Big(\int_a^{c-t} f(x)\,dx + \int_{c+t}^b f(x)\,dx \Big).$$

The notation is meant to remind us that this stands for something different than the usual integral.

Exercise 49.17. Show that if $\int_a^b f(x)\,dx$ is convergent in the sense of Definition 49.7, then $\mathrm{PV}\!\int_a^b f(x)\,dx = \int_a^b f(x)\,dx$.

Observe that with this definition, we have

$$\mathrm{PV}\!\int_{-1}^1 \frac{1}{x}\,dx = \lim_{t\to 0^+} \Big(\int_{-1}^{-t} \frac{1}{x}\,dx + \int_t^1 \frac{1}{x}\,dx \Big) = \lim_{t\to 0^+} 0 = 0,$$

consistent with our earlier intuition. However, there is a problem, as the following two examples illustrate.

Example 49.18.

$$\begin{aligned}
\mathrm{PV}\!\int_{-1}^1 \frac{2x+4}{(x^2+4x)^3}\,dx &= \lim_{t\to 0^+} \Big(\int_{-1}^{-t} \frac{2x+4}{(x^2+4x)^3}\,dx + \int_t^1 \frac{2x+4}{(x^2+4x)^3}\,dx \Big) \\
&= \lim_{t\to 0^+} \Big(\Big[-(x^2+4x)^{-2} \Big]_{-1}^{-t} + \Big[-(x^2+4x)^{-2} \Big]_t^1 \Big) \\
&= \lim_{t\to 0^+} \Big(\frac{1}{9} - \frac{1}{(t^2-4t)^2} + \frac{1}{(t^2+4t)^2} - \frac{1}{25} \Big) \\
&= \frac{1}{9} - \frac{1}{25} + \lim_{t\to 0^+} \frac{1}{t^2} \Big(\frac{1}{(t+4)^2} - \frac{1}{(t-4)^2} \Big) \\
&= \frac{16}{225} + \lim_{t\to 0^+} \frac{(t-4)^2 - (t+4)^2}{t^2(t^2-16)^2} = \frac{16}{225} + \lim_{t\to 0^+} \frac{-16}{t(t^2-16)^2} = -\infty.
\end{aligned}$$

Example 49.19. Evaluating the same integral using the substitution $u = x^2 + 4x$ gives

$$\mathrm{PV}\!\int_{-1}^1 \frac{2x+4}{(x^2+4x)^3}\,dx = \mathrm{PV}\!\int_{-3}^5 \frac{1}{u^3}\,du = \int_3^5 \frac{1}{u^3}\,du = \frac{1}{9} - \frac{1}{25} = \frac{16}{225}.$$

[41] This section will not appear on any tests, and largely follows an explanation I read on the website of Dave Rusin (University of Texas).

So we see that if we allow ourselves to evaluate integrals around a broader class of vertical asymptotes using the Cauchy principal value integral, then we need to come to terms with the fact that the substitution rule no longer works! One may reasonably conclude (as we do in this course) that this is too high a price to pay, and thus we will refrain from assigning finite values to any integrals that are divergent in the sense of Definition 49.7.

Remark 49.20. A similar phenomenon occurs for improper integrals of the form $\int_{-\infty}^{\infty} f(x)\,dx$. If the integral is convergent, then we have

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)\,dx &= \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{\infty} f(x)\,dx = \lim_{t\to\infty}\int_{-t}^{0} f(x)\,dx + \lim_{t\to\infty}\int_{0}^{t} f(x)\,dx \\ &= \lim_{t\to\infty}\left(\int_{-t}^{0} f(x)\,dx + \int_{0}^{t} f(x)\,dx\right) = \lim_{t\to\infty}\int_{-t}^{t} f(x)\,dx.\end{aligned}$$

In light of this, one might be tempted to define $\int_{-\infty}^{\infty} f(x)\,dx$ as $\lim_{t\to\infty}\int_{-t}^{t} f(x)\,dx$, provided the latter limit exists. However, **this turns out to be a bad idea**. Imagine that we adopt this definition. Then since $f(x) = x$ has $\int_{-t}^{t} x\,dx = \frac{1}{2}x^2|_{-t}^{t} = 0$ for all t, the limit exists and is equal to 0, so our new (bad!) definition would give $\int_{-\infty}^{\infty} x\,dx = 0$. But if we recall how integrals are supposed to behave, then we expect the following two properties to be true:

(1) shifting the graph to the left or right does not change the integral;
(2) shifting the graph up or down does change the integral.

Observe that shifting the graph of $f(x) = x$ one unit to the left has the same effect as shifting it one unit up. So according to the first rule, this shift should not change the value of the integral, but according to the second rule, it should change it! This contradiction can only be avoided by declaring that $\int_{-\infty}^{\infty} x\,dx$ is undefined (divergent), and indeed, using the true definition we observe that this improper integral is divergent because $\int_{0}^{\infty} x\,dx$ is divergent.

Review of integration strategies

Stewart §7.5, Spivak Ch. 19

This review is not included in a numbered lecture, but will be/was done during the hour preceding the first class test.

Now that we have learned several different tools for integration, it is worth reviewing them and describing an overall strategy.

Step 1: Check list of basic examples

The following list of integrals should be committed to memory, so that once we see one of these integrals appear, we know how to complete the solution. (To avoid cluttering up the display, we omit the constants of integration.)

$$\int x^n\,dx = \frac{x^{n+1}}{n+1} \qquad\qquad \int \frac{1}{x}\,dx = \ln|x|$$

$$\int e^x\,dx = e^x \qquad \int b^x\,dx = \frac{b^x}{\ln b}, \quad b > 0$$

$$\int \sin^x\,dx = -\cos x \qquad \int \cos x\,dx = \sin x$$

$$\int \sec^2 x\,dx = \tan x \qquad \int \csc^2 x\,dx = -\cot x$$

$$\int \sec x \tan x\,dx = \sec x \qquad \int \csc x \cot x\,dx = -\csc x$$

$$\int \sec x\,dx = \ln|\sec x + \tan x| \qquad \int \csc x\,dx = -\ln|\csc x + \cot x|$$

$$\int \tan x\,dx = \ln|\sec x| \qquad \int \cot x\,dx = \ln|\sin x|$$

$$\int \sinh x\,dx = \cosh x \qquad \int \cosh x\,dx = \sinh x$$

$$\int \frac{dx}{x^2+a^2} = \frac{1}{a}\tan^{-1}\frac{x}{a} \qquad \int \frac{dx}{\sqrt{a^2-x^2}} = \sin^{-1}\frac{x}{a}, \quad a > 0$$

$$\int \frac{dx}{x^2-a^2} = \frac{1}{2a}\ln\left|\frac{x-a}{x+a}\right| \qquad \int \frac{dx}{\sqrt{x^2\pm a^2}} = \ln|x + \sqrt{x^2 \pm a^2}|.$$

Remark 49.21. Some of these have alternate forms; for example Stewart's book lists the integral of $\csc x$ as $\ln|\csc x - \cot x|$. A short computation using properties of logarithms and the identity $\csc^2 x - \cot^2 x = 1$ shows that this agrees with the form here.

Step 2: Simplify if possible

If the integrand can be simplified using standard algebraic manipulations or trigonometric identities, this is the next thing to do.

Example 49.22.

$$\int (\sin x + \cos x)^2\,dx = \int (\sin^2 x + 2\sin x \cos x + \cos^2 x)\,dx = \int (1 + \sin 2x)\,dx.$$

Step 3: Make an obvious substitution, if there is one

If there is a clear choice for u such that du naturally appears in the integrand, then it is worth trying this substitution.

Example 49.23. $\displaystyle\int \frac{x}{x^2-1}\,dx$ has the derivative of the denominator in the numerator (up to a constant), so $u = x^2 - 1$ is natural and transforms the integral into $\displaystyle\frac{1}{2}\int \frac{1}{u}\,du$.

Step 4: Classify the integral as a type that we know how to deal with

There are four general classes of integrals that we have developed a procedure for dealing with by now.

(1) Trigonometric integrals such as $\int \sin^4 x \cos^3 x\,dx$, which can be handled using various substitutions and identities.

(2) Rational functions, which can be handled using partial fractions.
(3) Integrals of the form $\int f(x)g(x)\,dx$, where $g(x)$ is something we can integrate and $f(x)$ gets simpler after differentiating; the most important case is when f is a polynomial, but this also includes things like $f(x) = \ln x$ or $f(x) = \tan^{-1} x$. For integrals like this, integration by parts with $u = f(x)$ and $dv = g(x)\,dx$ is likely to be useful.
(4) Integrals involving quadratic polynomials inside square roots, for which an appropriate trigonometric substitution is often helpful.

Step 5: Get creative

Sometimes with a little more creativity we can find an algebraic manipulation or a substitution that helps, even if one was not obvious upon initial inspection.

Example 49.24.

$$\int \frac{1}{1-\sin x}\,dx = \int \frac{1+\sin x}{1-\sin^2 x}\,dx = \int \frac{1+\sin x}{\cos^2 x}\,dx$$
$$= \int (\sec^2 x + \sec x \tan x)\,dx = \tan x + \sec x + C.$$

Example 49.25. The substitution $u = \sqrt{x}$ has $x = u^2$, $dx = 2u\,du$, so

$$\int e^{\sqrt{x}}\,dx = \int 2ue^u\,du,$$

which can be integrated by parts.

Similarly, some functions that do not immediately look like rational functions can be transformed into rational functions via an appropriate substitution, and then integrated using partial fractions.

Example 49.26. Using the substitution $u = \sqrt{x+4}$, $x = u^2 - 4$, $dx = 2u\,du$, we have

$$\int \frac{\sqrt{x+4}}{x}\,dx = \int \frac{u}{u^2-4}2u\,du = 2\int \frac{u^2}{u^2-4}\,du = 2\int 1 + \frac{4}{u^2-4}\,du$$
$$= 2u + 2\int \frac{1}{u-2} - \frac{1}{u+2}\,du = 2u + 2\ln|u-2| - 2\ln|u+2| + C$$
$$= 2\sqrt{x+4} - \ln\frac{x+8-4\sqrt{x+4}}{x+8+4\sqrt{x+4}} + C.$$

where we use the computation

$$(u \pm 2)^2 = (\sqrt{x+4} \pm 2)^2 = x + 4 \pm 4\sqrt{x+4} + 4 = x + 8 \pm 4\sqrt{x+4},$$

and have omitted the computations to determine the partial fraction decomposition. Just to continue the fun, we point out that the last term on the first line can also be integrated with the substitution $u = 2\sec\theta$, so

$$\int \frac{4}{u^2-4}\,du = \int \frac{4}{4\tan^2\theta}2\sec\theta\tan\theta\,d\theta = 2\int \frac{\sec\theta}{\tan\theta}\,d\theta = 2\int \csc\theta\,d\theta,$$

which is a known integral. Turning θ back into u, and then into x, gives the same result as above.

Remark 49.27. It turns out that we can also integrate any rational function of trigonometric functions, such as $f(x) = \frac{\cos^2 x - \sin^3 x}{\tan x + \sec x}$, by using the *Weierstrass substitution* $t = \tan \frac{x}{2}$, which (after some work) yields $\cos x = \frac{1-t^2}{1+t^2}$, $\sin x = \frac{2t}{1+t^2}$, and $dx = \frac{2}{1+t^2}\,dt$, so that $\int f(x)\,dx = \int g(t)\,dt$ for some rational function g.

By now it should be clear that the task of finding formulas for antiderivatives – *symbolic integration* – is rather harder than the task of finding formulas for derivatives – *symbolic differentiation*. The latter task is fairly routine thanks to various results like the product rule, the chain rule, etc., which let us write down a formula for the derivative of any function that is written in terms of polynomials, radicals, exponentials, logarithms, and trigonometric functions. As we have seen, symbolic integration is an entirely different matter, and it turns out that there are some integrals that *cannot* be evaluated in terms of the 'elementary' functions we are used to dealing with; this include relatively innocuous-looking expressions such as $\int e^{-x^2}\,dx$ and $\int \frac{1}{\ln x}\,dx$.[42]

[42]To make this claim of impossibility rigorous, one needs to formulate clearly the class of functions that we consider, and then provide a proof of impossibility; this is beyond the scope of this course.

Part V. Applications of integration

Lecture 50 Arc length and the catenary

Stewart §8.1, Spivak exercise 13.25

50.1. A formula for arc length

We know how to compute the length of a straight line segment: it is simply the distance between the two endpoints (x_1, y_1) and (x_2, y_2) given by Pythagoras' formula

$$\text{distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

If we consider a "piecewise linear" curve that is a sequence of straight line segments connecting the points $P_0, P_1, \dots, P_n$, then we can similarly compute the total length of the curve as $\sum_{i=1}^{n} \text{distance}(P_{i-1}, P_i)$.

Given a more general curve in the plane, it is reasonable to approximate it by a piecewise linear curve, compute the length of the approximation, and then take a limit as the endpoints of the approximating line segments get closer and closer together. To make this more precise, suppose we consider the graph of $y = f(x)$ between $x = a$ and $x = b$. Then we might fix a large integer $n \in \mathbb{N}$; choose points in the interval $[a, b]$ by $x_i = a + i\Delta x$ for $0 \le i \le n$, where $\Delta x = \frac{b-a}{n}$; denote the point $(x_i, f(x_i))$ by P_i; and then declare the length of the curve to be

$$\text{length} = \lim_{n\to\infty} \sum_{i=1}^{n} \text{distance}(P_{i-1}, P_i). \tag{50.1}$$

This looks suspiciously similar to a limit of Riemann sums. Using Pythagoras' formula we get

$$\text{distance}(P_{i-1}, P_i) = \sqrt{(x_i - x_{i-1})^2 + (f(x_i) - f(x_{i-1}))^2}. \tag{50.2}$$

If f is continuous on $[a, b]$ and differentiable on (a, b), then the Mean Value Theorem says that for each $1 \le i \le n$ there is $x_i^* \in [x_{i-1}, x_i]$ such that

$$f(x_i) - f(x_{i-1}) = f'(x_i^*)(x_i - x_{i-1}).$$

Using this in (50.2) and recalling that $x_i - x_{i-1} = \Delta x$, we get

$$\text{distance}(P_{i-1}, P_i) = \sqrt{(\Delta x)^2 + f'(x_i^*)^2(\Delta x)^2} = \sqrt{1 + f'(x_i^*)^2} \cdot \Delta x.$$

Taking a sum over i from 1 to n, and then a limit as $n \to \infty$, we see that

$$\lim_{n\to\infty} \sum_{i=1}^{n} \text{distance}(P_{i-1}, P_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \sqrt{1 + f'(x_i^*)^2} \cdot \Delta x = \int_a^b \sqrt{1 + (f'(x))^2}\, dx$$

provided f' is continuous on (a, b). Thus we make the following definition: if f is continuously differentiable,[43] the *arc length* L of the curve $y = f(x)$ from $x = a$ to $x = b$

[43] This means that f is differentiable and that f' is continuous.

is

$$L = \text{length} = \int_a^b \sqrt{1 + (f'(x))^2}\, dx. \tag{50.3}$$

It is also often useful to define the following *arc length function*: given a continuously differentiable function $f\colon [a,b] \to \mathbb{R}$ and a value $x \in (a,b)$, the arc length of the section of curve from $(a, f(a))$ to $(x, f(x))$ is given by

$$s(x) = \int_a^x \sqrt{1 + (f'(t))^2}\, dt. \tag{50.4}$$

By the FTC, we have

$$s'(x) = \sqrt{1 + (f'(x))^2}. \tag{50.5}$$

Writing $y = f(x)$ and using Leibniz notation $f'(x) = \frac{dy}{dx}$, we can write (50.3) as

$$L = \int_a^b \sqrt{1 + \Big(\frac{dy}{dx}\Big)^2}\, dx,$$

and (50.5) can be rewritten as

$$\frac{ds}{dx} = \sqrt{1 + \Big(\frac{dy}{dx}\Big)^2}. \tag{50.6}$$

In an abuse of notation (since s, dx, and dy have no independent meaning), this is sometimes written as

$$(ds)^2 = (dx)^2 + (dy)^2,$$

which is an infinitesimal version of Pythagoras' formula (50.2).

50.2. Examples of arc length

Example 50.1. To find the arc length L of the parabola $y = x^2$ from $(0,0)$ to $(1,1)$, we use (50.3) with $a = 0$, $b = 1$, $f'(x) = 2x$ to write

$$\begin{aligned}
L &= \int_0^1 \sqrt{1 + (2x)^2}\, dx && \Big(u = 2x,\ dx = \frac{1}{2}du\Big) \\
&= \frac{1}{2}\int_0^2 \sqrt{1 + u^2}\, du && \Big(u = \tan\theta,\ du = \sec^2\theta\, d\theta\Big) \\
&= \frac{1}{2}\int_0^\alpha \sec^3\theta\, d\theta && (\tan\alpha = 2,\ \sec^2\alpha = 1 + \tan^2\alpha = 5).
\end{aligned}$$

Thus

$$\begin{aligned}
2L &= \int_0^\alpha \underbrace{(\sec\theta)}_{u}\underbrace{(\sec^2\theta)\, d\theta}_{dv} = [\underbrace{\sec\theta}_{u}\underbrace{\tan\theta}_{v}]_0^\alpha - \int_0^\alpha \sec\theta\tan^2\theta\, d\theta \\
&= \sec\alpha\tan\alpha - \sec 0\tan 0 - \int_0^\alpha \sec\theta(\sec^2\theta - 1)\, d\theta \\
&= 2\sqrt{5} - \int_0^\alpha \sec^3\theta\, d\theta + \int_0^\alpha \sec\theta\, d\theta = 2\sqrt{5} - 2L + \big[\ln|\sec\theta + \tan\theta|\big]_0^\alpha,
\end{aligned}$$

and solving for L gives

$$\begin{aligned} L &= \frac{1}{4}\Big(2\sqrt{5} + \ln|\sec\alpha + \tan\alpha| - \overbrace{\ln|\sec 0 + \tan 0|}^{=\ln|1+0|=0}\Big) \\ &= \frac{\sqrt{5}}{2} + \frac{\ln(\sqrt{5}+2)}{4}. \end{aligned}$$

We get a similar formula for arc length if x is written as a function of y; if $x = g(y)$ and the curve runs from $y = a$ to $y = b$, then the arc length is $\int_a^b \sqrt{1 + (g'(y))^2}\, dy$.

Example 50.2. Consider the curve $x^2 = y^3$ running from $(1, 1)$ to $(8, 4)$. If we write y as a function of x, then we get $y = x^{2/3}$ and the formula for arc length gives

$$L = \int_1^8 \sqrt{1 + \Big(\frac{2}{3}x^{-1/3}\Big)^2}\, dx = \int_1^8 \sqrt{1 + \frac{4}{9}x^{-2/3}}\, dx.$$

It is not at all clear how to evaluate this. On the other hand, if we write x as a function of y then we have $x = y^{3/2}$ (note that the curve is in the first quadrant so $x > 0$) and the arc length is

$$\begin{aligned} L &= \int_1^4 \sqrt{1 + \Big(\frac{3}{2}y^{1/2}\Big)^2}\, dy = \int_1^4 \sqrt{1 + \frac{9}{4}y}\, dy \qquad\qquad (u = 1 + \tfrac{9}{4}y,\ du = \tfrac{9}{4}\, dy) \\ &= \frac{4}{9}\int_{13/4}^{10} \sqrt{u}\, du = \frac{4}{9}\Big[\frac{2}{3}u^{3/2}\Big]_{13/4}^{10} = \frac{8}{27}\Big(10^{3/2} - \Big(\frac{13}{4}\Big)^{3/2}\Big) \\ &= \frac{1}{27}(80\sqrt{10} - 13\sqrt{13}). \end{aligned}$$

As is already apparent from the previous examples, the presence of the square root in the arc length formula often leads to a nasty integral.

Example 50.3. Consider the hyperbola $x^2 - y^2 = 1$. Let L denote the arc length from $(1, 0)$ to $(2, \sqrt{3})$. Writing $y = \sqrt{x^2 - 1}$ gives $\frac{dy}{dx} = \frac{x}{\sqrt{x^2-1}}$, so

$$L = \int_1^2 \sqrt{1 + \frac{x^2}{x^2 - 1}}\, dx = \int_1^2 \sqrt{\frac{2x^2 - 1}{x^2 - 1}}\, dx.$$

The presence of a quadratic inside a square root suggests a trigonometric substitution; but there are two quadratics in play here! Writing $x = \sec\theta$ gives $\sqrt{x^2 - 1} = \tan\theta$ and $dx = \sec\theta\tan\theta$, so

$$\int \sqrt{\frac{2x^2 - 1}{x^2 - 1}}\, dx = \int \frac{\sqrt{2\sec^2\theta - 1}}{\tan\theta} \cdot \sec\theta\tan\theta\, d\theta,$$

and it is not at all clear where to go from here, since none of the usual trigonometric identities help us simplify $\sqrt{2\sec^2\theta - 1}$. In fact it turns out that this integral cannot be evaluated in elementary terms using the functions that we have introduced so far, and so we cannot compute L exactly. Given this, our best bet would be to turn to numerical integration and use something like Simpson's rule to obtain an approximate value for the integral.

In fact, there is one more point to be made here. Note that the integrand $\sqrt{\frac{2x^2-1}{x^2-1}}$ has a vertical asymptote at $x = 1$, which was one of the limits of integration. Thus this is actually an *improper* integral! Since our discussion of numerical integration did not include any tools for dealing with improper integrals (and we would need to start by using a comparison theorem to check that this improper integral actually converges), we are better off writing $x = \sqrt{y^2+1}$ and working with the integral

$$L = \int_0^{\sqrt{3}} \sqrt{1 + \left(\frac{dx}{dy}\right)^2}\, dy = \int_0^{\sqrt{3}} \sqrt{1 + \frac{y^2}{y^2+1}}\, dy.$$

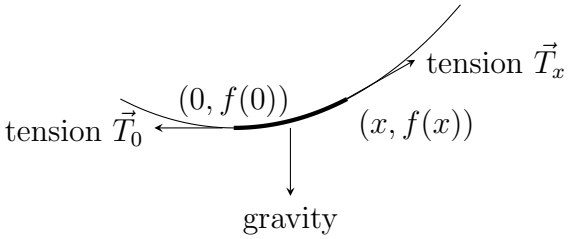
We still cannot evaluate this explicitly, but at least the integrand here is a bounded continuous function, and so we do not need to resort to improper integrals.

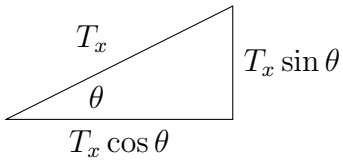
The appearance of the improper integral in the first expression comes because the curve has a vertical tangent line at $(1, 0)$, so the slope $\frac{dy}{dx}$ that appears in the arc length formula is infinite at this point. This is a phenomenon you should watch out for when computing arc lengths.

50.3. *The catenary

Consider a cable that is suspended at its endpoints and hangs freely in between them, such as a power line or telephone wire between two poles. What shape will it make?

Assume that the cable is relatively thin, so that it can be well-approximated by a curve $y = f(x)$, and that it is flexible, so that the tension at any point in the cable is in a direction tangent to the curve. Choose some point on the curve as the origin for x, and consider the part of the cable that lies between 0 and x. As shown in the picture, there are three forces acting on this segment of cable: tension pulling it to the left at the point $(0, f(0))$ (labeled $\vec{T}_0$), tension pulling it to the right at the point $(x, f(x))$ (labeled $\vec{T}_x$), and gravity pulling it downward. Let T_0 and T_x denote the magnitude of the tension forces; then since the tension at x points in the direction of the tangent line, which has slope $\tan\theta = f'(x)$, we see that the horizontal and vertical components of $\vec{T}_0$ are as shown in the picture at right.





Because the cable is not moving, the forces must all balance out, so $T_0 = T_x \cos\theta$, and $T_x \sin\theta$ equals the magnitude of the force due to gravity. This force is $m(x)g$, where g is the gravitational constant and $m(x)$ is the mass of the segment of cable between 0 and x. Assume that the cable has uniform density ρ (mass per unit length), so that $m(x) = \rho s(x)$, where $s(x)$ is the length of the section of cable from 0 to x. Then we have

$$T_x \cos\theta = T_0 \quad \text{and} \quad T_x \sin\theta = m(x)g = \rho g s(x).$$

Dividing these two equations gives

$$f'(x) = \tan\theta = \frac{T_x \sin\theta}{T_x \cos\theta} = \frac{\rho g}{T_0} s(x). \tag{50.7}$$

Our goal is to find a formula for the function f that allows us to write y as a function of x via $y = f(x)$. We will do this by first writing both x and y as functions of arc length s and as functions of a new variable t; this procedure of obtaining *parametrizations* for the curve is one that we will later return to and study in greater detail.

Using (50.6) and writing $a = \frac{T_0}{\rho g} > 0$ for a parameter that depends on the physical characteristics of the situation, we see that the function $x \mapsto s(x)$ satisfies

$$\frac{ds}{dx} = \sqrt{1 + \left(\frac{dy}{dx}\right)^2} = \sqrt{1 + \left(\frac{s}{a}\right)^2} = \frac{\sqrt{a^2 + s^2}}{a}, \tag{50.8}$$

where the second equality uses the fact that $\frac{dy}{dx} = f'(x) = \frac{s(x)}{a}$. Since the derivative of the inverse function $s \mapsto x(s)$ is the reciprocal of the derivative $s'(x)$, we conclude that

$$\frac{dx}{ds} = \frac{a}{\sqrt{a^2 + s^2}} \quad \Rightarrow \quad x(s) = \int \frac{a}{\sqrt{a^2 + s^2}}\, ds.$$

To evaluate this integral, we could use the trigonometric substitution $s = a\tan\theta$ and the identity $1 + \tan^2\theta = \sec^2\theta$, but it turns out to be simpler to use the substitution $s = a\sinh t$ and the identity $1 + \sinh^2 t = \cosh^2 t$:

$$x(s) = \int \frac{a}{\sqrt{a^2 + a^2\sinh^2 t}} \cdot a\cosh t\, dt = \int a\, dt = at + C.$$

Note that when $s = 0$ we have $x = 0$ and $t = \sinh^{-1}\frac{s}{a} = 0$, so the constant of integration is $C = 0$, and we have $t = x/a$.

To determine $y(s)$ we first use the chain rule to write

$$\frac{dy}{ds} = \frac{dy}{dx}\frac{dx}{ds} = \frac{s}{a} \cdot \frac{a}{\sqrt{a^2 + s^2}} \quad \Rightarrow \quad y(s) = \int \frac{s}{\sqrt{a^2 + s^2}}\, ds = \sqrt{a^2 + s^2} + b,$$

where b is a constant of integration. Since $\sqrt{a^2 + s^2} = \sqrt{a^2 + a^2\sinh^2 t} = a\cosh t$, we conclude that

$$y = f(x) = \sqrt{a^2 + s^2} + b = a\cosh(t) + b = a\cosh\left(\frac{x}{a}\right) + b.$$

Thus the curve formed by a hanging cable – called a *catenary* – is described by the hyperbolic cosine function. Note that here a, b are parameters determined by the physical characteristics of the situation, including the strength of gravity, the density of the cable, and the location of the two points at which it is suspended.

Lecture 51 Surface area

Stewart §8.2, Spivak appendix to Ch. 19

51.1. Surfaces of revolution

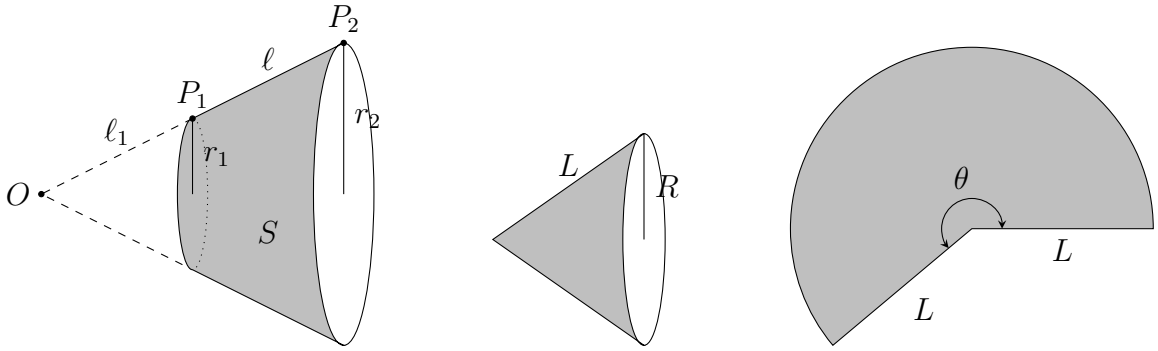
Suppose we are given a function $f\colon [a,b] \to \mathbb{R}$ whose graph describes a curve $\{(x, f(x)) : a \le x \le b\}$. The *surface of revolution* associated to this curve is the surface $S \subset \mathbb{R}^3$ that is obtained by rotating the curve around the x-axis. Observe that if (x, y, z) is a point on S, then the distance from (x, y, z) to $(x, 0, 0)$ must be equal to $f(x)$, and thus a precise description of S can be given by

$$S = \{(x, y, z) : \sqrt{y^2 + z^2} = f(x)\},$$

though we will not use this in what follows. Our goal in this section is to find a formula for the surface area of S.

As usual, the simplest case occurs when f is linear; $f(x) = mx+c$. If $m = 0$ so that f is constant, then the corresponding surface of revolution is a cylinder with radius c and depth $b - a$; this cylinder can be unrolled into a rectangle with the same surface area, whose dimensions are $(b - a) \times 2\pi c$, so the surface area of S is given by $2\pi c(b - a)$. We will find it convenient to write this as $2\pi r\ell$, where $r = c$ is the radius, and $\ell = b - a$ is the distance between $(a, f(a))$ and $(b, f(b))$ (since $f(a) = f(b)$).

If $m \neq 0$, then S is a truncated cone. To compute the surface area of S, write $r_1 = f(a)$ and $r_2 = f(b)$ for the radii of the circles that form the ends of the truncated cone. For concreteness, assume that $m > 0$ so that $r_1 < r_2$ (the case $m < 0$ is similar). Let $P_1 = (a, f(a))$ and $P_2 = (b, f(b))$ be the two endpoints of the line segment, and let ℓ be the distance between them. Let O be the point where the line $y = mx + c$ intersects the x-axis, and let ℓ_1 be the distance from O to P_1, as shown in the picture at left.



To find the surface area of S, we need to find the formula for the surface area of a cone. Consider the cone shown in the second picture, where the base has radius R and the diagonal side has length L. If we cut this cone along a line from the base to the tip and then unroll it, we obtain a shape such as the one shown in the third picture, which has the same area as the cone. The arc that forms the outer boundary has length θL by the formula for length of a circular arc; it also has length $2\pi R$ since this boundary is obtained by unrolling the circle at the cone's base. Thus we have $\theta L = 2\pi R$, and moreover, the area of this region is given by

$$\text{area} = \frac{\theta}{2\pi} \cdot \pi L^2 = \frac{1}{2}\theta L^2 = \frac{1}{2} \cdot \frac{2\pi R}{L} \cdot L^2 = \pi RL.$$

Returning to the surface area of S, observe that S is obtained by taking a cone with $L = \ell + \ell_1$ and $R = r_2$, and then removing from it a cone with $L = \ell_1$ and $R = r_1$. Thus

the surface area of S is

$$\text{area}(S) = \pi r_2(\ell + \ell_1) - \pi r_1 \ell_1 = \pi\big(\ell_1(r_2 - r_1) + \ell r_2\big).$$

Observe that $\frac{\ell_1}{r_1} = \frac{\ell_1+\ell}{r_2}$, so $\ell_1 r_2 = r_1\ell_1 + r_1\ell$, and thus $\ell_1(r_2 - r_1) = \ell r_1$, which gives

$$\text{area}(S) = \pi(\ell r_1 + \ell r_2) = \pi(r_1 + r_2)\ell.$$

Note that this agrees with the formula from above for the surface area of a cylinder when f is constant, since in this case we have $r_1 = r_2 = r$. In order to write everything directly in terms of the function f, we observe that

$$r_1 + r_2 = f(a) + f(b) = 2f(x^*), \quad \text{where } x^* = \frac{a+b}{2},$$

since $f(x) = mx + c$ is linear, and moreover

$$\ell = \sqrt{(b-a)^2 + (f(b) - f(a))^2} = \sqrt{(b-a)^2 + (m(b-a))^2} = (b-a)\sqrt{1+m^2}.$$

Thus we have proved the following.

Proposition 51.1. *If $f\colon [a,b] \to [0,\infty)$ is linear (in other words, its graph is a line segment), then the surface area of the corresponding surface of revolution is*

$$\text{area} = 2\pi f(x^*)\sqrt{1 + (f'(x^*))^2} \cdot (b-a),$$

where $x^* = \frac{a+b}{2}$ is the midpoint of $[a,b]$.

In light of Proposition 51.1, it is reasonable to define the surface area of a surface of revolution S for an *arbitrary* continuously differentiable function $f\colon [a,b] \to [0,\infty)$ by approximating the graph of f using a piecewise linear curve with n pieces, whose corresponding surface of revolution has an area that can be computed using the proposition, and then taking a limit as $n \to \infty$. Thus we define the surface area of S to be

$$\text{area}(S) = \lim_{n\to\infty} \sum_{i=1}^{n} 2\pi f(x_i^*)\sqrt{1 + f'(x_i^*)^2}\Delta x = \int_a^b 2\pi f(x)\sqrt{1 + f'(x)^2}\,dx, \tag{51.1}$$

where $\Delta x = \frac{b-a}{n}$, $x_i = a + i\Delta x$, and $x_i^* = \frac{1}{2}(x_{i-1} + x_i)$.

Using Leibniz notation, (51.1) can be written as $\int_a^b 2\pi y\sqrt{1 + (\frac{dy}{dx})^2}\,dx$, or even more compactly as $\int_a^b 2\pi y\,ds$, where ds is shorthand for $\sqrt{1 + (\frac{dy}{dx})^2}\,dx$, which is integrated to get arc length; this is a useful way to remember the formula for surface area.

Remark 51.2. The astute reader may notice that in Proposition 51.1, $f(x^*)$ and $f'(x^*)$ referred to the *linear* function f, while in (51.1) $f(x_i^*)$ and $f'(x_i^*)$ refer to the *original (nonlinear)* function f, rather than to its linear approximation. One can proceed as in the argument for arc length and use the MVT to produce x_i^* for which $f'(x_i^*)$ takes the value we expect, but even after doing this $f(x_i^*)$ need not be exactly equal to $\frac{1}{2}(f(x_{i-1}) + f(x_i))$. Thus to give a complete proof, one can use continuity of f to estimate the difference between these two quantities, and then prove that this difference goes to 0 as $n \to \infty$. We omit the details, but suggest this as a worthwhile exercise for the reader who wishes to see everything completely justified.

51.2. Examples of surface area

Example 51.3. Consider the surface of revolution obtained by rotating the curve $y = \sqrt{2-x}$ for $0 \le x \le 1$ around the x-axis. Then $\frac{dy}{dx} = -\frac{1}{2\sqrt{2-x}}$, so the surface area is

$$S = \int_0^1 2\pi y \sqrt{1 + \Big(\frac{dy}{dx}\Big)^2}\, dx = \int_0^1 2\pi\sqrt{2-x} \cdot \sqrt{1 + \frac{1}{4(2-x)}}\, dx$$
$$= 2\pi \int_0^1 \sqrt{2-x} \cdot \sqrt{\frac{9-4x}{4(2-x)}}\, dx = \pi \int_0^1 \sqrt{9-4x}\, dx.$$

Making the substitution $u = 9 - 4x$, $du = -4\,dx$ gives

$$S = -\frac{\pi}{4}\int_9^5 \sqrt{u}\, du = \frac{\pi}{4}\int_5^9 \sqrt{u}\, du = \frac{\pi}{4}\Big[\frac{2}{3}u^{3/2}\Big]_5^9 = \frac{\pi}{6}\big(9^{3/2} - 5^{3/2}\big).$$

Because the function $x \mapsto y = \sqrt{2-x}$ is 1-1 on $[0,1]$, we could also study this surface treating x as a function of y: solving for x gives $x = 2 - y^2$, and y ranges over the interval $[1, \sqrt{2}]$. To make this change of variables in the integral, we replace dx with $\frac{dx}{dy}\,dy$ (here we are using the substitution rule) and write

$$S = \int_1^{\sqrt{2}} 2\pi y \sqrt{1 + \Big(\frac{dy}{dx}\Big)^2} \cdot \frac{dx}{dy}\, dy = \int_1^{\sqrt{2}} 2\pi y \sqrt{\Big(\frac{dx}{dy}\Big)^2 + \Big(\frac{dy}{dx}\cdot\frac{dx}{dy}\Big)^2}\, dy.$$

By the rule for derivatives of inverse functions, we have $\frac{dy}{dx}\cdot\frac{dx}{dy} = 1$, and so this formula can be rewritten as

$$S = \int_1^{\sqrt{2}} 2\pi y \sqrt{1 + \Big(\frac{dx}{dy}\Big)^2}\, dy \tag{51.2}$$

Recall from our discussion of arc length that the symbol ds can be interpreted either as $\sqrt{1 + (\frac{dy}{dx})^2}\,dx$ or as $\sqrt{1 + (\frac{dx}{dy})^2}\,ds$; then the mnemonic formula $\int 2\pi y\, ds$ for surface area can reasonably be interpreted as standing for either

$$\int_a^b 2\pi y \sqrt{1 + \Big(\frac{dy}{dx}\Big)^2}\, dx \qquad \text{or} \qquad \int_c^d 2\pi y \sqrt{1 + \Big(\frac{dx}{dy}\Big)^2}\, dy, \tag{51.3}$$

where $[a,b]$ is the integral over which x ranges and $[c,d]$ is the integral over which y ranges.

Remark 51.4. Be careful to note that in both versions of the surface area formula (51.3), the first part of the integrand is $2\pi y$, regardless of whether we are integrating with respect to x or y. This part of the integrand represents the fact that the surface is constructed by rotation around the x-axis, so that y represents the distance from the axis and $2\pi y$ represents the circumference of the circle obtained by cutting a cross-section of the surface at a given value of x. The part of the integral that depends on which variable we integrate with respect to is the derivative appearing inside the square root.

Returning to Example 51.3, we see that the surface area can also be computed using (51.2) and the formula $\frac{dx}{dy} = -2y$:

$$S = \int_1^{\sqrt{2}} 2\pi y\sqrt{1+(-2y)^2}\,dy = 2\pi\int_1^{\sqrt{2}} y\sqrt{1+4y^2}\,dy \qquad (u = 1+4y^2,\ du = 8y\,dy)$$
$$= 2\pi\int_5^9 \frac{1}{8}\sqrt{u}\,du = \frac{\pi}{4}\int_5^9 \sqrt{u}\,du,$$

which is exactly the same integral we obtained the first time around (after making the substitution $u = 9-4x$).

Remark 51.5. One can also consider surfaces of revolution around the y-axis, and in this case the roles of x and y are reversed, so the general formula is $\int 2\pi x\,ds$.

Example 51.6. To find the surface area of a sphere of radius R, we can treat it as the surface of revolution around the y-axis of the curve $x = \sqrt{R^2-y^2}$ for $-R \le y \le R$, which has $\frac{dx}{dy} = -y/\sqrt{R^2-y^2}$, and we get

$$S = \int_{-R}^{R} 2\pi x\sqrt{1+\Big(\frac{dx}{dy}\Big)^2}\,dy = 2\pi\int_{-R}^{R}\sqrt{R^2-y^2}\cdot\sqrt{1+\frac{y^2}{R^2-y^2}}\,dy$$
$$= 2\pi\int_{-R}^{R}\sqrt{R^2-y^2}\cdot\sqrt{\frac{R^2}{R^2-y^2}}\,dy = 2\pi\int_{-R}^{R} R\,dy = 2\pi\Big[Ry\Big]_{-R}^{R} = 2\pi R(2R) = 4\pi R^2.$$

Example 51.7. Consider the curve $\{(x, \frac{1}{x}) : x \in [1,\infty)\}$, which has infinite length. The corresponding surface of revolution is called *Gabriel's horn*. Its surface area is given by the improper integral

$$S = \int_1^\infty 2\pi\cdot\frac{1}{x}\sqrt{1+\Big(-\frac{1}{x^2}\Big)^2}\,dx = \int_1^\infty \frac{2\pi}{x}\sqrt{1+\frac{1}{x^4}}\,dx. \tag{51.4}$$

Observe that for every $x \ge 1$, we have

$$\frac{2\pi}{x}\sqrt{1+\frac{1}{x^4}} \ge \frac{2\pi}{x} \ge \frac{1}{x},$$

and that we showed earlier that the improper integral $\int_1^\infty \frac{1}{x}\,dx$ is divergent. By the Comparison Theorem, the integral in (51.4) is divergent, which we interpret as meaning that Gabriel's horn has infinite surface area.

On the other hand, the *volume* of the region enclosed by Gabriel's horn is given by

$$V = \int_1^\infty \pi y^2\,dx = \int_1^\infty \frac{\pi}{x^2}\,dx = \lim_{t\to\infty}\Big[\frac{-\pi}{x}\Big]_1^t = \pi,$$

so this improper integral is convergent and the volume is finite.[44] This is a somewhat counter-intuitive state of affairs; can you explain it?

[44]You may also observe that if $P \subset \mathbb{R}^3$ is a plane containing the x-axis, then the corresponding cross-section of the enclosed region (its intersection with P) has infinite area, while this area is finite for any plane not containing the x-axis.

Lecture 52 *Hydrostatic force and pressure

Stewart §8.3

Here is an application from engineering. Suppose we have a dam that is holding back water, and we want to compute the total force that the water exerts on the dam; this is the *hydrostatic force*. The force per unit area at a given point is the *hydrostatic pressure*, and varies from point to point; near the surface of the water the pressure is relatively small, while deep down it is greater. Thus the force is obtained by integrating the pressure.

Imagine a small cube of water at depth d. If the water is motionless (we are at equilibrium) then all 6 faces of the cube experience the same force from the surrounding water; if it were not so, then the cube would move or deform. The top face experiences a downward force due to the column of water above it, which has mass ρAd, where ρ is the density of the fluid and A is the surface area of the top face of the cube. Thus the total force on the top face is $g\rho Ad$, where g is the gravitational constant, and thus the pressure in any given direction is force/area $= g\rho d$.

Example 52.1. Suppose that we consider a dam shaped like a trapezoid whose bottom and top edges are horizontal, with lengths 10 m and 18 m, respectively; suppose the total height of the dam is 16 m; and suppose that the water is 3/4 of the way to the top of the dam, so it is 12 m deep.

Let $w(x)$ denote the width of the dam at a depth x below the surface of the water. Then if we divide the interval $[0, 12]$ into n pieces $[x_{i-1}, x_i]$ of equal length $\Delta x = 12/n$, the strip of the dam between depths x_{i-1} and x_i is roughly a rectangle with width $w(x_i)$ and height Δx, so its area is $w(x_i)\Delta x$ and it experiences a hydrostatic force of pressure $\times$ area $= \rho g x \cdot w(x_i)\Delta x$. Summing up over all n strips and taking a limit as $n \to \infty$ gives a total force of

$$F = \lim_{n\to\infty} \sum_{i=1}^{n} \rho g x w(x_i)\Delta x = \int_0^{12} \rho g x w(x)\, dx. \tag{52.1}$$

In this case we see that $w(x) = ax + b$ for some $a, b \in \mathbb{R}$, which can be determined by using the fact that $w(12) = 10$ (at the deepest point) and $w(-4) = 18$ (at the top of the dam), so we have

$$12a + b = 10 \quad \text{and} \quad -4a + b = 18.$$

Subtracting the two equations gives $16a = -8$, so $a = -1/2$, and thus $b = 18 + 4a = 18 - 2 = 16$, which gives $w(x) = 16 - x/2$, and the total hydrostatic force on the dam is

$$\begin{aligned} F &= \int_0^{12} \rho g\Big(16x - \frac{x^2}{2}\Big)\, dx = \rho g\Big[8x^2 - \frac{x^3}{6}\Big]_0^{12} = \rho g\Big(8\cdot(12)^2 - \frac{(12)^3}{6}\Big) \\ &= \rho g \cdot 144 \cdot \Big(8 - \frac{12}{6}\Big) = \rho g \cdot 144 \cdot 6 = 864\rho g. \end{aligned}$$

Note that the number 864 represents $\int_0^{12} x w(x)\, dx$ and thus has units m^3. Using the values $g = 9.8\,\mathrm{m/s^2}$ and $\rho = 1000\,\mathrm{kg/m^3}$, we get

$$F \approx 8.47 \times 10^6\,\mathrm{N},$$

where the units of force are Newtons, $1\,\mathrm{N} = 1\,\mathrm{kg\,m/s^2}$.

Example 52.2. An undersea laboratory is built on the ocean floor where the water is 100 m deep. The end of the lab has the shape of a sine function, with width 10 m and height 5 m. How much hydrostatic force does the end of the lab experience?

Let y be the height above the ocean floor; then the depth of the water at any given point is $100 - y$, and the same arguments that lead to (52.1) show that the total force is

$$F = \int_0^5 \rho g(100 - y) w(y)\, dy,$$

where $w(y)$ is the width of the lab at height y. Taking the x-axis to be centred at the centre of the lab, the height of the lab at position x is given by $y(x) = 5\cos(\frac{\pi x}{10})$ (draw the graph of this function and observe that it has the height and width specified), and so if y is given, we have $x = \pm\frac{10}{\pi}\cos^{-1}(\frac{y}{5})$. The distance between these two x-coordinates is $\frac{20}{\pi}\cos^{-1}(\frac{y}{5})$, and this is our value for $w(y)$. Thus the total force is

$$\begin{aligned} F &= \int_0^5 \frac{20}{\pi}\rho g(100 - y)\cos^{-1}\left(\frac{y}{5}\right) dy \qquad\qquad (z = \tfrac{y}{5},\ dy = 5\,dz) \\ &= \frac{100}{\pi}\rho g \int_0^1 (100 - 5z)\cos^{-1}(z)\, dz, \end{aligned}$$

and we can integrate this using parts with $u = \cos^{-1} z$, $dv = (20 - z)\, dz$ to get

$$\begin{aligned} F &= \frac{500}{\pi}\rho g \int_0^1 \underbrace{\cos^{-1}(z)}_{u}\underbrace{(20 - z)\, dz}_{dv} \\ &= \frac{500}{\pi}\rho g\Big[\cos^{-1}(z)(20z - \tfrac{1}{2}z^2)\Big]_0^1 - \frac{500}{\pi}\rho g \int_0^1 \frac{20z - \frac{1}{2}z^2}{-\sqrt{1 - z^2}}\, dz. \end{aligned}$$

Observe that $\cos^{-1}(1) = 0$, so the first term vanishes at both $z = 0$ and $z = 1$, giving

$$F = \frac{500}{\pi}\rho g \int_0^1 \frac{20z - \frac{1}{2}z^2}{\sqrt{1 - z^2}}\, dz.$$

To evaluate the first part of the integral we observe that

$$\int_0^1 \frac{z}{\sqrt{1 - z^2}}\, dz = \Big[-\sqrt{1 - z^2}\Big]_0^1 = 0 - (-1) = 1.$$

For the second part, we put $z = \sin\theta$, $dz = \cos\theta\, d\theta$, $\sqrt{1 - z^2} = \cos\theta$, and get

$$\begin{aligned} \int_0^1 \frac{z^2}{\sqrt{1 - z^2}}\, dz &= \int_0^{\pi/2} \frac{\sin^2\theta}{\cos\theta}\cos\theta\, d\theta = \int_0^{\pi/2} \sin^2\theta\, d\theta \\ &= \int_0^{\pi/2} \frac{1 - \cos 2\theta}{2}\, d\theta = \frac{\pi}{4} - \Big[\frac{1}{4}\sin 2\theta\Big]_0^{\pi/2} = \frac{\pi}{4}. \end{aligned}$$

Putting it all together gives

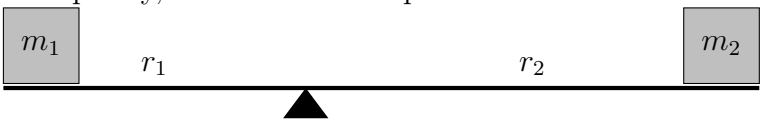
$$F = \frac{500}{\pi}\rho g\Big(20 \cdot 1 - \frac{1}{2}\cdot\frac{\pi}{4}\Big) = \rho g\Big(\frac{1000}{\pi} - \frac{125}{2}\Big) \approx 3.07 \times 10^8\,\mathrm{N}.$$

Lecture 53 Center of mass

Stewart §8.3

53.1. Point masses in one dimension

Suppose we have a rigid plank supported on a fulcrum, with two masses m_1 and m_2 placed on opposite sides of the fulcrum, at distances r_1 and r_2, as shown in the picture. For simplicity, assume that the plank is massless.



We want to determine conditions on m_1, m_2, r_1, r_2 such that the system balances; that is, if the masses are initially at rest, then they remain at rest.[45] Use a coordinate system in which the initial height of the masses is 0; then their initial potential energy is 0, and so is their initial kinetic energy. Let $\theta(t)$ be the angle made by the plank with the horizontal at time t, and let $v_1(t), v_2(t)$ be the velocities of the two masses at time t. Then the total energy at time t is

$$E(t) = \underbrace{\frac{1}{2}m_1v_1^2 + \frac{1}{2}m_2v_2^2}_{\text{kinetic energy}} + \underbrace{m_1g(-r_1\sin\theta) + m_2g(r_2\sin\theta)}_{\text{(gravitational) potential energy}}$$

By conservation of energy, we must have $E(t) = 0$ for all t. Observe that if $m_1r_1 = m_2r_2$, then the potential energy is 0 no matter what θ is. Thus the kinetic energy must also be 0, which means that $m_1v_1^2 + m_2v_2^2 = 0$, but this is only possible if $v_1 = v_2 = 0$ for all t. Thus we have proven the following.

Proposition 53.1 (Law of the lever)**.** *If $m_1r_1 = m_2r_2$ in the situation above, then the system is balanced and remains in equilibrium, motionless.*

Exercise 53.2. Prove that if $m_1r_1 > m_2r_2$, then the plank will rotate counterclockwise – m_1 will sink and m_2 will rise – and vice versa if $m_1r_1 < m_2r_2$.

Now suppose we have a finite set of masses $m_1, m_2, \ldots, m_n$ at locations $x_1, x_2, \ldots, x_n$ along the plank, and that the fulcrum is located at position $\bar{x}$. Note that these values can be either positive or negative, since we are not specifying which side of the fulcrum each mass lies on, and we do not require the fulcrum to lie at 0. In particular, the location of the mass m_i *relative to the fulcrum* is given not by x_i, but by $x_i - \bar{x}$, with a negative value indicating that the mass is to the left of the fulcrum, and a positive value indicating that it is to the right.

Repeating the same reasoning as before, we see that the system will be in equilibrium if and only if the values of $m_i, x_i, \bar{x}$ have the property that the change in potential energy is 0 no matter what value θ takes. If the plank is at angle θ, then mass m_i is at height $(x_i - \bar{x})\sin\theta$, and thus the total change in potential energy is

$$\sum_{i=1}^{n} m_i(x_i - \bar{x})\sin\theta.$$

[45] I learned the argument given here from a short write-up by Peter McLoughlin.

We need this to vanish for all θ; equivalently, we require that

$$0 = \sum_{i=1}^{n} m_i(x_i - \bar{x}) = \Big(\sum_{i=1}^{n} m_i x_i\Big) - \Big(\sum_{i=1}^{n} m_i\Big)\bar{x}.$$

Thus we have proved the following.

Proposition 53.3. *The plank with masses $m_1, \dots, m_n$ placed at positions $x_1, \dots, x_n$ is in equilibrium if and only if the fulcrum is placed at position*

$$\bar{x} = \frac{\sum_{i=1}^{n} m_i x_i}{\sum_{i=1}^{n} m_i}. \tag{53.1}$$

The point $\bar{x}$ where the fulcrum must be placed to ensure equilibrium is called the *center of mass* of the system; it is also called the *center of gravity* or the *centroid*. The numerator in (53.1) is called the *moment*[46] *of the system about the origin*, and represents the tendency that the system would have to rotate clockwise (if the moment is positive) or counterclockwise (if the moment is negative) **if** we were to place the fulcrum at the origin. The denominator in (53.1) is of course the total mass of the system.

53.2. Center of mass in two dimensions

Now suppose we have a system of masses $m_1, \dots, m_n$ located in the plane $\mathbb{R}^2$, at positions $(x_1, y_1), \dots, (x_n, y_n)$. We would like to find the point $(\bar{x}, \bar{y})$ with the property that if our masses are placed on a flat (massless) surface, which is then placed on a fulcrum located at $(\bar{x}, \bar{y})$, then the system would balance in equilibrium. This point $(\bar{x}, \bar{y})$ will again be called the center of mass, or centroid, of the system.

First imagine that we support the surface not on a fulcrum placed at a single point, but on a rod that is oriented parallel to the y-axis, so that rotation is only possible around this axis. Then the y-coordinates of the masses are irrelevant, and all that matters is their x-coordinates. As in the previous section, we see that

$$\bar{x} = \frac{M_y}{m} \qquad \text{where } M_y = \sum_{i=1}^{n} m_i x_i \text{ and } m = \sum_{i=1}^{n} m_i. \tag{53.2}$$

We call M_y the *moment around the y-axis*. A similar argument reveals that

$$\bar{y} = \frac{M_x}{m} \qquad \text{where } M_x = \sum_{i=1}^{n} m_i y_i \text{ and } m = \sum_{i=1}^{n} m_i, \tag{53.3}$$

where M_x is the *moment around the x-axis*.

Example 53.4. Suppose we place three small objects with masses 1, 2, and 4 at positions $(0, 1)$, $(1, 1)$, and $(2, 3)$, respectively. Then the two moments are

$$M_y = 1 \cdot 0 + 2 \cdot 1 + 4 \cdot 2 = 10 \quad \text{and} \quad M_x = 1 \cdot 1 + 2 \cdot 1 + 4 \cdot 3 = 15.$$

Since $m = 1 + 2 + 4 = 7$, we see that the center of mass is at $(\frac{10}{7}, \frac{15}{7})$.

[46] If we multiply the moment by the gravitational constant g, we get the *moment of force*, or *torque*.

Remark 53.5. If we have a set of masses with moments M_y and M_x, then (53.2) and (53.3) can be rewritten as

$$M_y = m\bar{x} \quad \text{and} \quad M_x = m\bar{y};$$

the moments of the entire set of masses are the same as the moments of a single point mass located at the centroid $(\bar{x}, \bar{y})$ with mass m. In other words, the moments are unchanged if we move all of the masses to the centroid.

Remark 53.6. Observe that if we have two sets S_1 and S_2 of masses, and compute their moments $M_y(S_1)$ and $M_y(S_2)$ independently, then the moment of the overall system comprising all the masses is given by $M_y(S_1 \cup S_2) = M_y(S_1) + M_y(S_2)$. A similar result holds for the the moments around the x-axis.

53.3. Continuous objects

Now we consider the continuous case – a plate with uniform density ρ. Let $R \subset \mathbb{R}^2$ be the region describing the shape of the plate, and let $C(R) \in \mathbb{R}^2$ denote the centroid of R; that is, the point at which a fulcrum must be placed in order for the plate to balance. As before, we have $C(R) = (M_y(R)/m, M_x(R)/m)$, where m is the total mass of the plate and $M_y(R)$, $M_x(R)$ are the moments of R around the y- and x-axes, respectively. The difference is that this time we do not have a formula for M_y and M_x; we must derive one. To do this, we assume that the centroid and moments obey the following principles.

(1) *Symmetry:* If R is symmetric around a line ℓ, then $C(R)$ lies on ℓ.
(2) *Replacement:* If all of the mass of R is moved to a single point located at $C(R)$, then the moments M_y and M_x are unchanged.
(3) *Additivity:* If R_1 and R_2 are disjoint regions, then $M_y(R_1 \cup R_2) = M_y(R_1) + M_y(R_2)$, and similarly for M_x.

Observe that the second and third principles are analogues of Remarks 53.5 and 53.6, respectively.

For simplicity we first assume that the plate is described by the set

$$R = \{(x, y) : a \le x \le b, 0 \le y \le f(x)\} = \bigcup_{x \in [a,b]} \{x\} \times [0, f(x)] \subset \mathbb{R}^2$$

for some function $f\colon [a, b] \to [0, \infty)$. As usual we approximate R by taking $n \in \mathbb{N}$ large, dividing $[a, b]$ into n intervals of length $\Delta x = (b-a)/n$ with endpoints $x_i = a + i\Delta x$, and considering the union of rectangles $R_i := [x_{i-1}, x_i] \times [0, f(\bar{x}_i)]$, where $\bar{x}_i = \frac{1}{2}(x_{i-1}, x_i)$. As long as f is continuous, it is reasonable to expect that

$$M_y(R) = \lim_{n\to\infty} M_y\Big(\bigcup_{i=1}^{n} R_i\Big) \quad \text{and} \quad M_x(R) = \lim_{n\to\infty} M_x\Big(\bigcup_{i=1}^{n} R_i\Big). \tag{53.4}$$

By the third principle above (additivity), we have

$$M_y\Big(\bigcup_{i=1}^{n} R_i\Big) = \sum_{i=1}^{n} M_y(R_i) \quad \text{and} \quad M_x\Big(\bigcup_{i=1}^{n} R_i\Big) = \sum_{i=1}^{n} M_x(R_i). \tag{53.5}$$

The centroid of R_i lies at $(\bar{x}_i, \frac{1}{2}f(\bar{x}_i))$ by the first principle above (symmetry), and the mass of R_i is $\rho f(\bar{x}_i)\Delta x$. Thus the second principle above (replacement) gives

$$M_y(R_i) = \rho \bar{x}_i f(\bar{x}_i)\Delta x \quad \text{and} \quad M_x(R_i) = \rho \cdot \frac{1}{2} f(\bar{x}_i)^2 \Delta x. \tag{53.6}$$

Combining (53.4)–(53.6) gives

$$M_y(R) = \lim_{n\to\infty} \sum_{i=1}^{n} \rho \bar{x}_i f(\bar{x}_i)\Delta x = \rho \int_a^b x f(x)\,dx,$$
$$M_x(R) = \lim_{n\to\infty} \sum_{i=1}^{n} \rho \cdot \frac{1}{2} f(\bar{x}_i)^2 \Delta x = \rho \int_a^b \frac{1}{2}\big(f(x)\big)^2\,dx.$$

Since $m = \rho \int_a^b f(x)\,dx$, we conclude that the centroid of R has coordinates given by

$$\bar{x} = \frac{\int_a^b x f(x)\,dx}{\int_a^b f(x)\,dx} \quad \text{and} \quad \bar{y} = \frac{\int_a^b \frac{1}{2} f(x)^2\,dx}{\int_a^b f(x)\,dx}. \tag{53.7}$$

Example 53.7. We find the centroid of a semicircular region R with radius r. For concreteness take the upper half of the circle centered at the origin. Since R is symmetric around the y-axis we immediately have $\bar{x} = 0$. For $\bar{y}$, we describe the region via $f(x) = \sqrt{r^2 - x^2}$ on $[-r, r]$ and observe that $\int_{-r}^{r} f(x)\,dx = \frac{1}{2}\pi r^2$ by the formula for circle area, so that (53.7) gives

$$\bar{y} = \frac{\int_{-r}^{r} \frac{1}{2} f(x)^2\,dx}{\int_{-r}^{r} f(x)\,dx} = \frac{1}{\pi r^2}\int_{-r}^{r} (r^2 - x^2)\,dx = \frac{2}{\pi r^2}\int_0^r (r^2 - x^2)\,dx$$
$$= \frac{2}{\pi r^2}\Big[r^2 x - \frac{1}{3}x^3\Big]_0^r = \frac{2}{\pi r^2}\Big(r^3 - \frac{1}{3}r^3\Big) = \frac{2}{\pi r^2}\cdot\frac{2}{3}r^3 = \frac{4r}{3\pi}.$$

Thus the centroid of the region is located at $(0, \frac{4r}{3\pi})$.

If we consider a more general region described as

$$R = \{(x, y) : x \in [a, b], y \in [g(x), f(x)]\}, \tag{53.8}$$

where $g, f\colon [a, b] \to \mathbb{R}$ are continuous functions with $g \le f$, then similar arguments give

$$\bar{x} = \frac{1}{A}\int_a^b x(f(x) - g(x))\,dx \quad \text{and} \quad \bar{y} = \frac{1}{A}\int_a^b \frac{1}{2}\big(f(x)^2 - g(x)^2\big)\,dx, \tag{53.9}$$

where $A = \int_a^b (f(x) - g(x))\,dx$ is the area of R.

53.4. Pappus's theorem

Theorem 53.8 (Pappus's theorem)**.** *Let R be a region in the plane that lies entirely to one side of some line ℓ, and let V be the volume of the solid of revolution formed by rotating R around the line ℓ. Let A be the area of R and let d be the distance traveled by the centroid of R as it revolves around ℓ. Then $V = Ad$.*

Proof. Without loss of generality, take ℓ to be the y-axis, and let R be given in terms of functions g, f as in (53.8).[47] Recall how we find volume by cylindrical shells:

[47] If R cannot be written in this form, you first need to decompose it as a finite union of such regions.

(1) the area of the annulus with inner radius p and outer radius q is $\pi q^2 - \pi p^2 = \pi(q^2 - p^2) = 2\pi m(q - p)$, where $m = \frac{p+q}{2}$;
(2) thus the volume of the cylindrical shell formed by rotating the rectangle $[x_{i-1}, x_i] \times [g(\bar{x}_i), f(\bar{x}_i)]$ around the y-axis is $2\pi\bar{x}_i(f(\bar{x}_i) - g(\bar{x}_i))$, where $\bar{x}_i$ is the midpoint of $[x_{i-1}, x_i]$;
(3) the volume of R is

$$V = \lim_{n\to\infty} \sum_{i=1}^{n} 2\pi\bar{x}_i(f(\bar{x}_i) - g(\bar{x}_i))\Delta x = \int_a^b 2\pi x(f(x) - g(x))\,dx,$$

where $\Delta x = (b - a)/n$ and $x_i = a + i\Delta x$.

Recalling the first half of (53.9), we have

$$V = 2\pi \int_a^b x(f(x) - g(x))\,dx = 2\pi A\bar{x},$$

and since $2\pi\bar{x}$ is the distance d traveled by the centroid as it rotates, this proves the theorem. □

Example 53.9. Consider a disc with center $(R, 0)$ and radius r, where $0 < r < R$. The centroid of the disc is its center (by the symmetry principle), and the corresponding solid of revolution is a torus, whose volume is

$$V = Ad = (\pi r^2)(2\pi R) = 2\pi^2 r^2 R.$$

Lecture 54 *Probability

Stewart §8.5

A *random variable* is a quantity that depends on some random factors. For example, any of the following could be described by a random variable:

- W = the sum of the numbers on a pair of dice after they are rolled;
- X = the number of students who come to class on a randomly selected day;
- Y = the height of a randomly selected person;
- Z = the amount of rainfall during a randomly selected week.

The first two examples above, W and X, are *discrete* random variables, meaning that we can make a list of all the possible values they can take, and then assign a probability to each individual value. The last two examples, Y and Z, are *continuous* random variables, meaning that they can take a continuum of values; instead of listing all possible values, we allow the value to be any real number. (Of course, some parts of the real line may have zero probability: both Y and Z will have probability 1 of being ≥ 0.)

The *probability distribution* of a random variable tells us the probabilities associated to the different values it can take. For a discrete random variable, we can describe the distribution by simply listing the probabilities associated to each of the possible values: for example, if W is the sum of the numbers on a pair of dice, then $\mathbb{P}(W = 2) = \frac{1}{36}$ because the $6 \times 6 = 36$ equally likely outcomes include exactly one that produces a sum of 2, and we can similarly list $\mathbb{P}(W = 3)$, $\mathbb{P}(W = 4)$, and so on.

For a continuous random variable X, we must do something else, since we cannot list all the possible values. Rather, we describe the distribution by a *probability density function*; this is a function $f\colon \mathbb{R} \to [0,\infty)$ with the property that

$$\underbrace{\mathbb{P}(a \le X \le b)}_{\text{probability that } a \le X \le b} = \int_a^b f(x)\,dx \text{ for every } a < b \in \mathbb{R}.$$

The interpretation of this is that if we make n independent observations of the random variable X, then the proportion of observations for which $a \le X \le b$ will converge to $\int_a^b f(x)\,dx$ as $n \to \infty$ (this is called the *law of large numbers*).

Probability density functions are required to have $f(x) \ge 0$ for all x, and to satisfy $\int_{-\infty}^{\infty} f(x)\,dx = 1$. The first condition guarantees that probabilities are always ≥ 0, and the second condition guarantees that the probability that *something* happens is equal to 1.

Example 54.1. An *exponentially distributed random variable* takes only positive values and has a probability density function (PDF) that decays exponentially as $x \to \infty$; that is, $f(x) = 0$ for $x < 0$, and there are $c, \lambda > 0$ such that $f(x) = ce^{-\lambda x}$ for $x \ge 0$. Random variables like this are often used to model *waiting time* phenomena in which X represents the amount of time until the next occurrence of a particular event, such as my dog barking at a car that drives past my house.

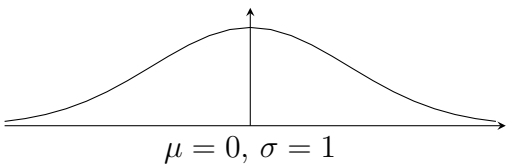
The value of λ reflects the rate at which the PDF decays; smaller λ means that X is more likely to take larger values, while larger λ means that it is more likely to take smaller values. We need to determine the value of c to guarantee that f is normalized: $\int_{-\infty}^{\infty} f(x)\,dx = 1$. From the definition of f we get

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^{\infty} f(x)\,dx = \lim_{t\to\infty} \int_0^t ce^{-\lambda x}\,dx = \lim_{t\to\infty} \Big[-\frac{c}{\lambda} e^{-\lambda x} \Big]_0^t = \frac{c}{\lambda}.$$

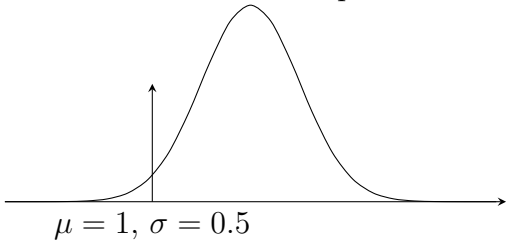
Thus we must put $c = \lambda$, obtaining a PDF of $f(x) = \lambda e^{-\lambda x}$. Then the probability that X lies in an interval $[a,b]$ for $a \ge 0$ is given by

$$\mathbb{P}(a \le X \le b) = \int_a^b \lambda e^{-\lambda x}\,dx = \Big[-e^{-\lambda x} \Big]_a^b = e^{-\lambda b} - e^{-\lambda a}.$$

Example 54.2. A *normally distributed random variable* can take both positive and negative values and has a PDF given by $f(x) = \frac{1}{A} e^{-(x-\mu)^2/(2\sigma^2)}$, where μ is the *mean* of the distribution, σ is the *standard deviation*, and $A = \int_{-\infty}^{\infty} e^{-(x-\mu)^2/(2\sigma^2)}\,dx$ is the normalizing constant that guarantees the property $\int_{-\infty}^{\infty} f(x)\,dx = 1$. This is also called a *Gaussian distribution* or sometimes informally a *bell curve* due to its shape.



$\mu = 0,\ \sigma = 1$



$\mu = 1,\ \sigma = 0.5$

It is possible to prove that $A = \sqrt{2\pi\sigma^2}$, but this requires tools that we have not yet developed (recall that we cannot find $\int e^{-x^2}\,dx$ explicitly). Note that varying μ has the effect of sliding the graph of f to the left or right; the graph is symmetric around the line $x = \mu$. Varying σ has the effect of squeezing or stretching it horizontally, so that when σ is small more of the area under the graph is concentrated closer to the line $x = \mu$, and when σ is large more area is located further away from this line. Thus σ quantifies how likely the value of X is to be close to the mean μ.

In the example of the normal distribution, the symmetry of the PDF makes it reasonable to interpret μ as an average, or mean, since for every range of values greater than μ, there is a range of values on the opposite side of μ that are achieved with equal probability. But how do we find the mean of an arbitrary random variable?

First recall that if we measure a random variable N times and record the results of the measurements as $X_1, \dots, X_N$, then the *observed* average value is

$$\bar{X} = \frac{1}{N}\sum_{j=1}^{N} X_j.$$

Suppose for a moment that we have a discrete random variable, which only takes values from a finite set $\{x_1, \dots, x_n\}$. Then for each i we can write k_i for the number of times that we see the value x_i appear in the list $(X_1, \dots, X_N)$, and obtain

$$\bar{X} = \frac{1}{N}\sum_{j=1}^{N} X_j = \frac{1}{N}\sum_{i=1}^{n} k_i x_i. \tag{54.1}$$

Now return to the case of a continuous random variable. Suppose we fix a large $t > 0$, a large $n \in \mathbb{N}$, and split the interval $[-t, t]$ into n intervals of length $\Delta x = 2t/n$ by putting $x_i = -t + i\Delta x$. If we measure the random variable N different times, we expect $\approx N\int_{x_{i-1}}^{x_i} f(x)\,dx \approx Nf(x_i)\Delta x$ of these measurements to lie in the interval $[x_{i-1}, x_i]$. Thus (54.1) gives

$$\text{average of } X \approx \frac{1}{N}\sum_{i=1}^{n} Nf(x_i)\Delta x \cdot x_i = \sum_{i=1}^{n} x_i f(x_i)\Delta x.$$

Once again we recognize this as a Riemann sum, whose limit as $n \to \infty$ is $\int_{-t}^{t} xf(x)\,dx$. Taking a limit as $t \to \infty$, we see that the average value (mean) of the random variable X with probability density function f is given by

$$\mu = \int_{-\infty}^{\infty} xf(x)\,dx. \tag{54.2}$$

Exercise 54.3. Use the symmetry of the normal distribution to show that this agrees with the use of the notation μ there.

Remark 54.4. In light of Remark 49.20, you should be mildly uneasy (at least) with our casual use of the relationship $\int_{-\infty}^{\infty} xf(x)\,dx = \lim_{t\to\infty}\int_{-t}^{t} f(x)\,dx$. This works fine provided the improper integral $\int_{-\infty}^{\infty} xf(x)\,dx$ is convergent; however, if the improper integral is divergent then (54.2) is invalid, and in fact we must say that in this case the mean does not exist!

Exercise 54.5. Find $c > 0$ such that $f(x) = \frac{c}{1+x^2}$ is a probability density function, and show that in this case the improper integral in (54.2) is divergent.

Remark 54.6. The mean μ is sometimes called the *first moment*. Observe that it is given by the same integral that we used to compute the moment around the y-axis of a region in $\mathbb{R}^2$. Since $\int_{-\infty}^{\infty} f(x)\,dx = 1$ for a PDF, this means that the centroid of the region under the graph of f lies on the line $x = \mu$.

In probability theory one also needs to study *higher moments* such as $\int_{-\infty}^{\infty} x^2 f(x)\,dx$, $\int_{-\infty}^{\infty} x^3 f(x)\,dx$, and so on. As with the mean, these integrals may or may not be convergent, depending on which probability density function we consider.

Example 54.7. For the exponential distribution given by $f(x) = \lambda e^{-\lambda x}$, the mean is

$$\begin{aligned}\mu &= \int_0^{\infty} \lambda x e^{-\lambda x}\,dx = \lim_{t\to\infty} \Big[\lambda x \cdot (-\lambda^{-1} e^{-\lambda x})\Big]_0^t - \int_0^t \lambda(-\lambda^{-1}e^{-\lambda x})\,dx \\ &= \lim_{t\to\infty} -te^{-\lambda t} + \int_0^t e^{-\lambda x}\,dx = \lim_{t\to\infty} \Big[-\frac{1}{\lambda} e^{-\lambda x}\Big]_0^t = \frac{1}{\lambda}.\end{aligned}$$

Part VI. Differential equations

Lecture 55 Ideas and examples

Stewart §9.1 and §9.2

55.1. Real-world problems modeled by DEs

When we write down a model describing some kind of real-world situation in which our goal is to determine a particular function f, we often end up with a *differential equation* (DE) that contains both f and some of its derivatives. For example, this occurred when we considered the hanging cable problem and discovered that the equation $y = f(x)$ describing the catenary could be determined by first finding the arc length function $s(x)$, which satisfies the equation (50.8): $\frac{ds}{dx} = \frac{1}{a}\sqrt{a^2 + s^2}$. We were able to solve this by rewriting it as $\frac{dx}{ds} = a/\sqrt{a^2 + s^2}$ and then integrating with respect to s; note that this represents the simplest sort of differential equation, where the function to be determined (in this case $x(s)$) appears only on the LHS in terms of its derivative, and thus can be found by computing a single integral. Most of the DEs we encounter from now on will be more involved than this.

One instructive example arose last semester when we studied population growth. If $P(t)$ represents the size of a particular population at time t, then the simplest model describing how P evolves in time simply accounts for the change due to reproduction and death:. Write $k_r > 0$ for the rate at which reproduction happens, so that the population increase due to reproduction in a short time interval Δt is $k_r P(t)\Delta t$, and $k_d > 0$ for the rate at which members of the population die, so that the decrease due to death in time Δt is $-k_d P(t)\Delta t$. Thus

$$\frac{dP}{dt} = \lim_{\Delta t \to 0} \frac{\Delta P(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{k_r P(t)\Delta t - k_d P(t)\Delta t}{\Delta t} = (k_r - k_d)P(t).$$

Writing $k := k_r - k_d$, we see that the population function satisfies the differential equation

$$\frac{dP}{dt} = kP. \tag{55.1}$$

If $k_r > k_d$ then $k > 0$ and the population grows; if $k_r < k_d$ then $k < 0$ and the population shrinks. We saw last semester that (55.1) can be solved by dividing both sides by P and using logarithmic derivatives:

$$\frac{d}{dt}\ln P = \frac{P'}{P} = \frac{kP}{P} = k \quad \Rightarrow \quad \ln P(t) = kt + C \quad \Rightarrow \quad P(t) = C_0 e^{kt},$$

where $C_0 = e^C$ is a constant of integration which can take any value in $(0, \infty)$. (Here we assume that the population is positive; if $P(t) = 0$ then there is nothing to model.) To determine C_0 we need to know the value of the population at some point in time: if we know the population at some time t_0, then we have

$$P(t_0) = C_0 e^{kt_0} \quad \Rightarrow \quad C_0 = P(t_0)e^{-kt_0} \quad \Rightarrow \quad P(t) = P(t_0)e^{-kt_0}e^{kt} = P(t_0)e^{k(t-t_0)}.$$

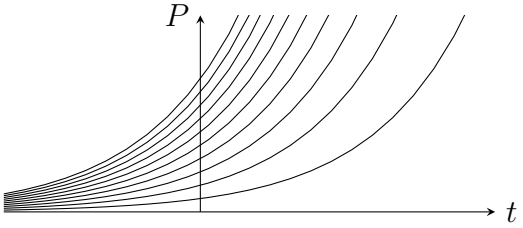
In particular, if we know the population at time 0 then we have

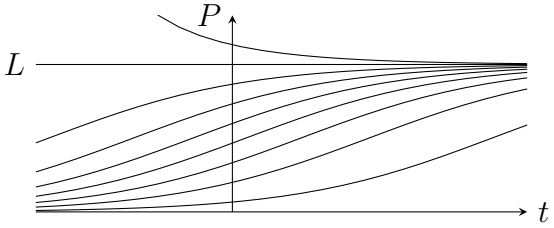
$$P(t) = P(0)e^{kt}.$$

The problem of finding $P(t)$ given that

(1) P satisfies (55.1) and
(2) $P(t_0)$ is known

is called an *initial value problem*. In an initial value problem, we expect to get a single function as the solution. If all we have is a differential equation but are not given the initial value, then we expect to get a whole family of solutions, such as $P(t) = C_0e^{kt}$; here the constant C_0 can be thought of as a parameter telling us which member of the family we are looking at. The picture at left shows some of the members of this family for the DE in (55.1) when $k > 0$.





Of course this model is not entirely realistic, because sooner or later the population will start to run out of resources and growth will slow. A more realistic model incorporates the *carrying capacity* of the environment in which the population lives, and has solutions with the shape shown in the right-hand figure. To describe it quantitatively, let L be the largest population that the environment can sustainably support; then we would like to have $P' \approx kP$ when P is small ($P \ll L$), while P'/P decreases for larger values of P, with P' becoming *negative* when $P > L$. This last requirement represents the idea that if the population is too large, then it will shrink towards the carrying capacity L. One DE that meets these requirements is the following *logistic DE* introduced by Verhulst in the 1840s:

$$\frac{dP}{dt} = kP\Big(1 - \frac{P}{L}\Big). \tag{55.2}$$

This is not quite so easy to solve as (55.1) was: dividing both sides by P does not help, because the RHS still contains P and so a straightforward integration does not solve the problem. We will see how to solve DEs like this in a few days. In the meantime, we can make some qualitative observations.

(1) $\frac{dP}{dt} = 0$ if and only if $P = 0$ or $P = L$. In particular, $P(t) = 0$ and $P(t) = L$ are both solutions of (55.2). Solutions such as these, where the function in question is constant, are called *equilibrium solutions*.
(2) $\frac{dP}{dt} > 0$ when $P \in (0, L)$, and $\frac{dP}{dt} < 0$ when $P \in (L, \infty)$. The picture suggests (and we will later prove) that $\lim_{t\to\infty} P(t) = L$ as long as the initial condition is positive.

Example 55.1. Consider a mass m attached to a spring, moving horizontally on a frictionless surface. Let $x(t)$ denote the displacement of the mass from its equilibrium position at time t. Then *Hooke's law* says that the spring exerts a force $F = -kx$ on

the mass, where $k > 0$ is a constant depending on the stiffness of the spring. Since $F = ma = m\ddot{x}$, the position function x satisfies the DE

$$\frac{d^2x}{dt^2} = -\frac{k}{m}x. \tag{55.3}$$

Exercise 55.2. Show that writing $\omega = \sqrt{\frac{k}{m}}$, the functions $x(t) = \sin(\omega t)$ and $x(t) = \cos(\omega t)$ are both solutions of (55.3). Can you think of any others?

Definition 55.3. The *order* of a differential equation is the order of the highest derivative that appears in the equation.

The population DEs (55.1) and (55.2) are first-order differential equations, while the spring equation (55.3) is second-order.

55.2. Explicit solutions using logarithms

The problem of finding the indefinite integral of a function f can be viewed as a differential equation $F' = f$, where the indefinite integral F is the solution of the DE. As we saw already, it is not always possible to find an elementary formula for the indefinite integral (such as when $f(x) = e^{-x^2}$) and thus one should not expect to always be able to write down an elementary formula for a differential equation. Indeed, in general the problem of solving differential equations is substantially more difficult than the problem of finding indefinite integrals, and there is no single technique that we can rely on to always lead us to the answer.

Remark 55.4. A significant item in the theory of differential equations is to determine whether or not a given DE even has a solution (*existence*), and if so, whether it is possible to have multiple solutions with the same initial conditions, or whether there is only one (*uniqueness*). Such existence and uniqueness results are not part of this course, however.

With that said, there are many classes of DEs for which it is possible to find a solution by reducing the problem to that of finding an indefinite integral. The population DE (55.1) illustrated this, and the technique used there works for any DE of the form

$$\frac{dy}{dx} = f(x)y, \tag{55.4}$$

where $f(x)$ is any given integrable function. Dividing both sides by y gives

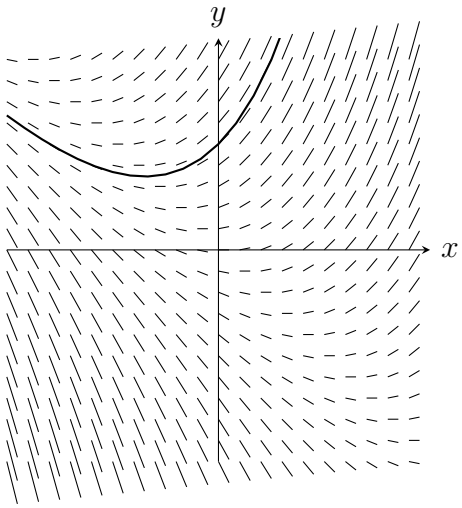
$$\frac{d}{dx}\log y = \frac{y'}{y} = f(x) \quad \Rightarrow \quad \log y(x) = \int f(x)\,dx \quad \Rightarrow \quad y(x) = e^{\int f(x)\,dx}.$$

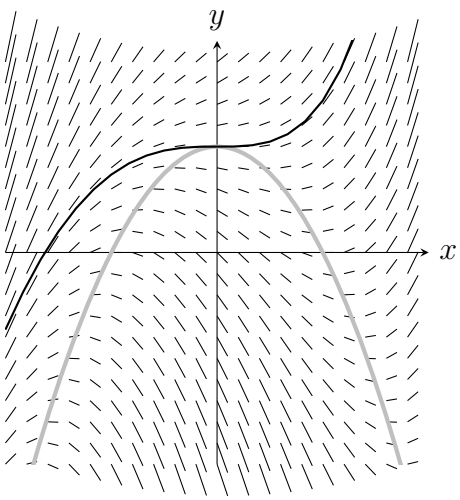
Example 55.5. To solve $y' = xy$ with $y(1) = 1$, we write it as $(\log y)' = x$, so $\log y = \frac{1}{2}x^2 + C$, and $\log y(1) = \log 1 = 0$ together with $\log y(1) = \frac{1}{2} + C$ gives $C = -\frac{1}{2}$, so the solution of the initial value problem is $y = e^{-1/2}e^{x^2/2}$.

55.3. *Qualitative solutions using direction fields

Suppose we are confronted with the differential equation $y' = x + y$. This does not immediately reduce to a simple integration like the examples we solved so far. But we can still at least sketch the general shape of the solutions by using a *direction field* (also

called a *slope field*), where at each point $(x, y) \in \mathbb{R}^2$ we put a short line segment with slope $x + y$, as shown in the first picture below. Then every solution of the DE will be tangent to these lines at all the points it passes through; the picture shows the specific solution with initial condition $y(0) = 1$.





This procedure works for every first-order DE of the form $y' = F(x, y)$; at each point (x, y) we put a short line segment with slope $F(x, y)$.

Example 55.6. The DE $y' = x^2 + y - 1$ has a direction field as in the second picture above. Observe that the points at which the direction field is horizontal can be found by solving $0 = y' = x^2 + y - 1$ to get $y = 1 - x^2$; this is the parabola in the picture. Below this parabola, solutions of the DE are decreasing; above it, solutions are increasing. The other curve in the picture is the solution with $y(0) = 1$.

Definition 55.7. A DE of the form $y' = F(x, y)$ is *autonomous* if the function F only depends on y, and not on x, so that it can actually be written as $y' = F(y)$. In the case when the independent variable is time, this can be thought of as "time-independence" of the system; the rule governing how y' is related to y does not change depending on t, but is the same for all time.

The two DEs above are not autonomous. The logistic DE $P' = kP(1 - \frac{P}{L})$ is autonomous. This has the consequence that its direction field looks the same if we shift it horizontally, and thus any solution curve remains a solution curve if we shift it left or right.

55.4. *Euler's method

As was the case when we computed definite integrals, there are situations in which it is better to take a numerical approach and try to find an approximate solution to an initial value problem (IVP). Such methods can become very sophisticated, but in this course we only consider the simplest one, called *Euler's method*.

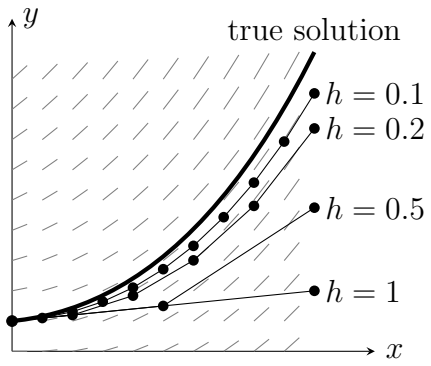
Roughly speaking, the idea of Euler's method is to move along the direction field in small steps of size h, where at each step we look at the direction field to see which way to move, then move that predetermined distance, and then look again at the direction field to get our instructions for the next step.

A little more precisely, the algorithm is this. Suppose we are given the IVP whose DE is $y' = F(x, y)$ and whose initial condition is $y(x_0) = y_0$. Fixing a step size h, we define (x_n, y_n) iteratively by

$$x_{n+1} = x_n + h, \qquad y_{n+1} = y_n + hF(x_n, y_n).$$

Thus the x-coordinate always increments by the step size, and the y-coordinate increments by the amount that it would change if F were constant and took the value that it does at (x_n, y_n).

The picture at right shows several applications of Euler's method to the initial value problem $y' = x + y$, $y(0) = 0.1$, with varying values of h. Observe that as h decreases it appears that the approximate solutions given by Euler's method are converging to the true solution. Whether this in fact occurs as $h \to 0^+$ is an important question in numerical analysis.



Lecture 56 *Separable differential equations

Stewart §9.3

56.1. Separable differential equations

Consider the first-order DE

$$\frac{dy}{dx} = \frac{x}{y}. \tag{56.1}$$

Based on our experience with the 'logarithm trick' for solving the DE $\frac{dy}{dx} = xy$ in Example 55.5, we might expect to get somewhere by multiplying both sides by y and writing $yy' = x$. In the previous example, the next step was to recognize that $\frac{y'}{y} = (\log y)'$. To proceed here, we need to replace $\log y$ with something that gives yy' upon differentiation by x.

After a little thought, you might realize that $\frac{d}{dx}(\frac{1}{2}y^2) = y\frac{dy}{dx}$, so (56.1) becomes $\frac{d}{dx}(\frac{1}{2}y^2)' = x$, or equivalently $\frac{d}{dx}(y^2) = 2x$, and integrating with respect to x gives $y^2 = x^2 + C$, so every solution of (56.1) has the form $y = \sqrt{x^2 + C}$ for some C. (Here we consider positive solutions; one could also consider negative solutions, but note that $y = 0$ is forbidden since y appears in the denominator of (56.1).)

To make this procedure into a more general strategy, let us replace the words "After a little thought" in the previous paragraph with the following more helpful argument:

after writing (56.1) as $y\frac{dy}{dx} = x$, integrate both sides with respect to x to obtain

$$\int y\frac{dy}{dx}\,dx = \int x\,dx.$$

By the substitution rule, the integral on the left-hand side can be rewritten as $\int y\,dy$, and thus we get

$$\int y\,dy = \int x\,dx,$$

which upon evaluation gives the same solution as before.

The general strategy, then, is this: given a first-order DE $\frac{dy}{dx} = F(x,y)$, we say that the equation is *separable* if the RHS can be written as $F(x,y) = g(x)f(y)$, where g depends only on x and f depends only on y. Then we have

$$\frac{dy}{dx} = g(x)f(y) \quad \Rightarrow \quad \frac{1}{f(y)}\frac{dy}{dx} = g(x) \quad \Rightarrow \quad \int \frac{dy}{f(y)} = \int \frac{1}{f(y)}\frac{dy}{dx}\,dx = \int g(x)\,dx,$$

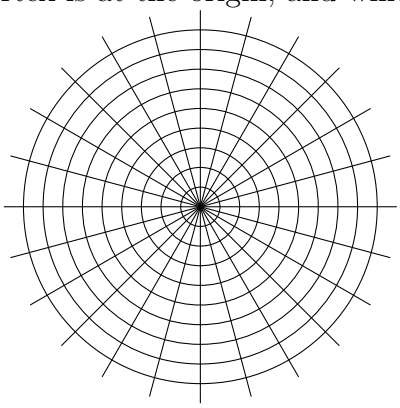
where the penultimate equality once again uses the substitution rule. Observe that our solutions of the DEs $y' = xy$ and $y' = x/y$ both used this strategy.

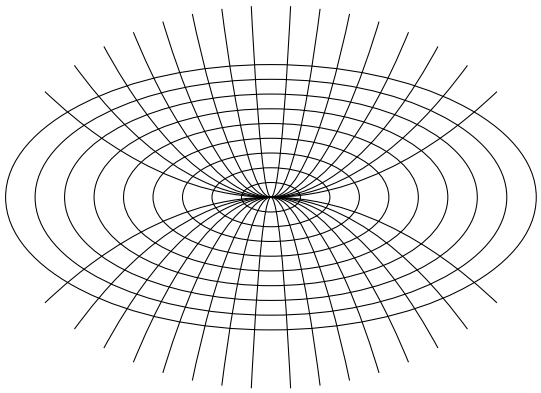
In general, evaluating the integrals is not quite the final step of the solution, because we still need to solve the resulting equation to find y in terms of x.

56.2. Orthogonal trajectories

Suppose we are given a family of curves, such as the set of lines through the origin in $\mathbb{R}^2$. It is occasionally of interest to find curves with the property that they intersect every curve in our original family at a right angle. For example, in an electrostatic field, the lines (curves) of force are always perpendicular to the lines (curves) of constant potential.

In the case of the set of lines through the origin, it is easy to see that the curves intersecting every such line orthogonally are just the circles centered at the origin, as shown in the picture at left. But what if we start with the family of parabolas whose vertex is at the origin, and which open up or down?





The family of parabolas just described comprises all the curves $y = kx^2$ where $k \in \mathbb{R}$. The picture at right suggests that the orthogonal trajectories for this family are ellipses. To confirm this, we first observe that if the point (x,y) lies on the parabola $y = kx^2$, then the slope of the parabola at this point is $2kx$. We can eliminate k by observing that $y = kx^2$ implies $k = y/x^2$, so the slope at this point is $2y/x$.

Now recall that two lines are perpendicular if and only if the product of their slopes is -1. Thus the slope of an orthogonal trajectory through (x, y) must be $-\frac{x}{2y}$ at this point. We conclude that a curve $x \mapsto y(x)$ describes an orthogonal trajectory if and only if it has the property that

$$\frac{dy}{dx} = -\frac{x}{2y}$$

everywhere. But this is a separable DE! So we can solve it by writing

$$2y\frac{dy}{dx} = -x \quad \Rightarrow \quad \int 2y\,dy = -\int x\,dx \quad \Rightarrow \quad y^2 = -\frac{1}{2}x^2 + C.$$

Thus the orthogonal trajectories to the family of parabolas are indeed the ellipses with equations $\frac{1}{2}x^2 + y^2 = C$.

56.3. Mixing problems

The following example gives another situation where a separable DE arises. Suppose mercury is leaking into a certain lake at a rate of γ g/min, and that water is flowing into the lake (from upstream) and out of the lake (downstream) at a rate of R L/min. (Since these rates are equal, the total volume of water in the lake remains constant.) Suppose also that at time $t = 0$, the lake is clean; there is no mercury in it. How much mercury is in the lake at time t?

Let V be the volume of the lake, which is constant. Let $y(t)$ be the mass of the mercury in the lake at time t, and let $\rho(t) = y(t)/V$ be the concentration. We make the simplifying assumption that mixing happens instantaneously, so that the concentration of mercury is the same throughout the lake. Then the rate at which mercury flows out of the lake is $\rho R = yR/V$ g/min, and since it flows in with rate γ g/min, we conclude that

$$\frac{dy}{dt} = \gamma - y\frac{R}{V}. \tag{56.2}$$

This is autonomous, and hence separable, so we can divide both sides by $\gamma - yR/V$ and then integrate, obtaining

$$\int \frac{dy}{\gamma - y\frac{R}{V}}\,dy = \int dt = t + C.$$

The integral on the LHS can be computed as follows:

$$\int \frac{dy}{\gamma - y\frac{R}{V}}\,dy = \frac{V}{R}\int \frac{dy}{\frac{\gamma V}{R} - y}\,dy = -\frac{V}{R}\ln\Big(\frac{\gamma V}{R} - y\Big),$$

and we conclude that

$$-\ln\Big(\frac{\gamma V}{R} - y\Big) = \frac{R}{V}t + C_1 \quad \Rightarrow \quad \frac{\gamma V}{R} - y = Ae^{-Rt/V},$$

so that the total amount of mercury in the lake at time t is given by

$$y(t) = \frac{\gamma V}{R} - Ae^{-Rt/V},$$

where A is a constant. To determine A we observe that at time $t = 0$ we have $0 = y = \frac{\gamma V}{R} - A$, so $A = \frac{\gamma V}{R}$, and we obtain

$$y(t) = \frac{\gamma V}{R}\Big(1 - e^{-Rt/V}\Big).$$

56.4. Solving the logistic model

The logistic DE $\frac{dP}{dt} = kP(1 - \frac{P}{L})$ from (55.2) is separable because the right-hand side does not depend on t, so we can divide both sides by $P(1 - \frac{P}{L})$ and then integrate:

$$\int \frac{dP}{P(1 - \frac{P}{L})} = \int k\, dt. \tag{56.3}$$

The integral on the RHS is easy. For the one on the left we use partial fractions to write

$$\begin{aligned}\int \frac{dP}{P(1 - \frac{P}{L})} = \int \frac{L}{P(L - P)}\, dP &= \int \Big(\frac{1}{P} + \frac{1}{L - P}\Big)\, dP \\ &= \ln P - \ln|L - P| = \ln \frac{P}{|L - P|},\end{aligned}$$

and thus (56.3) gives

$$\ln \frac{P}{|L - P|} = kt + C.$$

Taking the exponential of both sides gives

$$\frac{P}{|L - P|} = e^C e^{kt}.$$

Let $Q = e^C$ if $L > P$ and $Q = -e^C$ if $L < P$; then $\frac{P}{L-P} = Qe^{kt}$, and we can solve for P:

$$P = LQe^{kt} - PQe^{kt} \quad \Rightarrow \quad P(1 + Qe^{kt}) = LQe^{kt} \quad \Rightarrow \quad P = \frac{LQe^{kt}}{1 + Qe^{kt}} = \frac{L}{1 + Q^{-1}e^{-kt}}.$$

This gives the general solution of the logistic DE. To find a particular solution given an initial population P_0 at time 0, we observe that $P_0 = P(0) = L/(1 + Q^{-1})$, so $1 + Q^{-1} = L/P_0$, and thus $Q^{-1} = \frac{L}{P_0} - 1$. Thus it is convenient to write the solution of the IVP as

$$P(t) = \frac{L}{1 + Ae^{-kt}} \quad \text{where } A = \frac{L}{P_0} - 1 = \frac{L - P_0}{P_0}.$$

Remark 56.1. Recall that the logistic DE is autonomous; the RHS does not depend on the independent variable. The example above illustrates the general principle that *every* autonomous DE is separable, because it can be written as $\frac{dy}{dx} = f(y)$, and thus can in principle be solved by writing $\frac{1}{f(y)}\frac{dy}{dx} = 1$ and integrating to get $\int \frac{1}{f(y)}\, dy = x + C$. There are then two obstacles to turning this into a complete solution:

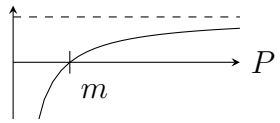
(1) the integral may be difficult or impossible to evaluate explicitly;
(2) it may be difficult or impossible to solve the resulting equation explicitly for y and write down a formula giving y in terms of x.

Lecture 57 *Other population models

Stewart §9.4

Beyond the logistic DE, there are other population models that are worth considering in certain situations. For example, suppose we expect that our population needs to be above a certain minimum size m to maintain itself, and that a population below this critical value will eventually die out. Then we might add another factor to the logistic DE that forces $\frac{dP}{dt}$ to be negative whenever $P < m$; we would like this factor to have the property that

- it is negative when $P < m$;
- it is positive when $P > m$;
- it is close to 1 for large values of P (when the population is well above the critical threshold, the original logistic DE should still be nearly accurate).



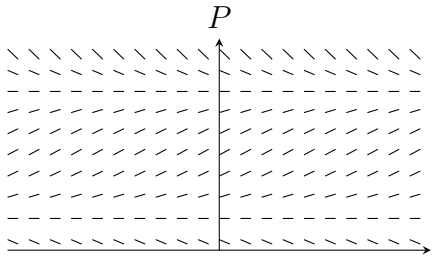
These suggest that its graph should have the general shape shown in the picture. An example of such a function is $(1 - \frac{m}{P})$, so we might multiply the RHS of the logistic DE by this factor and consider the DE

$$\frac{dP}{dt} = kP\Big(1 - \frac{P}{L}\Big)\Big(1 - \frac{m}{P}\Big). \tag{57.1}$$

We can rewrite the RHS as

$$\frac{dP}{dt} = \frac{k}{L}(L - P)(P - m).$$

To understand the behavior of this DE's solutions, we can draw its slope field.



This looks an awful lot like the slope field for the logistic DE:

(1) there are two equilibrium solutions, at $P = m$ and $P = L$;
(2) for $P \in (m, L)$, the population grows over time and appears to approach L;
(3) for $P > L$, the population decreases over time and appears to approach L.

The extra feature here is that there are positive values of P that are *below* the smaller equilibrium solution, and if the initial value of P lies in this range, then P decreases and eventually becomes 0, so the population goes extinct.

One could find an explicit solution of (57.1) by the same method as we used for the logistic DE, but we omit the details of this. Instead, we make the following observation: suppose that we write $y = P - m$ for the amount by which the population exceeds the critical threshold m. Then we have $P = y + m$ and can write

$$\frac{dy}{dt} = \frac{dP}{dt} = \frac{k}{L}(L - (y + m))y = \frac{k}{L}y(L - m - y) = \frac{k(L - m)}{L}y\Big(1 - \frac{y}{L - m}\Big).$$

But this means that y satisfies the original logistic DE! Granted, we need to change the parameters – the growth rate for y is $k(L-m)/L$ (instead of k) and the "carrying capacity" is $L-m$ (instead of L) – but this observation means that we can write any solution of (57.1) in terms of a solution for the logistic DE, and vice versa, so that in this sense the two problems are equivalent.

Remark 57.1. In fact, a similar change of variables (or substitution, if you prefer), can be used to turn any DE of the form $y' = ay^2 + by + c$ into the logistic DE, provided $b^2 - 4ac > 0$ so the DE has two equilibrium solutions.

So far, our population DEs have depended on parameters that affect the *quantitative* values of the solutions, but do not affect their *qualitative* form; that is, changing the parameters resulted in a new system that had the same number of equilibrium solutions, same overall description of types of solutions, etc. The next example is different.

Suppose $P(t)$ represents a population of fish that follows logistic growth but is also harvested at a constant rate c. Then the DE that it should satisfy is

$$\frac{dP}{dt} = kP\Big(1 - \frac{P}{L}\Big) - c. \tag{57.2}$$

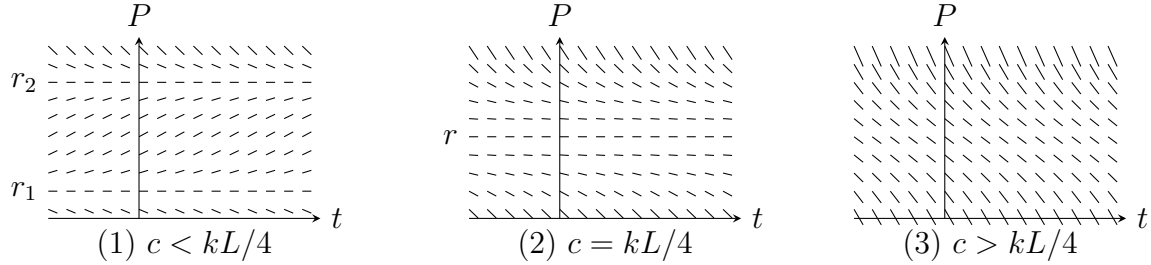
Again, we could solve this explicitly by dividing both sides by the quadratic on the RHS and then integrating, but instead of plunging blindly ahead with symbol manipulation, it is more instructive to take a moment and think about the overall picture. In particular, we want to understand for which values of P the RHS is positive, negative, and 0. Rewrite the DE as

$$\frac{dP}{dt} = -\frac{k}{L}P^2 + kP - c.$$

Observe that the RHS is a quadratic with discriminant[48] given by

$$k^2 - 4(-k/L)c = k^2 + 4ck/L = k(k + 4c/L).$$

Since $k > 0$, we see that the sign of the discriminant is the same as the sign of $k + 4c/L$. This is determined by how the harvesting rate c compares to $kL/4$. There are three cases, whose slope fields are shown in the pictures.



We describe these one by one.

(1) When $c < kL/4$, the discriminant $k(k + 4c/L)$ is positive and thus the quadratic $-\frac{k}{L}P^2 + kP - c$ has two real roots r_1 and r_2, which correspond to equilibrium solutions of (57.2). When P lies between these roots, the quadratic is positive so

[48]Recall that the *discriminant* of the quadratic polynomial $ax^2 + bx + c$ is $b^2 - 4ac$, which is the expression that appears under the square root in the quadratic formula for the roots of the polynomial. The polynomial has two real roots if the discriminant is positive, one if it is 0, and none if it is negative.

$\frac{dP}{dt} > 0$ and the population grows, converging to r_2. When $P > r_2$, the quadratic is negative and the population shrinks, again converging to r_2. When $P < r_1$, the population shrinks and eventually goes extinct.[49]

(2) When $c = kL/4$, the discriminant is 0 and thus the quadratic has exactly one real root r, so (57.2) has exactly one equilibrium solution $P = r$. When $P > r$ the quadratic is negative so the population shrinks, converging to r. When $P < r$ then quadratic is again negative and the population shrinks, then goes extinct.

(3) When $c > kL/4$, the discriminant is negative and the quadratic has no real roots. Thus no matter what value P takes, the quadratic is negative and the population shrinks, eventually going extinct.

The first case corresponds to a harvesting rate that is sustainable provided the initial population is between r_1 and r_2. The final case corresponds to an unsustainable harvesting rate that eventually wipes out the population. The second case is borderline and unstable; although $P = r$ is an equilibrium solution, any fluctuation below this population (due perhaps to some effects not included in the model) will eventually lead to extinction.

Remark 57.2. The phenomenon seen here, wherein the solutions of a DE change dramatically and exhibit qualitatively different behavior as a parameter (or family of parameters) is varied, is called a *bifurcation*, and we say that $kL/4$ is a *bifurcation value* for the parameter c. Such parameter values are extremely important in the study of differential equations and other models of real-world systems.

Lecture 58 *Linear differential equations

Stewart §9.5

58.1. Linear first-order DEs

Example 58.1. Consider the DE

$$xy' + y = 2x. \tag{58.1}$$

This is a first-order DE, but it is not written in a form where we can immediately determine if it is separable. To determine this, we need to solve for y' and get $y' = 2 - \frac{y}{x}$; since we cannot find functions $g(x)$ and $f(y)$ such that $g(x)f(y) = 2 - \frac{y}{x}$, this DE is not separable. So what do we do?

In the end, there is only one thing we know how to do: integrate. If we could integrate both sides of (58.1) with respect to x, then we might hope to once again end up with an equation that could be solved to determine y. To integrate the LHS, we can first use integration by parts with $u = x$, $v = y$ to write

$$\int \underbrace{x}_{u} \underbrace{y'\,dx}_{dv} = xy - \int y\,dx,$$

[49]Note that this looks just like the picture we gave for (57.1) above, and indeed, as suggested in Remark 57.1, the two DEs can be related by a change of variables.

and then obtain

$$\int (xy' + y)\,dx = \int xy'\,dx + \int y\,dx = xy, \tag{58.2}$$

so that

$$xy = \int 2x\,dx = x^2 + C,$$

and the solution of (58.1) is

$$y = x + \frac{C}{x}.$$

In retrospect it should not be surprising that the antiderivative in (58.2) has the form that it does; the LHS of (58.1) has two terms, one of which includes y' and the other of which includes y, so it is reasonable to expect that its antiderivative would have the form $R(x)y$ for some function R. Indeed, the product rule gives

$$(R(x)y)' = R(x)y' + R'(x)y,$$

and we see that (58.2) works because $R(x) = x$ has $R'(x) = 1$. Thus it would be reasonable to use this approach anytime we have a DE where

- the LHS has the form $R(x)y' + R'(x)y$ for some function $R(x)$, and
- the RHS depends only on x (not on y).

Now that we have a hammer, let's go looking for some nails; are there many DEs like this?

Definition 58.2. A *linear first-order differential equation* is a DE that can be written in the form

$$f(x)\frac{dy}{dx} + g(x)y = h(x) \tag{58.3}$$

for some functions f, g, h.

Taking $f(x) = x$, $g(x) = 1$, and $h(x) = 2x$ gives (58.1).

Remark 58.3. A linear first-order DE can always be rewritten in the form

$$\frac{dy}{dx} + P(x)y = Q(x) \tag{58.4}$$

by dividing both sides of (58.3) by $f(x)$ and writing $P(x) = g(x)/f(x)$ and $Q(x) = h(x)/f(x)$.

Example 58.4. The DE

$$x^2y' + 2xy = 1$$

is a linear first-order DE, for which we can use the approach described above: we want a function $R(x)$ for which the LHS is $(R(x)y)'$, and since $R(x) = x^2$ has $R'(x) = 2x$, we see that indeed we can rewrite the DE as

$$\frac{d}{dx}(x^2y) = 1,$$

and integrating gives

$$x^2y = x + C,$$

so that the solution is $y = \frac{1}{x} + \frac{C}{x^2}$.

In both of the examples we have done so far, the solution was to let $R(x)$ be the function in front of y'; however, this only worked because we got lucky (and because the examples were engineered to work out nicely). Indeed, in order for the linear first-order DE

$$f(x)y' + g(x)y = h(x)$$

to have a LHS that can be written as $(R(x)y)'$, we must have $R(x) = f(x)$ and $R'(x) = g(x)$; in other words, we must have $f'(x) = g(x)$. If the DE we are given does not have this property, then we need to do a little more work.

58.2. General solution to first-order linear DEs

Example 58.5. The DE

$$y' = x + y \tag{58.5}$$

appeared in §55.3, when we introduced direction fields to sketch the general shape of its solutions because we did not yet have the tools to solve it exactly. It is a linear first-order differential equation since we can rewrite it as

$$y' - y = x. \tag{58.6}$$

However, the LHS of this last equation cannot be written as $(R(x)y)'$, because we have $f(x) = 1$ and $g(x) = -1$, so $f'(x) \neq g(x)$. So what are we to do?

The solution is to observe that we can multiply the entire DE (58.6) by an *integrating factor* $I(x)$, which if we choose it correctly, will make the previous trick work out. So we rewrite (58.6) as

$$I(x)y' - I(x)y = I(x)x. \tag{58.7}$$

This is again a linear first-order DE with $f(x) = I(x)$, $g(x) = -I(x)$, and $h(x) = xI(x)$. We want to choose $I(x)$ so that $f'(x) = g(x)$; in other words, we need $I'(x) = -I(x)$. This is again a DE, but it is one we know how to solve! We can put $I(x) = e^{-x}$, and then (58.7) becomes

$$e^{-x}y' - e^{-x}y = xe^{-x}.$$

The LHS has antiderivative $e^{-x}y$, so we can integrate both sides with respect to x and get

$$e^{-x}y = \int xe^{-x}\,dx = -xe^{-x} + \int e^{-x}\,dx = -(x+1)e^{-x} + C.$$

Multiplying both sides by e^x gives the general solution

$$y = Ce^x - (x+1).$$

This technique works for any linear first-order DE as in (58.3). It is easiest if we first divide through by $f(x)$ to write the DE in the form (58.4), and then multiply through by a (not yet determined) integrating factor to obtain

$$I(x)y' + P(x)I(x)y = Q(x)I(x). \tag{58.8}$$

We want the LHS to be equal to $(I(x)y)'$, which is true if and only if I satisfies the DE

$$I'(x) = P(x)I(x).$$

We can solve this DE by dividing by $I(x)$ and then using logarithms:

$$\frac{I'}{I} = P \quad \Rightarrow \quad \ln I(x) = \int P(x)\,dx \quad \Rightarrow \quad I(x) = e^{\int P(x)\,dx}.$$

Then (58.8) gives

$$(Iy)' = Iy' + PIy = QI \quad \Rightarrow \quad Iy = \int QI\,dx \quad \Rightarrow \quad y = \frac{1}{I}\int QI\,dx.$$

This is a general procedure for solving linear first-order DEs. Observe that the process involves two indefinite integrals: one to find $\ln I$, and a second to find y. In the first of these, we can take the constant of integration to be any value we like; it is enough to take $\ln I$ to be *any* antiderivative of P. In the second integral, on the other hand, we need to include the constant of integration, because it is ultimately determined by the initial condition of the DE.

Example 58.6. Consider the DE

$$y' + 3x^2 y = 6x^2.$$

Multiplying through by an unknown integrating factor I gives

$$Iy' + 3x^2 Iy = 6x^2 I.$$

We want to choose I such that $I' = 3x^2 I$, so

$$\log I = \int 3x^2\,dx = x^3 \quad \Rightarrow \quad I = e^{x^3}.$$

Thus the second form of the DE gives

$$(e^{x^3}y)' = e^{x^3}y' + 3x^2 e^{x^3} y = 6x^2 e^{x^3},$$

and we conclude that

$$e^{x^3}y = \int 6x^2 e^{x^3}\,dx = 2e^{x^3} + C.$$

Thus the solution of the DE is

$$y = e^{-x^3}(2e^{x^3} + C) = 2 + Ce^{-x^3}.$$

58.3. Another solution of the logistic DE

We already solved the logistic DE $P' = kP(1 - P/L)$ in §56.4, but just for fun let's do it again, via a different approach. Let $P(t)$ be a solution of the logistic DE, and define a new function $y(t)$ by $y = 1/P$. Then we have

$$y' = -\frac{P'}{P^2} = -\frac{kP - \frac{k}{L}P^2}{P^2} = -\frac{k}{P} + \frac{k}{L} = -ky + \frac{k}{L},$$

so $y(t)$ is a solution of the first-order linear DE

$$y' + ky = \frac{k}{L}.$$

This can be solved by the method introduced in this lecture; multiplying by an integrating factor I gives

$$Iy' + Iky = \frac{k}{L}I, \tag{58.9}$$

and we want I to satisfy $I' = Ik$, so we choose $I(t) = e^{kt}$. Then the left-hand side of (58.9) is $\frac{d}{dt}(e^{kt}y)$, and integrating both sides of (58.9) gives

$$e^{kt}y = \int \frac{k}{L}e^{kt}\,dt = \frac{1}{L}e^{kt} + C.$$

Multiplying through by e^{-kt} gives

$$y = \frac{1}{L} + Ce^{-kt},$$

and since $y = 1/P$ we see that the solution to the logistic DE is given by

$$P(t) = \frac{1}{y(t)} = \frac{1}{\frac{1}{L} + Ce^{-kt}} = \frac{L}{1 + CLe^{-kt}},$$

which agrees with the solution in §56.4 (by putting $Q = (CL)^{-1}$).

Lecture 59 Coupled differential equations

Stewart §9.6

59.1. Predator-prey models

Before we leave our discussion of differential equations, we consider two more examples, starting with a population model. This time instead of considering a single population, we consider two populations that interact with each other as predator and prey.

For concreteness, let $R(t)$ represent the population of rabbits in a given area, and $W(t)$ the population of wolves. We suppose that if there were no wolves, then the rabbits would reproduce according to the simple population growth DE $\frac{dR}{dt} = kR$, where $k > 0$. On the other hand, if there were no rabbits, then the wolves would have no food source and their population would decay following the DE $\frac{dW}{dt} = -rW$, where again $r > 0$.

Each of these DEs is easy to solve on its own. Things get interesting (and harder!) when we consider the interaction between the two populations. If $R > 0$ and $W > 0$, then some of the rabbits will be eaten by wolves, which decreases $\frac{dR}{dt}$ and increases $\frac{dW}{dt}$. A reasonable assumption is that the contribution to the derivatives due to predation is proportional to RW, since this number represents the number of possible rabbit-wolf pairs. Thus we arrive at the *Lotka–Volterra equations*

$$\frac{dR}{dt} = kR - aRW, \qquad \frac{dW}{dt} = -rW + bRW, \tag{59.1}$$

where $a, b, k, r > 0$ are parameters determined by the physical characteristics of the populations, their environment, and their interactions.

Remark 59.1. The DE (59.1) is not a single DE, but rather two DEs coupled together. This kind of situation arises very often in real-world models, and has the potential to

increase the complexity of the situation tremendously. In particular, we should not expect to be able to write down explicit formulas for the solutions to such systems.

It turns out that autonomous systems of two DEs can be more or less completely understood at a qualitative level, similarly to our qualitative analysis of the various population models in §57, and the range of possible behaviors are very limited. However, with three or more DEs, there is the possibility of *chaotic* behavior, which has the appearance (in a way that can be made precise) of being nearly entirely random over long time scales, despite the fact that it is governed by deterministic equations.

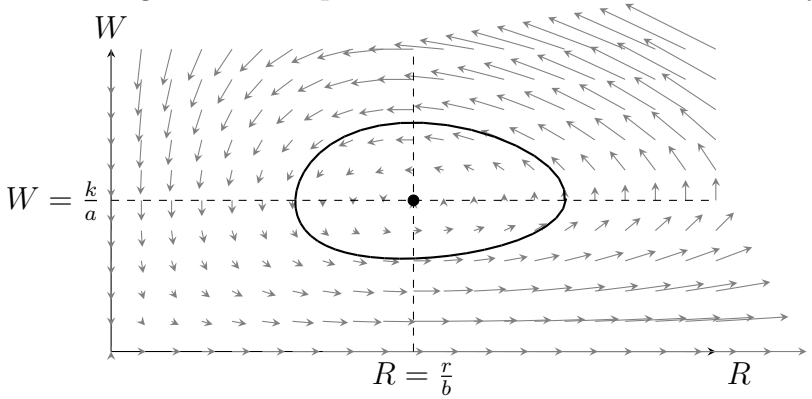
To understand the qualitative behavior of the Lotka–Volterra model, it is useful to find equilibrium solution(s), and more generally to find in which regions R and W are decreasing and increasing; we should also consider any special cases where the situation simplifies.

To find any equilibria, we see that

$$\frac{dR}{dt} = 0 \quad \Leftrightarrow \quad kR = aRW \quad \Leftrightarrow \quad R = 0 \text{ or } W = \frac{k}{a},$$
$$\frac{dW}{dt} = 0 \quad \Leftrightarrow \quad rW = bRW \quad \Leftrightarrow \quad W = 0 \text{ or } R = \frac{r}{b}.$$

Thus there are exactly two equilibrium solutions: the trivial solution where $R = W = 0$ (no rabbits, no wolves), and a nontrivial solution $W = k/a$, $R = r/b$.

To proceed further we draw an analogue of the direction field. The difference is that this time the line segment we draw at the point (R, W) has direction given by $(\frac{dR}{dt}, \frac{dW}{dt})$, and since this can be pointed in any direction (not just into the first or fourth quadrants, as was the case for our earlier direction fields) we will put arrowheads at the end of each of the line segments. The picture we obtain is called a *vector field*.



We refer to the region $\{(R, W) : R \geq 0, W \geq 0\}$ as the *phase space* of the system; each pair (R, W) represents a state that the system can be in. The horizontal line $W = k/a$ and the vertical line $R = r/b$ partition phase space into four regions. Observe that

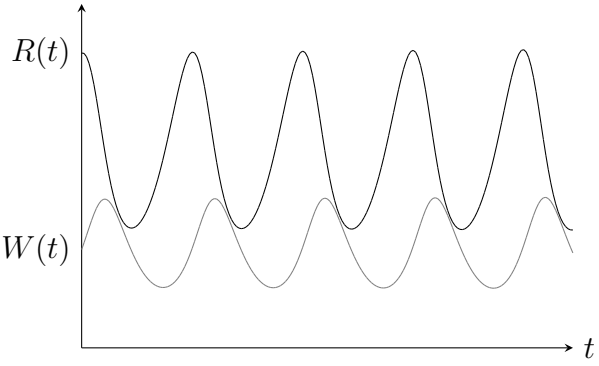
- if $W < \frac{k}{a}$ then $aW < k$, so $\frac{dR}{dt} = R(k - aW) > 0$ (rabbit population increases when wolf population is small);
- if $W > \frac{k}{a}$ then $\frac{dR}{dt} < 0$ (rabbit population decreases when wolf population is large);
- if $R < \frac{r}{b}$ then $bR < r$, so $\frac{dW}{dt} = W(bR - r) < 0$ (wolf population decreases when rabbit population is small);

- if $R > \frac{r}{b}$ then $\frac{dW}{dt} > 0$ (wolf population increases when rabbit population is large).

Thus in the lower left region, where W and R are both below the thresholds, we see that W is decreasing and R is increasing, and all the arrows point down and to the right. In the lower right region, they point up and to the right. In the upper right region, they point up and to the left, and in the upper left region they point down and to the left.

As time progresses, the point $(R(t), W(t))$ moves in a counterclockwise direction around the equilibrium solution $(\frac{r}{b}, \frac{k}{a})$. The picture shows a typical solution curve, computed numerically. The numerical computations suggest that the curve returns to its starting point and then repeats periodically. Is this actually what happens? After all, *a priori* it would be equally reasonable for the curve to spiral in towards the equilibrium point, or to spiral away from it. In fact, the curve is indeed closed, as the numerical evidence suggests, but we will set this question aside and move to other things.[50]

We make one final observation: if we graph the rabbit and wolf populations and superimpose the pictures, then the oscillatory behavior shown above leads to a picture reminiscent of sine and cosine: two oscillating functions whose phases are offset, with the peaks of one lagging behind the peaks of the other.



59.2. Systems with more than two variables

It turns out that for a system of two coupled autonomous differential equations, the only possible behaviors (from a qualitative point of view) are the ones we have encountered already; solutions can approach a fixed equilibrium solution as in the logistic DE, or diverge to infinity, or approach a periodic solution that oscillates endlessly and repeats itself exactly as in the Lotka–Volterra model. The precise theorem that describes all the possible behaviors is called the *Poincaré–Bendixson theorem*, and its details lie beyond the scope of this course; the basic idea is that solution curves cannot "get past" each other because everything lies in a two-dimensional space.

When we have more than two coupled DEs, on the other hand, life changes dramatically. In 1963, the meteorologist Edward Lorenz studied the following set of three

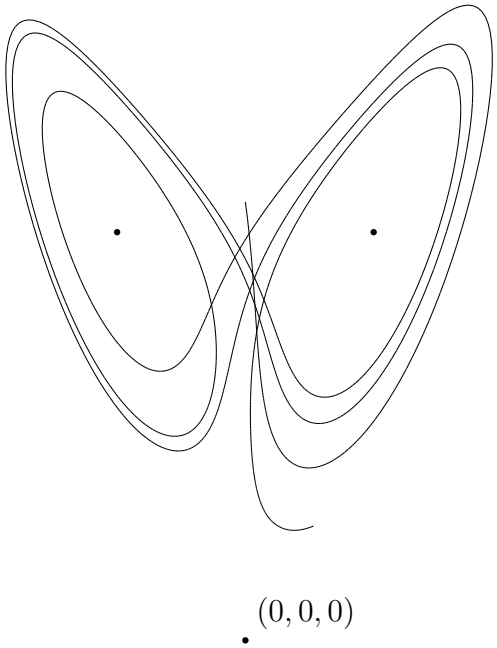
[50]The idea is to find a function $H(R, W)$ that depends on the size of both populations and that has the property that it does not change over time, so that the solution curve lies on a *level set* $\{(R, W) : H(R, W) = H_0\}$ for some value of H_0. It turns out that $H(R, W) = aW + bR - k \ln W - r \ln R$ does the job.

coupled DEs as part of a simplified model of atmospheric convection:

$$\begin{aligned} \frac{dx}{dt} &= \sigma(y-x), \\ \frac{dy}{dt} &= x(\rho-z)-y, \\ \frac{dz}{dt} &= xy-\beta z. \end{aligned} \tag{59.2}$$

Here x, y, z are three functions of t whose physical interpretations we omit, and σ, ρ, β are three real-valued parameters reflecting certain physical properties of the system being studied; Lorenz used the values $\sigma = 10$, $\beta = 8/3$, and $\rho = 28$.

It is not so difficult to find the equilibrium solutions here: if $\frac{dx}{dt} = \frac{dy}{dt} = \frac{dz}{dt} = 0$, then the first equation in (59.2) gives $y = x$, and the second becomes $x(\rho - z - 1) = 0$, so either $x = y = 0$ or $z = \rho - 1$. If $x = y = 0$ then the third equation gives $z = 0$, so one equilibrium solution is $x = y = z = 0$. If $z = \rho - 1$ then the third equation gives $x = y = \pm\sqrt{\beta z} = \pm\sqrt{\beta(\rho-1)}$, so there are two other equilibrium solutions. These three equilibria are shown in the picture at right, which also draws a single (numerically computed) solution of the system for some randomly chosen non-equilibrium initial condition. Observe that this solution does not appear to have any of the long-term behaviors described above: it does not approach an equilibrium solution, nor does it escape off to infinity, nor does it approach a periodic solution. Rather, it seems to spiral around one of the two nonzero equilibria for some time, then switches to spiral around the other, and so on in some manner that does not follow any readily discernible pattern.



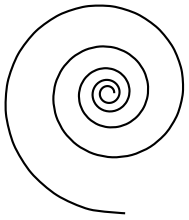
The butterfly-like object shown in the picture is sometimes called a *strange attractor* and is emblematic of the field of *chaos theory*; the Lorenz equations display the phenomenon of *sensitive dependence on initial conditions*, which is a mechanism by which systems that follow deterministic rules can still exhibit behavior that appears random. If you search online for animations of the Lorenz attractor, you should have no trouble finding videos showing how solution curves that start very close to each other can follow each other for a while and then very quickly diverge so that their behavior is quite different. This means that if you only know the *approximate* state of a system to start off with (which reflects the reality that any measurement we make includes some error), then as time progresses you lose information about what state the system is in, which can be interpreted as a 'growth of randomness'. This can be made more precise by studying entropy, decay of correlations, and other topics in the field of *dynamical systems*, but these lie well beyond the scope of this course.

Part VII. Parametric curves and polar coordinates

Lecture 60 — Parametric curves

Stewart §10.1, Spivak Chapter 12 appendix

Suppose we want to write an equation that describes the curve shown at right. Our usual approach to describing a curve by an equation is to write y as a function of x, or in some cases, x as a function of y. However, neither of these is an option here, since the curve fails both the vertical line test (so it cannot be written as the graph of $y = f(x)$) and the horizontal line test (so it cannot be written as the graph of $x = g(y)$).



In such situations, we can describe the curve by writing both x and y as functions of a new independent variable, instead of writing one as a function of the other. Thus we introduce a new variable t, called a *parameter*, and write $x = g(t)$, $y = f(t)$.

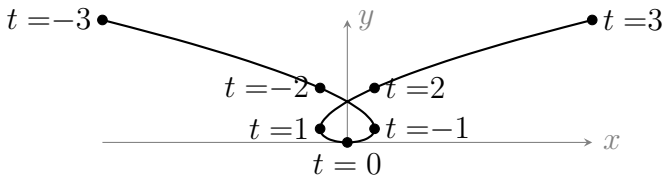
We have actually seen this situation several times already. It first appeared when we solved the catenary problem; although we ended up writing y as a function of x, an important intermediate step was to write both x and y as a function of arclength s, and also as a function of another parameter t. We also saw parametric curves appear in the last lecture on predator-prey models, where a solution of the system of differential equations was given by a curve written in terms of the parameter t, which represented time.

Example 60.1. The description of points on the unit circle as $x = \cos\theta$, $y = \sin\theta$ describes the circle as a parametric curve, where θ is the parameter.

As when graphing curves of the form $y = f(x)$, a useful approach to graphing a parametric curve is to make a table of values of t together with the corresponding values of x and y. For example, the curve shown below has the parametrization

$$x = t^3 - 3t, \qquad y = t^2, \tag{60.1}$$

and the table at right shows the values of x and y for integer values of t between -3 and 3; the corresponding points are marked on the curve.



t	-3	-2	-1	0	1	2	3
x	-18	-2	2	0	-2	2	18
y	9	4	1	0	1	4	9

Observe that one needs to be careful to connect the dots in the right order; based on the positions one might be tempted to connect the dot for $t = -2$ to the dot for $t = 1$, but this would give a very different shape to the curve.

The part of the curve shown in the picture above corresponds to parameter values lying in the interval $[-3, 3]$. When we restrict a curve to parameters $a \leq t \leq b$, the point $(x(a), y(a))$ is called the *initial point* of the curve, and $(x(b), y(b))$ is called the *terminal point*.

Remark 60.2. One should be careful to distinguish between a *curve*, which is a subset of $\mathbb{R}^2$, and a *parametric curve*, which is a curve equipped with a particular parametrization. The same curve can be equipped with many different parametrizations. For example, $x = t^2$, $y = t$, $-1 \leq t \leq 1$ describes an arc of a parabola opening to the right with vertex at the origin. This same curve is also described by $y = \cos t$, $x = \cos^2 t$. The difference between two parametrizations is analogous to the difference between two cars following the same road but with different (and varying) speeds.

As this example illustrates, any curve that can be described as the graph of a function (y in terms of x, or x in terms of y) can also be given as a parametric curve. The graph of $y = f(x)$ admits a parametrization $x = t$, $y = f(t)$, and the graph of $x = g(y)$ admits a parametrization $x = g(t)$, $y = t$. Thus our new technique is a more general one.

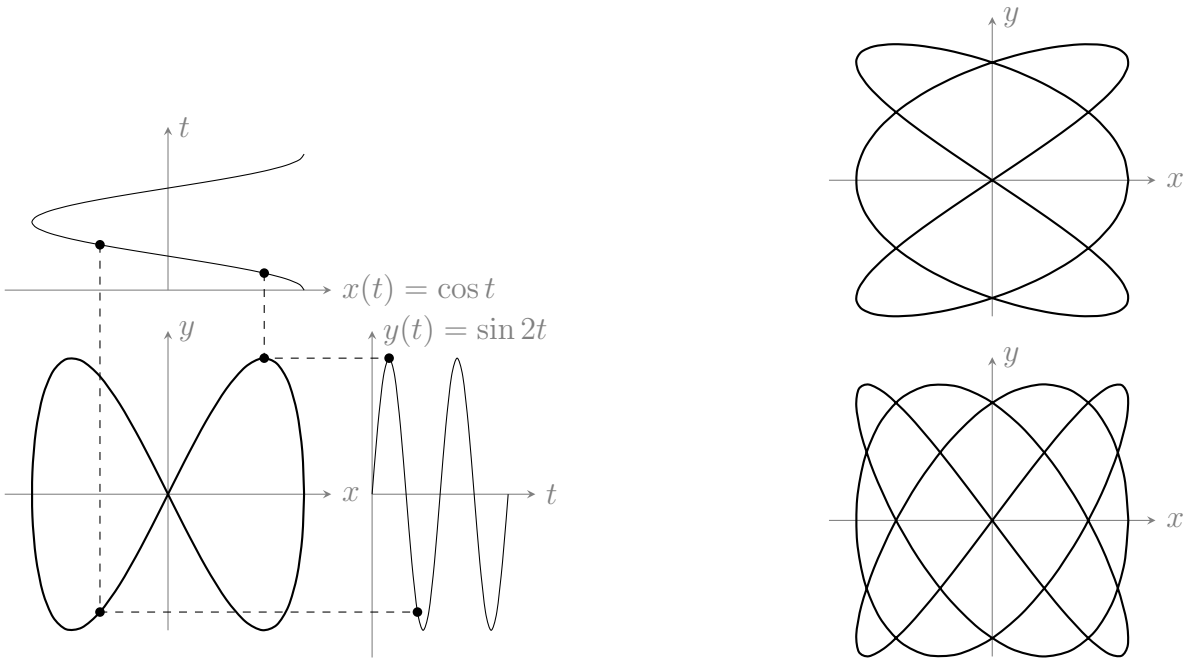


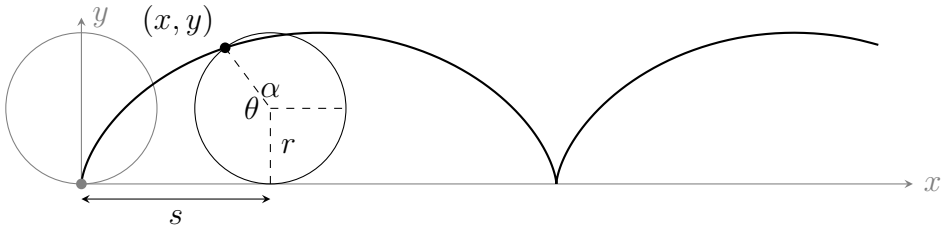
FIGURE 14. Lissajous figures

Example 60.3. The parametric curve $x = \cos t$, $y = \sin 2t$ has the appearance shown in the 'figure-eight' picture at left in Figure 14. Above and to the right of this curve are drawn the graphs of x and y with respect to the parameter t for $0 \leq t \leq 2\pi$, which covers the entire curve by periodicity. (Actually the t-axis in both graphs is compressed to save space.) Note that in the graph of $x(t)$, we plot t along the vertical axis and x along the horizontal axis.

This is an example of a *Lissajous figure*, a family of curves given by parametrizations $x = \cos(at)$, $y = \sin(bt)$ where $a, b \in \mathbb{N}$. (One can also add a phase shift by replacing at

with $at + c$.) The pictures at right in Figure 14 show Lissajous figures for $a = 3$, $b = 2$ (top) and $a = 3$, $b = 4$ (bottom).

Example 60.4. Here is a physical example that is easier to describe via a parametric curve. Consider a circle rolling along the ground; the curve traced out by a point on the circle is called a *cycloid*.



For simplicity, assume that when we start, the point we are interested in is on the ground, and take this as the origin. Now start rolling the circle to the right, and let (x, y) be the location of the point we marked. To write a parametric equation for the cycloid, we let r be the radius of the circle, and write θ for the angle through which it has rotated so far. Then the total horizontal distance s that the circle has rolled is equal to the arc length from the bottom of the circle to (x, y), so we have $s = r\theta$, and we see that the center of the circle is currently at $(r\theta, r)$. The displacement of (x, y) from the center can be given in terms of θ:

(1) if α denotes the angle from the positive horizontal to (x, y) as we move around the circle (see the picture), then $x = r\theta + r\cos\alpha$ and $y = r + r\sin\alpha$;
(2) $\alpha + \theta = \frac{3\pi}{2}$, so $\cos\alpha = \cos(\frac{3\pi}{2} - \theta) = \cos\frac{3\pi}{2}\cos\theta + \sin\frac{3\pi}{2}\sin\theta = -\sin\theta$, and $\sin\alpha = \sin(\frac{3\pi}{2} - \theta) = \sin\frac{3\pi}{2}\cos\theta - \cos\frac{3\pi}{2}\sin\theta = -\cos\theta$.

We obtain the following parametrization for the cycloid in terms of θ:

$$x = r\theta - r\sin\theta = r(\theta - \sin\theta), \qquad y = r - r\cos\theta = r(1 - \cos\theta). \tag{60.2}$$

Remark 60.5. It turns out that the cycloid arises in the solution of two questions of historical interest. One of these is the *brachistochrone* problem, which asks to find the curve connecting two points A and B in the plane along which an object will slide from A to B the fastest under the influence of gravity, without friction. It turns out that the answer is not a straight line, as one might initially expect; rather, it is an (upside-down) cycloid.

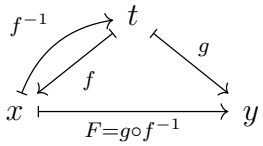
The *tautochrone* problem asks for a curve with the property that the time it takes an object to slide down to the lowest point of the curve is independent of the initial height. It turns out that once again, the solution is an inverted cycloid. The proofs of these facts, however, require tools from the *calculus of variations*, which is well beyond the scope of this course.

Lecture 61 Calculus with parametrizations

Stewart §10.2, Spivak Chapter 12 appendix

61.1. Slopes

Suppose we want to find the slope of a parametric curve $(x, y) = (f(t), g(t))$ at a given point. We can only do this if near this point, there is a differentiable function F such that $y = F(x)$ describes the curve. When is this possible? The following diagram is useful.



The parametrization lets us write y as a function of t, so we can write y as a function of x whenever we can write t as a function of x. This happens whenever the function f is invertible near the value of t that we are interested in; moreover, the chain rule gives the slope $F'(x)$ as $(f^{-1})'(x)g'(t)$.

Recall from last semester that the inverse function f^{-1} is defined and differentiable near $x = f(t)$ as long as $f'(t) \neq 0$, and that in this case we have $(f^{-1})'(f(t)) = 1/f'(t)$. Thus we have proved the following.

Proposition 61.1. *If $(x, y) = (f(t), g(t))$ is a parametric curve, where f, g are differentiable, and $t_0 \in \mathbb{R}$ is such that $f'(t_0) \neq 0$, then near t_0 we can write $y = F(x)$ for some differentiable function F, and the slope of the curve at $(f(t_0), g(t_0))$ is given by*

$$\left.\frac{dy}{dx}\right|_{x=f(t_0)} = F'(f(t_0)) = \frac{g'(t_0)}{f'(t_0)} = \frac{dy/dt|_{t=t_0}}{dx/dt|_{t=t_0}}. \tag{61.1}$$

If f and g are differentiable and $g'(t_0) = 0$ while $f'(t_0) \neq 0$, then at this point the curve has a horizontal tangent line; in other words, a horizontal tangent line occurs when $\frac{dy}{dt} = 0 \neq \frac{dx}{dt}$. A vertical tangent line occurs when $f'(t_0) = 0$ and $g'(t_0) \neq 0$; equivalently, when $\frac{dx}{dt} = 0 \neq \frac{dy}{dt}$.

When testing for horizontal or vertical tangent lines, the condition that the other derivative *not* vanish at t_0 is very important, as the following example shows.

Example 61.2. Let $x = t^3$ and $y = t^3$; then both $\frac{dx}{dt}$ and $\frac{dy}{dt}$ vanish when $t = 0$, but the tangent line is neither horizontal nor vertical here since the curve is just the graph of $y = x$.

61.2. Convexity

What about the second derivative? If we want to determine whether the curve is convex or concave, it is useful to compute $F''(x)$. To do this we use (61.1) to write $F'(x) = \frac{f'(t)}{g'(t)}$, and differentiating gives

$$F''(x) = \frac{d}{dx}F'(x) = \frac{d}{dx}\frac{g'(t)}{f'(t)} = \frac{dt}{dx}\frac{d}{dt}\frac{g'(t)}{f'(t)} = \frac{1}{dx/dt}\frac{d}{dt}\frac{g'(t)}{f'(t)} = \frac{1}{f'(t)}\frac{d}{dt}\frac{g'(t)}{f'(t)}, \tag{61.2}$$

where the third equality uses the chain rule, and the fourth uses the rule for derivatives of inverse functions. We can use the quotient rule to expand this as

$$F''(x) = \frac{f'(t)g''(t) - g'(t)f''(t)}{f'(t)^3},$$

but it is often easier to just work with the formula in (61.2), which can also be rewritten as

$$\frac{d^2y}{dx^2} = \frac{d}{dx}\frac{dy}{dx} = \frac{\frac{d}{dt}\frac{dy}{dx}}{\frac{dx}{dt}}. \tag{61.3}$$

Remark 61.3. Naive analogy with (61.1) might lead us to expect that the second derivative is given by $\frac{d^2y/dt^2}{d^2x/dt^2}$, since we may feel like we could "cancel the two appearances of dt^2", but we see from the above that this is not the case. This illustrates the dangers of treating higher derivatives as if they are fractions.

Example 61.4. Recall the parametric curve $x = t^3 - 3t$, $y = t^2$ from (60.1). This has vertical tangent lines when $0 = \frac{dx}{dt} = 3t^2 - 3$, so $t = \pm 1$; this corresponds to the points $(\mp 2, 1)$. Everywhere else we have $\frac{dx}{dt} \neq 0$ so we can use (61.1) to write

$$\frac{dy}{dx} = \frac{2t}{3t^2 - 3}.$$

We see that the only point with a horizontal tangent line occurs when $t = 0$, when the curve passes through the origin.

Note that the curve intersects itself where it crosses the y-axis; writing $x = 0$ gives $t = 0$ (at the origin) or $t^2 - 3 = 0$, so $t = \pm\sqrt{3}$, and both parameter values correspond to the point $(0, 3)$. Using $t = \sqrt{3}$, the slope is $2\sqrt{3}/6 = \sqrt{3}/3$; using $t = -\sqrt{3}$, the slope is $-\sqrt{3}/3$. These correspond to the tangent lines to the two different 'branches' of the curve passing through this point.

To determine concavity and convexity, we use (61.3) to write

$$\frac{d^2y}{dx^2} = \frac{\frac{d}{dt}\left(\frac{2t}{3t^2-3}\right)}{\frac{d}{dt}(t^3 - 3t)} = \frac{1}{3t^2 - 3} \cdot \frac{(3t^2 - 3) \cdot 2 - 2t(6t)}{(3t^2 - 3)^2} = \frac{-6t^2 - 6}{(3t^2 - 3)^3} = -\frac{2}{9}\Big(\frac{t^2 + 1}{(t^2 - 1)^3}\Big).$$

This never vanishes, so the graph has no inflection points. The second derivative is undefined at $t = \pm 1$, which makes sense because the first derivative is also undefined there. For $|t| < 1$ we have $\frac{d^2y}{dx^2} > 0$ and the graph is convex; for $|t| > 1$ we have $\frac{d^2y}{dx^2} < 0$ and the graph is concave.

Example 61.5. Consider the cycloid given by the parametrization $x = r(\theta - \sin\theta)$, $y = r(1 - \cos\theta)$, where $r > 0$ is the radius of the circle, and $\theta \in \mathbb{R}$ is the parameter. Then at a point (x, y) on the cycloid, the slope of the tangent line is given by

$$\frac{dy}{dx} = \frac{dy/d\theta}{dx/d\theta} = \frac{r\sin\theta}{r(1 - \cos\theta)} = \frac{\sin\theta}{1 - \cos\theta}.$$

It is tempting to immediately say "the tangent line is horizontal if and only if $\sin\theta = 0$". However, the full picture is a little more subtle, because when $\theta = 2n\pi$ for some $n \in \mathbb{Z}$, we have $\sin\theta = 0 = 1 - \cos\theta$, so both numerator and denominator vanish. This corresponds to the 'cusp' at the bottom of the cycloid, where the curve is not differentiable, although we can observe that $\lim_{\theta \to 2n\pi^\pm} \frac{\sin\theta}{1-\cos\theta} = \pm\infty$, which reflects the fact that the tangent line approaches vertical as (x, y) approaches a cusp.

The remaining values of θ for which $\sin\theta = 0$ are $\theta = (2n + 1)\pi$ for some $n \in \mathbb{Z}$, and in this case we have $1 - \cos\theta = 2$, so the slope is indeed horizontal; this corresponds to the highest point on each loop of the cycloid.

The only values of θ for which the denominator vanishes are $\theta = 2n\pi$, when $\cos\theta = 1$, and as we saw above we have $\frac{dy}{d\theta} = \frac{dx}{d\theta} = 0$ at these points.

Lecture 62 Geometry of parametric curves

Stewart §10.2, Spivak Chapter 12 appendix

62.1. Area

Consider the parametric curve $(x, y) = (f(t), g(t))$, where $\alpha \le t \le \beta$ and f, g are differentiable. Suppose that $g \ge 0$ everywhere and that f is increasing, so that writing $a = f(\alpha)$ and $b = f(\beta)$, the function $f\colon [\alpha, \beta] \to [a, b]$ is invertible. Then $F = g \circ f^{-1}$ gives y as a function of x:

$$y = g(t) = g(f^{-1}(x)) = (g \circ f^{-1})(x) = F(x).$$

We know that the area under the curve $y = F(x)$, where $a \le x \le b$, is given by $A = \int_a^b F(x)\,dx$. Using the substitution rule to write this integral in terms of t, which is related to x by $x = f(t)$, we have

$$A = \int_a^b F(x)\,dx = \int_a^b y\,dx = \int_\alpha^\beta y\frac{dx}{dt}\,dt = \int_\alpha^\beta g(t)f'(t)\,dt. \tag{62.1}$$

Example 62.1. The area under one loop of the cycloid is

$$\begin{aligned} A &= \int_0^{2\pi r} y\,dx = \int_0^{2\pi} r(1-\cos\theta)\big(r(1-\cos\theta)\big)\,d\theta = r^2\int_0^{2\pi}(1 - 2\cos\theta + \cos^2\theta)\,d\theta \\ &= r^2\int_0^{2\pi}\Big(1 - 2\cos\theta + \frac{1}{2}(1+\cos 2\theta)\Big)\,d\theta = r^2\Big[\frac{3}{2}\theta - 2\sin\theta + \frac{1}{4}\sin 2\theta\Big]_0^{2\pi} \\ &= r^2\cdot\frac{3}{2}\cdot 2\pi = 3\pi r^2. \end{aligned}$$

62.2. Arc length

As above, consider the parametric curve $(x, y) = (f(t), g(t))$ on the interval $t \in [\alpha, \beta]$, where f, g are differentiable. If $f' > 0$ everywhere, then we can find the arc length of this curve by following the procedure above and writing it as $y = F(x)$ where $F = g \circ f^{-1}$; then the substitution rule gives the arc length as

$$L = \int_a^b \sqrt{1 + F'(x)^2}\,dx = \int_a^b \sqrt{1 + \Big(\frac{dy}{dx}\Big)^2}\,dx = \int_\alpha^\beta \sqrt{1 + \Big(\frac{dy/dt}{dx/dt}\Big)^2}\cdot\frac{dx}{dt}\,dt,$$

and simplifying gives

$$L = \int_\alpha^\beta \sqrt{\Big(\frac{dx}{dt}\Big)^2 + \Big(\frac{dy}{dt}\Big)^2}\,dt. \tag{62.2}$$

The integrand in this final expression can be viewed as an infinitesimal version of the Pythagorean formula.

What if f' is not always positive? What if the curve fails to satisfy the vertical line test and cannot be written as $y = F(x)$? In this case we can still follow the procedure from Lecture 50.1 and consider polygonal approximations to the curve. Partitioning the interval $[\alpha, \beta]$ into n subintervals $[t_{i-1}, t_i]$ of equal length $\Delta t = (\beta - \alpha)/n$, where $t_i = \alpha + i\Delta t$ for $0 \le i \le n$, and writing $P_i = (f(t_i), g(t_i))$, we can once again declare the length of the curve to be given by (50.1), so that

$$L = \lim_{n\to\infty} \sum_{i=1}^{n} \text{distance}(P_{i-1}, P_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \sqrt{(f(t_i) - f(t_{i-1}))^2 + (g(t_i) - g(t_{i-1}))^2}$$

Applying the mean value theorem to f and g on $[t_{i-1}, t_i]$ gives t_i^* and t_i^{**} in this subinterval such that

$$f(t_i) - f(t_{i-1}) = f'(t_i^*)(t_i - t_{i-1}) = f'(t_i^*)\Delta t \quad \text{and} \quad g(t_i) - g(t_{i-1}) = g'(t_i^{**})\Delta t.$$

Thus we can compute the arc length as

$$\begin{aligned} L &= \lim_{n\to\infty} \sum_{i=1}^{n} \sqrt{(f'(t_i^*)\Delta t)^2 + (g'(t_i^{**})\Delta t)^2} = \lim_{n\to\infty} \sum_{i=1}^{n} \sqrt{(f'(t_i^*))^2 + (g'(t_i^{**}))^2} \cdot \Delta t \\ &= \int_\alpha^\beta \sqrt{f'(t)^2 + g'(t)^2}\, dt, \end{aligned}$$

where as in Remark 51.2 we observe that the expression on the first line is not quite a Riemann sum because f' and g' are evaluated at different values of t, but the sum nevertheless converges to the integral. Observe that the formula we obtained here is the same as the formula in (62.2). Thus we have the following.

Definition 62.2. If a curve C admits a parametrization $(x, y) = (f(t), g(t))$, $\alpha \le t \le \beta$, where f, g are differentiable and the curve C is traversed exactly once as t ranges from α to β, then the arc length of C is given by

$$L = \int_\alpha^\beta \sqrt{f'(t)^2 + g'(t)^2}\, dt = \int_\alpha^\beta \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt. \tag{62.3}$$

Exercise 62.3. Show that the arc length does not depend on the choice of parametrization; that is, show that if $(f_1(t), g_1(t))$ and $(f_2(t), g_2(t))$ are two parametrizations of the same curve, then they give the same value of L in (62.3).

Example 62.4. For the circle $x = \cos t$, $y = \sin t$, $t \in [0, 2\pi]$, we have $\frac{dx}{dt} = -\sin t$ and $\frac{dy}{dt} = \cos t$, so (62.3) gives the arc length

$$\int_0^{2\pi} \sqrt{\sin^2 t + \cos^2 t}\, dt = \int_0^{2\pi} 1\, dt = 2\pi,$$

as expected. If we reparametrize the circle as $x = \cos(t^2)$, $y = \sin(t^2)$, then $\frac{dx}{dt} = -2t\sin(t^2)$ and $\frac{dy}{dt} = 2t\cos(t^2)$, and we return to the starting point $(1, 0)$ when $t = \sqrt{2\pi}$, so (62.3) gives the arc length

$$\int_0^{\sqrt{2\pi}} \sqrt{4t^2\sin^2 t^2 + 4t^2\cos^2 t^2}\, dt = \int_0^{\sqrt{2\pi}} 2t\, dt = \Big[t^2\Big]_0^{\sqrt{2\pi}} = 2\pi,$$

which agrees with the earlier answer.

Example 62.5. The arc length of one loop of the cycloid, which has $\frac{dx}{d\theta} = r(1-\cos\theta)$ and $\frac{dy}{d\theta} = r\sin\theta$, is given by

$$\begin{aligned} L &= \int_0^{2\pi} \sqrt{r^2(1-\cos\theta)^2 + r^2\sin^2\theta}\, d\theta = \int_0^{2\pi} r\sqrt{1-2\cos\theta+\cos^2\theta+\sin^2\theta}\, d\theta \\ &= r\int_0^{2\pi} \sqrt{2-2\cos\theta}\, d\theta = r\int_0^{2\pi} \sqrt{4\sin^2\frac{\theta}{2}}\, d\theta = r\int_0^{2\pi} 2\sin\frac{\theta}{2}\, d\theta \\ &= r\Big[-4\cos\frac{\theta}{2}\Big]_0^{2\pi} = r\big(-4(-1)-(-4)(1)\big) = 8r. \end{aligned}$$

62.3. Surface area

Consider a parametric curve $(x,y) = (f(t), g(t))$ on the interval $t \in [\alpha,\beta]$, where f, g are differentiable and $g > 0$; let S be the surface area of the surface of revolution obtained by rotating this curve around the x-axis. Then similar arguments to those in Lecture 51.1 show that

$$S = \int_\alpha^\beta 2\pi y\, ds = \int_\alpha^\beta 2\pi y\sqrt{\Big(\frac{dx}{dt}\Big)^2 + \Big(\frac{dy}{dt}\Big)^2}\, dt.$$

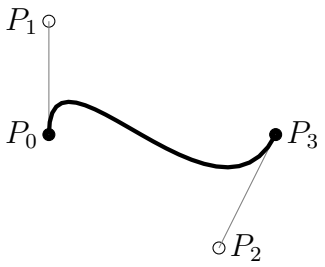
Example 62.6. The sphere of radius r is the surface of revolution for $x = r\cos t$, $y = r\sin t$, $t \in [0,\pi]$, so its surface area is

$$S = \int_0^\pi 2\pi r\sin t\sqrt{r^2\sin^2 t + r^2\cos^2 t}\, dt = \int_0^\pi 2\pi r^2\sin t\, dt = \Big[-2\pi r^2\cos t\Big]_0^\pi = 4\pi r^2.$$

This is a little simpler than the computation we did in Example 51.6.

62.4. Bezier curves

One useful application of parametric curves is given by *Bézier curves*, which are widely used in graphics, design, animation, and other related fields. A *cubic Bézier curve* is given by four *control points* P_0, P_1, P_2, P_3, as shown in the picture. Intuitively, the curve starts at P_0 and ends at P_3, with P_1 used to determine the tangent direction at P_0, and P_2 to determine the tangent direction at P_3.



A nice animation illustrating how to construct these curves can be found online at `https://www.jasondavies.com/animated-bezier/`, and an interactive applet that lets you see how the curve responds to changes in the locations of the four control points can be found at `https://www.desmos.com/calculator/cahqdxeshd`.

Lecture 63 — Polar coordinates

Stewart §10.3, Spivak Chapter 4 appendix 3

63.1. Rectangular and polar coordinates

We are accustomed to using a rectangular coordinate system[51] to describe points in the plane: the two real numbers x and y uniquely determine a point P in the plan as the point that you reach by starting at the origin, moving x units to the right, and moving y units up. Now we describe a new coordinate system, called *polar coordinates*.

Start by fixing a reference point, called the *pole* – usually we choose the origin. Fix an infinite ray starting at this point, called the *polar axis* – usually we choose the positive x-axis. Given real numbers r and θ, the *polar coordinates* (r, θ) describe a point P in the plane as follows:

(1) standing at the pole, face in the direction of the polar axis and then rotate θ radians counterclockwise;
(2) move a distance r in the direction you are now facing.

The point P is the point that you reach at the end of this procedure. To put it another way, the polar coordinates of P are the real numbers r and θ such that $r = |OP|$ is the distance from the origin to P, and θ is the angle from the positive x-axis to the line segment OP.

Remark 63.1. The numbers r and θ uniquely determine P, but (in sharp constrast to the situation with rectangular coordinates) the other direction requires some choice.

- When $r = 0$, *any* value of θ puts P at the origin.
- For $r > 0$, the angles θ and $\theta + 2\pi$ give the same point P. Thus θ is only determined up to multiples of 2π. We will often choose $\theta \in (-\pi, \pi]$, but one could just as easily choose $\theta \in [0, 2\pi)$, or any other half-open interval with length 2π.
- The second procedure described above, for obtaining r and θ from P, always returns a nonnegative value of r. However, the first procedure, for obtaining P from r and θ, makes sense even when r is negative, provided we interpret the second step for a negative value of r as meaning "move backwards by a distance of $|r|$". Then we see that the polar coordinates $(-r, \theta)$ and $(r, \theta+\pi)$ both correspond to the same point.

Exercise 63.2. Plot the points with polar coordinates $(1, \frac{\pi}{4})$, $(2, \frac{\pi}{2})$, $(3, -\frac{3\pi}{4})$, and $(4, \pi)$.

To compare rectangular and polar coordinates, observe that after rotating by an angle θ, we are standing at the origin and facing in the direction of the point on the unit circle with rectangular coordinates $(\cos\theta, \sin\theta)$. (Indeed, this is one definition of cos and sin.) Walking a distance r in this direction moves us to the point $(r\cos\theta, r\sin\theta)$. In other words, rectangular and polar coordinates are related by the equations

$$x = r\cos\theta, \qquad y = r\sin\theta. \tag{63.1}$$

These describe the first procedure above; converting polar coordinates to rectangular coordinates.

[51] Also called *Cartesian* coordinates, after René Descartes.

Example 63.3. The point with polar coordinates $(2, \frac{\pi}{3})$ has $r = 2$ and $\theta = \frac{\pi}{3}$, so its rectangular coordinates are

$$x = 2\cos\frac{\pi}{3} = 2 \cdot \frac{1}{2} = 1, \quad y = 2\sin\frac{\pi}{3} = 2 \cdot \frac{\sqrt{3}}{2} = \sqrt{3}.$$

Remark 63.4. The relationship (63.1) between polar coordinates and rectangular coordinates can be written in a single equation involving complex numbers. Recall that given a real number θ, the complex exponential function is $e^{i\theta} = \cos\theta + i\sin\theta$, and thus given $r \geq 0$ we have

$$re^{i\theta} = r\cos\theta + ir\sin\theta = x + iy,$$

where x, y are the real and imaginary parts, respectively, of the complex number $re^{i\theta}$. If $z = re^{i\theta}$, then the number r is called the *modulus* of z, and θ is called the *argument*. Observe that θ is only defined up to a multiple of 2π.

What about the other direction? If a point has rectangular coordinates (x, y), then squaring the two equations in (63.1) and adding them together gives

$$x^2 + y^2 = r^2\cos^2\theta + r^2\sin^2\theta = r^2.$$

If $x^2 + y^2 = 0$ then we must have $x = y = 0$, so the point is the origin and can be represented as $r = 0$, $\theta =$ any real number. If $x^2 + y^2 \neq 0$, then we can choose r to be the positive square root $r = \sqrt{x^2 + y^2}$ and convert (63.1) to

$$\cos\theta = \frac{x}{r}, \qquad \sin\theta = \frac{y}{r}.$$

Together these uniquely determine θ in the interval $(-\pi, \pi]$, or in any half-open interval of length 2π. Note that this restriction reflects the ambiguity mentioned in Remark 63.1 above: (r, θ) and $(r, \theta + 2\pi)$ represent the same point, because $\cos(\theta + 2\pi) = \cos\theta$ and $\sin(\theta + 2\pi) = \sin\theta$.

Dividing the two halves of (63.1) gives another useful formula,

$$\tan\theta = \frac{r\sin\theta}{r\cos\theta} = \frac{y}{x}.$$

Then θ is determined by any two of the three values $\cos\theta$, $\sin\theta$, $\tan\theta$.

It gets a little messy if we try to explicitly write down a formula for θ in terms of x, y, r using inverse trigonometric functions. One is tempted to simply write

$$\theta = \cos^{-1}\frac{x}{r};$$

however, in order to invert the cosine function, we must restrict it to an interval on which it is 1-1. The usual choice is $[0, \pi]$, but then we would always have $\sin\theta \geq 0$, and so we would need to choose $r < 0$ to represent points with $y < 0$. Thus we should actually follow a two-step procedure: first look at the sign of y to determine which branch of $\cos^{-1}$ to use, and then apply $\cos^{-1}$ to find θ. More precisely:

(1) if $y < 0$, restrict cos to $(-\pi, 0)$ and then invert, so that $\theta = \cos^{-1}\frac{x}{r} \in (-\pi, 0)$;
(2) if $y \geq 0$, restrict cos to $[0, \pi]$ and then invert, so that $\theta = \cos^{-1}\frac{x}{r} \in [0, \pi]$.

Another way of describing this is to take $\theta = \cos^{-1}\frac{x}{r} \in [0, \pi]$ and then check the sign of y: if $y \geq 0$, then leave θ as it is, and if $\theta < 0$, then replace θ by $-\theta$.

Exercise 63.5. Describe similar procedures for finding θ using $\sin^{-1}\frac{y}{r}$ and $\tan^{-1}\frac{y}{x}$; in both cases the inverse trigonometric function yields a value in $[-\frac{\pi}{2},\frac{\pi}{2}]$, and then we must look at the sign of x to determine whether θ is given by this value or by a related one.

Example 63.6. If P has rectangular coordinates $(1,-1)$, then $x=1$ and $y=-1$, so $r=\sqrt{x^2+y^2}=\sqrt{2}$, and thus

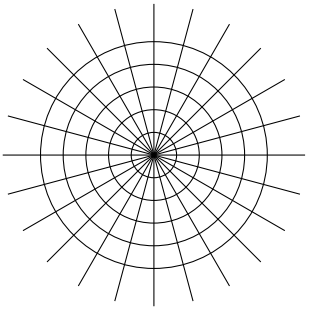
$$\cos^{-1}\frac{x}{r}=\cos^{-1}\frac{1}{\sqrt{2}}=\frac{\pi}{4}.$$

Since $y<0$, the point P lies below the x-axis and we have $\theta=-\cos^{-1}\frac{x}{r}=-\frac{\pi}{4}$.

63.2. Curves in polar coordinates

We know three ways to describe a curve in rectangular coordinates:

(1) *explicitly* as the graph of $y=f(x)$ or $x=g(y)$;
(2) *implicitly* as the solution set of $F(x,y)=0$;
(3) *parametrically* as $x=f(t)$, $y=g(t)$.

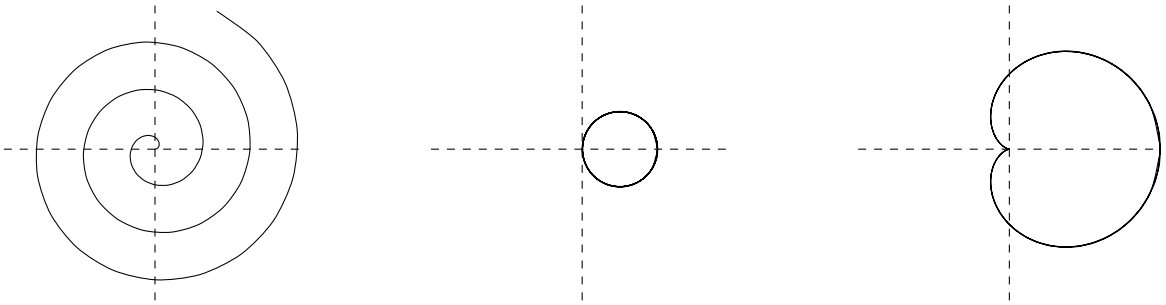


We can use each of these methods in polar coordinates as well. First consider the curves $r=c$ and $\theta=t$, where c,t are constants. These curves are shown in the picture at right. Given $c>0$, the equation $r=c$ describes the circle centered at the origin with radius c. Indeed, since $r=\sqrt{x^2+y^2}$ this formula can be rewritten in rectangular coordinates as the (implicit) formula $x^2+y^2=r^2=c^2$. Given $t\in\mathbb{R}$, we either have $\cos t=0$ (if t is an odd multiple of $\frac{\pi}{2}$), in which case $\sin t=\pm1$ and the curve is the y-axis ($x=0$, $y=\pm r$), or $\cos t\neq0$ in which case $\frac{y}{x}=\tan t$, so the curve is the line $y=(\tan t)x$. This shows that the curves of constant r are concentric circles around the origin, while curves of constant θ are lines through the origin.

More generally, a curve of the form $r=f(\theta)$ can be written parametrically as

$$x=f(\theta)\cos\theta,\quad y=f(\theta)\sin\theta. \tag{63.2}$$

The three pictures below illustrate the curves $r=\theta$, $r=\cos\theta$, and $r=1+\cos\theta$, which we discuss next.



Example 63.7. $r=\theta$ gives a spiral curve as shown in the left-hand picture; observe that increasing θ corresponds to moving around the origin in a counterclockwise direction, and that each time we cross the next axis (having increased θ by $\frac{\pi}{2}$) the value of r has increased and we are further from the origin.

Example 63.8. With the curve $r = \cos\theta$, we see that r decreases from 1 to 0 as θ goes from 0 to $\frac{\pi}{2}$. Then when θ goes over the interval $[\frac{\pi}{2}, \pi]$, where we might expect the curve to lie in the second quadrant, we have $\cos\theta \leq 0$, so in fact the curve lies in the fourth quadrant. Moreover, when $\theta = \pi$ we have $r = -1$ and thus $x = 1$, $y = 0$, which is where the curve starts at $\theta = 0$; thus the entire curve is covered by the parameter range $\theta \in [0, \pi]$. In fact, the curve is the circle with center at $(\frac{1}{2}, 0)$ (in rectangular coordinates) and radius $\frac{1}{2}$; to see this, observe that $x = r\cos\theta = r^2 = x^2 + y^2$, so this is the curve defined in rectangular coordinates by the implicit equation

$$0 = x^2 - x + y^2 = \left(x - \frac{1}{2}\right)^2 + y^2 - \frac{1}{4}.$$

Example 63.9. With $r = 1 + \cos\theta$, we have $r \geq 0$ for all θ, so the curve passes through all four quadrants. The value of r decreases from 2 to 0 as θ ranges from 0 to π; this is the top half of the curve shown. The bottom half of the curve corresponds to $\theta \in [\pi, 2\pi]$, when r increases from 0 back to 2. This curve is called the *cardioid* because of its heart-like shape.

Lecture 64 Calculus with polar coordinates

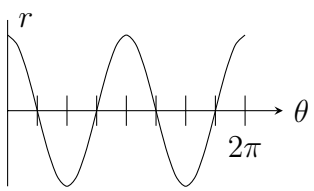
Stewart §10.4.

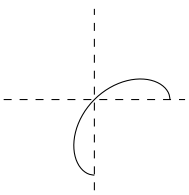
64.1. Slopes of tangent lines

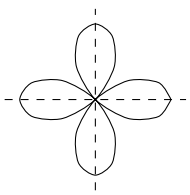
Consider a curve given in polar coordinates by the formula $r = f(\theta)$, where f is differentiable. Using the parametric representation of the curve in (63.2), we can compute the slope of the tangent line at any point by using Proposition 61.1 to write

$$\frac{dy}{dx} = \frac{dy/d\theta}{dx/d\theta} = \frac{f'(\theta)\sin\theta + f(\theta)\cos\theta}{f'(\theta)\cos\theta - f(\theta)\sin\theta} = \frac{\frac{dr}{d\theta}\sin\theta + r\cos\theta}{\frac{dr}{d\theta}\cos\theta - r\sin\theta}. \tag{64.1}$$

Example 64.1. Consider the curve with polar formula $r = \cos 2\theta$. The first picture below shows the graph of r as a function of θ where these are taken as *rectangular* coordinates; this is helpful in order to visualize how r decreases and increases as θ varies, which in turn lets us picture the graph. The second picture shows the graph of the curve on the interval $\theta \in [0, \frac{\pi}{2}]$. Observe how r decreases from 1 to 0 on $[0, \frac{\pi}{4}]$, and then to -1 on $[\frac{\pi}{4}, \frac{\pi}{2}]$, so that on this second interval the curve actually lies in the third quadrant. The third picture shows the complete curve, which consists of four copies of this first piece, each rotated by $\frac{\pi}{2}$ from the previous one.







Using (64.1), we see that the slope of the tangent line to the curve $r = \cos 2\theta$ is

$$\frac{dy}{dx} = \frac{-2\sin 2\theta \sin\theta + \cos 2\theta \cos\theta}{-2\sin 2\theta \cos\theta - \cos 2\theta \sin\theta} = \frac{-4\sin^2\theta\cos\theta + (1-2\sin^2\theta)\cos\theta}{-4\sin\theta\cos^2\theta - (2\cos^2\theta - 1)\sin\theta}$$
$$= \frac{\cos\theta(1-6\sin^2\theta)}{\sin\theta(1-6\cos^2\theta)}.$$

Considering $\theta \in [0, 2\pi)$ to get one full circuit around the curve, we see that the numerator vanishes if and only if $\cos\theta = 0$ or $\sin^2\theta = \frac{1}{6}$. The first possibility occurs at the values $\theta = \frac{\pi}{2}$ and $\theta = \frac{3\pi}{2}$, while the second occurs for one value of θ in each quadrant. Writing $\theta_0 = \sin^{-1}\frac{1}{\sqrt{6}} \in (0, \frac{\pi}{2})$ for the value of θ in this interval at which $\sin^2\theta = \frac{1}{6}$, we see that the four values at which this occurs are $\theta = \theta_0, \pi - \theta_0, \pi + \theta_0, 2\pi - \theta_0$. This gives six points at which the numerator vanishes.

Similarly, the denominator vanishes if and only if $\sin\theta = 0$ or $\cos^2\theta = \frac{1}{6}$. The first possibility occurs at $\theta = 0$ and $\theta = \pi$, while the second occurs at one point in each quadrant. Writing $\theta_1 = \cos^{-1}\frac{1}{\sqrt{6}} \in (0, \frac{\pi}{2})$ for the value of θ in this interval with $\cos^2\theta = \frac{1}{6}$, we see that the four values at which this occurs are $\theta_1, \pi - \theta_1, \pi + \theta_1, 2\pi - \theta_1$. This gives six points at which the denominator vanishes; observe that this does not include any of the points at which the numerator vanishes.

Since the denominator is nonzero everywhere that the numerator vanishes, we see that the tangent line is horizontal when $\theta = \frac{\pi}{2}, \frac{3\pi}{2}$, which corresponds to the points $(0, \pm 1)$, and when $\theta \in \{\theta_0, \pi - \theta_0, \pi + \theta_0, 2\pi - \theta_0\}$. At $\theta = \theta_0$ we have

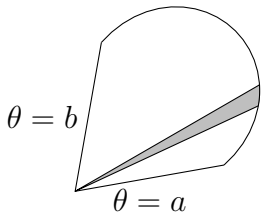
$$x = \cos 2\theta \cos\theta = (1 - 2\sin^2\theta)\cos\theta = \frac{2}{3}\sqrt{1 - \sin^2\theta_0} = \frac{2}{3}\sqrt{\frac{5}{6}},$$
$$y = \cos 2\theta \sin\theta = (1 - 2\sin^2\theta)\sin\theta = \frac{2}{3}\sqrt{\frac{1}{6}}.$$

Thus the four points with horizontal tangent lines are $(\pm\frac{2}{3}\sqrt{\frac{5}{6}}, \pm\frac{2}{3}\sqrt{\frac{1}{6}})$.

Similarly, since the numerator is nonzero everywhere that the denominator vanishes, the tangent line is vertical when $\theta \in \{0, \pi, \theta_1, \pi - \theta_1, \pi + \theta_1, 2\pi - \theta_1\}$, and these six points have coordinates $(\pm 1, 0)$ and $(\pm\frac{2}{3}\sqrt{\frac{1}{6}}, \pm\frac{2}{3}\sqrt{\frac{5}{6}})$.

64.2. Area in polar coordinates

We know that in rectangular coordinates, the region bounded by the curves $x = a$, $x = b$, $y = 0$, and $y = f(x)$ has area $\int_a^b f(x)\,dx$. What about polar coordinates? What is the area of the region bounded by the curves $\theta = a$, $\theta = b$, and $r = f(\theta)$?



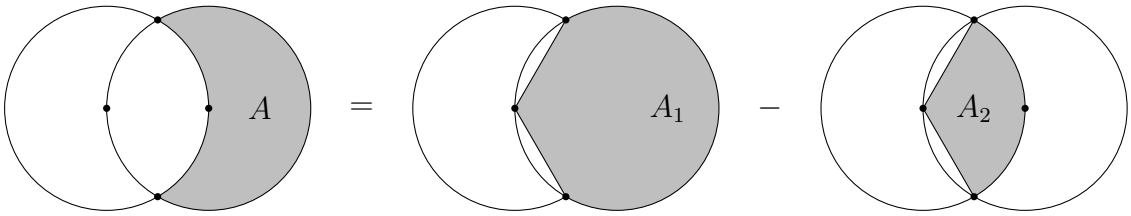
As usual, we fix a large $n \in \mathbb{N}$ and divide the parameter interval $[a, b]$ into n subintervals of equal length $\Delta\theta = (b - a)/n$, with endpoints $\theta_i = a + i\Delta\theta$. Then the ith interval $[\theta_{i-1}, \theta_i]$ determines a 'wedge' such as the one shown in the picture, whose area is $\approx \frac{1}{2}f(\theta_i^*)^2\Delta\theta$, where $\theta_i^* \in [\theta_{i-1}, \theta_i]$ and we remember that a sector of a circle with angle

θ and radius r has area $\frac{1}{2}\theta r^2$. Adding these areas together gives a Riemann sum, and taking a limit as $n \to \infty$ we see that the area of the region is given by

$$A = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{2} f(\theta_i^*)\Delta\theta = \int_a^b \frac{1}{2} f(\theta)^2 \, d\theta. \tag{64.2}$$

Example 64.2. The right-most leaf of the "clover" shape from Example 64.1 corresponds to the parameter interval $\theta \in [-\frac{\pi}{4}, \frac{\pi}{4}]$, so its area is

$$\begin{aligned} A &= \int_{-\pi/4}^{\pi/4} \frac{1}{2} r^2 \, d\theta = \int_{-\pi/4}^{\pi/4} \frac{1}{2} \cos^2 2\theta \, d\theta = \int_0^{\pi/4} \cos^2 2\theta \, d\theta \\ &= \int_0^{\pi/4} \frac{1}{2}(1 + \cos 4\theta) \, d\theta = \frac{1}{2}\Big[\theta + \frac{1}{4}\sin 4\theta\Big]_0^{\pi/4} = \frac{\pi}{8}. \end{aligned}$$



Example 64.3. Consider two circles with radius 1 whose centers are a distance 1 apart. What is the area of the region that lies outside one circle and inside the other?

Choose polar coordinates in which the first circle is centered at the origin, so its polar equation is $r = 1$. Recall that $r = \cos\theta$ gives a circle centered at $(\frac{1}{2}, 0)$ with radius $\frac{1}{2}$, so the second circle has polar equation $r = 2\cos\theta$. Observe that these circles intersect when $1 = r = 2\cos\theta$, so $\cos\theta = \frac{1}{2}$, which occurs when $\theta = \pm\frac{\pi}{3}$. As shown in the picture, the region in which we are interested in is given by the inequalities $-\frac{\pi}{3} \le \theta \le \frac{\pi}{3}$ and $1 \le r \le 2\cos\theta$. Its area is $A = A_1 - A_2$, where A_1 is the area of the region inside $r = 2\cos\theta$ and A_2 is the area of the region inside $r = 1$. Our area formula gives $A_1 = \int_{-\pi/3}^{\pi/3} \frac{1}{2}(2\cos\theta)^2 \, d\theta$ and $A_2 = \int_{-\pi/3}^{\pi/3} \frac{1}{2} \cdot 1 \, d\theta$, so we get

$$\begin{aligned} A &= \int_{-\pi/3}^{\pi/3} \frac{1}{2}(4\cos^2\theta - 1) \, d\theta = \int_0^{\pi/3} (4\cos^2\theta - 1) \, d\theta \\ &= \int_0^{\pi/3} (2\cos(2\theta) + 1) \, d\theta = \Big[\sin(2\theta) + \theta\Big]_0^{\pi/3} = \sin\frac{2\pi}{3} + \frac{\pi}{3} = \frac{\sqrt{3}}{2} + \frac{\pi}{3}. \end{aligned}$$

Remark 64.4. In the above example we found the intersection points of two curves $r = f(\theta)$ and $r = g(\theta)$ by finding the values of θ for which $f(\theta) = g(\theta)$. There is one caveat that comes with this process. Suppose we look for intersection points of the four-leaf clover $r = \cos 2\theta$ with the circle $r = \frac{1}{2}$. Solving $\cos 2\theta = \frac{1}{2}$ produces 4 solutions in $[0, 2\pi)$, but it is clear from the picture following Example 64.1 that the circle intersects the clover in 8 places. The other 4 intersections come from points where r is negative; in other words, they correspond to solutions of $f(\theta + \pi) = g(\theta)$. When we are dealing with curves for which r may take negative values, we must be on the alert for this phenomenon.

64.3. Arc length in polar coordinates

To find the arc length of a curve $r = f(\theta)$ given in polar coordinates, we can once again proceed by writing it as a parametric curve

$$x = f(\theta)\cos\theta, \quad y = f(\theta)\sin\theta,$$

so that

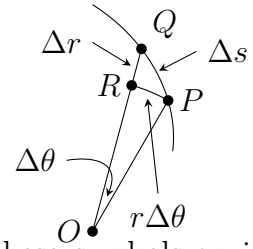
$$\frac{dx}{d\theta} = \frac{dr}{d\theta}\cos\theta - r\sin\theta, \quad \frac{dy}{d\theta} = \frac{dr}{d\theta}\sin\theta + r\cos\theta,$$

and the derivative of the arc length function s has square given by

$$\left(\frac{ds}{d\theta}\right)^2 = \left(\frac{dx}{d\theta}\right)^2 + \left(\frac{dy}{d\theta}\right)^2 = \left(\frac{dr}{d\theta}\right)^2\cos^2\theta - 2r\frac{dr}{d\theta}\cos\theta\sin\theta + r^2\sin^2\theta + \left(\frac{dr}{d\theta}\right)^2\sin^2\theta + 2r\frac{dr}{d\theta}\cos\theta\sin\theta + r^2\cos^2\theta,$$

which simplifies to

$$\left(\frac{ds}{d\theta}\right)^2 = \left(\frac{dr}{d\theta}\right)^2 + r^2. \tag{64.3}$$



As a mnemonic aid to remembering (64.3), we can multiply through by $(d\theta)^2$ to get

$$(ds)^2 = (dr)^2 + r^2(d\theta)^2, \tag{64.4}$$

where once again we have the caveat that we have not given these symbols an independent meaning. The formula (64.4) can be remembered by considering the diagram shown, in which P has polar coordinates (r, θ), Q has polar coordinates $(r+\Delta r, \theta+\Delta\theta)$, and R has polar coordinates $(r, \theta + \Delta\theta)$. Then the circular arc from P to R has length $r\Delta\theta$ and the line segment RQ has length Δr. The piece of curve from P to Q is not quite the hypotenuse of a right triangle with legs $r\Delta\theta$ and Δr, but it is very close to being this, and thus a good approximation to its length is given by

$$(\Delta s)^2 = (r\Delta\theta)^2 + (\Delta r)^2.$$

As P and Q get closer together, this approximation becomes better, and the meaning of (64.4) is that in the limit it gives exactly the integrand we need to compute arc length. In particular, using (64.3) we conclude that the arc length over the interval $a \le \theta \le b$ is

$$L = \int_a^b \sqrt{r^2 + \left(\frac{dr}{d\theta}\right)^2}\, d\theta. \tag{64.5}$$

Example 64.5. The curve $r = 2\cos\theta$ for $0 \le \theta \le \pi$ has arc length

$$L = \int_0^\pi \sqrt{(2\cos\theta)^2 + (-2\sin\theta)^2}\, d\theta = \int_0^\pi 2\, d\theta = 2\pi,$$

which is reassuring since this is a circle with radius 1.

Example 64.6. The arc length of the cardioid $r = 1 + \cos\theta$ $(0 \le \theta \le 2\pi)$ is

$$L = \int_0^{2\pi} \sqrt{(1+\cos\theta)^2 + (-\sin\theta)^2}\, d\theta = \int_0^{2\pi} \sqrt{1 + 2\cos\theta + \cos^2\theta + \sin^2\theta}\, d\theta$$

$$= \int_0^{2\pi} \sqrt{2+2\cos\theta}\,d\theta = \sqrt{2}\int_0^{2\pi} \sqrt{1+\cos\theta}\,d\theta.$$

Multiplying top and bottom by $\sqrt{1-\cos\theta}$ gives

$$\int_0^{2\pi} \sqrt{1+\cos\theta}\,d\theta = \int_0^{2\pi} \frac{\sqrt{(1+\cos\theta)(1-\cos\theta)}}{\sqrt{1-\cos\theta}}\,d\theta = \int_0^{2\pi} \frac{\sqrt{1-\cos^2\theta}}{\sqrt{1-\cos\theta}}\,d\theta$$
$$= \int_0^{2\pi} \frac{|\sin\theta|}{\sqrt{1-\cos\theta}}\,d\theta = \int_0^{\pi} \frac{\sin\theta}{\sqrt{1-\cos\theta}}\,d\theta + \int_\pi^{2\pi} \frac{-\sin\theta}{\sqrt{1-\cos\theta}}\,d\theta.$$

The substitution $u = 1-\cos\theta$ has $du = \sin\theta$ and thus

$$\int \frac{\sin\theta}{\sqrt{1-\cos\theta}}\,d\theta = \int u^{-1/2}\,du = 2\sqrt{u} + C = 2\sqrt{1-\cos\theta} + C.$$

Using this we can evaluate the above integrals and conclude that

$$\int_0^{2\pi} \sqrt{1+\cos\theta}\,d\theta = \Big[2\sqrt{1-\cos\theta}\Big]_0^{\pi} - \Big[2\sqrt{1-\cos\theta}\Big]_\pi^{2\pi} = 2\sqrt{2} - 0 - (0 - 2\sqrt{2}) = 4\sqrt{2}.$$

Thus the arc length of the cardioid is $L = \sqrt{2}\cdot 4\sqrt{2} = 8$.

An alternate method for evaluating $\int_0^{2\pi} \sqrt{1+\cos\theta}\,d\theta$ (instead of the algebraic trick we used) is to use the identity $\cos\theta = 2\cos^2\frac{\theta}{2} - 1$ to write

$$\int_0^{2\pi} \sqrt{1+\cos\theta}\,d\theta = \int_0^{2\pi} \sqrt{2\cos^2\frac{\theta}{2}}\,d\theta = \sqrt{2}\int_0^{2\pi} \Big|\cos\frac{\theta}{2}\Big|\,d\theta. \tag{64.6}$$

Since $\cos\frac{2\pi-\theta}{2} = \cos(\pi - \frac{\theta}{2}) = -\cos\frac{\theta}{2}$, we see that the function $\theta \mapsto |\cos\frac{\theta}{2}|$ is symmetric around the line $\theta = \pi$, and thus $\int_0^{\pi} |\cos\frac{\theta}{2}|\,d\theta = \int_\pi^{2\pi} |\cos\frac{\theta}{2}|\,d\theta$, so we conclude that

$$\int_0^{2\pi} \Big|\cos\frac{\theta}{2}\Big|\,d\theta = 2\int_0^{\pi} \Big|\cos\frac{\theta}{2}\Big|\,d\theta = 2\int_0^{\pi} \cos\frac{\theta}{2}\,d\theta = 2\Big[2\sin\frac{\theta}{2}\Big]_0^{\pi} = 4,$$

where the second equality uses the fact that $\cos\frac{\theta}{2} \geq 0$ for all $\theta \in [0,\pi]$. Together with (64.6) this once again gives $\int_0^{2\pi} \sqrt{1+\cos\theta}\,d\theta = 4\sqrt{2}$, thus $L = 8$.

Part VIII. Sequences and series

Lecture 65 Sequences

Stewart §11.1, Spivak Ch. 22

65.1. Sequences and limits

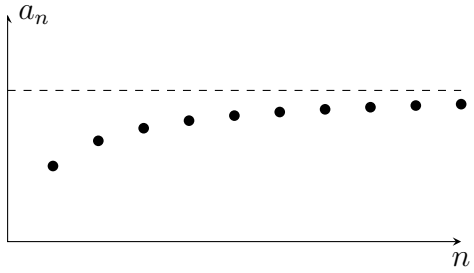
A *sequence* is a list of numbers $a_1, a_2, a_3, \dots$ in a given order (so $1, 2, 3, 4, \dots$ is a different sequence from $1, 3, 2, 4, \dots$); equivalently, a sequence is a function from $\mathbb{N}$ to $\mathbb{R}$. We refer to a_n as the *nth term* of the sequence. We will often write $\{a_n\}$, $\{a_n\}_{n=1}^{\infty}$, (a_n), or $(a_n)_{n=1}^{\infty}$ to refer to the sequence as a whole.

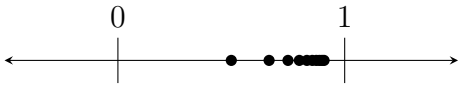
A sequence may or may not be given in terms of a nice formula: for example, $a_n = \frac{n}{n+1}$ has a nice explicit formula for each term, while the sequence

$$b_1 = 0, \quad b_{n+1} = 1 + \sqrt{b_n}$$

is defined *recursively*, and there is no simple formula for its nth term. Or we might consider the sequence whose nth term c_n is the nth digit of the decimal expansion of π, and then there is neither an explicit nor recursive formula that is readily available. A similar thing occurs with the sequence $2, 3, 5, 7, 11, 13, 17, \dots$, where the nth term p_n is the nth prime number.

We can plot a sequence $\{a_n\}$ by drawing a dot at each of the points (n, a_n); this is the graph of the function $\mathbb{N} \to \mathbb{R}$ defined by $n \mapsto a_n$. The first picture shows the result for $a_n = \frac{n}{n+1}$ (with the horizontal axis compressed to save space).





Another option is to draw a number line and put a dot at a_n for each value of n, as shown in the second picture. This second method has the advantage of providing a more compact representation, but the disadvantage that it loses all information about the *order* in which the terms of the sequence appear, since permuting these terms would result in the same picture; moreover, if several terms of the sequence are close together then it becomes difficult to distinguish them. In the end, we tend not to rely on graphical representations of sequences nearly as much as we do for functions $\mathbb{R} \to \mathbb{R}$, and so we will not use either of these methods that often.

Many of our basic definitions and theorems about limits for functions have analogues for sequences.

Definition 65.1. A sequence $\{a_n\}$ has a *limit* $L \in \mathbb{R}$ if for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that for every $n \geq N$, we have $|a_n - L| < \epsilon$. In this case we write $\lim_{n\to\infty} a_n = L$, or

sometimes "$a_n \to L$ as $n \to \infty$". If the sequence a_n has a limit, we say that *the sequence converges*. If it does not have a limit, we say that *the sequence diverges*.

The following simple fact will occasionally be useful.

Exercise 65.2. Prove that $\lim_{n\to\infty} a_{n+1} = \lim_{n\to\infty} a_n$ whenever the sequence converges.

Definition 65.3. One type of diverging sequence is worth particular mention. We write $\lim_{n\to\infty} a_n = \infty$ (or sometimes "$a_n \to \infty$ as $n \to \infty$") if for every $M > 0$ there exists $N \in \mathbb{N}$ such that for every $n \geq N$, we have $a_n \geq M$, and $\lim_{n\to\infty} a_n = -\infty$ (or $a_n \to -\infty$) if for every $M > 0$ there exists $N \in \mathbb{N}$ such that for every $n \geq N$ we have $a_n \leq -M$.

Exercise 65.4. Show that the sequence $x_n = (-1)^n$ is divergent.

All of the limit laws still work, just as they did for functions. Thus we have

$$\lim_{n\to\infty} \frac{n}{n+1} = \lim_{n\to\infty} \frac{1}{1+\frac{1}{n}} = \frac{1}{\lim_{n\to\infty}(1+\frac{1}{n})} = \frac{1}{1+\lim_{n\to\infty}\frac{1}{n}} = \frac{1}{1+0} = 1.$$

Similarly, the squeeze theorem still holds.

Theorem 65.5 (Squeeze theorem)**.** *Given three sequences satisfying $a_n \leq b_n \leq c_n$ for all n, if we have $\lim_{n\to\infty} a_n = \lim_{n\to\infty} c_n = L$, then $\lim_{n\to\infty} b_n = L$ as well.*

Proof. Exercise: recall the proof of the squeeze theorem for functions, and adapt it. Observe that as part of the proof, you must show that the sequence b_n converges. □

Proposition 65.6. *A sequence a_n converges to 0 if and only if $|a_n|$ also converges to 0.*

Proof. We have $-|a_n| \leq a_n \leq |a_n|$ for all n, so if $|a_n| \to 0$ then $-|a_n| \to 0$ by the limit laws, and the squeeze theorem implies that $a_n \to 0$ as well. The other direction is a short exercise using the definition. □

Theorem 65.7. *If a sequence $\{a_n\}$ and a function $f\colon \mathbb{R} \to \mathbb{R}$ are related by $a_n = f(n)$, and if moreover we have $\lim_{x\to\infty} f(x) = L$, then $\lim_{n\to\infty} a_n = L$.*

Proof. Exercise. □

Example 65.8. If $b_n = \frac{\ln n}{n}$, then we have $b_n = f(n)$ where $f(x) = \frac{\ln x}{x}$. Since $\ln x \to \infty$ as $x \to \infty$, we see that $\lim_{x\to\infty} f(x)$ has indeterminate form, and so l'Hospital's rule together with Theorem 65.7 gives

$$\lim_{n\to\infty} b_n = \lim_{x\to\infty} \frac{\ln x}{x} = \lim_{x\to\infty} \frac{1/x}{1} = 0.$$

Example 65.9. The sequence $(-1)^n$ whose terms are $-1, 1, -1, 1, -1, 1, \ldots$ diverges, but the sequence $\frac{(-1)^n}{n}$ whose terms are $-\frac{1}{n}, \frac{2}{n}, -\frac{3}{n}, \frac{4}{n}, \ldots$ converges to 0 by Proposition 65.6, since $\frac{1}{n} \to 0$.

Theorem 65.10. *If f is a function that is continuous at L, and a_n is a sequence in the domain of f such that $\lim_{n\to\infty} a_n = L$, then $\lim_{n\to\infty} f(a_n) = f(L)$.*

Proof. Exercise (use the definition of continuity). □

Example 65.11. Since $\frac{\pi}{n} \to 0$ as $n \to \infty$ and since $\theta \mapsto \sin\theta$ is continuous at 0, we have

$$\lim_{n\to\infty} \sin\frac{\pi}{n} = \sin\left(\lim_{n\to\infty}\frac{\pi}{n}\right) = \sin 0 = 0.$$

Example 65.12. Consider the sequence $a_n = \frac{n!}{n^n}$. The numerator and denominator both diverge to ∞, so this has indeterminate form, but we cannot use l'Hospital's rule without first finding some differentiable function $f(x)$ such that $f(n) = n!$. Since we do not know any such function,[52] we use a different argument, and observe that for every n we have

$$0 \le a_n = \frac{1}{n}\cdot\left(\frac{2}{n}\cdot\frac{3}{n}\cdots\frac{n}{n}\right) \le \frac{1}{n}.$$

Since $\frac{1}{n} \to 0$, the squeeze theorem implies that $\frac{n!}{n^n} \to 0$ as $n \to \infty$.

Example 65.13. Recall from our study of exponential functions that

$$\lim_{x\to\infty} a^x = \begin{cases} 0 & \text{if } 0 \le a < 1, \\ 1 & \text{if } a = 1, \\ \infty & \text{if } a > 1. \end{cases}$$

Using Theorem 65.7, this implies that given $r \ge 0$, the sequence r^n satisfies

$$\lim_{n\to\infty} r^n = \begin{cases} 0 & \text{if } 0 \le r < 1, \\ 1 & \text{if } r = 1, \\ \infty & \text{if } r > 1. \end{cases}$$

In particular, given any $r \in (-1, 1)$, we have

$$|r^n| = |r|^n \to 0 \quad \text{since } |r| \in [0, 1).$$

By Proposition 65.6, this implies that $r^n \to 0$. We conclude that $r^n \to 0$ for every $|r| < 1$, and $r^n \to 1$ when $r = 1$. For all other values of r, the sequence r^n diverges.

65.2. Monotonic sequences

A sequence a_n is called *increasing* if $a_{n+1} > a_n$ for every n, and *decreasing* if $a_{n+1} < a_n$ for every n. If one of these conditions holds, then the sequence is called *monotonic*.

Remark 65.14. If we weaken the condition to $a_{n+1} \ge a_n$ for all n, then we say that the sequence is *nondecreasing*. Similarly if $a_{n+1} \le a_n$ for all n, then the sequence is *nonincreasing*. You should be warned that some authors use "increasing" to mean "nondecreasing", and say "strictly increasing" when they mean $a_{n+1} > a_n$; similarly for "decreasing" and "strictly decreasing". Thus if you encounter the words "increasing" or "decreasing" when you read a piece of mathematics, it is worth checking to see in which sense the author is using them.

Example 65.15.

(1) The sequence 3, 3.1, 3.14, 3.141, 3.1415, 3.14159, …is increasing.[53]
(2) The sequence $a_n = \frac{1}{n}$ is decreasing, since $n + 1 > n$ implies $\frac{1}{n+1} < \frac{1}{n}$.

[52] In fact there is such a function, called the *gamma function*, but we have not studied this yet.

[53] Actually to make this completely true, we need to add two digits whenever we encounter a 0 in the decimal expansion of π; as given, the sequence is merely nondecreasing.

(3) The sequence $b_n = n$ is increasing.
(4) The sequence $c_n = (-1)^n$ is neither increasing nor decreasing.
(5) The sequence $d_n = \frac{n}{n^2+1}$ is decreasing. To prove this we can observe that $d_n = f(n)$ where $f(x) = \frac{x}{x^2+1}$ has derivative

$$f'(x) = \frac{(x^2+1)\cdot 1 - x\cdot 2x}{(x^2+1)^2} = \frac{1-x^2}{(x^2+1)^2} < 0 \quad \text{for all } x > 1$$

and thus is decreasing on $(1,\infty)$. Alternately we can use the direct computation

$$\begin{aligned} d_{n+1} - d_n &= \frac{n+1}{(n+1)^2+1} - \frac{n}{n^2+1} = \frac{(n+1)(n^2+1) - n(n^2+2n+2)}{(n^2+2n+2)(n^2+1)} \\ &= \frac{(n^3+n^2+n+1) - (n^3+2n^2+2n)}{(n^2+2n+2)(n^2+1)} = \frac{1-n-n^2}{(n^2+2n+2)(n^2+1)} < 0. \end{aligned}$$

Definition 65.16. A sequence $\{a_n\}$ is *bounded above* if there exists $M \in \mathbb{R}$ such that $a_n \le M$ for all $n \in \mathbb{N}$; in this case M is called an *upper bound* for the sequence.

Similarly, the sequence is *bounded below* if there exists $m \in \mathbb{R}$ such that $a_n \ge m$ for all $n \in \mathbb{N}$; in this case m is a *lower bound* for the sequence.

We say that $\{a_n\}$ is *bounded* if it is bounded above and bounded below.

Exercise 65.17. Show that $\{a_n\}$ is bounded if and only if $\{|a_n|\}$ is bounded above.

Example 65.18.

(1) The sequence 3, 3.1, 3.14, 3.141, 3.1415, …is bounded; 3 is a lower bound, and π is an upper bound.
(2) The sequence $a_n = \frac{1}{n}$ is bounded; 0 is a lower bound, and 1 is an upper bound.
(3) The sequence $b_n = n$ is bounded below by 1, but is not bounded above.
(4) The sequence $c_n = (-1)^n$ is bounded; -1 is a lower bound, and 1 is an upper bound.
(5) The sequence $d_n = \frac{n}{n^2+1}$ is bounded; 0 is a lower bound, and $d_1 = \frac{1}{2}$ is an upper bound because the sequence is decreasing.

Observe that in each of these cases, the lower and upper bounds that are quoted are in fact optimal. For example, -1 is also a lower bound for the sequence $a_n = \frac{1}{n}$, but it seems better to use the (larger) lower bound 0, since this carries more information. Similarly, 2 is an upper bound for this sequence, but the upper bound 1 is in some sense better. This line of thinking motivates the following definition.

Definition 65.19. A real number M is the *least upper bound* for a sequence $\{a_n\}$ if

- M is an upper bound ($a_n \le M$ for all $n \in \mathbb{N}$), and
- no number smaller than M is an upper bound (for every $L < M$, there is $n \in \mathbb{N}$ such that $a_n > L$).

In this case we also call M the *supremum* of the sequence, and write $M = \sup_n a_n$.

Similarly, m is the *greatest lower bound* for $\{a_n\}$ if

- m is a lower bound ($a_n \ge m$ for all $n \in \mathbb{N}$), and
- no number larger than m is an upper bound (for every $\ell > m$, there is $n \in \mathbb{N}$ such that $a_n < \ell$).

In this case we also call m the *infimum* of the sequence, and write $m = \inf_n a_n$.

It is easy to identify the supremum or infimum when it occurs as a term in the sequence; in the example above, this was the case for the infimums of the increasing sequences 3, 3.1, 3.141, …and $b_n = n$, and for the supremums of the decreasing sequences $a_n = \frac{1}{n}$ and $d_n = \frac{n}{n^2+1}$. It was also the case for $c_n = (-1)^n$, where every term is either ± 1.

When the supremum or infimum does not occur as a term in the sequence, we rely on the following fundamental property of the real numbers.[54]

Least Upper Bound Property. *If $\{a_n\}$ is a sequence of real numbers that is bounded above, then it has a least upper bound in the real numbers. Similarly, if $\{a_n\}$ is a sequence of real numbers that is bounded below, then it has a greatest lower bound in the real numbers.*

Remark 65.20. The Least Upper Bound Property is not a theorem that we are going to prove; rather, it is a fundamental property of the real numbers, which we assume as an axiom. Later in your mathematical career, you will learn how to *construct* the real numbers in such a way that this property is satisfied. For now we content ourselves with the observation that this property fails dramatically if we work with the rational numbers instead of the real numbers. Indeed, the first sequence in Example 65.18 is a sequence of rational numbers that admits a rational upper bound (4 will work) but does *not* have a least upper bound in the rational numbers (because π is irrational).

Theorem 65.21 (Monotone Convergence Theorem)**.** *If a_n is a nondecreasing sequence that is bounded above, then it converges to its supremum. Similarly, if b_n is a nonincreasing sequence that is bounded below, then it converges to its infimum.*

Proof. Let M be the least upper bound of the sequence a_n. Then for every $\epsilon > 0$, the numbers $M - \epsilon$ is not an upper bound (by the definition of least upper bound), so there is some $N \in \mathbb{N}$ such that $a_N > M - \epsilon$. But since the sequence is nondecreasing, this implies that for every $n \geq N$ we have $M - \epsilon < a_N \leq a_n \leq M$, which verifies the definition of a limit and proves the first half of the theorem. The second half follows by observing that $a_n = -b_n$ is nondecreasing and is bounded above. □

Observe that the first, second, and last sequences in Example 65.18 illustrate the theorem; the first sequence converges to its supremum π, while the second and last sequences converge to their infimum 0.

Example 65.22. Define a sequence a_n recursively by $a_1 = 1$, $a_{n+1} = \frac{1}{2}(a_n + 2)$. We claim that $a_n \leq 2$ for all n, and that a_n is nondecreasing. Observe that if $a_n \leq 2$, then $a_{n+1} \leq 2$ as well, so the first claim follows by induction since $a_1 = 1 < 2$. Moreover, if $a_n \leq 2$, then $a_{n+1} = \frac{1}{2}(a_n + 2) \geq \frac{1}{2}(a_n + a_n) = a_n$, so a_n is nondecreasing. By the Monotone Convergence Theorem, $L = \lim_{n\to\infty} a_n$ exists. Thus we have

$$L = \lim_{n\to\infty} a_{n+1} = \lim_{n\to\infty} \frac{1}{2}(a_n + 2) = \frac{1}{2}\left(\left(\lim_{n\to\infty} a_n\right) + 2\right) = \frac{1}{2}(L + 2),$$

and solving for L gives $L = 2$. Thus $a_n \to 2$ as $n \to \infty$.

[54]We state the property for sequences, but in fact it, and the definitions of infimum and supremum above, are valid for any subset of $\mathbb{R}$.

Lecture 66 Summing an infinite series

Stewart §11.2, Spivak Ch. 23

66.1. Convergence and divergence

Suppose we want to add together all of the terms of a sequence $a_1, a_2, a_3, \ldots$. We refer to this as an *infinite series* (often just *series*) and write

$$a_1 + a_2 + a_3 + \cdots + a_n + \cdots = \sum_{n=1}^{\infty} a_n = \sum a_n,$$

where the last notation is a shorthand that we will usually avoid, preferring to write the bounds of summation explicitly to avoid confusion.

Does this notion make sense? What does it mean to add infinitely many numbers together? Certainly we feel as though we run into trouble if we try to compute $1 + 2 + 3 + 4 + \cdots$. On the other hand, if we are confronted with the sum $\sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots$, then we may reasonably observe that the first n terms in the sum admit the explicit formula

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^n} = 1 - \frac{1}{2^n},$$

which can easily be proved by induction. The RHS converges to 1 as $n \to \infty$, so it is reasonable to say that the infinite sum $\sum_{n=1}^{\infty} \frac{1}{2^n}$ also converges to 1.

Definition 66.1. Given a sequence $\{a_n\}$, the corresponding series is $\sum_{n=1}^{\infty} a_n$. The *partial sums* of the series are the numbers $S_n = \sum_{k=1}^{n} a_k$. If the sequence of partial sums converges to a real number S, then we say that the series $\sum a_n$ is *convergent*, and write $\sum_{n=1}^{\infty} a_n = S$; we call S the *sum* of the series. If the sequence of partial sums does not converge, we say that the series is *divergent*.

A good way of remembering this is by the notation

$$\sum_{n=1}^{\infty} a_n = \lim_{N \to \infty} \sum_{n=1}^{N} a_n,$$

which is clearly analogous to the way we dealt with improper integrals:

$$\int_1^{\infty} f(x)\,dx = \lim_{t \to \infty} \int_1^{t} f(x)\,dx.$$

We will develop the relationship between infinite series and improper integrals further in a little while.

66.2. Geometric series

Example 66.2. A *geometric series* is a series of the form

$$a + ar + ar^2 + ar^3 + \cdots = \sum_{n=1}^{\infty} ar^{n-1},$$

where $a, r \in \mathbb{R}$. If $r = 1$ then clearly this series diverges since the nth partial sum is $S_n = an$. When $r \neq 1$, we can write the nth partial sum explicitly by observing that

$$S_n = a + ar + ar^2 + \cdots + ar^{n-1},$$
$$rS_n = ar + ar^2 + ar^3 + \cdots + ar^n.$$

Subtracting these two gives

$$S_n - rS_n = a - ar^n \quad \Rightarrow \quad S_n = \Big(\frac{1 - r^n}{1 - r}\Big)a.$$

This diverges if $|r| \geq 1$, while if $|r| < 1$ then we have

$$\lim_{n\to\infty} S_n = \Big(\frac{1 - \lim_{n\to\infty} r^n}{1 - r}\Big)a = \frac{a}{1 - r}.$$

The result of this example is important enough to be worth stating as a theorem.

Theorem 66.3. *The geometric series $\sum_{n=1}^{\infty} ar^{n-1}$ is convergent if and only if $|r| < 1$, and in this case the sum is $\frac{a}{1-r}$.*

Example 66.4. $\sum_{n=1}^{\infty} 2^{2n}3^{1-n} = \sum_{n=1}^{\infty} \frac{4^n}{3^n} \cdot 3$ diverges because it is a geometric series with $r = \frac{4}{3}$.

Example 66.5. The repeating decimal $3.2\overline{41} = 3.2414141414141\ldots$ can be written using a geometric series:

$$3.2\overline{41} = 3.2 + \frac{41}{10^3} + \frac{41}{10^5} + \frac{41}{10^7} + \cdots = 3.2 + \frac{41}{10^3}\sum_{n=1}^{\infty}(10^{-2})^{n-1}$$
$$= \frac{32}{10} + \frac{41}{10^3} \cdot \frac{1}{1 - \frac{1}{100}} = \frac{32}{10} + \frac{41}{10 \cdot 99} = \frac{3209}{990}.$$

It is also worth highlighting the case of a geometric series with $a = 1$: given any $|x| < 1$, we have

$$\sum_{n=0}^{\infty} x^n = \sum_{n=1}^{\infty} x^{n-1} = \frac{1}{1 - x}.$$

This is our first example of a *power series* representation of a function, which we will spend more time on later.

66.3. Other examples

Example 66.6. Consider the series

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \cdots.$$

To compute the partial sums and determine convergence or divergence, we can use the observation that

$$\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1},$$

and thus

$$S_n = \sum_{k=1}^{n} \frac{1}{k(k+1)} = \sum_{k=1}^{n} \Big(\frac{1}{k} - \frac{1}{k+1}\Big)$$
$$= \Big(1 - \frac{1}{2}\Big) + \Big(\frac{1}{2} - \frac{1}{3}\Big) + \Big(\frac{1}{3} - \frac{1}{4}\Big) + \cdots + \Big(\frac{1}{n} - \frac{1}{n+1}\Big) = 1 - \frac{1}{n+1}.$$

The sum $\sum_{k=1}^{n}(\frac{1}{k} - \frac{1}{k+1})$ is called a *telescoping sum* because it collapses into the short easy-to-handle expression $1 - \frac{1}{n+1}$. We now see that $S_n \to 1$ as $n \to \infty$, so the series is convergent and the infinite sum is 1.

Example 66.7. The series $\sum_{n=1}^{\infty} \frac{1}{n}$ is called the *harmonic series*. We claim that it is divergent. To prove this, observe that

$$S_1 = 1, \quad S_2 = 1 + \frac{1}{2}, \quad S_4 = 1 + \frac{1}{2} + \underbrace{\frac{1}{3} + \frac{1}{4}}_{>2\cdot\frac{1}{4}=\frac{1}{2}} > 1 + 2 \cdot \frac{1}{2},$$

and that similar estimates are available for S_8, S_{16}, etc.:

$$S_8 = S_4 + \underbrace{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}}_{>4\cdot\frac{1}{8}=\frac{1}{2}} > S_4 + \frac{1}{2} > 1 + 3 \cdot \frac{1}{2},$$

$$S_{16} = S_8 + \sum_{n=9}^{16} \frac{1}{n} > S_8 + \sum_{n=9}^{16} \frac{1}{16} = S_8 + \frac{1}{2} > 1 + 4 \cdot \frac{1}{2}.$$

In general, we have

$$S_{2^{n+1}} = S_{2^n} + \sum_{k=2^n+1}^{2^{n+1}} \frac{1}{k} > S_{2^n} + \sum_{k=2^n+1}^{2^{n+1}} \frac{1}{2^{n+1}} = S_{2^n} + 2^n \cdot \frac{1}{2^{n+1}} = S_{2^n} + \frac{1}{2},$$

and it follows by induction that for every n we have $S_{2^n} > 1 + \frac{n}{2}$. Since the RHS goes to ∞ as $n \to \infty$, we conclude that the partial sums diverge, hence the harmonic series diverges.

Remark 66.8. Writing $N = 2^n$, the lower bound above gives $\sum_{k=1}^{N} \frac{1}{k} > \frac{n}{2} = \frac{1}{2}\log_2 N$. In fact a better estimate is $\sum_{k=1}^{N} \frac{1}{k} \approx \ln N$, but this takes a little more work.

66.4. Basic theorems

Theorem 66.9. *If the series $\sum_{n=1}^{\infty} a_n$ is convergent, then the sequence of terms a_n converges to 0.*

Proof. As usual, let $S_n = \sum_{k=1}^{n} a_k$, and observe that $a_n = S_n - S_{n-1}$. If the series is convergent then $L = \lim_{n\to\infty} S_n$ exists, and by Exercise 65.2 and the limit laws we have

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} (S_n - S_{n-1}) = \lim_{n\to\infty} S_n - \lim_{n\to\infty} S_{n-1} = L - L = 0. \qquad \square$$

Remark 66.10. It is worth reiterating that every series has two sequences associated to it: the sequence of terms, which we often denote a_n, and the sequence of partial sums which we often denote S_n. Theorem 66.9 says that if S_n converges, then a_n converges to 0.

Remark 66.11. The converse of this theorem is not true; $a_n \to 0$ does not guarantee that the series converges, as the example of the harmonic series illustrates.

Corollary 66.12. *If a_n is a divergent sequence, or a convergent sequence whose limit is not equal to 0, then the corresponding series $\sum_{n=1}^{\infty} a_n$ is divergent.*

Example 66.13. $\sum_{n=1}^{\infty} \frac{2n^2}{n^2+1}$ diverges because $\frac{2n^2}{n^2+1} \to 2$.

Theorem 66.14. *If the series $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ both converge, then so do the series $\sum_{n=1}^{\infty}(a_n + b_n)$, $\sum_{n=1}^{\infty}(a_n - b_n)$, and $\sum_{n=1}^{\infty} ca_n$, where $c \in \mathbb{R}$. Moreover, we have*

$$\sum_{n=1}^{\infty}(a_n \pm b_n) = \Big(\sum_{n=1}^{\infty} a_n\Big) \pm \Big(\sum_{n=1}^{\infty} b_n\Big), \qquad \sum_{n=1}^{\infty}(ca_n) = c\Big(\sum_{n=1}^{\infty} a_n\Big).$$

Proof. We prove the result for addition and leave the others as exercises. Observe that the partial sums $S_n = \sum_{k=1}^{n}(a_k + b_k)$ satisfy

$$S_n = \Big(\sum_{k=1}^{n} a_k\Big) + \Big(\sum_{k=1}^{n} b_k\Big),$$

and the two sequences of partial sums on the RHS converge by assumption, so the limit law for addition gives

$$\lim_{n\to\infty} S_n = \lim_{n\to\infty}\Big(\sum_{k=1}^{n} a_k\Big) + \lim_{n\to\infty}\Big(\sum_{k=1}^{n} b_k\Big) = \sum_{n=1}^{\infty} a_n + \sum_{n=1}^{\infty} b_n.$$

The other two results are similar, using the corresponding limit laws. □

Example 66.15.

$$\sum_{n=1}^{\infty}\Big(\frac{2}{n(n+1)} + \frac{3}{2^n}\Big) = 2\sum_{n=1}^{\infty}\frac{1}{n(n+1)} + \frac{3}{2}\sum_{n=1}^{\infty}\Big(\frac{1}{2}\Big)^{n-1} = 2\cdot 1 + \frac{3/2}{1-\frac{1}{2}} = 2 + 3 = 5.$$

We conclude with one more general observation: convergence only depends on the 'tail' of the series, and is not affected if we change finitely many terms. The following exercise makes this precise.

Exercise 66.16. Let $\sum a_n$ and $\sum b_n$ be two series with the property that there exists $N \in \mathbb{N}$ such that $a_n = b_n$ for all $n \geq N$; in other words, we can obtain (b_n) from (a_n) by changing finitely many terms. Show that $\sum a_n$ converges if and only if $\sum b_n$ converges.

For example, if $a_n = \frac{1}{n}$ for $n < 1000$ and $a_n = 2^{-n}$ for $n \geq 1000$, then $\sum a_n$ converges even though the first part (the first 1000 terms) looks like the (divergent) harmonic series, because we can obtain a_n from the (convergent) geometric series $\sum 2^{-n}$ by changing finitely many terms.

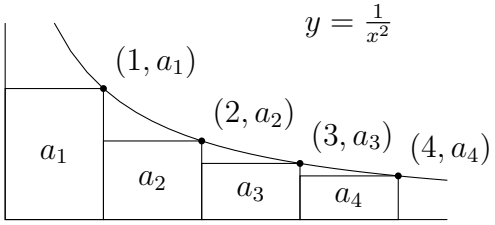
Lecture 67 The integral test

Stewart §11.3, Spivak Ch. 23

67.1. Some examples and a theorem

In Example 66.6 we showed that the series $\sum \frac{1}{n(n+1)} = \sum \frac{1}{n^2+n}$ converges. What about the series $\sum_{n=1}^{\infty} \frac{1}{n^2}$? In this case we have no nice formula for $S_n = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{n^2}$, so it is not clear how to check whether the partial sums converge.

One approach is to observe that $a_n = \frac{1}{n^2} = f(n)$, where $f(x) = \frac{1}{x^2}$, and the integral of $f(x)$ is easy to compute; then we need to compare $\sum_{k=1}^{n} \frac{1}{k^2}$ and $\int_1^n \frac{1}{x^2}\,dx$. The picture at right shows how to do this. The rectangles shown have areas $a_1, a_2, \ldots$, and they all lie underneath the graph of $\frac{1}{x^2}$. In particular, we see that for any $n \geq 2$, the region covered by the rectangles with areas $a_2, a_3, \ldots, a_n$ lies inside the region underneath the graph between $x = 1$ and $x = n$, so we have



$$a_2 + a_3 + \cdots + a_n \leq \int_1^n \frac{1}{x^2}\,dx = \Big[-\frac{1}{x}\Big]_1^n = 1 - \frac{1}{n}.$$

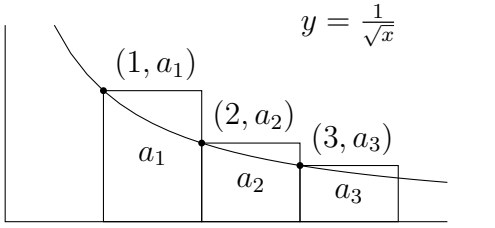
We conclude that the partial sums of the series satisfy

$$S_n = a_1 + a_2 + a_3 + \cdots + a_n \leq 2 - \frac{1}{n} \leq 2$$

for all n, so the sequence of partial sums is bounded above. Since all the terms a_n are positive, the sequence S_n is also increasing, and thus by the monotone convergence theorem it converges.

Remark 67.1. Note that the above argument gives us no information about the actual value of the infinite sum $\sum_{n=1}^{\infty} \frac{1}{n^2}$, merely that it converges. In fact one can prove that the value is $\frac{\pi^2}{6}$, but this takes significantly more work than we will enter into here.

Now consider another example, the series $\sum_{n=1}^{\infty} \frac{1}{\sqrt{n}}$. In this case the above argument does not yield an upper bound for the partial sums, because $\int_1^n \frac{1}{\sqrt{x}}\,dx = [2\sqrt{x}]_1^n = 2(\sqrt{n} - 1) \to \infty$. This suggests that perhaps we should try to prove that this series is divergent. And indeed, by modifying the above picture slightly, sliding each rectangle one unit to the right so that their tops lie *above* the graph of the function, we obtain the bound



$$S_n = a_1 + a_2 + a_3 + \cdots + a_n \geq \int_1^{n+1} \frac{1}{\sqrt{x}}\,dx = 2(\sqrt{n+1} - 1).$$

The RHS diverges to ∞ as $n \to \infty$, so we conclude that the series $\sum_{n=1}^{\infty} \frac{1}{\sqrt{n}}$ diverges.

The arguments used in these two examples lead to the following result.

Theorem 67.2 (Integral test for series)**.** *Consider the series $\sum a_n$. Suppose that $f\colon [1, \infty) \to [0, \infty)$ is a continuous, nonnegative, nonincreasing function such that $f(n) = a_n$ for all $n \in \mathbb{N}$. Then $\sum a_n$ converges if and only if the improper integral $\int_1^{\infty} f(x)\,dx$ converges.*

We will prove this theorem, together with some more detailed estimates, in Proposition 67.6 below. First we point out a couple applications.

Example 67.3. Given $p \in \mathbb{R}$, the *p-series* $\sum_{n=1}^{\infty} \frac{1}{n^p}$ converges if and only if $p > 1$. To see this, observe that for all $p \le 0$, the terms do not converge to 0, so the series diverges by Corollary 66.12. For $p > 0$, the function $f(x) = x^{-p}$ is continuous, positive, and decreasing on $[1, \infty)$, so by the integral test the series converges if and only if $\int_1^\infty x^{-p}\,dx$ converges, which occurs if and only if $p > 1$.

Example 67.4. To check convergence of $\sum \frac{\ln n}{n^2}$, we attempt to use the integral test with $f(x) = \frac{\ln x}{x^2}$. Differentiating to check whether the function is decreasing, we see that

$$f'(x) = \frac{x^2 \cdot \frac{1}{x} - (\ln x)2x}{x^4} = \frac{1 - 2\ln x}{x^3},$$

which is < 0 for all $x > \sqrt{e}$. Thus the function is decreasing on $(\sqrt{e}, \infty)$; this is not quite what Theorem 67.2 asked for, but as we will see below, it turns out to be enough, and the integral test still works. We can compute the integral by parts:

$$\int_1^t \frac{\ln x}{x^2}\,dx = \Big[-\frac{\ln x}{x}\Big]_1^t + \int_1^t \frac{1}{x^2}\,dx = \Big[-\frac{\ln x}{x} - \frac{1}{x}\Big]_1^t = -\frac{1+\ln t}{t} - (-1) = 1 - \frac{1+\ln t}{t}.$$

This converges to 1 as $t \to \infty$, so the series is convergent as well.

The hypothesis that f is nonincreasing is vital, as the following exercise shows.

Exercise 67.5. Define a function f by setting $f(n) = 0$ and $f(n + \frac{1}{2}) = 1$ for all integers n, and then connecting these points on the graph with straight lines, so that $f(n+t) = f(n-t) = 2t$ for $t \in [0, \frac{1}{2}]$. Sketch the graph of f and show that $\int_1^\infty f(x)\,dx$ diverges but $\sum_{n=1}^{\infty} f(n)$ converges.

67.2. Estimating the remainder

It is often important to understand how quickly a series converges to its limit S, by estimating the *remainder*

$$R_n := S - S_n = \sum_{k=1}^{\infty} a_k - \sum_{k=1}^{n} a_k = \lim_{N\to\infty} \sum_{k=1}^{N} a_k - \sum_{k=1}^{n} a_k = \lim_{N\to\infty} \sum_{k=n+1}^{N} a_k = \sum_{k=n+1}^{\infty} a_k.$$

Proposition 67.6. *Given a series $\sum a_n$ and a natural number $n \in \mathbb{N}$, suppose that $f\colon [n, \infty) \to [0, \infty)$ is a continuous, nonnegative, nonincreasing function such that $f(k) = a_k$ for all $k \in \mathbb{N}$. Then for every $N > n$, we have*

$$\int_{n+1}^{N+1} f(x)\,dx \le \sum_{k=n+1}^{N} a_k \le \int_n^N f(x)\,dx. \tag{67.1}$$

In particular, $\sum_{k=1}^{\infty} a_k$ converges if and only if the improper integral $\int_n^\infty f(x)\,dx$ converges, and in this case we have

$$\int_{n+1}^{\infty} f(x)\,dx \le \sum_{k=n+1}^{\infty} a_k \le \int_n^\infty f(x)\,dx. \tag{67.2}$$

Proof. For the lower bound in both cases, we let $g(x) = f(\lfloor x \rfloor)$, so that $g(x) = a_k$ for all $x \in [k, k+1)$. Then $g(x) \geq f(x)$ for all x because x is nondecreasing, and thus

$$\sum_{k=n+1}^{N} a_k = \int_{n+1}^{N+1} g(x)\,dx \geq \int_{n+1}^{N+1} f(x)\,dx.$$

Since $f \geq 0$, the only way for the improper integral to diverge is if it goes to ∞, and thus a divergent improper integral leads to a divergent sequence of partial sums, which proves one half of the claim following (67.1). For the other inequality, and the other half of the claim, let $g(x) = f(\lceil x \rceil)$, so that $g(x) = a_k$ for all $x \in (k-1, k]$, and observe that $g(x) \leq f(x)$ for all x because x is nondecreasing. Thus

$$\sum_{k=n+1}^{N} a_k = \int_{n}^{N} g(x)\,dx \leq \int_{n}^{N} f(x)\,dx.$$

If the improper integral converges, then since $f \geq 0$ we have $\int_n^N f(x)\,dx < \int_n^\infty f(x)\,dx$ for all N, and thus the partial sums $\sum_{k=n+1}^{N} a_k$ form a nondecreasing sequence that is bounded above, which implies convergence. In this case the estimates in (67.2) follow from (67.1) by taking a limit as $N \to \infty$. □

Observe that the integral test as formulated in Theorem 67.2 is a specific case of this proposition.

Example 67.7. Consider the series $\sum \frac{1}{n^2}$, and suppose we wish to find how many terms it takes for the partial sum to get within $\frac{1}{100}$ of the limit. Using (67.2) we see that

$$R_n \leq \int_n^\infty \frac{1}{x^2}\,dx = \lim_{t\to\infty} \Big[-\frac{1}{x} \Big]_n^t = \lim_{t\to\infty} \Big(\frac{1}{n} - \frac{1}{t} \Big) = \frac{1}{n}.$$

Thus we get the desired error estimate when $\frac{1}{n} \leq \frac{1}{100}$, so we need to take $n = 100$ terms.

Note that we could get a better approximation to the limit in Example 67.7 by adding one of the integrals from (67.2) to the partial sum. Indeed, under the assumptions of Proposition 67.6, the quantity $\big(\sum_{k=1}^{n} a_k\big) + \int_n^\infty f(x)\,dx$ is generally a better approximation to $\sum_{k=1}^{\infty} a_k$ than the partial sum is on its own, and we can bound the error between the approximation and the true value as follows:

$$\begin{aligned} \Big| \sum_{k=1}^{\infty} a_k - \Big(\sum_{k=1}^{n} a_k + \int_n^\infty f(x)\,dx \Big) \Big| &= \Big| \sum_{k=n+1}^{\infty} a_k - \int_n^\infty f(x)\,dx \Big| \\ &\leq \Big| \int_{n+1}^\infty f(x)\,dx - \int_n^\infty f(x)\,dx \Big| = \int_n^{n+1} f(x)\,dx. \end{aligned}$$

In Example 67.7, we see that this error bound is equal to

$$\int_n^{n+1} f(x)\,dx = \int_n^{n+1} \frac{1}{x^2}\,dx = -\frac{1}{x}\Big|_n^{n+1} = \frac{1}{n} - \frac{1}{n+1} = \frac{1}{n(n+1)} < \frac{1}{n^2},$$

and so to get within $\frac{1}{100}$ of the limit we could use $n = 10$ and then add the improper integral (which we can calculate explicitly in this case).

Lecture 68 Comparison tests and alternating series

Stewart §11.4 and §11.5, Spivak Ch. 23

68.1. Comparison tests

We know that the geometric series $\sum_{n=1}^{\infty} \frac{1}{2^n}$ converges. What about $\sum_{n=1}^{\infty} \frac{1}{2^n+1}$? This is not a geometric series but looks similar enough that we might expect similar convergence behavior. And indeed, if we compare the partial sums $S_n = \sum_{k=1}^{n} \frac{1}{2^k+1}$ to the partial sums $T_n = \sum_{k=1}^{n} \frac{1}{2^k}$, we can observe that the inequality $\frac{1}{2^k+1} \le \frac{1}{2^k}$ immediately implies that

$$S_n = \sum_{k=1}^{n} \frac{1}{2^k+1} \le \sum_{k=1}^{n} \frac{1}{2^k} = T_n \le T := \lim_{n\to\infty} T_n,$$

where the inequality $T_n \le T$ uses the fact that the terms $\frac{1}{2^n}$ are nonnegative so the sequence of partial sums T_n is nondecreasing. The partial sums S_n are nondecreasing for the same reason, and they are bounded above by T, so the monotone convergence theorem implies that $\sum \frac{1}{2^n+1}$ is convergent.

This argument is worth codifying. Note the analogy between the following result and Theorem 49.14 for improper integrals.

Theorem 68.1 (Comparison test)**.** *Consider two series $\sum a_n$ and $\sum b_n$ whose terms are all nonnegative: $a_n, b_n \ge 0$.*

(1) If $\sum b_n$ is convergent and $a_n \le b_n$ for all n, then $\sum a_n$ is convergent.
(2) If $\sum b_n$ is divergent and $a_n \ge b_n$ for all n, then $\sum a_n$ is divergent.

Proof. Consider the partial sums $S_n = \sum_{k=1}^{n} a_k$ and $T_n = \sum_{k=1}^{n} b_k$. For the first claim, we have $T = \lim_{n\to\infty} T_n$, so $S_n \le T_n \le T$ for all n, and thus S_n is a bounded monotonic sequence, which therefore converges. The second half of the theorem follows from the first half by taking a contrapositive and reversing the roles of a_n and b_n. □

Exercise 68.2. Prove that the theorem remains true if the inequalities are only assumed to hold for all *sufficiently large* n. That is, in part (1) we can replace the assumption that $a_n \le b_n$ for all n with the assumption that there exists $N \in \mathbb{N}$ such that $a_n \le b_n$ for all $n \ge N$, and similarly in part (2).

Example 68.3. The series $\sum \frac{\ln n}{n}$ has $\frac{\ln n}{n} \ge \frac{1}{n}$ for all $n \ge 3$; since the harmonic series $\sum \frac{1}{n}$ diverges, the series $\sum \frac{\ln n}{n}$ diverges as well.

Example 68.4. Consider the series $\sum \frac{\ln n}{n^2}$. We saw in Example 67.4 that this converges by the integral test. We can also prove this using the comparison test. Recall that $\lim_{n\to\infty} \frac{\ln n}{\sqrt{n}} = 0$, and thus $\ln n \le \sqrt{n}$ for all sufficiently large n. For all such n we have

$$\frac{\ln n}{n^2} \le \frac{\sqrt{n}}{n^2} = n^{\frac{1}{2}-2} = n^{-3/2} = \frac{1}{n^{3/2}}.$$

The p-series $\sum 1/n^{3/2}$ is convergent (since $\frac{3}{2} > 1$), so the comparison test shows that $\sum \frac{\ln n}{n^2}$ is convergent as well.

As these examples show, it is often the case that a series can be compared to either a p-series or a geometric series, and so these are usually the first candidates that you should consider.

Remark 68.5. As in Proposition 67.6 and (67.2), one can use the comparison test to estimate the remainder in a convergent sum: if $a_n \le b_n$ for all $n \ge N$, then the remainder term for $\sum a_n$ is bounded above the the remainder term for $\sum b_n$.

Sometimes it is easier to compare two sequences asymptotically than it is to go term-by-term. The following result shows that this is enough to study convergence.

Theorem 68.6 (Limit comparison test)**.** *Consider two series $\sum a_n$ and $\sum b_n$ whose terms are all positive: $a_n, b_n > 0$. Suppose that there is a real number $c > 0$ such that $\lim_{n\to\infty} \frac{a_n}{b_n} = c$. Then $\sum a_n$ is convergent if and only if $\sum b_n$ is convergent.*

Proof. By the assumption that $\frac{a_n}{b_n} \to c > 0$, there exists $N \in \mathbb{N}$ such that for all $n \ge N$ we have $\frac{c}{2} < \frac{a_n}{b_n} < 2c$, or equivalently, $\frac{c}{2} b_n < a_n < 2cb_n$. By Theorem 66.14, $\sum \frac{c}{2} b_n$ and $\sum 2cb_n$ converge if and only if $\sum b_n$ converges. Thus convergence of $\sum b_n$ implies convergence of $\sum 2cb_n$, and hence convergence of $\sum a_n$ by the comparison test. Similarly, convergence of $\sum a_n$ implies convergence of $\sum \frac{c}{2} b_n$ by the comparison test, and hence convergence of $\sum b_n$. □

Example 68.7. The series $\sum \frac{1}{2^n - 1}$ converges by applying the limit comparison test with the reference series $\sum \frac{1}{2^n}$, which is a convergent geometric series: observe that

$$\lim_{n\to\infty} \frac{1/(2^n - 1)}{1/2^n} = \lim_{n\to\infty} \frac{1}{1 - 2^{-n}} = 1.$$

Example 68.8. Consider the series

$$\sum_{n=1}^{\infty} \frac{2n^2 + 3n}{\sqrt{5 + n^5}}.$$

To determine what series to compare this to, observe that for large n we have

$$\frac{2n^2 + 3n}{\sqrt{5 + n^5}} \approx \frac{2n^2}{n^{5/2}} = \frac{2}{n^{1/2}}.$$

Since the p-series $\sum n^{-1/2}$ is divergent, we can prove that $\sum \frac{2n^2+3n}{\sqrt{5+n^5}}$ is divergent by observing that

$$\lim_{n\to\infty} \frac{(2n^2 + 3n)/(\sqrt{5 + n^5})}{n^{-1/2}} = \lim_{n\to\infty} \frac{2n^2 + 3n}{n^{-1/2}\sqrt{5 + n^5}} \cdot \frac{n^{-2}}{n^{-2}} = \lim_{n\to\infty} \frac{2 + 3n^{-1}}{\sqrt{5n^{-5} + 1}} = 2,$$

and then applying the limit comparison test.

68.2. Alternating series

The integral test and comparison tests only apply to series with nonnegative entries. It is also sometimes important to understand series with both positive and negative entries. The simplest class of series like this is the following.

Definition 68.9. A series $\sum a_n$ is *alternating* if its terms alternate between positive and negative, so that writing $b_n = |a_n|$, we have $a_n = (-1)^n b_n$ for every n. We also call the series alternating if we have $a_n = (-1)^{n-1} b_n$ for every n.

The following theorem says that for alternating series, the condition in Theorem 66.9 is actually sufficient for convergence of the series, in sharp distinction to what happens for more general series.

Theorem 68.10. *If $\sum a_n$ is an alternating series for which $b_n = |a_n|$ is a nonincreasing sequence ($b_{n+1} \le b_n$ for all n) that converges to 0 ($\lim_{n\to\infty} b_n = 0$), then $\sum a_n$ converges.*

Proof. Suppose that $a_n = (-1)^{n-1} b_n$ (the case with $(-1)^n$ is similar), so that the series is

$$b_1 - b_2 + b_3 - b_4 + b_5 - b_6 + \cdots .$$

Then the even partial sums satisfy

$$S_{2n+2} = S_{2n} + a_{2n+1} + a_{2n+2} = S_{2n} + b_{2n+1} - b_{2n+2} \ge S_{2n},$$

where the last inequality uses the fact that $b_{2n+2} \le b_{2n}$. This shows that the sequence of even partial sums is nondecreasing. Moreover, for every n we have

$$\begin{aligned} S_{2n} &= b_1 - b_2 + b_3 - b_4 + b_5 - \cdots + b_{2n-1} - b_{2n} \\ &= b_1 - (b_2 - b_3) - (b_4 - b_5) - \cdots - (b_{2n-2} - b_{2n-1}) - b_{2n} \le b_1, \end{aligned} \tag{68.1}$$

where the last inequality uses the fact that each term in brackets is nonnegative (since b_k is nonincreasing). By the monotone convergence theorem, the sequence of even partial sums S_{2n} converges to some limit S. Since $b_k \to 0$, we also have

$$\lim_{n\to\infty} S_{2n+1} = \lim_{n\to\infty} (S_{2n} + b_{2n+1}) = \lim_{n\to\infty} S_{2n} + \lim_{n\to\infty} b_{2n+1} = S + 0 = S.$$

We leave it as an exercise to show that the two results $\lim_{n\to\infty} S_{2n} = S$ and $\lim_{n\to\infty} S_{2n+1} = S$ together imply $\lim_{n\to\infty} S_n = S$, which completes the proof. □

Example 68.11. Although the harmonic series $\sum \frac{1}{n}$ diverges, the *alternating* harmonic series $\sum \frac{(-1)^{n-1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \cdots$ converges.

Theorem 68.12. *Under the conditions of Theorem 68.10, if $S = \sum_{n=1}^{\infty} a_n$, then the remainder term $R_n = S - S_n$ satisfies $|R_n| \le |a_{n+1}|$ for all n.*

Proof. Observe that the odd partial sums and the even partial sums converge to S from different sides, so S is always between S_n and S_{n+1}. In particular,

$$|S - S_n| \le |S_{n+1} - S_n| = |a_{n+1}|.$$ □

Lecture 69 Absolute convergence, ratio and root tests

Stewart §11.6, Spivak Ch. 23

69.1. Absolute convergence

Now that we are discussing series with both positive and negative terms, the following definition becomes important.

Definition 69.1. A series $\sum a_n$ is *absolutely convergent* if $\sum |a_n|$ is convergent.

Theorem 69.2. *If $\sum a_n$ is absolutely convergent, then it is convergent.*

Proof. For every n, we have $0 \leq a_n + |a_n| \leq 2|a_n|$, so $\sum(a_n + |a_n|)$ is convergent by the comparison test. Then $\sum a_n = \sum((a_n + |a_n|) - |a_n|)$ is convergent by Theorem 66.14 as the difference of two convergent series. □

Definition 69.3. A series is *conditionally convergent* if it is convergent, but not absolutely convergent.

Example 69.4. The alternating series $\sum \frac{(-1)^{n-1}}{n^2}$ is absolutely convergent, while $\sum \frac{(-1)^{n-1}}{n}$ is only conditionally convergent.

Example 69.5. $\sum \frac{\cos n}{n^2}$ is absolutely convergent by the comparison test, since $|\frac{\cos n}{n^2}| \leq \frac{1}{n^2}$ and the series $\sum \frac{1}{n^2}$ is convergent.

The crucial difference between absolute and conditional convergence is the way in which rearrangements of a series behave. We are used to the idea that rearranging the terms in a sum does not change its value. The following two exercises ask you to prove that this continues to be true for an absolutely convergent infinite series.

Exercise 69.6. Let $\sum a_n$ be a series whose terms are all nonnegative ($a_n \geq 0$). Prove that the value of the infinite sum is the supremum of all the partial sums, and use this fact to deduce that $\sum a_n$ remains the same no matter what order the terms of the series are written in.

Exercise 69.7. Show that given any series $\sum a_n$, there are sequences $b_n, c_n \geq 0$ such that the $a_n = b_n - c_n$ for all n. (Hint: one of b_n, c_n should be $|a_n|$, and the other should be 0.) Then show that $\sum a_n$ is absolutely convergent if and only if $\sum b_n$ and $\sum c_n$ are both convergent, and use the previous exercise to deduce that for an absolutely convergent series, the value of the infinite sum is unchanged by rearranging the terms of the series.

When a series is only conditionally convergent, the story changes dramatically, as the following example illustrates: let S be the sum of the alternating harmonic series, so

$$S = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \frac{1}{9} - \frac{1}{10} + \cdots .$$

Multiplying every term by $\frac{1}{2}$, Theorem 66.14 gives

$$\frac{1}{2}S = 0 + \frac{1}{2} + 0 - \frac{1}{4} + 0 + \frac{1}{6} + 0 - \frac{1}{8} + 0 + \frac{1}{10} + \cdots .$$

Adding these two sequences and using Theorem 66.14 again gives

$$\frac{3}{2}S = 1 + 0 + \frac{1}{3} - \frac{1}{2} + \frac{1}{5} + 0 + \frac{1}{7} - \frac{1}{4} + \frac{1}{9} + 0 + \cdots .$$

Observe that every odd-numbered term in this last series is the same as the corresponding term in the first series, so all of the terms $\frac{1}{2n+1}$ appear exactly once. Every term $-\frac{1}{2n}$ appears exactly once in the last series also, but they are 'stretched out' further by placing 0's in between. Thus this last series is a rearrangement of the first one, and both series are convergent, but their sums are different!

In fact, the story is even stranger; it is possible to show that if $\sum a_n$ is a conditionally convergent series, then for *any* $r \in \mathbb{R}$ there is a conditionally convergent series $\sum b_n$ with sum r that is a rearrangement of $\sum a_n$. Thus conditionally convergent series must be treated with a certain amount of caution.

69.2. Ratio test

In light of the previous discussion, it is useful to be able to determine when a series is absolutely convergent.

Theorem 69.8 (Ratio test)**.** *Consider a series $\sum a_n$ with nonzero terms, and let $L = \lim_{n\to\infty} |\frac{a_{n+1}}{a_n}|$ if the limit exists.*

(1) If $L < 1$, then $\sum a_n$ is absolutely convergent.
(2) If $L > 1$, or if the limit is ∞, then $\sum a_n$ is divergent.
(3) If $L = 1$, or if the limit does not exist, then the ratio test is inconclusive and gives no information.

Proof. For the first part, choose $r \in (L, 1)$; then there is $N \in \mathbb{N}$ such that $|\frac{a_{n+1}}{a_n}| < r$ for all $n \geq N$. This gives $|a_{n+1}| < r|a_n|$, and iterating gives $|a_{N+k}| < |a_N|r^k$ for all $k \geq 1$. Since $\sum_{k=1}^{\infty} |a_N|r^k$ is convergent (a geometric series with $|r| < 1$), the comparison test implies that $\sum |a_n|$ is convergent as well.

For the second part, $L > 1$ implies that there is N such that $|a_{n+1}| > |a_n|$ for all $n \geq N$, so $\lim a_n \neq 0$, and by Corollary 66.12 the series is divergent. □

Example 69.9. Consider the series $\sum (-1)^n \frac{n^3}{2^n}$. The limit in the ratio test is

$$L = \lim_{n\to\infty} \frac{(n+1)^3/2^{n+1}}{n^3/2^n} = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^3 \cdot \frac{1}{2} = \frac{1}{2} < 1,$$

so the series is absolutely convergent.

Example 69.10. The series $\sum \frac{2^n}{n!}$ is absolutely convergent because

$$\lim_{n\to\infty} \frac{2^{n+1}/(n+1)!}{2^n/n!} = \lim_{n\to\infty} \frac{2}{n+1} = 0.$$

Observe that the ratio test does not catch all convergent series. Indeed, although $\sum \frac{1}{n^2}$ is convergent, the ratio test gives

$$L = \lim_{n\to\infty} \frac{1/(n+1)^2}{1/n^2} = \lim_{n\to\infty} \frac{n^2}{(n+1)^2} = \lim_{n\to\infty} \frac{1}{(1+\frac{1}{n})^2} = 1,$$

and thus is inconclusive.

69.3. Root test

Theorem 69.11 (Root test)**.** *Consider a series $\sum a_n$, and let $L = \lim_{n\to\infty} \sqrt[n]{|a_n|}$ if the limit exists.*

(1) If $L < 1$, then $\sum a_n$ is absolutely convergent.
(2) If $L > 1$, or if the limit is ∞, then $\sum a_n$ is divergent.
(3) If $L = 1$, or if the limit does not exist, then the ratio test is inconclusive and gives no information.

Proof. In the first case, once again choose $r \in (L, 1)$, so there is $N \in \mathbb{N}$ such that for all $n \geq N$, we have $\sqrt[n]{|a_n|} = |a_n|^{1/n} < r$. Raising both sides to the nth power gives $|a_n| < r^n$, and since $\sum r^n$ converges, the comparison test implies that $\sum |a_n|$ converges as well. In the second case, a similar argument shows that $|a_n| \to \infty$, so $\sum a_n$ diverges. □

Example 69.12. The series $\sum(\frac{n}{2n+1})^n$ is absolutely convergent, because

$$\lim_{n\to\infty} \sqrt[n]{\left(\frac{n}{2n+1}\right)^n} = \lim_{n\to\infty} \frac{n}{2n+1} = \frac{1}{2} < 1.$$

Although the ratio test would work in this example, it would be rather messier to carry out the computations. (Try it!) In some other cases, the root test may work where the ratio test fails.

Exercise 69.13. Let $a_n = \frac{1}{3^n}$ when n is odd, and $a_n = \frac{2}{3^n}$ when n is even. Use the root test to prove that $\sum a_n$ converges. Show that the limit in the ratio test does not exist.

On the other hand, if the ratio test fails because the limit is equal to 1, then the root test will not work either.

Exercise 69.14. Prove that if the limit in the ratio test exists, then the limit in the root test exists as well, and the two limits are the same.

This last exercise implies that whenever the ratio test works, the root test would also work, although the computations might be harder (they could also be easier). If $L = 1$ in either the ratio test or the root test, then neither of the tests will determine convergence.[55]

Lecture 70 Power series

Stewart §11.8, Spivak Ch. 24

Definition 70.1. A *power series* is a series of the form $\sum_{n=0}^{\infty} c_n x^n$, where $c_n \in \mathbb{R}$ are constants, called the *coefficients* of the power series, and $x \in \mathbb{R}$ is a variable. For a given value of x, a power series becomes a series in the sense we have been studying so far,

[55] If you read the preceding passage carefully, though, you will see that it is possible to have $L = 1$ in the root test while the limit in the ratio test does not exist.

and can be either convergent or divergent. The *domain* of the power series is the set of all x for which the power series converges. When x lies in this domain, we write

$$f(x) = \sum_{n=0}^{\infty} c_n x^n = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \cdots$$

for the function determined by the power series.

Example 70.2. We already saw from the formula for the sum of a geometric series that the function $\frac{1}{1-x}$ can be represented by the power series

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n = 1 + x + x^2 + x^3 + \cdots,$$

where the coefficients are $c_n = 1$, and the series converges iff $|x| < 1$.

It is possible for the domain to be all of $\mathbb{R}$.

Example 70.3. Let $f(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$. Then for all $x \in \mathbb{R}$, the terms $a_n = \frac{x^n}{n!}$ satisfy

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{x^{n+1}/(n+1)!}{x^n/n!}\right| = \left|\frac{x}{n+1}\right| \to 0 \text{ as } n \to \infty$$

and thus the series is convergent by the ratio test.[56]

It is also possible for the domain to be a single point.

Example 70.4. Let $f(x) = \sum_{n=0}^{\infty} n! x^n$. Then the series converges for $x = 0$ because all terms are 0, but for $x \neq 0$ the terms $a_n = n!x^n$ satisfy

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{(n+1)!x^{n+1}}{n!x^n}\right| = |(n+1)x| \to \infty \text{ as } n \to \infty$$

and thus the series is divergent by the ratio test.

One can also center the series at a point $a \neq 0$ by replacing x^n with $(x-a)^n$.

Example 70.5. Consider the series

$$\sum_{n=1}^{\infty} \frac{(x-2)^n}{n}.$$

For a given $x \in \mathbb{R}$, the ratio of successive terms (in absolute value) is

$$\left|\frac{(x-2)^{n+1}/(n+1)}{(x-2)^n/n}\right| = \left|\frac{(x-2)n}{n+1}\right| \to |x-2|.$$

Thus by the root test the series converges when $|x-2| < 1$ (that is, when $1 < x < 3$), and diverges when $|x-2| > 1$. At $x = 1$ it converges (alternating harmonic series) and at $x = 3$ it diverges (harmonic series).

Theorem 70.6. *Given a power series $\sum_{n=0}^{\infty} c_n(x-a)^n$, one of the following three things happens.*

(1) The series converges when $x = a$ and diverges for all $x \neq a$.

[56] Compare this to Example 69.10, which did this same calculation in the case $x = 2$.

(2) *The series converges absolutely for all* $x \in \mathbb{R}$.
(3) *There exists* $R > 0$ *such that the series converges absolutely when* $|x - a| < R$ *and diverges when* $|x - a| > R$.

Definition 70.7. The number R from Theorem 70.6 is called the *radius of convergence* of the power series.

Before proving Theorem 70.6, we observe that while this result gives absolute convergence in the interior of the interval, it is silent on what happens at the endpoints, where we can have either convergence or divergence.

Example 70.8. Fixing any $p \in \mathbb{R}$, we see that the power series $\sum_{n=0}^{\infty} n^{-p}x^n$ has the property that

$$\left|\frac{(n+1)^{-p}x^{n+1}}{n^{-p}x^n}\right| = \left|\left(1+\frac{1}{n}\right)^{-p}x\right| \to |x| \text{ as } n \to \infty,$$

and thus by the ratio test its radius of convergence is $R = 1$. The behavior at the endpoints $x = \pm 1$ depends on p.

(1) For $p = 0$, $\sum x^n$ diverges at both endpoints.
(2) For $p = 1$, $\sum \frac{x^n}{n}$ converges conditionally when $x = -1$, and diverges when $x = 1$.
(3) For $p = 2$, $\sum \frac{x^n}{n^2}$ converges absolutely at both endpoints.

The following exercises illustrate the remaining possible behaviors at the endpoints.

Exercise 70.9. Prove that $\sum(-2x)^n/n$ has radius of convergence $1/2$, converges conditionally at $x = 1/2$, and diverges at $x = -1/2$.

Exercise 70.10. Prove that the power series

$$x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \frac{x^9}{9} - \cdots$$

has radius of convergence 1 and converges conditionally at both endpoints.

Exercise 70.11. Prove that if a power series converges absolutely at one endpoint of its interval of convergence, then it converges absolutely at the other endpoint as well.

Now we prove Theorem 70.6, starting with a lemma.

Lemma 70.12. *Let* c_n *be a sequence of coefficients and* $r \in \mathbb{R}$ *a real number such that* $\sum c_n r^n$ *converges. Then given any* $a, x \in \mathbb{R}$ *with* $|x - a| < |r|$*, the series* $\sum c_n(x-a)^n$ *converges absolutely.*

Proof. Convergence of $\sum c_n r^n$ implies that $c_n r^n \to 0$, so there is $N \in \mathbb{N}$ such that $|c_n r^n| < 1$ for all $n \geq N$. For such n we then have

$$|c_n(x-a)^n| = |c_n r^n| \cdot \left|\frac{(x-a)^n}{r^n}\right| < \left|\frac{x-a}{r}\right|^n,$$

and thus $\sum |c_n(x-a)^n|$ converges by the comparison test, using the geometric series $\sum |\frac{x-a}{r}|^n$, which is convergent because $|x - a| < |r|$. □

Proof of Theorem 70.6. Consider the set $A = \{r \geq 0 : \sum c_n r^n \text{ converges}\}$. This is nonempty because $0 \in A$. If it is unbounded then for every $x \in \mathbb{R}$ there exists $r \in A$ such

that $r > |x-a|$, and thus by Lemma 70.12, the series $\sum c_n(x-a)^n$ converges at x. This puts us in case (2).

Now suppose that A is bounded, and let $R = \sup A$ be its least upper bound. Then for every $x \in \mathbb{R}$ with $|x-a| > R$, we see that $\sum c_n(x-a)^n$ diverges, otherwise Lemma 70.12 would imply that $\sum c_n r^n$ converges for some $r \in (R, |x-a|)$, contradicting the claim that R is an upper bound for A. If $R = 0$, then we are in case (1); the series diverges for all $x \neq a$. If $R > 0$, then for every x with $|x-a| < R$ we can choose $r \in A$ with $|x-a| < r$ (since R is the *least* upper bound) and use Lemma 70.12 to deduce that $\sum c_n(x-a)^n$ converges absolutely. This puts us in case (3). □

Theorem 70.6 tells us that every power series converges on an interval (which could be a single point, or all of $\mathbb{R}$), and diverges on its complement. The radius of convergence can often – but not always – be determined by using either the ratio test or the root text.

Theorem 70.13 (Ratio test for radius of convergence)**.** *Suppose $\sum c_n(x-a)^n$ is a power series for which the limit $L = \lim_{n\to\infty} |c_{n+1}/c_n|$ exists. Then the radius of convergence is $R = 1/L$. If $L = 0$ then the radius of convergence is ∞; if $L = \infty$ then the radius of convergence is* 0*.*

Proof. We apply the ratio test. Given $x \in \mathbb{R}$, we have

$$\lim_{n\to\infty} \left| \frac{c_{n+1}(x-a)^{n+1}}{c_n(x-a)^n} \right| = L|x-a|.$$

This is < 1 when $|x-a| < 1/L$, implying absolute convergence, and > 1 when $|x-a| > 1/L$, implying divergence. If $L = 0$ then the limit is always 0, giving absolute convergence, and if $L = \infty$ then the limit is ∞ for all $x \neq a$, giving divergence. □

Theorem 70.14 (Root test for radius of convergence)**.** *Suppose $\sum c_n(x-a)^n$ is a power series for which the limit $L = \lim_{n\to\infty} |c_n|^{1/n}$ exists. Then the radius of convergence is $R = 1/L$. If $L = 0$ then the radius of convergence is ∞; if $L = \infty$ then the radius of convergence is* 0*.*

Proof. This is exactly the same as the previous proof except we use the following computation:

$$\lim_{n\to\infty} |c_n(x-a)^n|^{1/n} = |x-a| \lim_{n\to\infty} |c_n|^{1/n} = L|x-a|.$$ □

Lecture 71 — Calculus with power series

Stewart §11.9, Spivak Ch. 24

Of the various elementary functions that we have encountered so far, polynomials are among the easiest to work with; they can be added, subtracted, and multiplied relatively easily, and differentiation and integration are also straightforward using the rules

$$\frac{d}{dx}x^n = nx^{n-1} \quad \text{and} \quad \int x^n\,dx = \frac{x^{n+1}}{n+1} + C.$$

Since polynomials are finite sums of expressions like these, they can be manipulated without incident. For power series, which are *infinite* sums, more care is needed, as indicated by the results about rearrangements of conditionally convergent series. The following theorem says that differentation and integration work as expected.

Theorem 71.1. *Let $\sum_{n=0}^{\infty} c_n(x-a)^n$ be a power series with radius of convergence $R > 0$, and let $f\colon (a-R, a+R) \to \mathbb{R}$ be the function defined by $f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n$. Then f is continuously differentiable on this interval, and therefore also integrable; moreover, we can represent f' and $\int f$ by the following power series:*

$$f'(x) = \sum_{n=1}^{\infty} nc_n(x-a)^{n-1},$$
$$\int f(x)\,dx = C + \sum_{n=0}^{\infty} c_n \frac{(x-a)^{n+1}}{n+1}.$$

Both of these power series also have radius of convergence R.

Proof. The proof that these power series have the same radius of convergence R can be given by a mild modification of Lemma 70.12, which we leave as an exercise.

The proof that they actually give the derivative and integral of f is more difficult.[57] For simplicity we prove the result for the derivative, with $a = 0$. The result for the integral is a corollary, and the proof for other values of a is the same; one simply needs to replace x with $x - a$ everywhere that it appears. Thus we need to show that if we define two functions $f, g\colon (-R, R) \to \mathbb{R}$ by the power series

$$f(x) = \sum_{n=0}^{\infty} c_n x^n \quad \text{and} \quad g(x) = \sum_{n=1}^{\infty} nc_n x^{n-1},$$

then $f'(x) = g(x)$ for all $x \in (-R, R)$. By definition of the derivative, we have

$$f'(x) = \lim_{y\to x} \frac{1}{y-x} \sum_{n=1}^{\infty} c_n(y^n - x^n) = \lim_{y\to x} \sum_{n=1}^{\infty} c_n(y^{n-1} + xy^{n-2} + x^2y^{n-3} + \cdots + x^{n-2}y + x^{n-1}),$$

where we used the factorization $y^n - x^n = (y-x)(y^{n-1} + xy^{n-2} + \cdots + x^{n-1})$. The expression in brackets can be written as $\sum_{j=0}^{n-1} y^j x^{n-1-j}$, and we can write nx^{n-1} as $\sum_{j=0}^{n-1} x^{n-1}$, so we conclude that

$$f'(x) - g(x) = \lim_{y\to x} \sum_{n=1}^{\infty} c_n \Big(\sum_{j=0}^{n-1} (y^j x^{n-1-j} - x^{n-1}) \Big). \tag{71.1}$$

We must show that this quantity vanishes. It is tempting to move the limit inside the sums and observe that $\lim_{y\to x} y^j x^{n-1-j} - x^{n-1} = 0$; however, while this would be allowed

[57] I learned the proof here from a blog post by Tim Gowers at `https://gowers.wordpress.com/2014/02/22/differentiating-power-series/` – the "more common" proof of this theorem uses the concept of *uniform convergence* of a sequence of functions, which is beyond the scope of this course.

if the sums were both finite, it is *not* always allowed for infinite sums. Indeed, the infinite sum is itself a limit, and we could more properly write the above equation as

$$f'(x) - g(x) = \lim_{y\to x} \lim_{N\to\infty} \sum_{n=1}^{N} c_n \Big(\sum_{j=0}^{n-1} (y^j x^{n-1-j} - x^{n-1}) \Big).$$

Thus before we can pass $\lim_{y\to x}$ inside the sums, we would need to interchange the order of the limits. This is an issue that has not arisen for us so far, and that we will not treat in any detail, save by issuing this warning: be very, very careful if anyone tries to sell you a computation in which the order of two limits are interchanged. Sometimes it is valid, and sometimes it is not; in this course we have not developed the tools to tell the difference.

Instead, we will estimate the magnitude of the inner sum as follows:

$$\Big| \sum_{j=0}^{n-1} (y^j x^{n-1-j} - x^{n-1}) \Big| \le \sum_{j=0}^{n-1} |x|^{n-1-j} |y^j - x^j| = \sum_{j=0}^{n-1} |x|^{n-1-j} |y-x| \Big| \sum_{i=0}^{j-1} y^i x^{j-1-i} \Big|.$$

Here the last equality once again uses the factorization for $y^j - x^j$ that we used before. Recalling that R is the radius of convergence of the power series and $|x| < R$, fix r such that $|x| < r < R$, and choose y close enough to x that $|y| < r$. Then we have $|\sum_{i=0}^{j-1} y^i x^{j-1-i}| \le \sum_{i=0}^{j-1} r^{j-1} = j r^{j-1}$, and the above estimate gives

$$\Big| \sum_{j=0}^{n-1} (y^j x^{n-1-j} - x^{n-1}) \Big| \le \sum_{j=0}^{n-1} r^{n-1-j} |y-x| \cdot j r^{j-1} = \sum_{j=0}^{n-1} j r^{n-2} |y-x| = \frac{n(n-1)}{2} r^{n-2} |y-x|.$$

Returning to the expressions in (71.1), we see that

$$\Big| \sum_{n=1}^{\infty} c_n \Big(\sum_{j=0}^{n-1} (y^j x^{n-1-j} - x^{n-1}) \Big) \Big| \le \sum_{n=1}^{\infty} c_n \cdot \frac{n(n-1)}{2} r^{n-2} |y-x|$$

and thus

$$|f'(x) - g(x)| \le \lim_{y\to x} |y-x| \sum_{n=1}^{\infty} \frac{n(n-1)}{2} c_n r^{n-2} = 0,$$

provided the last sum converges. The fact that it converges for $r \in (0, R)$ is a consequence of the exercise at the beginning of this proof, since this is the power series that "should" represent $f''(r)$, and you were asked to prove in that exercise that formal term-by-term differentiation does not change the radius of convergence. □

Example 71.2. We have seen that on $(-1, 1)$, the function $f(x) = \frac{1}{1-x}$ is represented by the power series

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n = 1 + x + x^2 + x^3 + \cdots. \tag{71.2}$$

Differentiating and integrating term-by-term, as in Theorem 71.1, we see that

$$f'(x) = \sum_{n=1}^{\infty} n x^{n-1} \quad \text{and} \quad \int f(x)\, dx = C + \sum_{n=0}^{\infty} \frac{x^{n+1}}{n+1}.$$

Since $f'(x) = \frac{1}{(1-x)^2}$, we obtain the new power series representation

$$\frac{1}{(1-x)^2} = \sum_{n=1}^{\infty} nx^{n-1} = 1 + 2x + 3x^2 + 4x^3 + 5x^4 + \cdots . \tag{71.3}$$

Similarly, since $\int f(x)\,dx = -\ln(1-x) + C$, we obtain

$$\ln(1-x) = C - \sum_{n=0}^{\infty} \frac{x^{n+1}}{n+1}.$$

When $x = 0$ the LHS vanishes, so the constant of integration is $C = 0$, and we have the power series representation

$$\ln(1-x) = -\sum_{n=1}^{\infty} \frac{x^n}{n} = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \cdots . \tag{71.4}$$

Remark 71.3. It is possible to show that (71.4) remains valid not just on the interval $(-1, 1)$, but also at the endpoint $x = -1$; see the extra credit problems on the homework for an outline of the proof of *Abel's theorem*, which establishes this fact.[58] Observe that then this series gives

$$\ln 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \cdots ,$$

so that $\ln 2$ is the value of the sum of the alternating harmonic series.

Example 71.4. Replacing x in (71.2) by $(-x^2)$, we obtain the power series representation

$$\frac{1}{1+x^2} = \frac{1}{1-(-x^2)} = \sum_{n=0}^{\infty}(-x^2)^n = \sum_{n=0}^{\infty}(-1)^n x^{2n} = 1 - x^2 + x^4 - x^6 + x^8 - \cdots ,$$

which is valid on the interval $(-1, 1)$. Integrating gives

$$\tan^{-1}(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots ,$$

where the constant of integration is 0 because $\tan^{-1}(0) = 0$. This is valid on the interval $(-1, 1)$ (though we observe that the function $\tan^{-1}(x)$ is defined on all of $\mathbb{R}$). Once again, Abel's theorem can be used to extend its validity to include $x = 1$, and we obtain the following formula:

$$\frac{\pi}{4} = \tan^{-1}(1) = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots .$$

Example 71.5. The *Bessel function of order* 0 is given by the power series

$$J_0(x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{4^n (n!)^2}.$$

[58] At $x = 1$, the series diverges, which is consistent with the fact that $\ln 0$ is undefined.

Using the root test one can check that it converges for all $x \in \mathbb{R}$. Using Theorem 71.1 one can compute its derivative:

$$J_0'(x) = \sum_{n=1}^{\infty} \frac{(-1)^n 2n x^{2n-1}}{4^n (n!)^2}$$

Similarly one can compute J_0'', and verify that J_0 is a solution of the differential equation

$$x^2 f''(x) + x f'(x) + x^2 f(x) = 0, \tag{71.5}$$

which arises (among other places) when one studies the shape of a vibrating drumhead.

The DE in (71.5) does not admit a closed form solution in terms of functions we have studied earlier, so this last example illustrates the utility of power series as a tool. Even in situations where a problem can be solved using other methods, the power series is often easier to compute: for example, we could compute $\int \frac{1}{1+x^4}\,dx$ using partial fractions, but it is fairly long and tedious to do so, while using power series we quickly get

$$\begin{aligned}\int \frac{1}{1+x^4}\,dx &= \int (1 - x^4 + x^8 - x^{12} + x^{16} - \cdots)\,dx \\ &= C + x - \frac{x^5}{5} + \frac{x^9}{9} - \frac{x^{13}}{13} + \cdots .\end{aligned}$$

Of course it may then be difficult or impossible to translate this power series back into a closed form for the integral, but in many cases the power series is just as useful, especially if what we are after is a numerical approximation.

Lecture 72 — Taylor and Maclaurin series

Stewart §11.10, Spivak Ch. 24

72.1. Obtaining coefficients from higher derivatives

It is natural to ask whether a given function can be represented by a power series, and if so, how the coefficients of that series can be found. In light of Theorem 71.1, we see that any function represented by a power series needs to be at least differentiable on the interior of the interval of convergence; thus we cannot expect to represent $f(x) = |x|$ by a power series around 0.

Moreover, since a power series representation for f gives a power series for f' with the same radius of convergence, we see that f' must be differentiable as well. Continuing this line of reasoning, every derivative $f^{(n)}$ must exist. Is this enough? It turns out that the answer is no.

Exercise 72.1. Prove that the function

$$f(x) = \begin{cases} e^{-1/x^2} & x > 0, \\ 0 & x \le 0 \end{cases}$$

has derivatives of all orders at 0, but is not given by a power series in any open interval containing 0.

A function that has derivatives of all orders is called *smooth*. A function that is given by a convergent power series is called *analytic*. Analytic functions are smooth, but not every smooth function is analytic. For the moment, we address the question of how to find the coefficients of the power series, *assuming that f is indeed represented by a power series*. To this end, suppose that near $a \in \mathbb{R}$, a function f is given by a power series

$$f(x) = c_0 + c_1(x-a) + c_2(x-a)^2 + c_3(x-a)^3 + (\text{terms containing } (x-a)^4).$$

Then its derivative is given by

$$f'(x) = c_1 + 2c_2(x-a) + 3c_3(x-a)^2 + (\text{terms containing } (x-a)^3),$$

its second derivative is given by

$$f''(x) = 2c_2 + 2\cdot 3\cdot c_3(x-a) + (\text{terms containing } (x-a)^2),$$

and in general, its nth derivative is given by

$$f^{(n)}(x) = n!c_n + (\text{terms containing } (x-a)).$$

Since any term containing $(x-a)$ vanishes when we put $x = a$, we conclude that

$$f(a) = c_0, \quad f'(a) = c_1, \quad f''(a) = 2c_2, \quad \dots \quad f^{(n)}(a) = n!c_n.$$

Thus we can recover the coefficients c_n from the values of the higher derivatives of f at a. We have proved the following theorem.

Theorem 72.2. *If f has a power series representation $\sum_{n=0}^{\infty} c_n(x-a)^n$, with radius of convergence $R > 0$, then we have*

$$c_n = \frac{f^{(n)}(a)}{n!} \text{ for all } n = 0, 1, 2, \dots.$$

Thus for every $x \in (a-R, a+R)$, we have

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n. \tag{72.1}$$

The power series in (72.1) is called the *Taylor series* for f. In the case when $a = 0$, it is also called the *Maclaurin series*:

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

Example 72.3. For the exponential function $f(x) = e^x$, we have $f^{(n)}(x) = e^x$ for every $n = 0, 1, 2, \dots$, and thus $c_n = \frac{f^{(n)}(0)}{n!} = \frac{1}{n!}$. Thus the Maclaurin series for e^x (the Taylor series around 0) is

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots. \tag{72.2}$$

Observe that for every $x \in \mathbb{R}$, we have

$$\frac{x^{n+1}/(n+1)!}{x^n/(n!)} = \frac{x}{n+1} \to 0 \text{ as } n \to \infty,$$

so the series converges absolutely by the ratio test. Thus the radius of convergence is $R = \infty$.

What the above example does *not* immediately tell us is whether or not the power series in (72.2) actually converges to e^x. Could it converge to something else instead? We will investigate this question in the next section.

72.2. Approximation by polynomials

Definition 72.4. Given $n \in \mathbb{N}$, the nth *Taylor polynomial* for $f(x)$ around a is the nth partial sum of the Taylor series:

$$T_n(x) := \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k. \tag{72.3}$$

Thus the Taylor series converges to $f(x)$ at x if and only if $T_n(x) \to f(x)$ as $n \to \infty$. Equivalently, writing $R_n(x) = f(x) - T_n(x)$ for the nth *remainder* term at x, we have convergence if and only if $R_n(x) \to 0$. Exercise 72.1 gives an example where this does not occur. Can we give conditions under which it does occur?

Theorem 72.5 (Taylor's inequality)**.** *Let $f\colon (a-d, a+d) \to \mathbb{R}$ be $n+1$ times differentiable, and suppose that $|f^{(n+1)}(x)| \le M$ for all $x \in (a-d, a+d)$. Then for every x in this interval, the nth remainder term $R_n(x) = f(x) - T_n(x)$ satisfies*

$$|R_n(x)| \le \frac{M}{(n+1)!}|x-a|^{n+1}.$$

Proof. We prove this by induction in n when $x \in (a, a+d)$; the proof for $x \in (a-d, a)$ is similar. First consider the case $n = 0$. In this case we have $T_0(x) = f(a)$ for all x, so $R_0(x) = f(x) - f(a)$, and by assumption $|f'(x)| \le M$ for all $x \in (a, a+d)$, so

$$|R_0(x)| = |f(x) - f(a)| = \Big|\int_a^x f'(t)\,dt\Big| \le \int_a^x |f'(t)|\,dt \le \int_a^x M\,dt = M(x-a).$$

Now suppose that $n \ge 1$ and that the result holds for $n-1$. Then observe that since $R_n(x) = f(x) - T_n(x)$ by definition, we have $R_n'(x) = f'(x) - T_n'(x)$. We claim that T_n' is the degree $(n-1)$ Taylor polynomial for f': indeed, differentiating (48.1) gives

$$T_n'(x) = \sum_{k=1}^{n} \frac{f^{(k)}(a)}{(k-1)!}(x-a)^{k-1} = \sum_{j=0}^{n-1} \frac{f^{(j+1)}(a)}{j!}(x-a)^j,$$

and since $f^{(j+1)}(a) = (f')^{(j)}(a)$, this proves the claim. Moreover, we have $|(f')^{(n)}(x)| = |f^{(n+1)}(x)| \le M$ for all $x \in (a, a+d)$, so by the inductive hypothesis we obtain

$$|R_n'(x)| \le \frac{M}{n!}|x-a|^n.$$

Integrating this gives

$$|R_n(x)| = \Big|\int_a^x R_n'(t)\,dt\Big| \le \int_a^x |R_n'(t)|\,dt \le \int_a^x \frac{M}{n!}(t-a)^n\,dt = \Big[\frac{M(t-a)^{n+1}}{(n+1)!}\Big]_a^x$$

This last expression is equal to $M(x-a)^{n+1}/(n+1)!$, which proves the theorem. $\square$

Remark 72.6. In fact, one can prove the following more explicit formulas for the remainder term:

$$R_n(x) = \frac{1}{n!}\int_a^x (x-t)^n f^{(n+1)}(t)\,dt,$$
$$R_n(x) = \frac{f^{(n+1)}(t)}{(n+1)!}(x-a)^{n+1} \text{ for some } t \text{ between } x \text{ and } a.$$

Observe that each of these implies Taylor's inequality. Note also that the second of these reduces to the Mean Value Theorem in the case $n = 0$.

Returning to the case of $f(x) = e^x$, we see from Taylor's inequality that for every $|x| \le d$ and every $n \in \mathbb{N}$, we have

$$|e^x - T_n(x)| \le \frac{e^d}{(n+1)!}|x|^{n+1} \to 0 \text{ as } n \to \infty,$$

where the convergence to 0 follows because $\sum \frac{x^n}{n!}$ converges. This proves that the Maclaurin series $\sum \frac{x^n}{n!}$ does indeed converge to e^x, so that we can write

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \text{ for all } x \in \mathbb{R}. \tag{72.4}$$

In particular, putting $x = 1$ gives the following infinite series for e:

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = 1 + 1 + \frac{1}{2} + \frac{1}{3!} + \frac{1}{4!} + \cdots.$$

Observe also that if we differentiate (72.4) term-by-term, we get the same power series, consistent with the fact that $\frac{d}{dx}e^x = e^x$.

Example 72.7. We could also take the Taylor series of e^x around another point; for example, with $a = 1$ we see that $f^{(n)}(a) = e^a = e^1 = e$ for all n, and thus

$$e^x = \sum_{n=0}^{\infty} \frac{e}{n!}(x-1)^n,$$

where the argument that the remainder terms go to 0 is similar to the one given above.

Example 72.8. For $f(x) = \sin x$, we have

$$f'(x) = \cos x, \quad f''(x) = -\sin x, \quad f'''(x) = -\cos x,$$

and in general,

$$f^{(n)}(x) = \begin{cases} \sin x & \text{if } n \equiv 0 \pmod 4, \\ \cos x & \text{if } n \equiv 1 \pmod 4, \\ -\sin x & \text{if } n \equiv 2 \pmod 4, \\ -\cos x & \text{if } n \equiv 3 \pmod 4. \end{cases}$$

Thus the nth derivatives at 0 are $0, 1, 0, -1, 0, 1, 0, -1, \ldots$, and the Maclaurin series is

$$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!}.$$

Since all derivatives have absolute value ≤ 1 for every x, we can take $M = 1$ in Taylor's inequality and obtain

$$|R_n(x)| \leq \frac{|x|^{n+1}}{(n+1)!} \to 0,$$

which proves that the Maclaurin series converges to $\sin x$ for every $x \in \mathbb{R}$, and thus

$$\sin x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots . \tag{72.5}$$

By making a similar argument for $f(x) = \cos x$, or by differentiating (72.5) term-by-term and applying Theorem 71.1, we get

$$\cos x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots . \tag{72.6}$$

Remark 72.9. Using (72.4), (72.5), and (72.6), we see that writing i for a (complex) square root of -1, we have

$$\begin{aligned} e^{ix} &= 1 + (ix) + \frac{(ix)^2}{2!} + \frac{(ix)^3}{3!} + \frac{(ix)^4}{4!} + \frac{(ix)^5}{5!} + \frac{(ix)^6}{6!} + \frac{(ix)^7}{7!} + \frac{(ix)^8}{8!} + \cdots \\ &= 1 + ix - \frac{x^2}{2!} - i\frac{x^3}{3!} + \frac{x^4}{4!} + i\frac{x^5}{5!} - \frac{x^6}{6!} - i\frac{x^7}{7!} + \frac{x^8}{8!} + \cdots \\ &= \Big(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} + \cdots\Big) + i\Big(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots\Big) \\ &= \cos x + i \sin x, \end{aligned}$$

where the equality in the third line uses the fact that Taylor series converge *absolutely* on the interior of the interval of convergence, and thus we can rearrange the series without changing the sum.

72.3. Binomial series

Recall from the Binomial Theorem that given a positive integer n and any real number x, we can write

$$(1+x)^n = 1 + nx + \binom{n}{2}x^2 + \cdots + \binom{n}{n-2}x^{n-2} + nx^{n-1} + x^n = \sum_{k=0}^{n} \binom{n}{k} x^k. \tag{72.7}$$

Writing $f(x) = (1+x)^n$, we see that for $0 \leq k \leq n$, we have

$$f^{(k)}(x) = \frac{d^k}{dx^k}(1+x)^n = n(n-1)\cdots(n-k+1)(1+x)^{n-k},$$

and so $f^{(k)}(0) = n(n-1)\cdots(n-k+1) = n!/(n-k)!$. When $k > n$ we have $f^{(k)}(0) = 0$, and so the Maclaurin series for $(1+x)^n$ is

$$\sum_{k=0}^{\infty} \frac{1}{k!} f^{(k)}(0) x^k = \sum_{k=0}^{n} \frac{1}{k!} \frac{n!}{(n-k)!} x^k = \sum_{k=0}^{n} \binom{n}{k} x^k,$$

which recovers the formula in (72.7). But we can compute the Maclaurin series even if n is *not* an integer; consider the function $f(x) = (1+x)^\alpha$ for an arbitrary real number α. Then for any $k \geq 0$, we have

$$f^{(k)}(x) = \frac{d^k}{dx^k}(1+x)^\alpha = \alpha(\alpha-1)\cdots(\alpha-k)(1+x)^{\alpha-k}.$$

Note that if α happens to be a positive integer, then this expression vanishes whenever $k \geq \alpha$. Evaluating at $x = 0$ gives

$$f^{(k)}(0) = \alpha(\alpha-1)\cdots(\alpha-k),$$

and so the Maclaurin series for $(1+x)^\alpha$ is

$$\sum_{k=0}^{\infty} \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!}x^k. \tag{72.8}$$

Extending the notation from the integer case, we write

$$\binom{\alpha}{k} := \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!} \tag{72.9}$$

and refer to these numbers as *binomial coefficients*. Once again we see that if α is a positive integer, then $\binom{\alpha}{k} = 0$ for all $k > \alpha$, while for $0 \leq k \leq \alpha$ the formula in (72.9) reduces to our usual definition of binomial coefficients $\frac{\alpha!}{k!(\alpha-k)!}$.

When α is *not* a positive integer, we can determine the radius of convergence of the power series in (72.8) by applying the ratio test (Theorem 70.13):

$$\left|\frac{\binom{\alpha}{k+1}}{\binom{\alpha}{k}}\right| = \left|\frac{\alpha(\alpha-1)\cdots(\alpha-k)}{(k+1)!}\frac{k!}{\alpha(\alpha-1)\cdots(\alpha-k+1)}\right| = \left|\frac{\alpha-k}{k+1}\right| \to 1$$

as $k \to \infty$, and thus the radius of convergence is 1. (Can you determine when it does and does not converge at the endpoints ± 1?)

Theorem 72.10. *For every $\alpha \in \mathbb{R}$ and $|x| < 1$, we have $(1+x)^\alpha = \sum_{k=0}^{\infty}\binom{\alpha}{k}x^k$.*

Proof. (This proof was omitted in the lecture.) Let $g(x) = \sum_{k=0}^{\infty}\binom{\alpha}{k}x^k$. Our goal is to write a differential equation involving g' and g that we can solve to find a closed-form expression for g, which will turn out to be $(1+x)^\alpha$.

Differentiating g term-by-term and using Theorem 71.1, we see that

$$\begin{aligned} g'(x) &= \sum_{k=1}^{\infty}\binom{\alpha}{k}kx^{k-1} = \sum_{k=1}^{\infty}\frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!}kx^{k-1} \\ &= \sum_{k=1}^{\infty}\frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{(k-1)!}x^{k-1} = \sum_{j=0}^{\infty}\frac{\alpha(\alpha-1)\cdots(\alpha-j)}{j!}x^j, \end{aligned}$$

where in the last step we reindexed the sum by putting $j = k-1$. Multiplying the second-to-last series by x gives

$$xg'(x) = \sum_{k=1}^{\infty}\frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{(k-1)!}x^k,$$

and adding these two formulas (renaming both indices to i for consistency) gives

$$\begin{aligned} g'(x) + xg'(x) &= \alpha + \sum_{i=1}^{\infty} \Big(\frac{\alpha(\alpha-1)\cdots(\alpha-i)}{i!} + \frac{\alpha(\alpha-1)\cdots(\alpha-i+1)}{(i-1)!} \Big) x^i \\ &= \alpha + \sum_{i=1}^{\infty} \frac{\alpha(\alpha-1)\cdots(\alpha-i+1)}{(i-1)!} \Big(\frac{\alpha-i}{i} + 1 \Big) x^i \\ &= \alpha + \alpha \sum_{i=1}^{\infty} \frac{\alpha(\alpha-1)\cdots(\alpha-i+1)}{i!} x^i = \alpha g(x). \end{aligned}$$

Thus we have proved that

$$g'(x) = \frac{\alpha g(x)}{1+x},$$

and we conclude that

$$\frac{d}{dx} \ln g(x) = \frac{g'(x)}{g(x)} = \frac{\alpha}{1+x}.$$

Using the fact that $g(0) = 1$, we obtain

$$\ln g(x) = \ln g(0) + \int_0^x \frac{\alpha}{1+t}\, dt = \alpha \ln(1+t) \Big|_0^x = \alpha \ln(1+x).$$

Taking exponentials gives $g(x) = (1+x)^\alpha$ and completes the proof. □

72.4. Power series arithmetic

If we want to find the Maclaurin series for $f(x) = e^x \sin x$, we could proceed by computing all of its derivatives and using Theorem 72.2. However, if we want to avoid using the product rule over and over again, there is another way. We already know the power series representations of e^x and $\sin x$, and it turns out (though we will not prove it) that multiplying these series as though they were polynomials gives the power series representation of their product:

$$\begin{aligned} e^x \sin x &= \Big(1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \cdots\Big)\Big(x - \frac{1}{6}x^3 + \cdots\Big) \\ &= \Big(x + x^2 + \frac{1}{2}x^3 + \frac{1}{6}x^4 + \cdots\Big) + \Big(-\frac{1}{6}x^3 - \frac{1}{6}x^4 - \frac{1}{12}x^5 - \frac{1}{36}x^6 - \cdots\Big) + \cdots \\ &= x + x^2 + \frac{1}{3}x^3 + \cdots, \end{aligned}$$

where each instance of $\cdots$ indicates terms of degree 4 or higher. A similar computation can be done when we divide power series, using a long-division-type procedure analogous to polynomial long division; this can be used, for example, to find the Maclaurin series for $\tan x = \frac{\sin x}{\cos x}$.

Review of convergence for series

Stewart §11.7, Spivak Ch. 23

This review is not included in a numbered lecture, but will be done in the lecture hour preceding the third test.

When determining convergence or divergence of a given series, the first fundamental fact to keep in mind is Corollary 66.12: if the sequence of terms a_n does not converge to 0, then the series $\sum a_n$ diverges.

If the sequence of terms does go to 0 – that is, if $\lim_{n\to\infty} a_n = 0$ – then the series $\sum a_n$ might converge and might diverge. If all the terms are ≥ 0, then it is reasonable to think of this convergence/divergence as being determined by whether the terms a_n go to 0 "quickly enough".

Keep in mind the harmonic series $\sum \frac{1}{n}$ as a reminder that convergence to 0 of the *sequence* of terms does not imply convergence of the *series* (which requires convergence of the sequence of partial sums); this is an example where the terms go to 0 slowly enough that the series diverges.

Two classes of series are especially important to keep in mind.

- Given a real number p, the corresponding *p-series* is $\sum \frac{1}{n^p}$. The series converges if $p > 1$ and diverges if $p \leq 1$. (Note that the terms go to 0 for every $p > 0$.)
- Given real numbers a, r, the corresponding *geometric series* is $\sum ar^{n-1}$. Assuming $a \neq 0$, the series converges if $|r| < 1$ and diverges if $|r| \geq 1$. (Note that the terms go to 0 if and only if $|r| < 1$.)

For a more general series with terms $a_n \geq 0$ that go to 0, the question of "do the terms go to 0 quickly enough for the series to converge" can often be answered by comparing to one of these two kinds of series and using the Comparison Test or the Limit Comparison Test. Intuitively, one can think of ar^{n-1} as going to 0 *exponentially quickly*, and $\frac{1}{n^p}$ as going to 0 *polynomially quickly with degree p*. Then one way to verify that the terms a_n go to 0 "quickly enough for the series to converge" is to relate them to a sequence that goes to 0 either exponentially quickly, or polynomially quickly with degree > 1.[59]

For series with some negative terms, the situation is a little more subtle and there are two ways that convergence can happen.

- *Absolute convergence:* The terms a_n go to 0 quickly enough that $\sum |a_n|$ converges.
- *Conditional convergence:* The terms go to 0 slowly enough that $\sum |a_n|$ diverges, but there is enough cancellation between positive and negative terms that the series $\sum a_n$ still converges.

For examples of conditional convergence, we can look to alternating series of the form $\sum (-1)^n b_n$, where $b_n \geq 0$. The Alternating Series Test says that such a series converges if and only if $b_n \to 0$ (so the naive divergence test is actually a necessary and sufficient condition in this case), and so if we choose $b_n \to 0$ with $\sum b_n$ divergent, then $\sum (-1)^n b_n$ is conditionally convergent. This includes the alternating harmonic series $\sum (-1)^n \frac{1}{n}$.

Three other convergence tests are worth keeping in mind.

(1) If the terms a_n can be written as $a_n = f(n)$ where f is a continuous nonnegative nonincreasing function and we can determine the convergence or divergence of the improper integral $\int_1^\infty f(x)\,dx$, then the integral test can be applied.
(2) If the terms a_n contain factorials or other products, it is often useful to use the ratio test.
(3) If the terms a_n contain an nth power, it is often useful to use the root test.

[59] This is not an exhaustive list of the different rates with which a sequence can go to 0, but exponential and polynomial rates are the most important.

Example 72.11. $\sum \frac{n-1}{2n+1}$ diverges by Corollary 66.12 because

$$\lim_{n\to\infty} \frac{n-1}{2n+1} = \lim_{n\to\infty} \frac{1-\frac{1}{n}}{2+\frac{1}{n}} = \frac{1}{2} \neq 0.$$

Example 72.12. $\sum \frac{\sqrt{n^3+1}}{3n^3+4n^2+2}$ has terms on the same order of magnitude as $n^{3/2}/n^3 = n^{-3/2}$, so we use the limit comparison test and observe that

$$\lim_{n\to\infty} \frac{\sqrt{n^3+1}}{3n^3+4n^2+2} \div n^{-3/2} = \lim_{n\to\infty} \frac{n^{-3/2}\sqrt{n^3+1}}{n^{-3}(3n^3+4n^2+2)} = \lim_{n\to\infty} \frac{\sqrt{1+n^{-3}}}{3+4n^{-1}+2n^{-3}} = \frac{1}{3}.$$

Since $\sum n^{-3/2}$ is a convergent p-series, the limit comparison test implies that the original series converges as well.

Example 72.13. Consider $\sum ne^{-n^2}$. The function xe^{-x^2} can be integrated by the substitution $u = -x^2$ and is decreasing as soon as $\frac{d}{dx}(xe^{-x^2}) = e^{-x^2} - 2x^2e^{-x^2} < 0$, which is true for all $x > 1$; since $\int_1^t xe^{-x^2}\,dx = [-\frac{1}{2}e^{-x^2}]_1^t = \frac{1}{2}(e^{-1} - e^{-t}) \to \frac{1}{2e}$ as $t \to \infty$, the integral test tells us that $\sum ne^{-n^2}$ is convergent. This fact can also be proved using the ratio test, the root test, or the comparison test (using a geometric series as reference); try it!

For power series of the form $\sum_{n=0}^{\infty} c_n(x-a)^n$, there is always a *radius of convergence* R such that the series converges absolutely on $(a-R, a+R)$ and diverges when $|x-a| > R$. It is possible to have $R = 0$ or $R = \infty$; the power series $\sum \frac{x^n}{n!}$ is an important example with $R = \infty$.

When $x \in (a-R, a+R)$ so that $|x-a| < R$, the proof of absolute convergence goes by comparing the series to a geometric series (exponential behavior). At the endpoints $x = a \pm R$, we typically have to compare to a p-series (polynomial behavior) and there are multiple possibilities (each of the following examples has $R = 1$):

(1) absolute convergence at both endpoints ($\sum \frac{1}{n^2}x^n$);
(2) conditional convergence at both endpoints ($\sum \frac{(-1)^n}{2n+1}x^{2n+1}$);
(3) conditional convergence at one endpoint and divergence at the other ($\sum \frac{1}{n}x^n$);
(4) divergence at both endpoints ($\sum x^n$).

The radius of convergence can often be determined using a version of the Ratio or Root Tests. On the interval $(a-R, a+R)$, the power series defines a function f whose derivative and integral can be written as power series (with the same radius of convergence) using analogues of the familiar formulas for polynomials. Writing down the power series representations of the higher-order derivatives $f^{(n)}(x)$ and evaluating them at a reveals that when f admits a power series representation, it must be given by its *Taylor series* $\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$. The partial sums of this series are the *Taylor polynomials*, and the difference between a function and its Taylor polynomial can be controlled by *Taylor's inequality* provided we have a good upper bound on $|f^{(n+1)}(x)|$.

We saw power series representations of e^x, $\sin x$, $\cos x$, $\frac{1}{1-x}$, $\ln(1-x)$, $\frac{1}{1+x^2}$, $\tan^{-1} x$, and $(1+x)^\alpha$ for $\alpha \in \mathbb{R}$; these were obtained with Taylor series, with the geometric series formula, and using differentiation and integration from known formulas.

Part IX. Conic sections, planetary motion

Lecture 73 Parabolas

Stewart §10.5, Spivak Chapter 4 appendix 2

73.1. Different descriptions

You may have encountered various ways to describe parabolas, or various properties that these curves have. Let us start our discussion of conic sections by recalling some (five!) of these descriptions; see Figure 15 below.

(1) *Analytic geometry – equation.* A parabola is the graph of the function $y = x^2$, or more generally $y = ax^2 + bx + c$, where $a, b, c \in \mathbb{R}$ are arbitrary parameters with $a \neq 0$. This gives a parabola that opens up (if $a > 0$) or down (if $a < 0$). For parabolas opening left and write we write $x = ay^2 + by + c$.

(2) *Physics – dynamics.* A projectile moving without air resistance in a uniform gravitational field flies along a parabola. Thus if we throw a ball, a parabola describes its flight path.

(3) *Physics – optics, acoustics.* A parabola has a distinguished point called the *focus* with the property that if a light bulb is placed at the focus and emits beams of light in all directions, then when these beams are reflected off of the parabola, they all become parallel to each other; see the first picture below. This is used in building headlights for cars. If the direction of the arrows is reversed this principle means that parallel incoming lines are all reflected to a single point (the focus), which is useful in building satellite dishes, radio telescopes, and parabolic microphones.

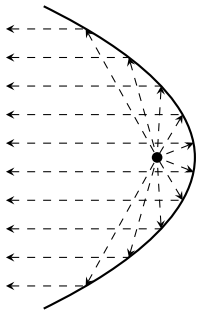
(4) *Two-dimensional geometry – focus and directrix.* A parabola has a point F, called the *focus*, and a line ℓ, called the *directrix*, with the property that given any point Q on the parabola, if R is the closest point to Q on the directrix ℓ, then $|QF| = |QR|$; see the second picture below. Notice that the point lying halfway between F and ℓ is on the parabola; this point is called the *vertex*.

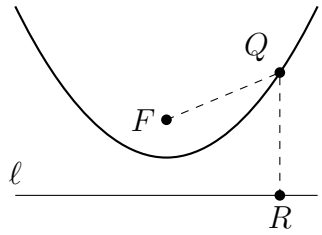
(5) *Three-dimensional geometry – cross-section of cone.* Given a plane $\mathbf{P}$ and a cone C in three-dimensional space, if $\mathbf{P}$ is parallel to one of the lines containing the vertex of C, then the cross-section $\mathbf{P} \cap C$ is a parabola, as shown in the third picture below.

At first glance, it is not at all clear why the five different descriptions in the list above should all determine the same curve. Why should they be equivalent?

73.2. Analytic geometry and projectile dynamics

We have already seen one equivalence: the first two descriptions are equivalent because if a projectile has constant horizontal velocity $v \neq 0$, initial vertical velocity w, and is subject to constant downward acceleration g, then its position (x, y) as a function of





picture to be added

FIGURE 15. Different representations of a parabola.

time t satisfies

$$\dot{x} = v, \quad \dot{y}(0) = w, \quad \ddot{y} = -g.$$

Integrating gives

$$\dot{y}(t) = \dot{y}(0) + \int_0^t \ddot{y}(\tau)\, d\tau = w - gt.$$

If the initial position is (x_0, y_0), then we have

$$x(t) = x_0 + vt, \quad y(t) = y_0 + \int_0^t \dot{y}(\tau)\, d\tau = y_0 + \int_0^t (w - g\tau)\, d\tau = y_0 + wt - \frac{g}{2}t^2.$$

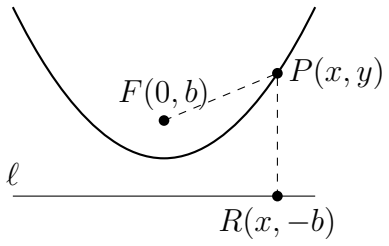
This gives the trajectory as a parametric curve. To write y as a function of x we solve the first equation and get $t = (x - x_0)/v$ (recall that we assumed $v \neq 0$, so the projectile is not simply moving straight up and down), and deduce that

$$y = y_0 + w \cdot \frac{x - x_0}{v} - \frac{g}{2} \cdot \frac{(x - x_0)^2}{v^2}.$$

Thus y is a quadratic function of x, so the projectile follows a parabola.

73.3. Analytic geometry and plane Euclidean geometry

Now we show that the first and fourth descriptions from the list above are equivalent; that is, a curve with the focus-directrix property described there is in fact given as the graph of a quadratic polynomial.



Suppose C is a curve in the plane that has the focus-directrix property; that is, there is a point F and a line ℓ (not containing F) such that a point P in the plane lies on the curve C if and only if the distance $|PF|$ is equal to the distance from P to ℓ. We choose a coordinate system in which F lies on the positive y-axis and the origin is halfway between F and ℓ; thus $F = (0, b)$ and ℓ is given by the equation $y = -b$. Consider a point P with coordinates (x, y). Then the closest point on ℓ to P is the point R that lies directly beneath P, which has coordinates $(x, -b)$. The focus-directrix property says that P lies on C if and only if $|PF| = |PR|$, or equivalently, $|PF|^2 = |PR|^2$. Observe that

$$|PF|^2 = (x - 0)^2 + (y - b)^2 = x^2 + y^2 - 2by + b^2,$$

$$|PR|^2 = (y-(-b))^2 = (y+b)^2 = y^2+2by+b^2.$$

Thus P lies on C if and only if

$$x^2+y^2-2by+b^2 = y^2+2by+b^2 \quad \Leftrightarrow \quad x^2 = 4by \quad \Leftrightarrow \quad y = \frac{1}{4b}x^2.$$

In other words, C is the graph of the quadratic $y = ax^2$, where $a = \frac{1}{4b}$.

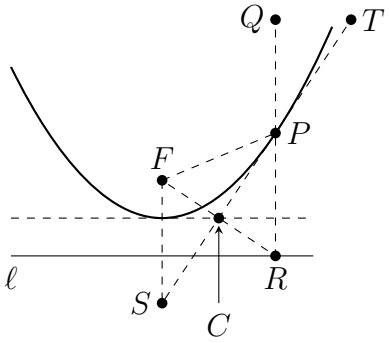
More generally, if C is a parabola with focus $F = (p,q)$ and directrix $y = r$ for some $p, q, r \in \mathbb{R}$ with $r \neq q$, then writing $k = (q+r)/2$, we can do a horizontal translation by p and a vertical translation by k to move F to $(0,b)$ and ℓ to $y = -b$, where $b = (q-r)/2$. The argument above gives the formula for the translated curve, so the original curve is the graph of

$$y-k = \frac{1}{4b}(x-p)^2 \quad \Leftrightarrow \quad y = \frac{q+r}{2} + \frac{1}{2(q-r)}(x-p)^2.$$

Remark 73.1. If we consider a parabola with a vertical directrix, then we interchange the roles of x and y in the above computations. If we consider a parabola with a directrix that is neither horizontal nor vertical, then the coordinates become more complicated, at least as long as we use a rectangular coordinate system, and we will not pursue this any further here.

73.4. Focus-directrix property and reflection property

Now we prove that the focus-directrix property implies the property that lines emanating from the focus are reflected to parallel lines, or equivalently, that an incoming line perpendicular to the directrix is reflected to the focus.



Consider such a line QP, and imagine a beam of light traveling along this line. When it reaches the point P on the parabola, what does it do? The law of reflection says that its outgoing angle is equal to its incoming angle. But angle with what? Whenever we discuss the angle that a line makes with a curve (or that two curves make with each other), what we mean is the angle that is made with the *tangent line* to the curve. In other words, if TS is the tangent line to the parabola at P, then the incoming beam is reflected towards the focus if and only if $\measuredangle FPS = \measuredangle QPT$. Since $\measuredangle QPT = \measuredangle RPS$, we conclude that

> *in order to prove that the incoming beam along QP is reflected towards the point F, it suffices to prove that the line bisecting the angle $\measuredangle FPR$ is the tangent line to the parabola at P.*

At this point one might expect that we should introduce some coordinates and use the description of the parabola in terms of the graph of a quadratic polynomial, since describing a tangent line involves computing a derivative. But in fact, we can get a little more mileage out of a purely geometric approach.

Let ℓ' be the line bisecting $\measuredangle FPR$, and let S be the point where ℓ' intersects the vertical line through F (in the picture, T also lies on ℓ'). Then $\measuredangle RPC = \measuredangle FPC$

and $|PF| = |PR|$ by the focus-directrix property, so the triangles FPC and RPC are congruent. In particular, C is the midpoint of FR, and ℓ' and FR are perpendicular.

Recall a fundamental property of perpendicular bisectors: ℓ' is the set of points in the plane that are the same distance from both F and R. If a point is on the same side of ℓ' as F is, then it is closer to F than it is to R, and vice versa. In particular, if X is *any* point on the parabola, then we have

$$|XF| = \text{distance from } X \text{ to } \ell \leq |XR|.$$

Moreover, the latter inequality is strict unless X lies directly above R; that is, unless $X = P$. This means that P is the only point where the parabola intersects ℓ', and that every other point on the parabola lies above ℓ'. Then the proof that ℓ' is the tangent line to the parabola at P, and thus that the focus-directrix property implies the reflection property, is completed by the following exercise.

Exercise 73.2. Let $I \subset \mathbb{R}$ be an open interval and suppose that $f\colon I \to \mathbb{R}$ is differentiable at a point $a \in I$. Suppose moreover that $y = mx + b$ is a line in the plane with the property that $f(a) = ma + b$, and $f(x) > mx + b$ for all $x \neq a$. Prove that $f'(a) = m$, so that in particular this line is the tangent line to the graph of f at a.

Remark 73.3. Without the assumption that f is differentiable at a, the conclusion of the exercise could fail; consider the absolute value function $f(x) = |x|$ and $a = 0$.

73.5. Dandelin spheres

The only remaining property to consider is the one that gives *conic sections* their names: a parabola is the cross-section obtained by intersecting a cone with a plane that is parallel to one of the lines that makes up the edge of the cone. For this we use a beautiful and elegant geometric argument discovered by the 19th century Belgian mathematician Germinal Dandelin.

In the following it is useful to keep Figure 16 in mind; we orient the cone so that it opens straight up, and consider the curve formed by intersecting the cone with a plane $\mathbf{P}$. Now imagine that we drop a tiny sphere – like a small scoop of ice cream – into the cone. When it comes to rest near the bottom of the sphere, it will be tangent to it along a horizontal circle. If we increase the size of the sphere – changing our analogy, we may imagine that the sphere is a balloon that we inflate – then the sphere, and its circle of tangency, will rise higher on the cone. When the sphere is very small, it will lie entirely beneath the plane $\mathbf{P}$. When it is sufficiently large, some points of it will lie above $\mathbf{P}$. By the Intermediate Value Theorem, for some size of the sphere, it intersects the plane $\mathbf{P}$ in exactly one point.[60] Then one can deduce from Exercise 73.2 that $\mathbf{P}$ is tangent to the sphere. The sphere obtained by this process is the *Dandelin sphere* associated to the cone and to the plane $\mathbf{P}$; we denote it by S.

Let $F = S \cap \mathbf{P}$; this point will be the focus. Let C_1 be the circle of points where S intersects (and is tangent to) the cone. Since the picture is symmetric under rotation around a vertical axis, this lies in a horizontal plane $\mathbf{H}$. Let $\ell = \mathbf{P} \cap \mathbf{H}$; this line will be the directrix. Given a point Q on the intersection of $\mathbf{P}$ and the cone, we must show that $|QF| = |Q\ell|$.

[60] It is a good exercise to make this statement a little more formal by writing down a continuous function that vanishes precisely when there is exactly one point of intersection.

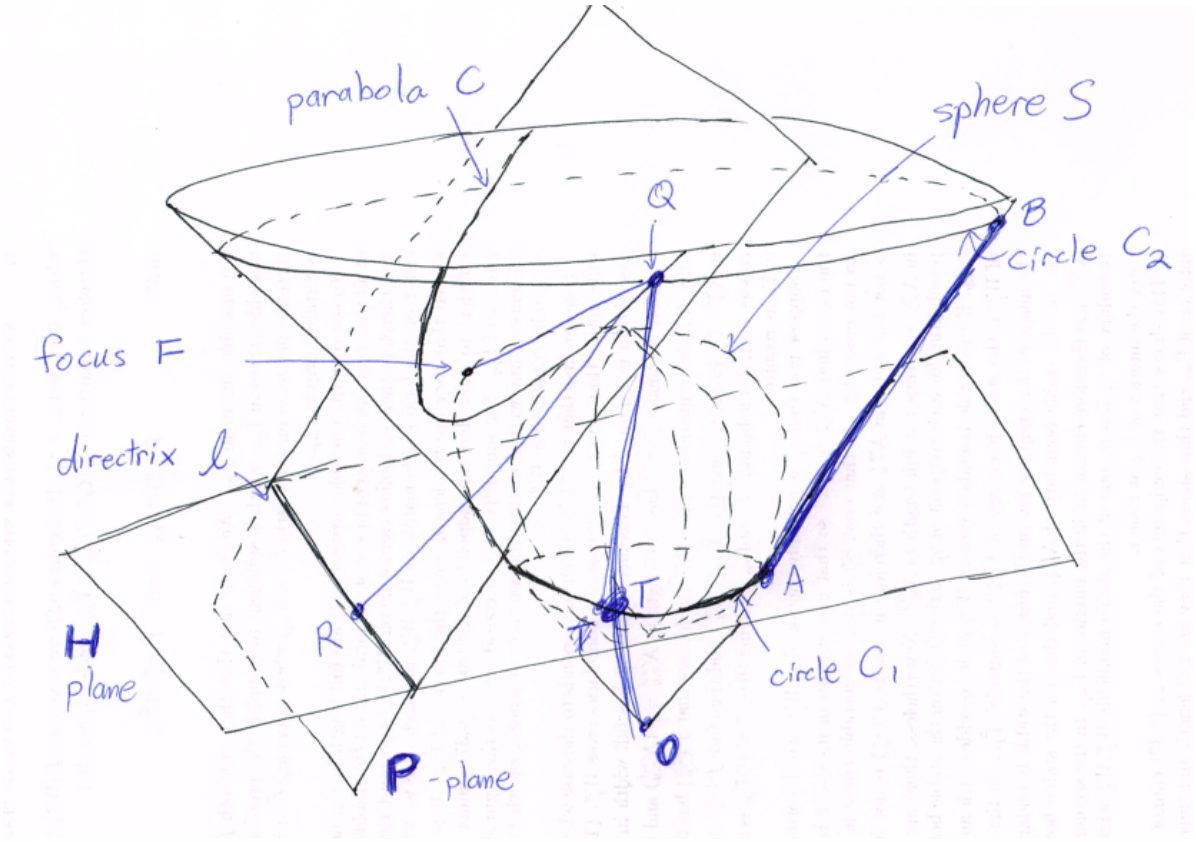


FIGURE 16. The Dandelin sphere for a parabola.

Start by drawing the line QO, where we recall that O is the vertex of the cone, and let T be the point where this line intersects $\mathbf{H}$. This line is tangent to S, as is the line QF; thus if we write C for the center of S (not pictured), we see that $\angle QFC$ and $\angle QTC$ are both right angles, since a tangent line to a sphere is perpendicular to the radius at that point. Now Pythagoras gives

$$|QF|^2 = |QC|^2 - |CF|^2 = |QC|^2 - |CT|^2 = |QT|^2, \tag{73.1}$$

where the second equality uses the fact that CF and CT are radii of the sphere. In fact the computation in (73.1) proves the following general lemma, which is useful to record here for future reference.

Lemma 73.4. *Given a sphere S and a point X outside of the sphere, if XY and XZ are tangent to the sphere at Y and Z, respectively, then $|XY| = |XZ|$.*

So far we have not actually assumed that $\mathbf{P}$ is parallel to a line generating the cone, and so everything we have said applies to *any* plane intersecting the cone, a fact thats will come in useful in the next lecture when we consider more general conic sections.

Now we add the assumption that the cone has a generating line parallel to $\mathbf{P}$; this is OA in the picture, where A is chosen to lie on C_1. Let C_2 be the circle in which the horizontal plane through Q intersects the cone, and let B denote the point where C_2 intersects the line OA. Then the line segment AB is obtained from QT by rotating

around the vertical axis, so

$$|QT| = |AB| = |QR| = |Q\ell|$$

where R is the point on ℓ that is closest to Q, and the second equality follows because QR and BA are parallel line segments running between the same two horizontal planes (the planes containing C_1 and C_2). Combining this with (73.1) shows that $|QF| = |QT| = |Q\ell|$, and thus the point F and the line ℓ satisfy the focus-directrix property for the intersection of $\mathbf{P}$ with the cone.

Lecture 74 Ellipses (and hyperbolas)

Stewart §10.6

74.1. Focus-directrix description of conics, and polar coordinates

Now suppose we take a cross-section of a cone with an *arbitrary* plane $\mathbf{P}$, which is not assumed to be parallel to any of the sides of the cone; see Figure 17. As before, take the axis of revolution of the cone to be vertical, and consider the Dandelin sphere S that is tangent to both the cone and the plane $\mathbf{P}$, and lies below the plane. Let F be the point where S intersects $\mathbf{P}$. As long as $\mathbf{P}$ is not horizontal (and in this case $\mathbf{P}$ intersects the cone in a circle, which we understand), $\mathbf{P}$ intersects $\mathbf{H}$ in a line ℓ.

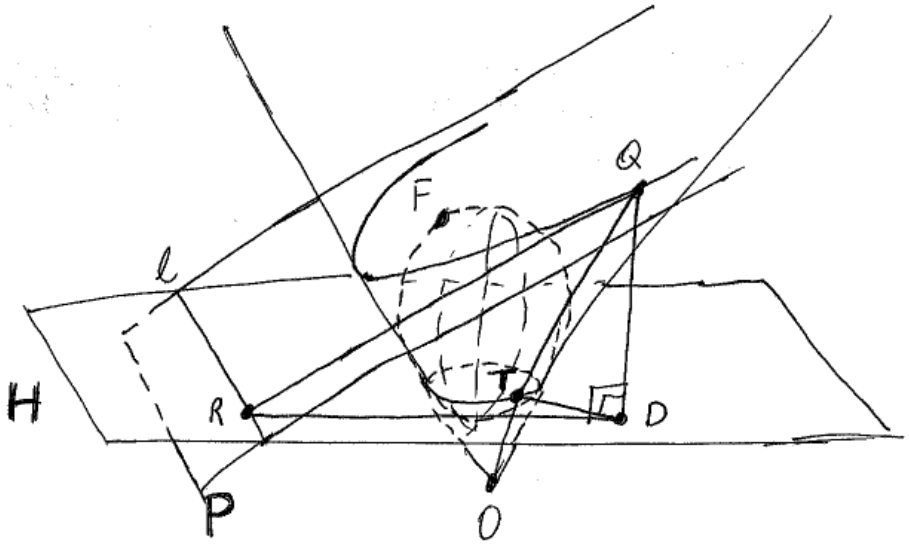


FIGURE 17. A Dandelin sphere for a general conic section.

We want to describe the curve in which $\mathbf{P}$ intersects the cone. Consider an arbitrary point Q on this curve. As before, Lemma 73.4 gives $|QF| = |QT|$, where T is the point in which the line QO intersects $\mathbf{H}$. And once again, we can choose a point R on the line ℓ such that $|Q\ell| = |QR|$. However, since $\mathbf{P}$ is not assumed to be parallel to any of the sides of the cone, we can no longer deduce that $|QT|$ and $|QR|$ are the same. Instead, we

can compare both of these lengths to $|QD|$, where D is the point in $\mathbf{H}$ that lies directly below Q, so that in particular QD is vertical and $\angle QDR$, $\angle QDT$ are right angles. Then elementary trigonometry gives

$$\sin \measuredangle QTD = \frac{|QD|}{|QT|} \quad \text{and} \quad \sin \measuredangle QRD = \frac{|QD|}{|QR|}.$$

Observe that $\alpha = \measuredangle QTD$ is the angle that measures how wide or narrow the cone is, and does not depend on the specific choice of Q. Similarly, $\beta = \measuredangle QRD$ is the angle in which the planes $\mathbf{P}$ and $\mathbf{H}$ intersect, and once again is independent of Q. Thus we have

$$\frac{|QF|}{|Q\ell|} = \frac{|QT|}{|QR|} = \frac{|QD|/\sin\alpha}{|QD|/\sin\beta} = \frac{\sin\beta}{\sin\alpha}.$$

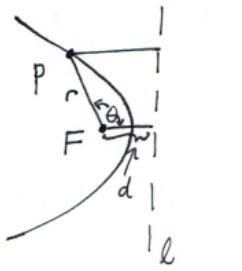
We have proved the following result.

Theorem 74.1. *Consider a cone obtained as follows: take a line through the origin that makes an angle α with the horizontal plane, and rotate it around the vertical axis. Let C be a curve obtained by intersecting this cone with a plane $\mathbf{P}$ that makes a nonzero angle β with the horizontal. Let S be the Dandelin sphere for this cone and plane, and let $\mathbf{H}$ be the horizontal plane through the circle in which S intersects the cone. Let $\ell = \mathbf{H} \cap \mathbf{P}$ and $F = S \cap \mathbf{P}$. Then the curve C can be described via the following focus-directrix property: a point $Q \in \mathbf{P}$ lies on the curve C if and only if*

$$|QF| = e|Q\ell|, \quad \textit{where } e = \frac{\sin\beta}{\sin\alpha}. \tag{74.1}$$

The number $e > 0$ is called the *eccentricity* of the conic section C. When $\beta < \alpha$, we have $0 < e < 1$ and the curve C is an *ellipse*. When $\beta = \alpha$, we have $e = 1$ and the curve C is a *parabola*. When $\beta > \alpha$, we have $e > 1$ and the curve C is a *hyperbola*.[61]

We can use (74.1) to write a formula for C in polar coordinates. Put the focus F at the origin and let the directrix ℓ be the line $x = d$. Then given a point P at polar coordinates (r, θ), we have $|PF| = r$, while the x-coordinate of P is $r\cos\theta$, so $|P\ell| = d - r\cos\theta$. Thus P satisfies $|PF| = e|P\ell|$ if and only if $r = ed - er\cos\theta$. Solving for r, we see that in polar coordinates, this conic section is the graph of



$$r = \frac{ed}{1 + e\cos\theta}. \tag{74.2}$$

Remark 74.2. Choosing a directrix $x = -d$ gives the related equation $r = ed/(1 - e\cos\theta)$, and choosing a horizontal directrix $y = \pm d$ has the effect of replacing cos with sin.

Another way of writing (74.2) is $r = R/(1 + e\cos\theta)$, where we no longer specify the distance to the directrix explicitly. Then putting $e = 0$ gives $r = R$, which is the polar equation of a circle, so we see that a circle is a conic section with eccentricity 0. (Note that this has no focus-directrix characterization, since the directrix would need to be "at infinity".)

[61] In fact, in this case C is one branch of a hyperbola; one typically considers also the reflection of the cone below the origin, so that the hyperbola has a corresponding branch in the lower half-space.

74.2. Focus-focus description of ellipses, and rectangular coordinates

The ellipse also has a description not in terms of a focus and directrix, but in terms of *two* foci (plural of focus). This is illustrated in the picture at right, which for the moment is borrowed from Apostol's textbook (until I manage to produce one of my own). When the plane is not parallel to the generator of the cone, it actually has *two* Dandelin spheres, one below and one above. Writing F_1 and F_2 for the two points in which these spheres intersect the plane, we see that a point P on the ellipse has the property (by Lemma 73.4) that $|PF_1| = |PA_1|$ and $|PF_2| = |PA_2|$, where A_1 and A_2 are the points in which the line PO intersects the horizontal circles corresponding to the two spheres. But then $|PF_1|+|PF_2| = |PA_1|+|PA_2| = |A_1A_2|$, and this last quantity does not depend on P (by another application of Lemma 73.4, as in the proof for the parabola). Summarizing the result of this argument, we have proved the following.

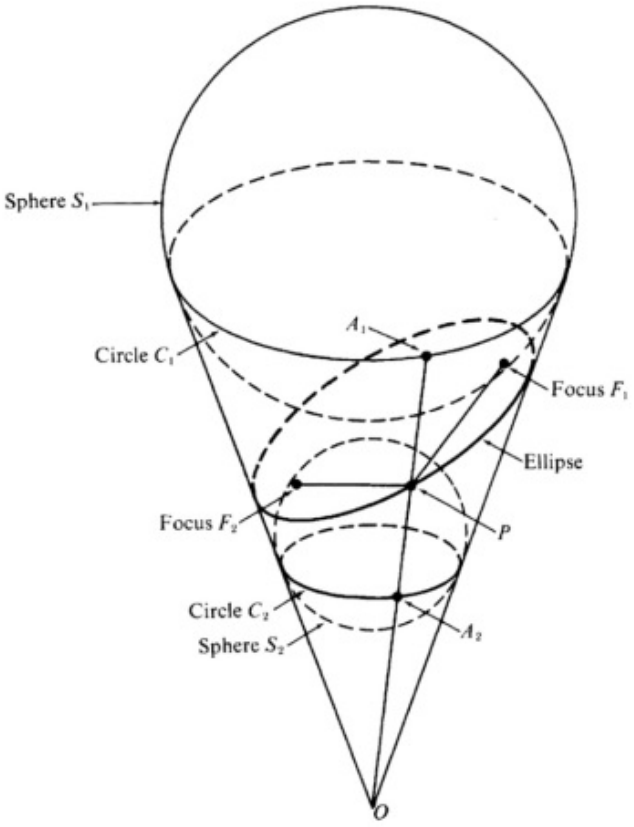


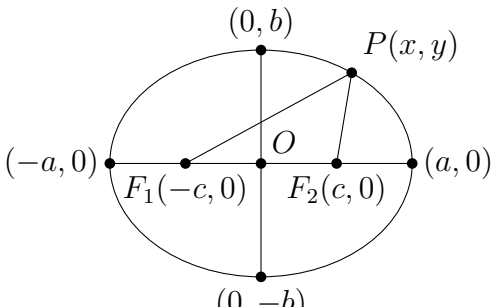
FIGURE 2.11 The ice-cream-cone proof.

Theorem 74.3. *If C is an ellipse (a conic section with eccentricity $e < 1$), then there are two points F_1, F_2, called* foci*, and a real number $r > 0$, such that a point P lies on C if and only if $|PF_1| + |PF_2| = r$.*

Note that if $F_1 = F_2$, so that the two foci coincide, then this condition reduces to the statement that the distance from P to the (single) focus is constant, which gives a circle.

Theorem 74.3 can be used to write an equation for an ellipse with a given eccentricity in rectangular coordinates. Instead of putting the origin at a focus, as we did with polar coordinates, we put the two foci on the x-axis with the origin at their midpoint, so that the foci F_1 and F_2 have coordinates $(\pm c, 0)$. Let $(a, 0)$ be the right-most point of the ellipse; by symmetry the left-most point is $(-a, 0)$. These two points are called the *vertices* of the ellipse. The line segment between the two vertices is called the *major axis*, and the line segment from the origin to one of the vertices is the *semimajor axis*. Similarly, the line between the top and bottom points $(0, \pm b)$ is the *minor axis*.

Observe that the vertex $Q = (a, 0)$ has $|QF_1| = a + c$ and $|QF_2| = a - c$, so the sum of the distances to the foci is $|QF_1| + |QF_2| = 2a$. Since the sum of the distances to the foci is



constant for all points on the ellipse, we see that the ellipse is the set of point P such that $|PF_1| + |PF_2| = 2a$; in other words, the sum of the distances to the foci must always be equal to the length of the major axis. We also observe that when $P = (0, b)$, we have $|PF_1| = |PF_2| = \sqrt{b^2 + c^2}$, so each of these is equal to a, and in particular, a, b, c are related by

$$a^2 = b^2 + c^2. \tag{74.3}$$

Now suppose P has coordinates (x, y). Using the Pythagorean formula to write $|PF_1| = \sqrt{(x+c)^2 + y^2}$ and $|PF_2| = \sqrt{(x-c)^2 + y^2}$, we see that the ellipse is the set of points (x, y) such that

$$\sqrt{(x+c)^2 + y^2} + \sqrt{(x-c)^2 + y^2} = 2a.$$

Isolating the first square root and then squaring both sides, this is equivalent to

$$\begin{aligned}(x+c)^2 + y^2 &= (\sqrt{(x-c)^2 + y^2} + 2a)^2 \\ &= (x-c)^2 + y^2 + 4a\sqrt{(x-c)^2 + y^2} + 4a^2.\end{aligned}$$

Expanding both sides gives

$$x^2 + 2cx + c^2 + y^2 = x^2 - 2cx + c^2 + y^2 + 4a\sqrt{(x-c)^2 + y^2} + 4a^2,$$

and after simplifying and isolating the square root we obtain

$$4cx - 4a^2 = 4a\sqrt{(x-c)^2 + y^2}.$$

Dividing by 4 and squaring both sides gives

$$\begin{aligned}(cx - a^2)^2 &= a^2\big((x-c)^2 + y^2\big), \\ c^2x^2 - 2a^2cx + a^4 &= a^2(x^2 - 2cx + c^2 + y^2) \\ &= a^2x^2 - 2a^2cx + a^2c^2 + a^2y^2, \\ a^4 - a^2c^2 &= (a^2 - c^2)x^2 + a^2y^2.\end{aligned}$$

Recalling from (74.3) that $a^2 - c^2 = b^2$, this is equivalent to

$$a^2b^2 = b^2x^2 + a^2y^2,$$

and dividing through by a^2b^2 gives the equation for the ellipse as

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1. \tag{74.4}$$

Observe that when $a = b = r$ this becomes the familiar equation $x^2 + y^2 = r^2$ for a circle with radius r.

As with the parabola, we can shift this equation to describe an ellipse at other locations in the plane. If the ellipse has foci which lie on the same horizontal or vertical line, with midpoint (h, k), and if the lengths of the horizontal and vertical axes of the ellipse are $2a$ and $2b$, respectively, then the equation of the ellipse is

$$\frac{(x-h)^2}{a^2} + \frac{(y-k)^2}{b^2} = 1. \tag{74.5}$$

74.3. Hyperbolas

The results for ellipses in the previous section have analogues for hyperbolas. We omit the details here, and merely mention the conclusions: the hyperbola corresponding to two foci F_1 and F_2 and a real number $r > 0$ is the set of all points P in the plane such that

$$\big||PF_1| - |PF_2|\big| = r. \tag{74.6}$$

If $r \geq |F_1F_2|$, then the only way to satisfy (74.6) is if P is on the line ℓ through F_1 and F_2 but does not lie between them. This is a degenerate case that we ignore, so we assume that $0 < r < |F_1F_2|$. In this case the hyperbola contains two points on ℓ, which lie between F_1 and F_2.

If we work in rectangular coordinates where the foci are at $(\pm c, 0)$, and the points $(\pm a, 0)$ are on the hyperbola (as before, we call these the *vertices*), then $a < c$ by the previous paragraph. Writing $b^2 = c^2 - a^2$ for convenience, a similar computation to the one in the previous section gives the equation for the hyperbola as

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1. \tag{74.7}$$

If the foci lie on the y-axis then the roles of x, y are reversed. As in (74.5), this can be shifted to put the hyperbola elsewhere in the plane.

One feature specific to hyperbolas is worth mentioning. As x^2 gets large, y^2 must also get large, and the right-hand side of (74.7) becomes insignificant in comparison. Without this RHS, the equation would be $y = \pm\frac{b}{a}x$. These lines are the *asymptotes* of the hyperbola.

Exercise 74.4. Prove that as $x \to \infty$, the corresponding positive value of y (such that (x, y) lies on the hyperbola) has the property that the distance from (x, y) to the line $y = \frac{b}{a}x$ approaches 0.

74.4. List of characterizations

Our discussion of conic sections can be summarized by the following list, which gives equivalent ways of characterizing these curves.

(1) *Cross-section of a cone and a plane.* If α and β are the angles that the cone and plane, respectively, make with the horizontal, then $e = \sin\beta/\sin\alpha$ is the *eccentricity* of the resulting conic. $e = 0$ gives a circle, $0 < e < 1$ gives an ellipse, $e = 1$ gives a parabola, and $e > 1$ gives a hyperbola.

(2) *Focus-directrix.* If the plane is not horizontal (the conic is not a circle), then the conic is described by a focus (point) F and a directrix (line) ℓ as the set of points Q in the plane such that $|QF| = e|Q\ell|$. The focus and directrix can be found using a Dandelin sphere.

(3) *Polar coordinates.* If we choose a polar coordinate system with origin at the focus and such that the directrix is vertical, then the curve is given in polar coordinates as the graph of $r = R/(1 + e\cos\theta)$, where $R > 0$ is a constant and e is the eccentricity. When $e > 0$ we have $R = ed$, where d is the distance from the focus to the directrix.

(4) *Focus-focus.* If the curve is an ellipse then there are two foci F_1 and F_2 such that the conic is the set of points Q in the plane such that $|QF_1| + |QF_2|$ is equal to the length of the major axis. These two foci can be found using two Dandelin spheres. A similar characterization is available for hyperbolas (replacing sum with difference), but not for parabolas.

(5) *Rectangular coordinates.* Choosing a rectangular coordinate system with origin at the midpoint of the foci (for ellipses and hyperbolas) or at the midpoint of the focus and the directrix (for parabolas), the curve takes the familiar form $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ (ellipse), $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ (hyperbola), or $y = ax^2$ (parabola), or possibly with x and y reversed depending on which orientation we choose.

(6) *Reflection property.* For a parabola, the lines emanating from the focus in all directions are reflected off of the parabola into a family of parallel lines. We proved this, and we leave the following laws for ellipses and hyperbolas as exercises.
   - For an ellipse, lines from one focus are reflected towards the other focus.
   - For a hyperbola, lines directed *towards* one focus, but with the hyperbola in the way, are reflected towards the other focus.

(7) *Motion in a gravitational field.* We proved that in a gravitational field that points uniformly downwards, an object moving without air resistance follows a parabola. Next we will turn our attention to movement in a gravitation field directed towards a single fixed point (the sun) that obeys an inverse square law, and show that the resulting trajectories are always conic sections.

Lecture 75 — Kepler and Newton

Spivak Ch. 17

In the early 1600's, the German astronomer Johannes Kepler formulated the following three laws of planetary motion.

(1) *Elliptical motion*: Planets move in ellipses, with the sun at one focus.
(2) *Equal areas in equal times*: For a given planet, the area swept out by the line from the planet to the sun depends only on the amount of time elapsed, and not on when we start recording.
(3) *Harmonic law*: The ratio $(\text{major axis})^3/(\text{period})^2$ is the same for all planets.

Kepler's work was based on extensive observations and computations, and did not offer an explanation for *why* these laws should be true. An explanation of the mechanism behind the laws would have to wait for the work of Isaac Newton, who began developing the ideas of calculus in the 1660's, both at Cambridge and during a period of isolation in 1665-1666 when the university was closed due to an epidemic of the bubonic plague. Eventually Newton developed a theory of physics that he used to derive Kepler's laws in his *Principia*, published in 1687. The two crucial laws are the following.

- *Newton's second law*: The force F acting on an object, and its resulting acceleration a, are related by $F = ma$, where m is the mass of the object.

- *Law of universal gravitation*: Given two objects with masses M and m, each object attracts the other with force GMm/r^2, where r is the distance between the objects and G is a universal constant.

In the setting we are interested in, M denotes the mass of the sun, and m denotes the mass of the planet that we study. Although the planet moves in 3-dimensional space, its orbit is contained in a single 2-dimensional plane, so we will describe its position using both polar coordinates (r, θ) and rectangular coordinates (x, y). We will write $c(t)$ for the its position at time t. We will also write $\dot{c}(t) = \frac{d}{dt}c(t)$ for its velocity at time t and $\ddot{c}(t) = \frac{d^2}{dt^2}c(t)$ for its acceleration. Observe that because the object moves in a 2-dimensional plane, its velocity and acceleration are given by not just a magnitude, but also a direction; that is, they are *vectors* in this plane, as is the force F. You will study these further in a later calculus course. For the time being we merely observe that we can use rectangular coordinates to write

$$\dot{c} = (\dot{x}, \dot{y}) \quad \text{and} \quad \ddot{c} = (\ddot{x}, \ddot{y}),$$

where x and y are the coordinate functions describing the position c; both x and y are functions of time t, and the notation above represents their first and second derivatives with respect to t. We will similarly write $\dot{r}, \dot{\theta}, \ddot{r}, \ddot{\theta}$ for the first and second derivatives of the polar coordinates of $c(t)$ with respect to t.

With our notation established, let us begin our analysis. For simplicity we assume that $m \ll M$ and ignore the motion of the sun, assuming instead that the location of the sun is fixed; we will use this as the origin of our coordinate system.[62] We also ignore the effect of any other planets, asteroids, comets, etc., that may be lurking in the vicinity.[63] Under these assumptions, Newton's laws imply that a planet at position (r, θ) (in polar coordinates) has acceleration $a = GM/r^2$, directed along the line from the planet towards the sun. We prove the following three theorems, which demonstrate that this implies Kepler's laws. All three theorems use Newton's second law $F = ma$, but the first two theorems do not require the full strength of the law of universal gravitation.

Theorem 75.1. *Suppose that an object moves according to Newton's second law $F = ma$, and that F depends only on the object's current position (r, θ). (We do not yet assume the law of universal gravitation.) Then the object's orbit satisfies Kepler's second law (equal areas in equal times) if and only if $F(r, \theta)$ always points along the line connecting the object to the origin. In this case, there is a constant K such that $r^2\dot{\theta} = K$ for all times t, and the area swept out during any interval of time with length T is equal to $\frac{1}{2}KT$. Moreover, writing $a(t) := \ddot{r} - r(\dot{\theta})^2 = \ddot{r} - K^2/r^3$, we have*

$$\ddot{c} = (a(t)\cos\theta, a(t)\sin\theta). \tag{75.1}$$

[62] For a completely precise treatment, this assumption should be removed, and we should put the origin at the center of mass of the sun-planet system.

[63] This seems reasonable since the gravity exerted by these objects is extremely small relative to the gravity exerted by the sun. However, the cumulative effect of these perturbations over long (long!) periods of time can be substantial, and the question of asymptotic stability of the solar system remains extremely difficult; this was one of the questions that led to the development of the part of the theory of dynamical systems that is popularly known as *chaos theory*.

Informally, this says that an object moving in a force field has the property of "equal areas in equal times" if and only if the force field is *central* (always points along the line to the origin). Thus Newton's laws imply Kepler's second law.

Theorem 75.2. *Suppose an object moves in a central force field following an inverse square law, meaning that $\ddot{c}$ has magnitude Q/r^2 for some constant Q, and always points towards the origin. Then the object moves along a conic section with one focus at the origin, whose equation in polar coordinates is*

$$r = \frac{K^2/Q}{1 + e\cos(\theta + \alpha)} \tag{75.2}$$

for some constant e, where K is the constant value of $r^2\dot{\theta}$ provided by Theorem 75.1. In particular, if the object's orbit is periodic, then it moves along an ellipse (or a circle).

In fact, Theorem 75.2 is also an 'if and only if' – if every object moving in a central force field moves along a conic section, then the force satisfies an inverse square law. We will not prove this direction, however, and will content ourselves with the direction stated, which shows that Newton's laws imply Kepler's first law.

Theorem 75.3. *Under the conditions of Theorem 75.2, Kepler's third law is satisfied if and only if the constant Q is the same for all planets.*

This theorem shows that Kepler's third law holds if and only if the gravitational constant G is truly universal.

Proof of Theorem 75.1. First we determine how to write Kepler's second law in terms of r and θ. By (64.2), the area swept out by the curve c from time t_1 to t_2 is $\int_{\theta(t_1)}^{\theta(t_2)} \frac{1}{2}r^2\,d\theta$, where in the integral we consider r as a function of θ. Using the substitution rule to write the integral in terms of t, we see that the area is

$$\int_{t_1}^{t_2} \frac{1}{2} r(t)^2 \dot{\theta}(t)\,dt. \tag{75.3}$$

Thus Kepler's second law – equal areas in equal times – is true if and only if $r^2\dot{\theta}$ is constant.

Now we look at the acceleration $\ddot{c}$, since this points in the same direction as the force. Writing $c = (x, y) = (r\cos\theta, r\sin\theta)$ and differentiating coordinate-wise gives

$$\begin{aligned} \dot{x} &= \dot{r}\cos\theta - r\dot{\theta}\sin\theta, \\ \dot{y} &= \dot{r}\sin\theta + r\dot{\theta}\cos\theta. \end{aligned} \tag{75.4}$$

Differentiating a second time gives

(75.5)

$$\ddot{x} = \ddot{r}\cos\theta - 2\dot{r}\dot{\theta}\sin\theta - r\ddot{\theta}\sin\theta - r(\dot{\theta})^2\cos\theta = (\ddot{r} - r(\dot{\theta})^2)\cos\theta - (2\dot{r}\dot{\theta} + r\ddot{\theta})\sin\theta,$$

$$\ddot{y} = \ddot{r}\sin\theta + 2\dot{r}\dot{\theta}\cos\theta + r\ddot{\theta}\cos\theta - r(\dot{\theta})^2\sin\theta = (\ddot{r} - r(\dot{\theta})^2)\sin\theta + (2\dot{r}\dot{\theta} + r\ddot{\theta})\cos\theta.$$

Thus if we plot $\ddot{c} = (\ddot{x}, \ddot{y})$ in the plane, we can reach it by first moving a distance $(\ddot{r} - r(\dot{\theta})^2)$ in the direction of $(\cos\theta, \sin\theta)$, which is the direction of the line between the origin and the object, and then moving a distance of $(2\dot{r}\dot{\theta} + r\ddot{\theta})$ in the direction of $(-\sin\theta, \cos\theta)$. Observe that this second motion is at right angles to the direction of the

first motion, and thus $\ddot{c}$ points along the line to the origin if and only if $2\dot{r}\dot{\theta} + r\ddot{\theta} = 0$ (that is, if and only if our second motion had no distance).

Compare this to the criterion for Kepler's second law, that $r^2\dot{\theta}$ is constant. Differentiating $r^2\dot{\theta}$ w.r.t. t gives

$$\frac{d}{dt}(r^2\dot{\theta}) = 2r\dot{r}\dot{\theta} + r^2\ddot{\theta} = r(2\dot{r}\dot{\theta} + r\ddot{\theta}).$$

This shows that $r^2\dot{\theta}$ is constant if and only if $2\dot{r}\dot{\theta} + r\ddot{\theta} = 0$ at all times when the object is not at the origin, which shows that Kepler's second law holds if and only if the force always points along the line connecting the object to the origin. We saw already that $r^2\dot{\theta}$ is constant in this case, and then (75.1) follows immediately from (75.5). □

Proof of Theorem 75.2. If acceleration always has magnitude Q/r^2 and points towards the origin, then from (75.1) we have

$$\ddot{r} - \frac{K^2}{r^3} = -\frac{Q}{r^2} \quad \Rightarrow \quad \frac{d^2r}{dt^2} = \frac{K^2}{r^3} - \frac{Q}{r^2}. \tag{75.6}$$

A first this looks like a separable equation – the RHS depends only on r, which is what we want to find – so we might try dividing both sides by the RHS and then integrating. But the LHS is a *second* derivative, not a first derivative! So this is not actually a first-order separable DE like the ones we encountered earlier, and our techniques from before do not work.

Instead we need to reformulate things a little bit. Instead of writing r as a function of t, we consider r as a function of θ, and write (75.6) to obtain a DE in terms of $\frac{d}{d\theta}$, not $\frac{d}{dt}$. We can do this using the chain rule, but we need to be careful because a second derivative is involved. Recall from Theorem 75.1 that $\frac{d\theta}{dt} = \dot{\theta} = \frac{K}{r^2}$, and thus

$$\frac{d^2r}{dt^2} = \frac{d}{dt}\Big(\frac{dr}{dt}\Big) = \frac{d\theta}{dt}\frac{d}{d\theta}\Big(\frac{dr}{d\theta}\frac{d\theta}{dt}\Big) = \frac{K}{r^2}\frac{d}{d\theta}\Big(\frac{K}{r^2}\frac{dr}{d\theta}\Big),$$

where the first equality is the definition of second derivative, the second equality is two applications of the chain rule, and the third equality uses the formula for $\dot{\theta}$. Comparing this to (75.6) gives

$$\frac{K^2}{r^2}\frac{d}{d\theta}\Big(\frac{1}{r^2}\frac{dr}{d\theta}\Big) = \frac{K^2}{r^3} - \frac{Q}{r^2} \quad \Rightarrow \quad \frac{d}{d\theta}\Big(\frac{1}{r^2}\frac{dr}{d\theta}\Big) = \frac{1}{r} - \frac{Q}{K^2}.$$

We could expand the left-hand side using the product rule, but the resulting DE would not fit into any of the categories that we have a good procedure for solving at this point. Instead, the way forward is to make the observation that

$$\frac{d}{d\theta}\frac{1}{r} = -\frac{1}{r^2}\frac{dr}{d\theta},$$

and so the DE can be rewritten as

$$\frac{d}{d\theta}\Big(-\frac{d}{d\theta}\frac{1}{r}\Big) = \frac{1}{r} - \frac{Q}{K^2} \quad \Rightarrow \quad \frac{d^2}{d\theta^2}\frac{1}{r} = -\frac{1}{r} + \frac{Q}{K^2}.$$

Let $f(\theta) = \frac{1}{r} - \frac{Q}{K^2}$, and observe that $\frac{d^2}{d\theta^2}f(\theta) = \frac{d^2}{d\theta^2}\frac{1}{r}$; thus

$$\frac{d^2}{d\theta^2}f(\theta) = -f(\theta).$$

This is a DE that we can solve; the general solution is

$$f(\theta) = B\cos(\theta + \alpha),$$

where B, α are constants of integration determined by the initial values of f and $\frac{df}{d\theta}$. Thus

$$\frac{1}{r} - \frac{Q}{K^2} = B\cos(\theta + \alpha),$$

and solving for r gives

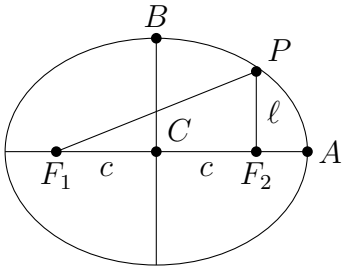
$$r = \frac{1}{\frac{Q}{K^2} + B\cos(\theta + \alpha)} = \frac{K^2/Q}{1 + \frac{BK^2}{Q}\cos(\theta + \alpha)}. \tag{75.7}$$

Writing $e = BK^2/Q$ gives (75.2), and we recall that this is the polar equation for a conic section. Observe that Q represents the strength of the central force divided by the mass of the object, while the parameters B, K, α are determined by the initial position and velocity. If these are such that the eccentricity is < 1, then the orbit is periodic and thus is an ellipse. □

Proof of Theorem 75.3. Given a planet in an elliptical orbit as in Theorem 75.2, we need to relate the period of the orbit to the length of the major axis and to the constant Q. First observe that by Theorem 75.1, if we write T for the amount of time it takes the planet to complete one revolution (its period), then the area of the ellipse is $A = \frac{1}{2}KT$, where K is a constant (that may be different for different planets). On the other hand, we have $A = \pi ab$, where a, b are the lengths of the semimajor and semiminor axes, so

$$KT = 2\pi ab. \tag{75.8}$$

This equation involves the period T and the major axis $2a$, which moves us in the direction of Kepler's third law. However, it also involves the constant K and the semiminor axis b, which do not appear in that law, and does *not* involve the constant Q, which does appear there. So we need a way to relate K, b, and Q. This is provided by recalling the equation (75.2) that gives the orbit in polar coordinates, and using some elementary geometry of ellipses.



Referring to the picture shown, suppose the origin is at the focus F_2, and let ℓ be the length of the line segment F_2P, which is perpendicular to the semi-major axis CA (hence parallel to the semi-minor axis CB). We will relate ℓ to K, b, and Q. First observe that $r = r(\theta)$ is minimized at the point A, so $\cos(\theta + \alpha)$ must be maximized at this point. Moving from A to P increases θ by $\pi/2$ and thus $\cos(\theta + \alpha) = 0$ at P; using this value in (75.2) gives $r = K^2/Q$ here, and we conclude that

$$\ell = K^2/Q. \tag{75.9}$$

At the same time, recall that the semi-major axis $a = |CA|$ has the property that $|PF_1| + |PF_2| = |AF_1| + |AF_2| = 2a$ by the focus-focus characterization of an ellipse, and thus

$$2a = \ell + \sqrt{(2c)^2 + \ell^2} \quad \Rightarrow \quad (2c)^2 + \ell^2 = (2a - \ell)^2 = 4a^2 - 4a\ell + \ell^2.$$

Subtracting ℓ^2 from both sides gives $4c^2 = 4a^2 - 4a\ell$, and simplifying gives

$$a\ell = a^2 - c^2. \tag{75.10}$$

We proved in (74.3) that $a^2 = b^2 + c^2$; using this together with (75.9) and (75.10) gives

$$b^2 = a^2 - c^2 = a\ell = aK^2/Q. \tag{75.11}$$

Squaring (75.8) and using (75.11) gives

$$K^2T^2 = 4\pi^2a^2b^2 = 4\pi^2a^3\frac{K^2}{Q}.$$

Solving for Q gives

$$Q = 4\pi^2\frac{a^3}{T^2}, \tag{75.12}$$

which proves the theorem and establishes Kepler's third law as a consequence of the law of universal gravitation. □