12 The logarithm, exponential, and trigonometric functions

We know what x^n means for any real x and natural number n. But what does, for example, $x^{\sqrt{2}}$ mean? Even when x is a natural number, there is no obvious intuitive interpretation, and the situation is far more unclear when we ask about, say, $\sqrt{3}^{\sqrt{2}}$.

The goal of this section is to introduce the logarithm and exponential functions, which allow us to interpret sensible and unambiguous expressions of the form x^y for arbitrary reals x, y.

12.1 Informal introduction

For natural number n, define $f_n : (0,\infty) \to \mathbb{R}$ by $f_n(x) = x^n$.[185] We already know alot about f_n: it has domain $(0,\infty)$, range $(0,\infty)$, is increasing on its domain, is continuous everywhere and differentiable everywhere, and has derivative $f_n'(x) = nx^{n-1}$.

By the results of the section on inverse functions, we know that f_n has an inverse, call it g_n, which has domain $(0,\infty)$, range $(0,\infty)$, is increasing on its domain, is continuous everywhere. We denote $g_n(x)$ by $x^{1/n}$, and also refer to g_n as $f_{1/n}$. Since $f_n'(x) \neq 0$ for any x, $f_{1/n}$ is differentiable everywhere, with derivative

$$f_{1/n}'(x) = \frac{1}{f_n'(x^{1/n})} = \frac{1}{n}x^{\frac{1}{n}-1}\text{[186]}.$$

For any positive rational $r = m/n$ (with $m, n \in \mathbb{N}$) we can define f_r by $f_r(x) = (x^{1/n})^m$. It is a straightforward check that this is in fact well defined, that is, that the value of $f_r(x)$ doesn't depend on the particular choice of m, n. This amounts to using the axioms of the real numbers to check that if $m/n = p/q$ with $m, n, p, q \in \mathbb{N}$ then $(x^{1/n})^m = (x^{1/q})^p$. Again, f_r has domain $(0,\infty)$, range $(0,\infty)$, is increasing on its domain, and is continuous everywhere. By basic properties of the derivative, it is differentiable everywhere, and an application of the chain rule yields

$$f_r'(x) = m(x^{1/n})^{m-1} \times \frac{1}{n}x^{\frac{1}{n}-1} = rx^{r-1}.$$

Denote by x^r the value $f_r(x)$.

For negative rational r, we can define a function f_r by $f_r(x) = 1/f_{-r}(x)$, and again denote by x^r the value $f_r(x)$. Again, f_r has domain $(0,\infty)$, range $(0,\infty)$, but now is decreasing on

[185] The natural domain of $f(x) = x^n$ for $n \in \mathbb{N}$ is $\mathbb{R}$, but we choose to restrict to the domain on $(0,\infty)$. The reason for this is that f is increasing on this domain, for *every* n (whereas on its natural domain, f is only increasing for odd n); we will avoid a lot of annoyance by focussing exclusively on non-negative inputs to the power function, and we wouldn't gain much by trying to extend to negative inputs.

[186] This is jumping the gun a little bit. Really, $f_{1/n}'(x) = 1/(n(x^{1/n})^{n-1})$. With the definition we will give in a moment for x^r for positive rational r, this becomes $1/(n(x^{(n-1)/n}))$, and with the definition we will give a moment later for x^r for negative rational r, this becomes $(1/n)x^{(1/n)-1}$. But technically we need those later definitions to jump to the final answer.

its domain. It is continuous everywhere, and by the chain rule (or the reciprocal rule, or the quotient rule), it is easily seen to satisfy $f_r'(x) = rx^{r-1}$.

The conclusion of all this is we can define, for *every* rational r, a function f_r that acts as a "raising to the power r" function, i.e., $f_r(x) = x^r$, that can be applied to any positive x. This function has the properties that

- f_r has domain $(0, \infty)$ and range $(0, \infty)$ (unless $r = 0$, in which case it has range $\{1\}$; recall that $x^0 = 1$ for all $x \neq 0$);
- f_r is increasing if $r > 0$, decreasing if $r < 0$, and constant if $r = 0$;
- f_r is continuous everywhere; and
- f_r is differentiable everywhere, with derivative $f_r'(x) = rx^{r-1}$.

Moreover, f_r agrees with our usual notion of "raising to the power r" where r is a natural number.

But what can we do to make sense of x^a when a is *not* rational? One approach is through the completeness axiom: for $x > 1$ and rational numbers $0 < r_1 < r_2$ it can be checked that $x^{r_1} < x^{r_2}$. It follows that for $x > 1$ and real $a > 0$, the set $A = \{x^r : 0 < r < a\}$ is bounded above (by $x^{r'}$ for any $r' > a$), It's also non-empty (obviously), so by completeness, $\sup A$ exists. We could then *declare* x^a to be $\sup A$. Similarly, for $0 < x < 1$, we could declare x^a to be $\inf\{x^r : r < a\}$, and (of course) declare 1^a to be 1. Then, for $a < 0$, we could declare x^a to be $1/x^{-a}$.All this certainly defines, for each real a, a function $f_a : (0, \infty) \to (0, \infty)$.

It's an easy check that this agrees with the previous definition when a is rational. It is far from easy to check that when a is irrational, f_a is still continuous and differentiable, with derivative $f_a'(x) = ax^{a-1}$, that is increasing when $a > 0$ and decreasing when $a < 0$, and that it has range $(0, \infty)$.

It is also far from easy to check that f_a satisfies all the properties that we would expect of the "raising to the power a" function, properties such as

- $x^{a+b} = x^a x^b$

and

- $(x^a)^b = x^{ab}$;

its not even all that straightforward to verify these properties for rational a, b.

So, we'll take an alternate approach to defining the power function for general exponents. Instead of constructing a function that seems like it should work, and then verifying that the properties we want to hold do actually hold, we'll list those properties, considering them as axioms, and then try to argue that there exists a unique function that satisfies those properties.

Actually, we will address a related, but slightly different, question:

Fix a real number $a > 0$. For real x, what does a^x mean?

(The difference here is that we now are considering the base to be fixed, and we are varying the exponent, whereas before we were considering the exponent to be fixed, and we were varying the base.[187])

One approach follows the lines we described earlier:

- Set $a^0 = 1$ and set $a^n = aa^{n-1}$ for $n \in \mathbb{N}$.
- For $n \in \mathbb{N}$ define $a^{1/n}$ via the intermediate value theorem, as we did last semester.
- For positive rational $r = m/n$, set $a^r = (a^{1/m})^n$ (after checking that this is well-defined, i.e., doesn't depend on the choice of representation of $r = m/n$ with $m, n \in \mathbb{N}$).
- For negative rational r, set $a^r = 1/(a^{-r})$.

This defines a^x for rational x, and a series of tedious inductions, together with lots of algebraic manipulation, verifies the relation

$$a^{x+y} = a^x a^y \quad \text{for all } x, y \in \mathbb{Q}. \quad (\star)$$

Then, for general real x, we can define

$$a^x = \begin{cases} \sup\{a^r : r \in \mathbb{Q}, r < x\} & \text{if } a > 1 \\ \inf\{a^r : r \in \mathbb{Q}, r < x\} & \text{if } a < 1 \\ 1 & \text{if } a = 1. \end{cases}$$

It is a long and intricate exercise that for each $a > 0$, this yields a continuous function that satisfies $(\star)$ (in fact, this gives the unique such continuous function that extends the given definition of a^r for rational r, as we discuss in a moment).

Instead of taking this approach, we'll take an axiomatic approach. Fix $a > 0$, with $a \neq 1$[188]. Let $\exp_a : \mathbb{R} \to \mathbb{R}$ be a function (an as-yet unknown function) that captures the notion of "raising a base a to a power"; that is, $\exp_a(x)$ is a sensible interpretation of a^x for all real x. Here are the properties that we want $\exp_a$ to satisfy:

- $\exp_a(1) = a$ and $e_a(0) = 1$ (a normalizing property);
- for all real x, y, $\exp_a(x + y) = \exp_a(x)\exp_a(x)$ (this property, together with the normalizing property, and a lot of induction, is what's need to ensure that $\exp_a(r) = a^r$ for rational r, where a^r is defined in the natural way that we described above);

[187] Of course, it will amount to the same thing in the end — we'll end up define u^v for any $u > 0$ and any real v.

[188] If $a = 1$, there is an obvious choice for $\exp_a$, namely $\exp_a(x) = 1$ for all x.

- $\exp_a$ is continuous on its domain (this condition ensures that $\exp_a(x) = a^x$ for all real x, where a^x is defined via the completeness axiom as described above. Indeed, let $f : \mathbb{Q} \to \mathbb{R}$ be the function $f(r) = a^r$, with the natural definition, and let $g : \mathbb{Q} \to \mathbb{R}$ be the function $g(r) = \exp_a^r$. Suppose both f and g extend to continuous functions on the whole real line. By the first two properties, $f - g$ is identically 0 on the rationals; but it is also continuous on the reals. A simple argument based on the density of the rationals quickly gives that $f - g$ is identically 0 on the reals, i.e., that $f = g$[189]);

- $\exp_a$ is differentiable; and

- $\exp_a$ is monotone (these last two properties will allow us to derive non-obvious properties that $\exp_a$ must also satisfy, if it satisfies the ones listed above; from these we will get our actual explicit expression for $\exp_a$).

Assuming such a function exists, here is what its derivative must look like:

$$\begin{aligned}
\exp_a'(x) &= \lim_{h\to 0} \frac{\exp_a(x+h) - \exp_a(h)}{h} \\
&= \exp_a(x) \lim_{h\to 0} \frac{\exp_a(h) - 1}{h} \\
&= \exp_a(x)\exp_a'(0).
\end{aligned}$$

And here is what the derivative of its inverse must look like:

$$\begin{aligned}
(\exp_a^{-1})'(x) &= \frac{1}{\exp_a'(\exp_a^{-1}(x))} \\
&= \frac{1}{\exp_a'(0)\exp_a(\exp_a^{-1}(x))} \\
&= \frac{1}{\exp_a'(0)x}.
\end{aligned}$$

Now the fundamental theorem of calculus (part 1) tells us that if we define (for any specific constant c)

$$\exp_a^{-1}(x) = \int_c^x \frac{dt}{\exp_a'(0)t}$$

then we indeed have $(\exp_a^{-1})'(x) = 1/(\exp_a'(0)x)$. We should probably choose $c = 1$, because with the above definition, we have $\exp_a^{-1}(c) = 0$ so $\exp_a(0) = c$, and we want $\exp_a(0) = 1$.

So we have been led to defining not $\exp_a$, but rather $\exp_a^{-1}$; and we have been led to the definition

$$\exp_a^{-1}(x) = \int_1^x \frac{dt}{\exp_a'(0)t}.$$

A problem with this definition is that we don't know what $\exp_a'(0)$ is. So as it stands, the definition is somewhat circular.

[189]This is an example of the general result: if $f, g : \mathbb{R} \to \mathbb{R}$ are continuous, and agree on a dense set, then they agree everywhere.

Here's a solution: presumably, there is a base a for which $\exp_a'(0) = 1$. For that special base, we have a completely explicit candidate for $\exp_a^{-1}$, namely

$$\exp_a^{-1}(x) = \int_1^x \frac{dt}{t}.$$

All this was hypothetical — *if* there is a function $\exp_a$ satisfying all of the required properties, then (at least for one special, undetermined as-yet a), its inverse has the simple explicit expression given above. In the next section we start the whole process over, and put it on firm foundations. It will go quickly, because we now know where to start from — with the integral $\int_1^x dt/t$.

12.2 Defining the logarithm and exponential functions

Motivated by the discussion in the previous section, we now define the (natural) logarithm function.

Definition of logarithm The function log[190] is defined by

$$\log(x) = \int_1^x \frac{dt}{t}.$$

Because the function $f(t) = 1/t$ is continuous and bounded on every interval of the form $[\varepsilon, N]$ (for arbitrarily small $\varepsilon > 0$ and arbitrarily large $N > 0$), it follows that the natural domain of log is (at least) $(0, \infty)$. In fact, this is the full natural domain, because it is easy to check (in a manner similar to how we checked that $\int_1^\infty dt/t$ diverges) that $\int_0^1 dt/t$ diverges.[191]

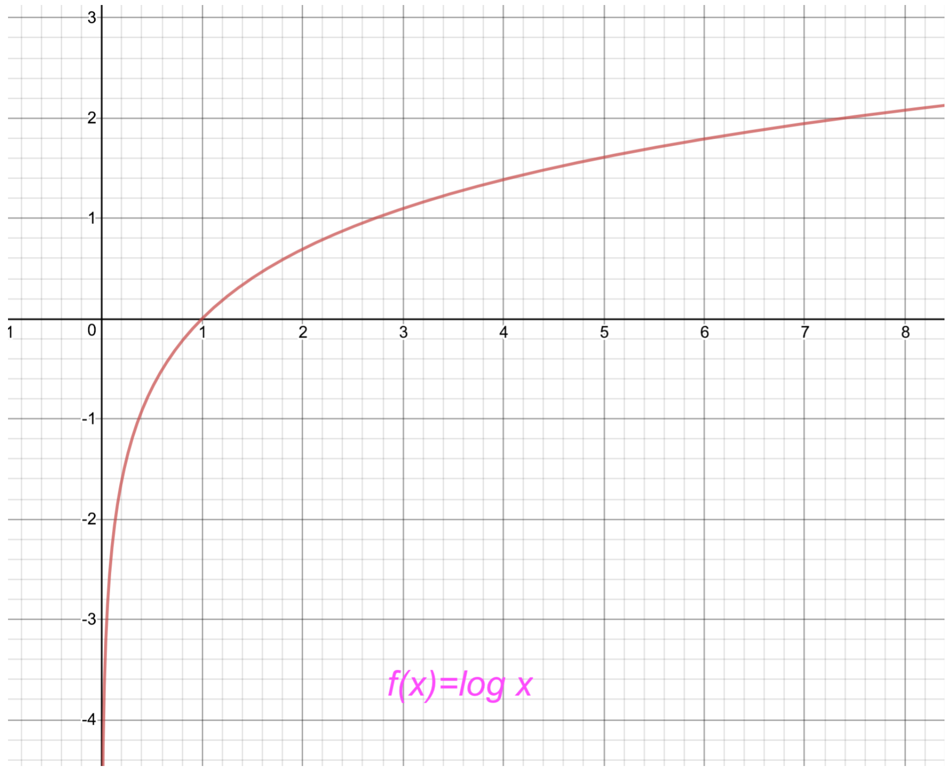
Moreover, since $1/t$ is non-negative, log is increasing on its domain. It takes the value 0 once, at $x = 1$. Because $\int_1^N dt/t$ can be made arbitrarily large by choosing N large enough, and $\int_1^\varepsilon$ can be made arbitrarily small (large and negative) by choosing $\varepsilon > 0$ close enough to zero, it follows that the range of log is $(-\infty, \infty)$. Specifically,

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

Because it is defined as the integral of an integrable function, log is continuous on its whole domain, and because $f(t) = 1/t$ is itself continuous, log is moreover differentiable everywhere, with derivative $\log'(x) = 1/x$ and second derivative $\log''(x) = -1/x^2$. Since this latter is always negative, log is concave on its whole domain.

We are now in a good position to sketch the graph of log:

[190]The name ln is also sometimes used for this function, but this notation is far more commonly seen in calculus textbook that in scientific papers.
[191]Consider the integral on the intervals $[1/2, 1]$, $[1/4, 1/2]$, $[1/8, 1/4$, et cetera.



Because log is increasing, it has an inverse, which we denote by exp (for "exponential"). From our general discussion of inverse functions, together with the properties we have just established about log, we immediately get that exp is continuous and increasing, has domain $\mathbb{R}$ and range $(0, \infty)$, and satisfies

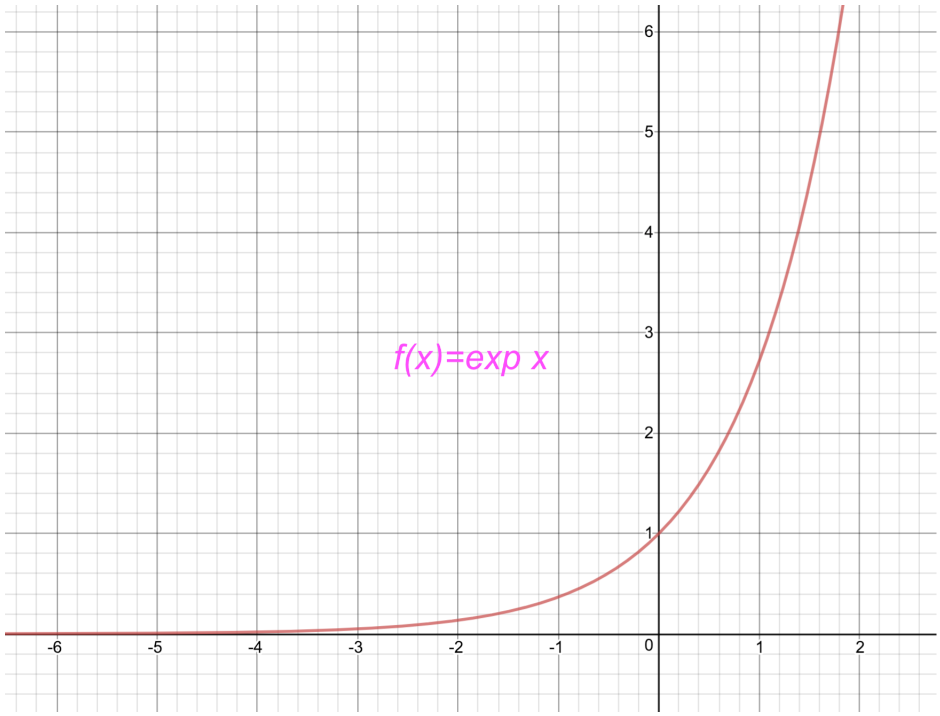
- $\lim_{x \to -\infty} \exp(x) = 0$
- $\lim_{x \to \infty} \exp(x) = +\infty$.

Because the derivative of *log* is never 0, the derivative of exp exists at all points in its domain, and we have

$$\exp'(x) = (\log^{-1})'(x) = \frac{1}{\log'(\log^{-1}(x))} = \log^{-1}(x) = \exp(x).$$

So $\exp''(x) = \exp(x) > 0$, and exp is convex. Since $\log(1) = 0$ we get $\exp(0) = 1$. We are now in a good position to sketch the graph of exp[192]:

[192]Of course, we could have also obtained the graph of exp by reflecting the graph of log across $x = y$



There is some unique number $\alpha > 1$ that has the property

$$\int_1^\alpha \frac{dt}{t} = 1.$$

We call this number e. The two basic properties of e, the first of which is the defining relation (reframed in the language of the log function), and the second of which is an immediate consequence of the first:

$$\log(e) = 1 \quad \text{and} \quad \exp(1) = e.$$

The number e is ubiquitous in mathematics. It has numerous "equivalent definitions", such as

- $\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n$
- $\lim_{n\to\infty} \sum_{k=1}^n \frac{1}{k!}$
- $\lim_{n\to\infty} \frac{n^n}{\sqrt[n]{n!}}$.

Hopefully our approach to defining e shows why it is natural: it attempting to define a function e_a that could sensibly serve as an interpretation for a^x, we discovered that e_a should satisfy

$$\exp_a^{-1}(x) = \int_1^x \frac{dt}{\exp_a'(0)t}.$$

The number e turns out to be the unique choice of a for which $\exp_a'(0) = 1$, leading to a particularly clean definition.

We can use the definition of e to give a numerical estimate. First, we show that $e > 2.7182$. To do this, we need to show that $\int_1^{2.7182} dt/t < 1$. Dividing the interval $[1, 2.7182]$ into n equal subintervals, we use that $1/t$ is decreasing to get

$$\int_1^{2.7182} \frac{dt}{t} \leq \sum_{i=1}^{n} M_i(t_i - t_{i-1}) = \frac{1.7182}{n} \sum_{i=1}^{n} \frac{1}{1 + \frac{1.7182(i-1)}{n}}.$$

A `Mathematica` calculation shows that when $n = 100,000$ the right-hand side above is $0.99997\cdots$. Next, we show that $e < 2.7183$. To do this, we need to show that $\int_1^{2.7183} dt/t > 1$. Using the same approach as before, we have

$$\int_1^{2.7183} \frac{dt}{t} \geq \sum_{i=1}^{n} m_i(t_i - t_{i-1}) = \frac{1.7183}{n} \sum_{i=1}^{n} \frac{1}{1 + \frac{1.7183i}{n}}.$$

A `Mathematica` calculation shows that when $n = 100,000$ the right-hand side above is $1.000001\cdots$. So we have the bounds

$$2.7182 < e < 2.7183.$$

We now derive the key property of the log function.

Theorem 12.1. *For all a, b in the domain of* log*,*

$$\log(ab) = \log(a) + \log(b)$$

and

$$\log(a/b) = \log(a) - \log(b).$$

Proof: To prove that $\log(ab) = \log(a) + \log(b)$ we want to show that for all $a, b \in (0, \infty)$,

$$\int_1^a \frac{dt}{t} + \int_1^b \frac{dt}{t} = \int_1^{ab} \frac{dt}{t}.$$

By the basic properties of integration, this is equivalent to

$$\int_1^a \frac{dt}{t} = \int_b^{ab} \frac{dt}{t}.$$

Define $G(x) = \int_1^x \frac{dt}{t}$, so $G'(x) = 1/x$, and $H(x) = \int_b^{xb} \frac{dt}{t}$, so $H'(x) = (1/xb)b = 1/x$. Since $G'(x) = H'(x)$ for all x, we have that $G - H$ is constant. Since $G(1) = H(1) = 0$, that constant is 0, so $G = H$, and in particular $G(a) = H(a)$, which is what we wanted to show.

For the second identity we have

$$\log a = \log(a/b)b = \log(a/b) + \log b$$

(using the result we have just proven), so

$$\log(a/b) = \log a - \log b.$$

□

Translating this to the exponential function, we obtain the following corollary.

Corollary 12.2. *For all* $a, b \in \mathbb{R}$,

$$\exp(a+b) = \exp(a)\exp(b)$$

and

$$\exp(a-b) = \exp(a)/\exp(b).$$

Proof: Let a', b' be such that $\log(a') = a$, $\log(b') = b$. We have

$$\exp(a+b) = \exp(\log(a') + \log(b')) = \exp(\log(a'b')) = a'b' = \exp(a)\exp(b)$$

(since $a = \exp(x)$, $b = \exp(y)$). Similarly

$$\exp(a-b) = \exp(\log(a') - \log(b')) = \exp(\log(a'/b')) = a'/b' = \exp(a)/\exp(b).$$

□

Both Theorem 12.1 and Corollary 12.2 can be extended by induction: for $a_1, \ldots, a_n \in (0, \infty)$,

$$\log(a_1 \cdot a_2 \cdot \cdots \cdot a_n) = \log a_1 + \log a_2 + \cdots + \log a_n$$

and for $a_1, \ldots, a_n \in \mathbb{R}$,

$$\exp(a_1 + a_2 + \cdots + a_n) = \exp a_1 \cdot \exp a_2 \cdot \cdots \cdot \exp a_n.$$

Recall that we set out to find, for each $a > 0$, a function $\exp_a : \mathbb{R} \to \mathbb{R}$ that captures the notion of "base a raised to a power", and we decided that such a function should be continuous, differentiable, invertible, and satisfy $\exp_a(0) = 1$, $\exp_a(1) = a$, and $\exp_a(x+y) = \exp_a(x)\exp_a(y)$ for all real x, y. Looking back on what we have done so far, we see that that function exp satisfies these conditions for the specific value $a = e \approx 2.7128\ldots$. It therefore makes sense to make the following definition.

Definition of e raised to the power x For real x, e^x means $\exp x$.

This agrees with the natural definition of e^x, for rational x, given earlier. Recall that specified that

- first $e^0 = 1$ ($\exp(0) = 1$), and for $n \in \mathbb{N}$, $e^n = e \cdot e^{n-1}$ ($\exp n = \exp(1 + (n-1)) = \exp(1)\exp(n-1)$).
- It then specified that for $n \in \mathbb{N}$, $e^{1/n}$ is that unique positive number such that $(e^{1/n})^n = e$; but $(\exp(1/n))^n = \exp(1/n) \cdot \cdots \cdot \exp(1/n) = \exp(1/n + \cdots + 1/n) = \exp(1) = e$, and $\exp(1/n) > 0$, so $\exp(1/n)$ is indeed that unique positive number that $e^{1/n}$ was defined to be.
- It then specified that $e^{m/n} = (e^{1/n})^m$; but also $\exp(m/n) = \exp((1/n) + \cdots + (1/n))$ (where there are m summands, all $1/n$), and this equals $\exp(1/n) \cdot \cdots \exp(1/n) = (\exp(1/n))^m$.

- Finally, it specified that $e^r = 1/e^{-r}$ for $r < 0$ and rational; but $1/\exp(-r) = \exp(0)/\exp(-r) = \exp(0 - -(r)) = \exp(r)$.

So $\exp(x)$ agrees with the natural definition of e^x for all rational x; and since there is at most one continuous function on the reals that agrees with e^x on the rationals, that fact that exp is a continuous function on the reals makes it the only sensible choice for an interpretation of e^x.

What about defining a^x, for arbitrary $a > 0$? To start the process, we need the identity

$$\log(a^x) = x \log a$$

for *rational* x, which we can verify from previously established properties. For $x > 0$ with $x = m/n$, $m, n \in \mathbb{N}$ we have

$$\log(a^x) = \log(a^{m/n}) = \log((a^{1/n})^m) = m \log(a^{1/n}),$$

with the last equality following from $\log(a_1 \cdot \cdots \cdot a_n) = \log a_1 + \cdots + \log a_n$, applied with $n = m$ and $a_i = a^{1/n}$. Also, from the same identity we get

$$\log(a) = \log\left((a^{1/n})^n\right) = n \log a^{1/n},$$

so $\log a^{1/n} = n \log a$, and $m \log(a^{1/n}) = (m/n) \log a$. From this we get $\log a^x = x \log a$.

If $a < 0$ then

$$\log a^x = \log(1/a^{-x}) = \log 1 - \log a^{-x} = 0 - (-x) \log a = x \log a.$$

So, for *rational* x we have $\log a^x = x \log a$ or

$$a^x = e^{x \log a}.$$

This suggests a very obvious choice for a^x, for arbitrary real x.

Definition of a raised to the power x For $a > 0$, and real x, a^x means $\exp(x \log a)$ (or $e^{x \log a}$).

Clearly the function that sends x to $\exp(x \log a)$ is continuous — it is the composition of the continuous function "multiply by $\log a$" with the continuous function exp — and by the discussion above it agrees with the natural definition of a^x for $x \in \mathbb{Q}$; so it is the only sensible choice for a^x for arbitrary real x.

Notice that the fundamental relation for logarithms,

$$\log(a^x) = x \log a \quad \text{for all } a > 0 \text{ and all real } x$$

follows now *by definition* of a^x.

Here are some of the basic properties of the a^x function, all of which follow from just unravelling the definition, and using properties of exp and log:

- $a^0 = e^{0\log a} = e^0 = 1$ and $a^1 = e^{1\log a} = e^{\log a} = a$;
- for any reals x, y, $a^{x+y} = e^{(x+y)\log a} = e^{x\log a + y\log a} = e^{x\log a}e^{y\log a} = a^x a^y$ (these two properties are the ones we desired of a^x, along with the continuity we have already established);
- $(a^b)^c = e^{c\log(a^b)} = e^{(bc)\log a} = a^{bc}$; and
- if $\exp_a(x) = a^x = e^{x\log a}$ then $\exp_a'(x) = (\log a)a^x$ and so if $a > 1$ then $\exp_a$ is increasing (from 0 to ∞), while if $a < 1$ it is decreasing (from ∞ to 0), and in either case it is invertible; and furthermore, since $\exp_a''(x) = (\log a)^2 a^x > 0$, $\exp_a$ is concave.

We denote the inverse of $\exp_a$ by $\log_a$, so $\log_a(x) = y$ means $a^y = x$. There is an easy translation between $\log_a$ and log, namely:

$$\log_a x = \frac{\log x}{\log a}$$

(if $\log_a x = y$, then $a^y = x$, so $e^{y\log a} = x$, so $y\log a = \log x$). This allows us, for example, to quickly deduce that

$$\log_a'(x) = \frac{1}{x\log a}.$$

There are some basic algebraic identities that $\log_a$ satisfies, but we won't bother mentioning them here; in general when working with $\log_a$ or $\exp_a$ it is best to translate back to log and exp to do algebraic manipulations.

As an example of this, consider, for fixed $a \in \mathbb{R}$, the power function $f_a : (0, \infty) \to \mathbb{R}$ given by $f_a(x) = x^a$. (We have previously only considered the power function for rational a). We have $f_a(x) = e^{a\log x}$, and so

$$f_a'(x) = e^{a\log x} \cdot a \cdot \frac{1}{x} = ae^{a\log x} \cdot x^{-1} = ae^{a\log x}e^{-\log x} = ae^{(a-1)\log x} = ax^{a-1};$$

and so the usual rule for differentiating the power function holds, even when the exponent is an arbitrary real.

We end this discussion of the exponential and logarithm functions by presenting two theorems, one of which is a nice result with a short proof, that will not get used again, and the other of which is a slightly more technical result with a longer proof, that will be very useful to us later.

Theorem 12.3. *Suppose $f : \mathbb{R} \to \mathbb{R}$ is differentiable at all x, and $f' = f$. Then there is a constant c such that $f(x) = ce^x$ for all x. In particular, if $f(0) = 1$ then $f(x) = e^x$.*

Proof: Consider the function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(x) = f(x)/e^x$. This is differentiable everywhere, with derivative

$$g'(x) = \frac{f(x)e^x - f'(x)e^x}{e^{2x}} = 0$$

(since $f'(x) = f(x)$ for all x). So $g(x) = c$ for all x, for some constant c. In other words, $f(x) = ce^x$. And if $f(0) = 1$, then $1 = ce^0$ so $c = 1$ and $f(x) = e^x$. □

Theorem 12.4. *For each fixed $n \in \mathbb{N}$,*

$$\lim_{x \to \infty} \frac{e^x}{x^n} = \infty.$$

Essentially this says that the exponential function grows faster than any polynomial. Spivak gives a fairly delicate proof of this, but there are lots of proofs; one, using L'Hôpital's rule, appears in homework. The proof we give uses derivatives, and needs the following lemma.

Lemma 12.5. *If $f, g : [a, \infty)$ are both differentiable, with $f(a) = g(a)$ and $f'(x) \geq g'(x)$ for all x, then $f(x) \geq g(x)$ for all x. In particular, if $g(x) \to \infty$ as $x \to \infty$ then also $f(x) \to \infty$ as $x \to \infty$.*

Proof: We use the mean value theorem. Suppose there was some $b > a$ with $f(b) < g(b)$. Then, applying the mean value theorem to the function $h = f - g$ on $[a, b]$, we have that there is some $c \in (a, b)$ with

$$f'(c) - g'(c) = h'(c) = \frac{h(b) - h(a)}{b - a} = \frac{f(b) - g(b)}{b - a} < 0,$$

so $f'(c) < g'(c)$, a contradiction. □

Proof (of Theorem 12.4): Set $f(x) = e^x/x^n$. We have

$$f'(x) = \frac{e^x}{x^n}\left(1 - \frac{n}{x}\right).$$

This is positive for $x > n$, so f is increasing on $[n, \infty)$, and in particular that means that $f(x) \geq f(n) = e^n/n^n$ for $x \geq n$. It follows that for $x \in [n, \infty)$ we have

$$f'(x) \geq \frac{e^n}{n^n}\left(1 - \frac{n}{x}\right).$$

In particular, for $x \geq 2n$ we have

$$f'(x) \geq \frac{e^n}{2n^n} = c_n,$$

where c_n is some positive constant.

Now let $g(x)$ be the linear function with slope c_n that passes through the point $(2n, f(2n))$, that is,

$$g(x) = c_n(x - 2n) + f(2n).$$

On the interval $[2n, \infty)$ the conditions of Lemma 12.5 are satisfied by f and g, so, since $g(x) \to \infty$ as $x \to \infty$, we conclude that $f(x) \to \infty$ as $x \to \infty$, as claimed. □

Almost the same proof can be used to show that every exponential function (with base at least one) grows faster than every power function:

for $a > 1$ and $b < \infty$, $\lim_{x\to\infty} \frac{a^x}{x^b} = \infty$.

For example, 1.00000001^x grows faster than $x^{1,000,000,000}$ (though you have to look at very large values of x to see this!)

12.3 The trigonometric functions sin and cos

In this section, we use the integral to formally define the trigonometric functions sin (*sine*) and cos (*cosine*), and establish all their properties.

Recall first our provisional definition of sin and cos:

Provisional definition of sin and cos The points reached on unit circle centered at the origin, starting from $(1, 0)$, after traveling a distance θ, measured counter-clockwise, is $(\cos\theta, \sin\theta)$.

Knowing that the area of the unit circle is π, and that the circumference is 2π, we would get exactly the same thing if we said that the point P on the unit circle $x^2 + y^2 = 1$ has coordinates $(\cos\theta, \sin\theta)$ when $\theta/2$ is *area* of circle sector between $(1, 0)$ and P — at least, as long as P is in upper half plane (so $0 \leq \theta \leq \pi$).

To start the formal definition, we *define* π to be the "area" of the unit circle, or more specifically to be twice the area of that part of the unit circle $x^2 + y^2 = 1$ that lies in the upper half plane.

Definition of π

$$\pi = 2\int_{-1}^{1} \sqrt{1 - x^2}\, dx.$$

Using upper and lower Darboux sums for the partition $(-1, -4/5, -3/5, 0, 3/5, 4/5, 1)$ we get the very rough estimates

$$2.4 \leq \pi \leq 3.52.$$

Next, we set up a function $A(x)$ that captures the notion of the area of circle sector between $(1, 0)$ and $P = (x, \sqrt{1 - x^2})$, where P is in the upper half plane, that is, $-1 \leq x \leq 1$. For $0 \leq x \leq 1$ we have

$$A(x) = \frac{x\sqrt{1 - x^2}}{2} + \int_x^1 \sqrt{1 - t^2}\, dt$$

and for $-1 \leq x \leq 0$,

$$A(x) = \int_x^1 \sqrt{1 - t^2}\, dt - \frac{(-x)\sqrt{1 - x^2}}{2}$$

So in fact for every $x \in [-1, 1]$ we have

$$A(x) = \frac{x\sqrt{1 - x^2}}{2} + \int_x^1 \sqrt{1 - t^2}\, dt.$$

A is a continuous function on $[-1, 1]$, and it is differentiable on $(-1, 1)$, with derivative

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

This derivative is never 0, and in fact is always negative on $(-1, 1)$, so A is decreasing on $[-1, 1]$. It follows that the range of A is $[A(1), A(-1)] = [0, \pi/2]$. All this says that A has an inverse $A^{-1} : [0, \pi/2] \to [-1, 1]$ which is decreasing.

Following our informal definition of sin and cos, we want that for $0 \leq \theta \leq \pi$, $(\cos(\theta), \sin(\theta)$ is the point P on the circle $x^2 + y^2 = 1$ for which the area of the circle sector between $(1, 0)$ and P equals $\theta/2$. That is, we want $A(\cos\theta) = \theta/2$, or $\cos\theta = A^{-1}(\theta/2)$ (note that this seems to be sensible: as θ goes from 0 to π, $\theta/2$ goes from 0 to $\pi/2$, exactly the domain of A^{-1}).

Initial definition of $\cos$ **and** $\sin$ Define $\cos : [0, \pi] \to [-1, 1]$ by

$$\cos\theta = A^{-1}(\theta/2).$$

Define $\sin : [0, \pi] \to [0, 1]$ by

$$\sin\theta = \sqrt{1 - \cos^2\theta}.$$

Observe that, since A is differentiable on $(-1, 1)$ with derivative never 0, A^{-1} is differentiable on $(0, \pi)$, and for $\theta \in (0, \pi)$ cos is differentiable, with Have

$$\begin{aligned}
\cos'\theta &= (A^{-1})'(\theta/2) \\
&= \frac{1}{2A'(A^{-1}(\theta/2))} \\
&= -\sqrt{1 - A^{-1}(\theta/2))^2} \\
&= -\sin\theta.
\end{aligned}$$

Now differentiating the equation $\sin^2\theta + \cos^2\theta = 1$ get, on $(0, \pi)$,

$$\sin'\theta = \cos\theta.$$

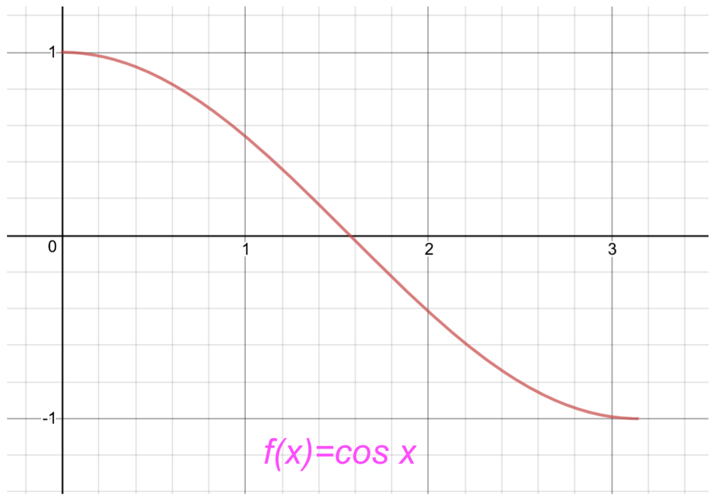
We are now in a position to sketch a reasonable graph of cos on the interval $[0, \pi]$. We have

- $\cos 0 = A^{-1}(0) = 1$
- $\cos\pi = A^{-1}(\pi/2) = -1$
- $\cos' = -\sin < 0$ on $(0, \pi)$, so cos decreasing
- cos is continuous, so by the intermediate value theorem there is an $m \in (0, \pi)$ with $\cos m = 0$. We have $A^{-1}(m/2) = 0$, so

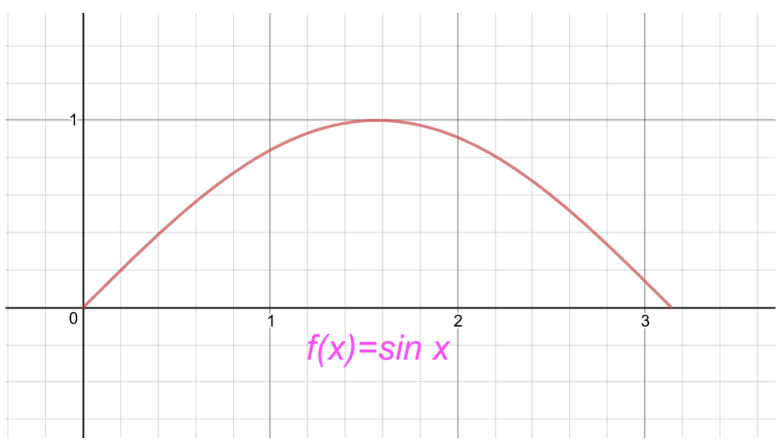
$$m = 2A(0) = \int_0^1 \sqrt{1-t^2}\,dt = 2\int_0^1 \sqrt{1-t^2}\,dt = \pi/2.$$

- $\cos'' = -\cos$, so cos is concave on $[0, \pi/2]$ and convex on $[\pi/2, \pi]$.
- $\cos'(\pi/2) = -\sin(\pi/2) = -\sqrt{1 - \cos^2(\pi/2)} = -1$, so as the graph crosses the x-axis it has slope -1.
- As $\theta \to 0^+$ we have $\cos'(\theta) = -\sin(\theta) = -\sqrt{1 - \cos^2(\theta)} \to 0$, and similarly as $\theta \to \pi^-$ we have $\cos'(\theta) \to 0$, so the graph is flat near 0 and π.

We sketch the graph of cos below, on the interval $[0, \pi]$.



The same reasoning can be used to sketch a graph of sin (again on $[0, \pi]$); this is left as an exercise.



We now extend sin and cos to the whole real line.

- We begin with the interval $[\pi, 2\pi]$. For θ in this range, the x-coordinate of the point that is distance θ from $(1, 0)$, is the same as the x-coordinate of the point that is distance θ' from $(1, 0)$, where $\theta' = 2\pi - \theta$. This motivates the definition that for $\theta \in [\pi, 2\pi]$,

$$\cos\theta = \cos(2\pi - \theta).$$

- Similarly for $\theta \in [\pi, 2\pi]$ we set

$$\sin\theta = -\sin(2\pi - \theta).$$

Observe that since cos is continuous on $[0, \pi]$, with $\lim_{x\to\pi^-} \cos x = \cos\pi = -1$, it follows that cos is continuous on $[\pi, 2\pi]$, with

$$\lim_{x\to\pi^+} \cos x = \lim_{x\to\pi^+} \cos(2\pi - x) = \lim_{x\to\pi^-} \cos x = \cos\pi = -1.$$

But from this it follows that actually cos is continuous on $[0, 2\pi]$. Similarly it can be argued that sin is continuous on $[0, 2\pi]$.

The relation $\cos^2\theta + \sin^2\theta = 1$ for $\theta \in [\pi, 2\pi]$ follows immediately from the same relation for $\theta \in [0, \pi]$, so in fact it too holds for all $\theta \in [0, 2\pi]$.

Finally, we turn to differentiability. Since $\cos' = -\sin$ on $(0, \pi)$, we have for $\theta \in (\pi, 2\pi)$ that

$$\cos'(\theta) = \cos'(2\pi - \theta) = -\sin(2\pi - \theta) \times (-1) = \sin(2\pi - \theta) = -\sin\theta.$$

What about $\cos'\pi$? We utilize the following lemma.

Lemma 12.6. *Suppose*

- *f is continuous at a,*
- *f' exists near a, and*
- *$\lim_{x\to a} f'(x) = L$ exists.*

Then $f'(a) = L$.

Proof: For $h > 0$, by the mean value theorem there's $\alpha_h \in (a, a+h)$ with

$$\frac{f(a+h) - f(a)}{h} = f'(\alpha_h).$$

As $h \to 0^+$ we have $\alpha_h \to a$, and so $f'(\alpha_h) \to L$. This shows that f is differentiable from above at a, with derivative L.

A similar argument applies for $h < 0$. □

We apply this lemma with $f = \cos$ and $a = \pi$. We've shown that cos is continuous at π, and it's differentiable near π. It's derivative near π is $-\sin$, which approaches 0 near π, so we conclude that cos is differentiable at π, with derivative 0 — which is $-\sin\pi$, so in fact $\cos' = -\sin$ on all of $(0, 2\pi)$. Similarly we can argue $\sin' = \cos$ on $[0, 2\pi]$.

- Finally, we extend both cos, sin periodically to $\mathbb{R}$, via

$$\cos\theta = \cos\theta', \quad \sin\theta = \sin\theta'$$

for $\theta = 2k\pi + \theta'$, $0 \le \theta' \le 2\pi$, $k \in \mathbb{Z}$.

That

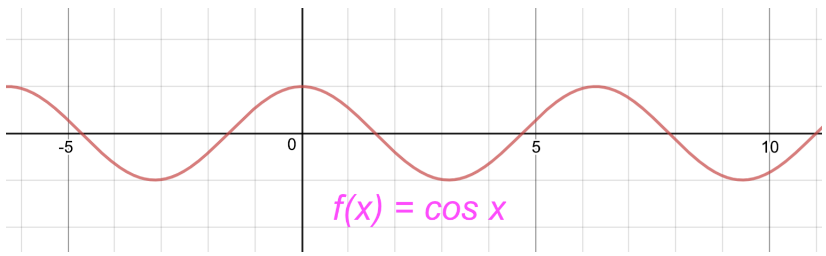
$$\sin^2\theta + \cos^2\theta = 1$$

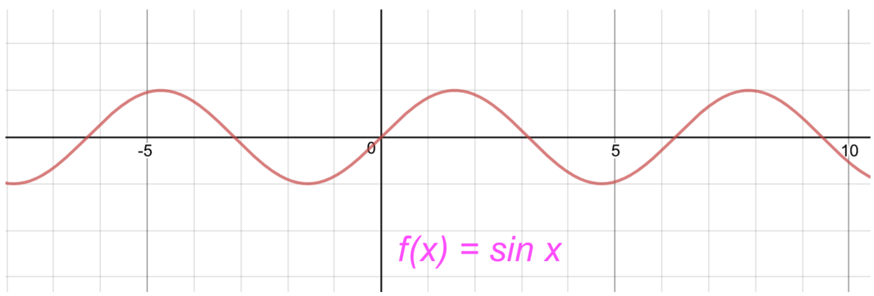
for all θ, follows almost immediately from the same relation for $\theta \in [0, 2\pi]$. That

$$\cos'(\theta) = -\sin(\theta), \quad \sin'(\theta) = \cos(\theta)$$

for all θ, follows exactly as this relation extended from $(0, \pi)$ to $(0, 2\pi)$ (via Lemma 12.6).

Here are the graphs of cos and sin on their full domains:





We make a digression here, to give another application of Lemma 12.6, that will be useful later. Consider the function f defined by

$$f(x) = \begin{cases} e^{-1/x^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

We claim that f is continuous and differentiable, and that $f'(0) = 0$. Away from 0, the function is clearly continuous and differentiable arbitrarily many times (also known as *infinitely differentiable*).

To show continuity at 0, we need to establish $\lim_{x\to 0} e^{-1/x^2} = 0$. Recall that we have proven that

$$\lim_{x\to 0^+} g(x) = \lim_{y\to\infty} g(1/y) \quad \text{and} \quad \lim_{x\to 0^-} g(x) = \lim_{y\to-\infty} g(1/y).$$

So to show $\lim_{x\to 0} e^{-1/x^2} = 0$, it suffices to show that

$$\lim_{y\to\infty} e^{-y^2} = 0 \quad \text{and} \quad \lim_{y\to-\infty} e^{-y^2}.$$

Since $e^y \to \infty$ as $y \to \infty$, and $y^2 > y$ for all large y, it follows that $e^{y^2} \to \infty$ as $y \to \infty$, so that indeed $\lim_{y\to\infty} e^{-y^2} = 0$; that $\lim_{y\to-\infty} e^{-y^2}$ is established similarly (see below for details in a similar case). This shows that f is continuous at 0.

For differentiability, we use Lemma 12.6. Here, f is continuous at 0, and differentiable near 0. So to establish that f is differentiable at 0, with derivate 0, if suffices to show that

$$\lim_{x\to 0} \frac{2e^{-1/x^2}}{x^3} = 0$$

(note that $f'(x) = 2e^{-1/x^2}/x^3$ if $x \neq 0$).

As before, this limit is implied by

$$\lim_{y\to\infty} \frac{y^3}{e^{y^2}} = 0 \quad \text{and} \quad \lim_{y\to-\infty} \frac{y^3}{e^{y^2}} = 0.$$

Since $e^{y^2} > e^y$ for all large positive y, and since $y^3/e^y \to 0$ as $y \to \infty$ (a basic estimate that we proved in class), it follows that $\lim_{y\to\infty} \frac{y^3}{e^{y^2}} = 0$. For the negative limit, notice that

$$\lim_{y\to-\infty} \frac{y^3}{e^{y^2}} = \lim_{z\to\infty} \frac{(-z)^3}{e^{(-z)^2}} = -\lim_{z\to\infty} \frac{z^3}{e^{z^2}} = -0 = 0.$$

So we conclude that f is differentiable at 0, with derivative 0.

In fact, we can do more: f is k times differentiable for every natural number k, and $f^{(k)}(0) = 0$. To see this we will make use of the fact that away from 0, the kth derivative of f has the following form:

$$f^{(k)}(x) = P_k(1/x)e^{-1/x^2}$$

where P_k is a polynomial. This fact can be proven by induction on k. Indeed, for $k = 1$ we have already shown it (in part (a)), with specifically the polynomial being $P_1(z) = 2z^3$.

For $k > 1$, suppose that $f^{(k-1)}(x) = P_{k-1}(1/x)e^{-1/x^2}$ where P_{k-1} is a polynomial. We then have

$$\begin{aligned} f^{(k)}(x) &= P_{k-1}(1/x)e^{-1/x^2}\left(2/x^3\right) + e^{-1/x^2}P'_{k-1}(1/x)\left(-1/x^2\right) \\ &= \left(\frac{2}{x^3}P_{k-1}(1/x) - \frac{P'_{k-1}(1/x)}{x^2}\right)e^{-1/x^2}, \end{aligned}$$

so that indeed $f^{(k)}(x) = P_k(1/x)e^{-1/x^2}$ with P_k the polynomial given by $P_k(z) = 2z^3P_{k-1} - z^2P'_{k-1}(z)$. This completes the induction.

We will also make use of the fact that if P is a polynomial, then $\lim_{x\to 0} P(1/x)e^{-1/x^2} = 0$. Indeed, to show this it suffices to show

$$\lim_{y\to\infty} P(y)/e^{y^2} = 0 \quad \text{and} \quad \lim_{y\to-\infty} P(y)/e^{y^2} = 0.$$

The first of these follows immediately from $e^{y^2} > e^y$ for large (positive y) and the fact that y^k/e^y goes to 0 as $y \to \infty$ for any natural number k; then the second follows from the first on observing that

$$\lim_{y\to-\infty} P(y)/e^{y^2} = \lim_{z\to\infty} P(-z)/e^{z^2} = 0.$$

We now prove that the predicate $p(k)$: "f is k times differentiable and $f^{(k)}(0) = 0$" is true for all natural numbers k, by induction on k, with the base case $k = 1$ having been proven earlier.

For the induction step, suppose that f is k times differentiable and $f^{(k)}(0) = 0$. Let g be the kth derivative of f. By induction $g(0) = 0$, and by the calculations done above g approaches 0 near 0, so that g is continuous at 0. But now again by the calculations done above, g' approaches 0 near 0 (away from zero, g' is the $(k+1)$st derivative of f calculated above), so by Lemma 12.6, g' exists at 0 and takes value 0 there. This completes the induction.

The function f is as flat as it possible can be at 0 — its value, and the values of all its derivatives, are 0. And yet the function is *not* the zero function everywhere. We will return to this when we talk about Taylor polynomials.

Returning to trigonometric functions: it is now immediate that $\sin''(\theta) = -\sin(\theta)$ and that $\cos''(\theta) = -\cos(\theta)$ for all θ, that is, that $\sin, \cos$ are both solutions to the differential equation $f'' + f = 0$. Just as exp was (essentially) characterized by the differential equation $f' = f$ (Theorem 12.3), it turns out that sin and cos are (essentially) characterized by the differential equation $f'' + f = 0$. Unlike Theorem 12.3, which is nice but will not get used again this year, the following analogous theorem will have an immediate and important pay-off.

Theorem 12.7. *Suppose $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable at all x, that $f'' + f = 0$, and that $f(0) = a$ and $f'(0) = b$. Then $f(x) = a\cos x + b\sin x$ for all x.*

Proof: We begin with the special case $a = b = 0$. Since $f'' + f = 0$, we have $f'f'' + f'f = 0$,[193] and so $((f')^2 + f^2)' = 0$ and so $(f')^2 + f^2 = C$ for some constant C. Evaluating the left-hand side at 0, we find that $C = 0$, so f^2, $(f')^2$ are both zero, and in particular $f = 0$, as claimed.

Now for general a, b, set $g = f - a\cos - b\sin$. We have $g'' + g = 0$, $g(0) = 0$, $g'(0) = 0$, and so $g = 0$. This says that $f = a\cos x - b\sin x$, as claimed. □

The immediate pay-off is that the addition formulae for sin and cos are now almost immediate.

Theorem 12.8. *For all x, y,*

- $\sin(x + y) = \sin x \cos y + \sin y \cos x$ *and*
- $\cos(x + y) = \cos x \cos y - \sin x \sin y$.

Proof: We just prove the first identity; the second is similar. For each *fixed* y, the function $f(x) = \sin(x + y)$ (a function of x only) satisfies $f'' + f = 0$, $f(0) = \sin y$, and $f'(0) = \cos y$, so by Theorem 12.7 we have $\sin(x + y) = f(x) = \sin y \cos x + \cos y \sin x$. □

This allows us to calculate some particular values of the functions sin and cos. The first we will use fairly soon, so we derive that fully; the others we will never use, so are left as exercises.

[193] So ... this is a rabbit-out-of-a-hat proof.

- $\sin \pi/4 = \cos \pi/4 = \sqrt{2}/2$. We have $0 = \cos \pi/2 = \cos(\pi/4 + \pi/4) = \cos^2(\pi/4) - \sin^2(\pi/4)$, so $\cos^2(\pi/4) = \sin^2(\pi/4)$. Since both are positive, we have $\cos(\pi/4) = \sin(\pi/4) > 0$. Now from $\cos^2(\pi/4) + \sin^2(\pi/4) = 1$ we get $2\cos^2(\pi/4) = 1$ or $\cos(\pi/4) = \sqrt{2}/2$.

- $\sin \pi/6 = \cos \pi/3 = 1/2$.

- $\sin \pi/3 = \cos \pi/6 = \sqrt{3}/2$.

Another consequence of Theorems 12.7 and 12.8 is that we can use it to verify some properties of sin and cos that appears obvious from the graphs of the two functions, but up until now would have been quite hard to prove.

Theorem 12.9. *1. The graph of* sin *is a shift of the graph of* cos*; specifically, for all x,* $\sin(x + \pi/2) = \cos(x)$.

2. sin *is an odd function; that is, for all x,* $\sin(-x) = -\sin(x)$.

3. cos *is an even function; that is, for all x,* $\cos(-x) = \cos(x)$.

Proof: For item 1 we have, using Theorem 12.8 and some special values of sin and cos that come out of the definition,

$$\sin(x + \pi/2) = \sin(x)\cos(\pi/2) + \cos(x)\sin(\pi/2) = \cos(x).$$

For item 2, consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = \sin(-x)$. We have $f'(x) = -\cos(-x)$ and $f''(x) = -\sin(-x)$, so f satisfies the equation $f'' + f = 0$. It follows from Theorem 12.8 that for all x,

$$f(x) = f(0)\cos x + f'(0)\sin x = 0 \cdot \cos x - 1 \cdot \sin x = -\sin x.$$

But since $f(x) = \sin(-x)$ for all x, we immediately get that $\sin(-x) = -\sin(x)$ for all x.

The proof of item 3 is similar to that of item 2, and is omitted. □

12.4 The other trigonometric functions

Having defined sin and cos, we can now define some auxiliary trigonometric functions. The most important of these is the tangent function.

Definition of tan The *tangent* function $\tan : \mathbb{R} \setminus \{(n + 1/2)\pi : n \in \mathbb{Z}\}$[194] is defined by

$$\tan\theta = \frac{\sin\theta}{\cos\theta}.$$

[194]Note that the domain is precisely those points where $\cos \neq 0$.

Since sin and cos are periodic with period 2π ($\sin(x+2\pi) = \sin x$ for all x), it is clear that tan is also periodic with period 2π. But in fact, tan has period π: using the angle sum formulae for $\sin, \cos$, along with $\sin\pi = 0$, $\cos\pi = -1$ we get

$$\tan(x+\pi) = \frac{\sin(x+\pi)}{\cos(x+\pi)} = \frac{-\sin x}{-\cos x} = \tan x.$$

To understand the tan function, then, it suffices to examine it on the interval $(-\pi/2, \pi/2)$. On this interval it is continuous and differentiable, with (by the quotient rule)

$$\tan'(x) = \frac{(\cos x)(\cos x) - (\sin x)(-\sin x)}{\cos^2 x} = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = \frac{1}{\cos^2 x}.$$

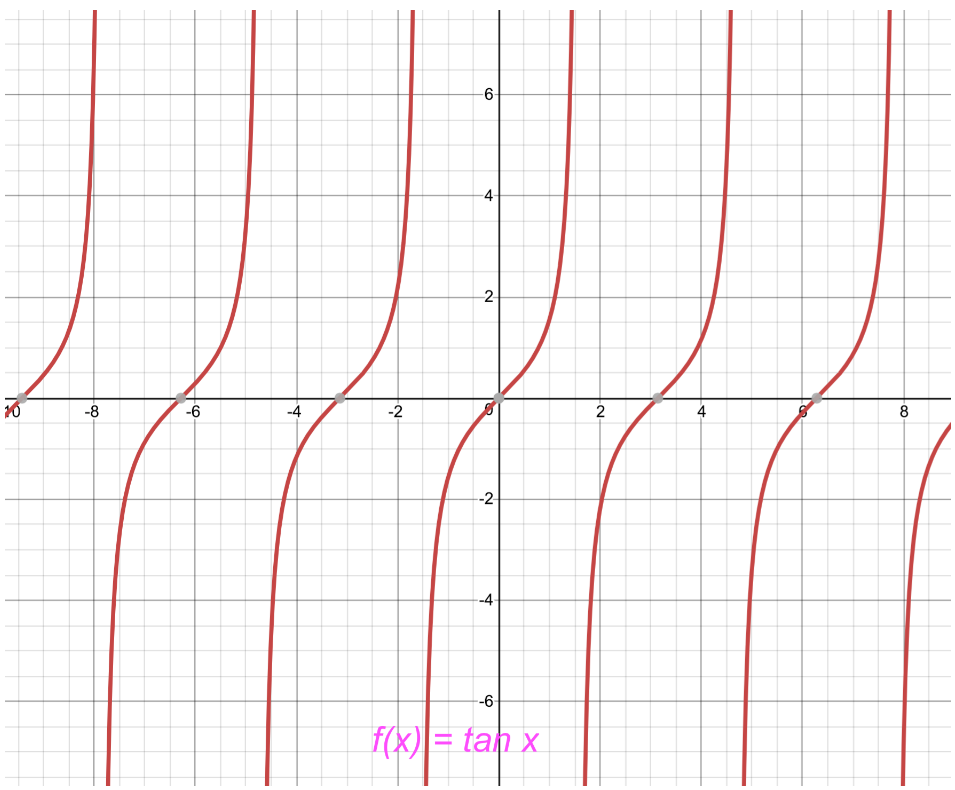
Since this is positive, tan is increasing on $(-\pi/2, \pi/2)$. From our knowledge of sin and cos, we have

$$\lim_{x\to\pi/2^-} \tan x = +\infty, \qquad \lim_{x\to-\pi/2^+} \tan x = -\infty.$$

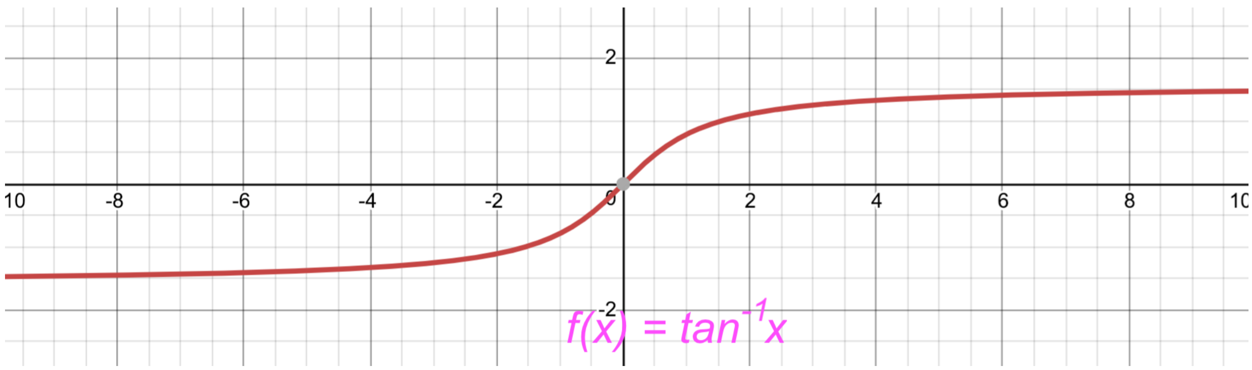
Also,

$$\tan''(x) = \frac{2\sin x}{\cos^3 x},$$

which is positive for $x \in [0, \pi/2)$ (so tan is convex on that interval) and negative for $x \in (-\pi/2, 0]$ (so tan is concave on that interval). Finally noting that $\tan 0 = 0$, we have enough information to produce an accurate graph tan:



tan is clearly not invertible, but it becomes invertible if it is restricted to the domain $(\pi/2, \pi/2)$ (on that interval it is monotone increasing from $-\infty$ to ∞). We define the function $\tan^{-1} : \mathbb{R} \to (\pi/2, \pi/2)$[195] to be the inverse of the function $\tan : (\pi/2, \pi/2) \to \mathbb{R}$ (the restriction of tan to the domain $(\pi/2, \pi/2)$). That is, for each real x, $\tan^{-1}(x)$ is defined to be the unique $\theta \in (\pi/2, \pi/2)$ such that $\tan\theta = x$. From the graph of tan we easily get the graph of $\tan^{-1}$:



Notice that it is monotone increasing on the whole real line, but bounded.

We now compute the derivative of $\tan^{-1}$, which turns out to be $1/(1+x^2)$[196]. We will use the identity

$$\tan^2 x + 1 = \frac{1}{\cos^2 x},$$

valid on $(-\pi/2, \pi/2)$, which follows immediately from $\sin^2 x + \cos^2 x = 1$. We have

$$\begin{aligned}(\tan^{-1})'(x) &= \frac{1}{\tan'(\tan^{-1}(x))} \\ &= \cos^2(\tan^{-1}(x)) \\ &= \frac{1}{1+\tan^2(\tan^{-1}(x))} \\ &= \frac{1}{1+x^2}.\end{aligned}$$

Note that this tallies with the graph of $\tan^{-1}$:

- $(\tan^{-1})'$ is positive, the function is increasing;
- $(\tan^{-1})'' = -2x/(1+x^2)^2$, which is negative for negative x (where the function is concave), and positive for positive x (where the function is convex);
- $\lim_{x\to\pm\infty}(\tan^{-1})'(x) = 0$, and the graph is flat at $\pm\infty$.

[195] Sometimes called "arctan".

[196] The appearance of such a simple, rational function, as the derivative of $\tan^{-1}$, should not be surprising; recall that cos was defined as the inverse of a function that very clearly has a rational function as its derivative.

The relation $(\tan^{-1})'(x) = 1/(1+x^2)$ leads to an integral relation. Define $F : \mathbb{R} \to (-\pi/2, \pi/2)$ by $F(x) = \int_0^x dt/(1+t^2)$. Since by the fundamental theorem of calculus $F'(x) = 1/(1+x^2)$, we have that $F(x) = \tan^{-1}(x) + C$ for some constant C; and setting $x = 0$ we get $C = 0$. So:

$$\tan^{-1}(x) = \int_0^x \frac{dt}{1+t^2}.$$

Recall that previously we showed, by comparison with $1/t^2$, that $\int_0^\infty dt/(1+t^2)$ exists; now we give a value to that integral:

$$\int_0^\infty \frac{dt}{1+t^2} = \lim_{x\to\infty} \int_0^\infty \frac{dt}{1+t^2} = \lim_{x\to\infty} \left(\tan^{-1}(x) - \tan^{-1}(0)\right) = \frac{\pi}{2}.$$

The main point of all of this discussion of $\tan^{-1}$, though, is the following. We have shown $\cos(\pi/4) = \sin(\pi/4) = \sqrt{2}/2$, so $\tan(\pi/4) = 1$. In other words,

$$\int_0^1 \frac{dt}{1+t^2} = \frac{\pi}{4}.$$

There is a way to estimate $\int_0^1 dt/(1+t^2)$ that does not require trigonometric functions. We have, for each natural number n that is divisible by 4, at for each $t \geq 0$,

$$1 - t^2 + t^4 - t^6 + t^8 - \cdots - t^{n-2} \leq \frac{1}{1+t^2} \leq 1 - t^2 + t^4 - t^6 + t^8 - \cdots + t^n.$$

Indeed, if we multiply across by $1 + t^2$, this becomes

$$1 - t^n \leq 1 \leq 1 + t^{n+2},$$

which is clearly true. It follows that

$$\int_0^1 \left(1 - t^2 + t^4 - t^6 + t^8 - \cdots - t^{n-2}\right) dt \leq \int_0^1 \frac{dt}{1+t^2} \leq \int_0^1 \left(1 - t^2 + t^4 - t^6 + t^8 - \cdots + t^n\right) dt.$$

We know that the integral in the middle is $\pi/4$. The integral on the right can easily be evaluated by the fundamental theorem of calculus:

$$\int_0^1 \left(1 - t^2 + t^4 - t^6 + t^8 - \cdots + t^n\right) dt = 1 - \frac{1}{3} + \frac{1}{5} - \cdots + \frac{1}{n+1},$$

while

$$\int_0^1 \left(1 - t^2 + t^4 - t^6 + t^8 - \cdots - t^{n-2}\right) = 1 - \frac{1}{3} + \frac{1}{5} - \cdots - \frac{1}{n-1}.$$

Combining, we get that for any n that is a multiple of 4,

$$1 - \frac{1}{3} + \frac{1}{5} - \cdots - \frac{1}{n-1} \leq \frac{\pi}{4} \leq 1 - \frac{1}{3} + \frac{1}{5} - \cdots + \frac{1}{n+1}.$$ [197]

[197] When we come to learn about infinite series, we will see that this translates to the sum

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots.$$

This is known variously as *Leibniz formula for* π, or as *Gregory's series*.

The difference between the right- and left-hand sides of this series of inequalities is $1/(n+1)$, which goes to 0 as n goes to infinity. This says that for all $\varepsilon > 0$ we can find an interval $[a, b]$, with rational endpoints and of length at most ε, inside which $\pi/4$ must lie; and of course, by using

$$4 - \frac{4}{3} + \frac{4}{5} - \cdots - \frac{4}{n-1} \leq \pi \leq 4 - \frac{4}{3} + \frac{4}{5} - \cdots + \frac{4}{n+1}$$

we can pin down π itself into a window of arbitrarily small width. For example, taking $n = 10,000$ we find that $\pi \in [3.14154, 3.14165]$.

This is nice, though not very efficient. But we can do better. From $\tan^{-1}(x) = \int_0^x dt/(1+t^2)$ we can use exactly the same argument to conclude that for positive x and n a multiple of 4, we have

$$x - \frac{x^3}{3} + \cdots - \frac{x^{n-1}}{n-1} \leq \tan^{-1}(x) \leq x - \frac{x^3}{3} + \cdots + \frac{x^{n+1}}{n+1}.$$

The difference between the right- and left-hand sides here is $x^{n+1}/(n+1)$, which goes very quickly to 0 as n grows, as long as $0 < x < 1$ (in contrast to $1/(n+1)$, which goes to 0 very slowly). So if we had some expression for π involving $\tan^{-1}(x)$ for small x, could get more accurate estimates for π more quickly.

Many such expressions are known. The most famous of them[198] is

$$\pi = 16 \tan^{-1} \frac{1}{5} - 4 \tan^{-1} \frac{1}{239}$$

(whose proof is left as an exercise).

This leads to the following bounds for π: with $x = 1/5$ and $y = 1/239$,

$$\pi \leq 16 \left(x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + \frac{x^{4n+1}}{4n+1} \right) - 4 \left(y - \frac{y^3}{3} + \frac{y^5}{5} - \cdots - \frac{y^{4n-1}}{4n-1} \right)$$

and

$$\pi \geq 16 \left(x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + \frac{x^{4n-1}}{4n-1} \right) - 4 \left(y - \frac{y^3}{3} + \frac{y^5}{5} - \cdots - \frac{y^{4n+1}}{4n+1} \right).$$

At $n = 5$ this already leads to

$$3.14159265358979 \leq \pi \leq 3.14159265358980,$$

accurate to 12 decimal places!

There are three other (somewhat) commonly encountered trigonometric functions, that are (essentially) the reciprocals of sin, cos and tan[199]. We mention them here for completeness, without really delving too deeply into them.

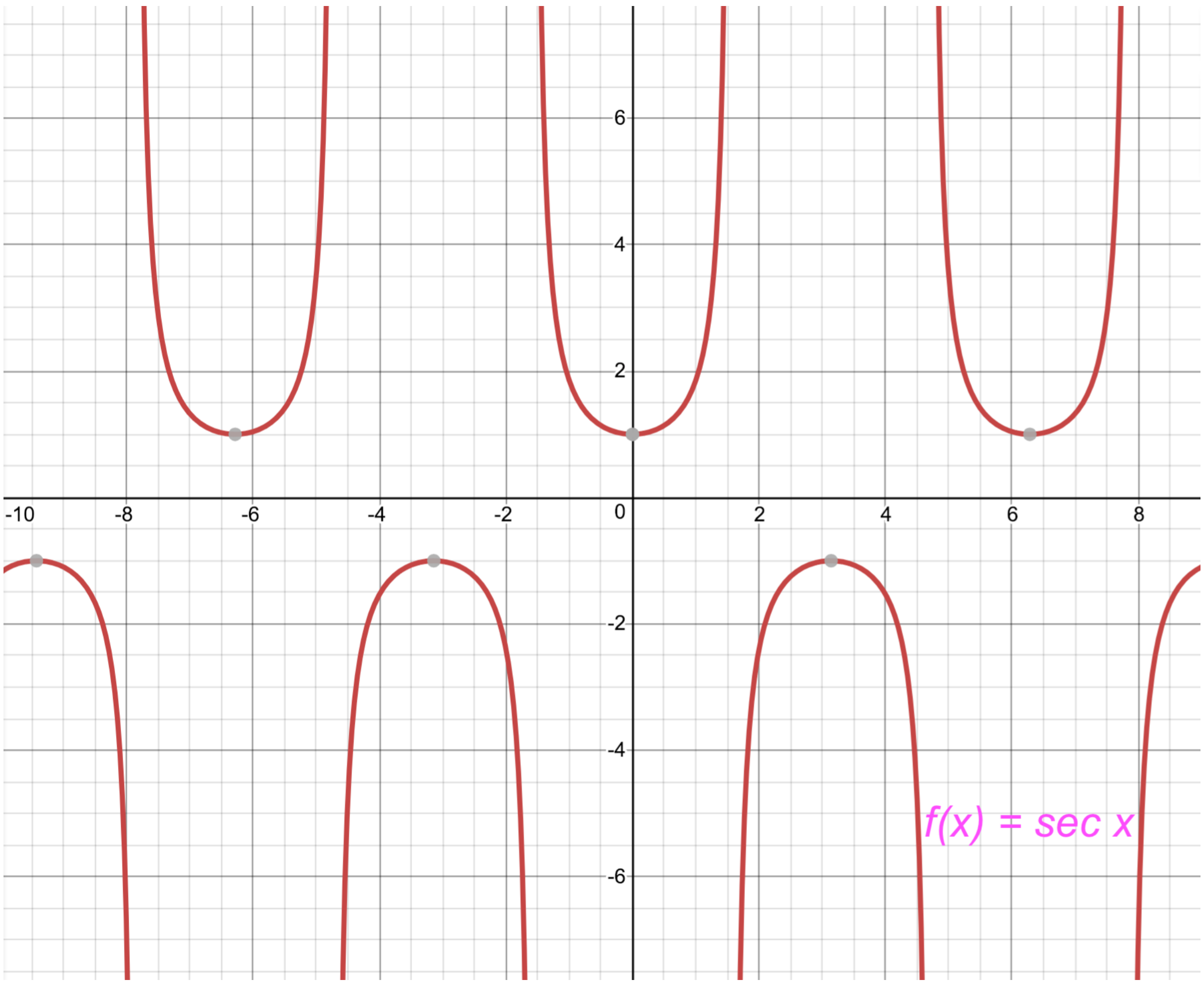
[198] Discovered in 1706 by J. Machin, and hence referred to as *Machin* or *Machin-like formulae*. Machin used his formula to calculate π to 100 decimal places in 1706. Today, much more elaborate Machin-like formulae are known, that allow π to be rapidly calculated to unfathomably many decimal places — see for example https://en.wikipedia.org/wiki/Machin-like_formula for many examples.

[199] Not exactly. tan is defined as sin / cos, while cot is defined as cos / sin. But cot is *not* the reciprocal of tan. Why not?

Definition of sec**, the secant function** $\sec : \mathbb{R} \setminus \{(n+1/2)\pi : n \in \mathbb{Z}\} \to (-\infty, -1] \cup [1, \infty)$ is defined by

$$\sec x = \frac{1}{\cos x}.$$

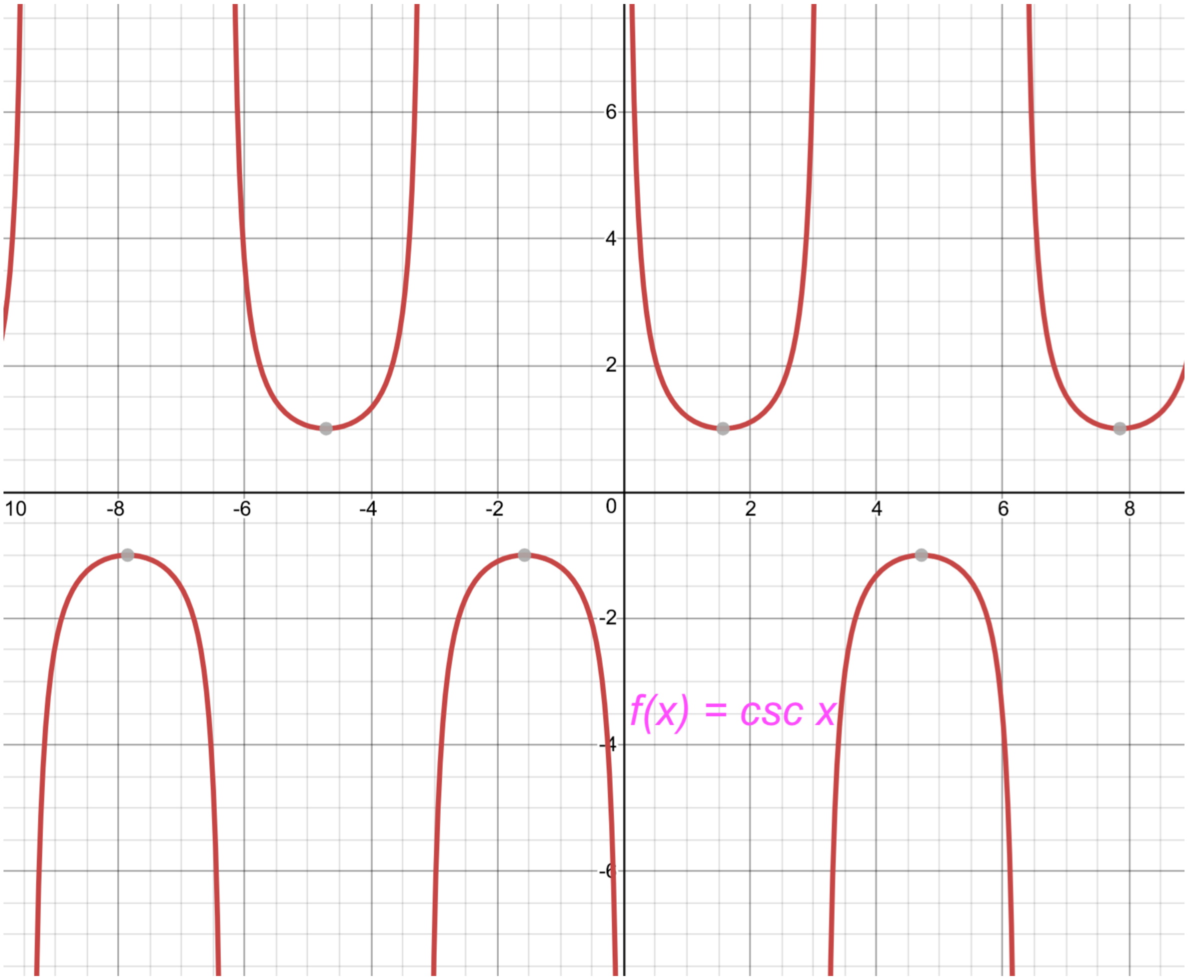
Here is a graph of sec:



Definition of csc**, the cosecant function** $\csc : \mathbb{R} \setminus \{n\pi : n \in \mathbb{Z}\} \to (-\infty, -1] \cup [1, \infty)$ is defined by

$$\csc x = \frac{1}{\sin x}.$$

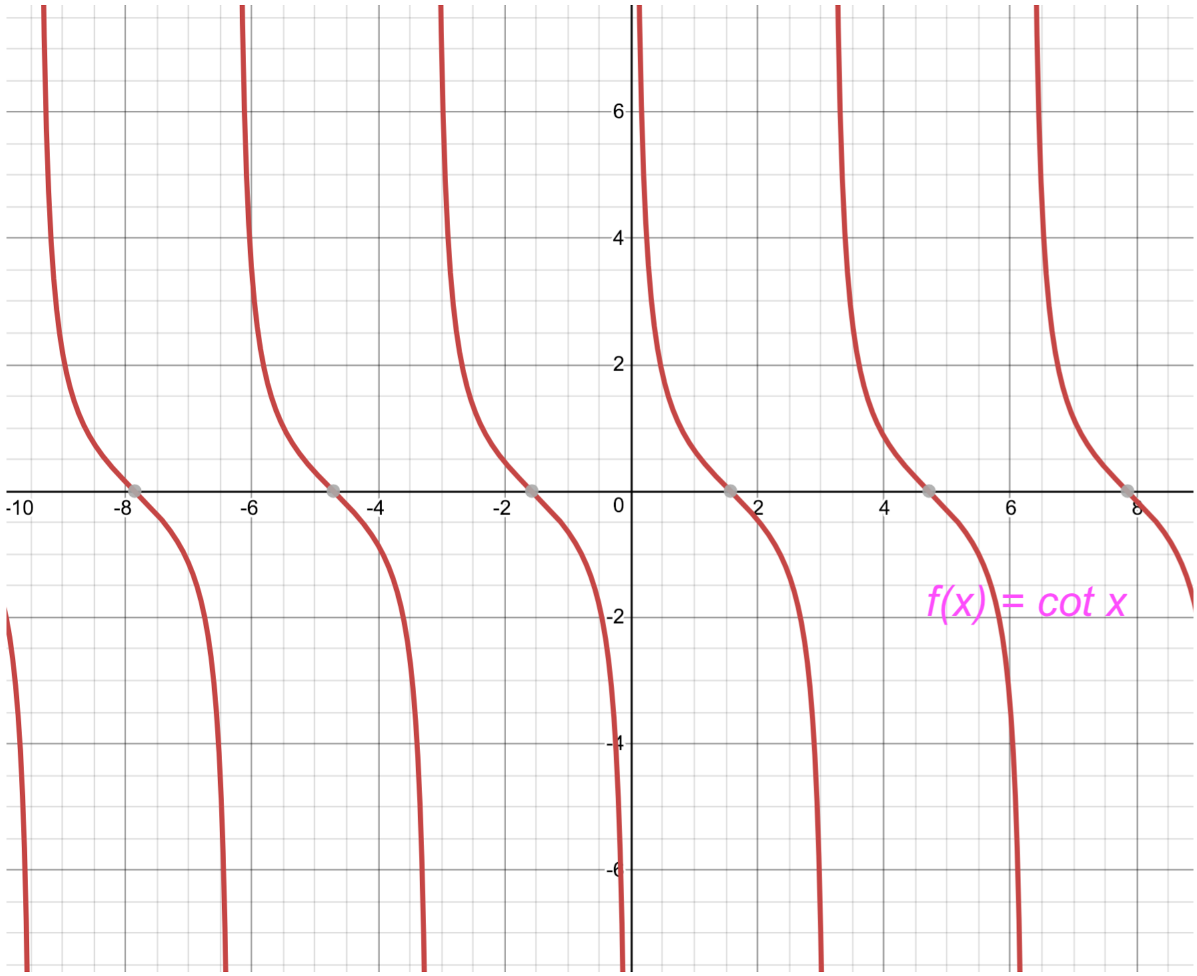
Here is a graph of csc:



Definition of $\cot$**, the cotangent function** $\cot : \mathbb{R} \setminus \{n\pi : n \in \mathbb{Z}\} \to (-\infty, -1] \cup [1, \infty)$ is defined by

$$\cot x = \frac{\cos x}{\sin x}.$$

Here is a graph of cot:



All these functions are easily differentiated. It's not worth writing down any of the derivatives (at least at the moment); but it is worth noting that since $\tan' = 1/\cos^2$, we have in this language that

$$\tan' = \sec^2,$$

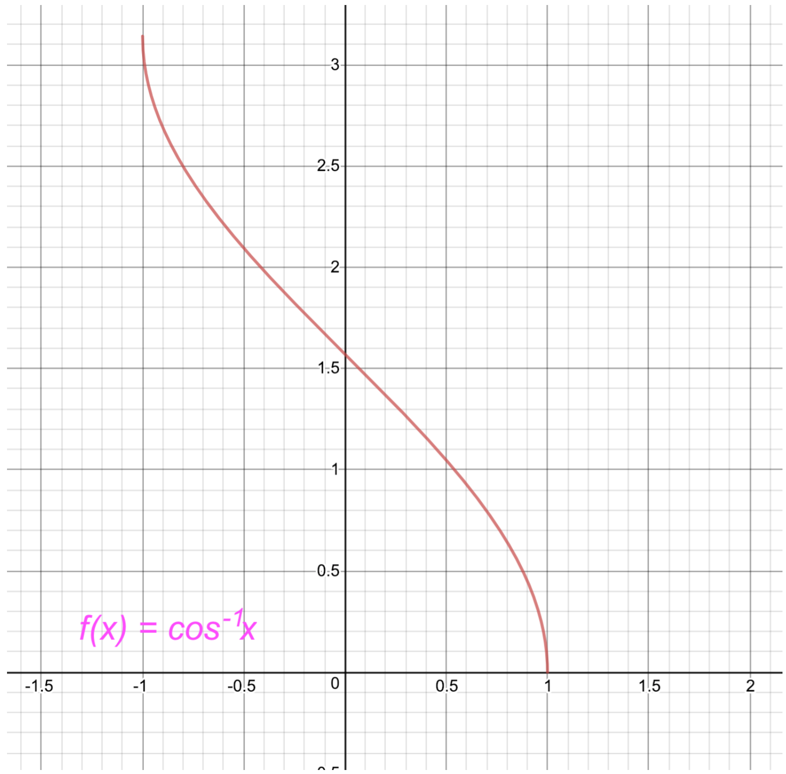
and also that since $\tan^2 + 1 = 1/\cos^2$, we have in this language that

$$\tan^2 + 1 = \sec^2.$$

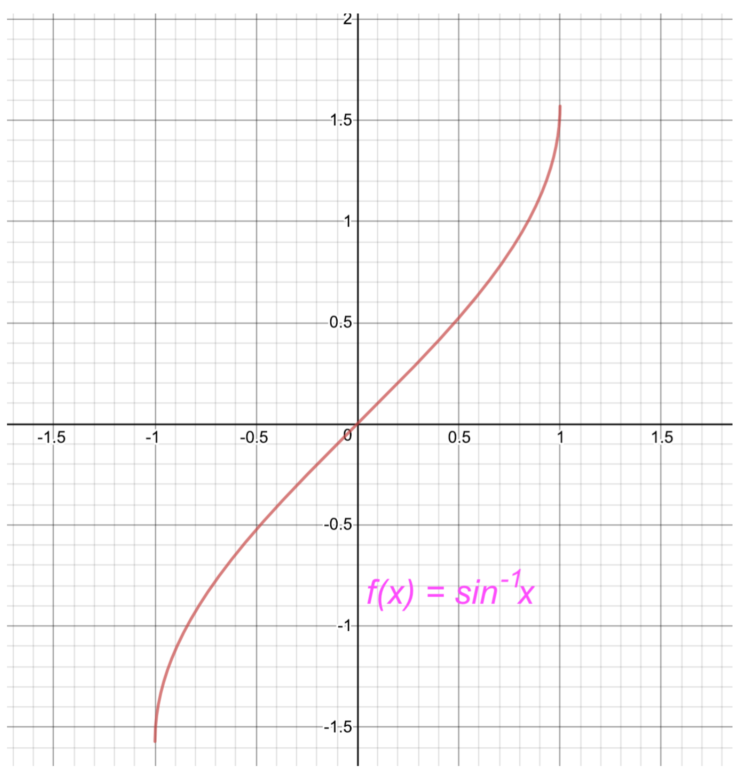
We have already discussed the inverse of the tangent function. Just like tan, none of the other trigonometric functions are invertible on their full domains, but they become invertible if suitably restricted. We only discuss here the inverses of cos, sin and sec.

It is standard to restrict cos to the interval $[0, \pi]$, and sin to the interval $[-\pi/2, \pi/2]$, to define their inverses. Formally we define the function $\cos^{-1} : [-1, 1] \to [0, \pi]$[200] to be the inverse of the function $\cos : [0, \pi] \to [-1, 1]$ (the restriction of cos to the domain $[0, \pi]$). That is, for each $x \in [-1, 1]$, $\cos^{-1}(x)$ is defined to be the unique $\theta \in [0, \pi]$ such that $\cos \theta = x$. From the graph of cos we easily get the graph of $\cos^{-1}$:

[200] Sometimes called "arccos".



And we define the function $\sin^{-1} : [-1, 1] \to [-\pi/2, \pi/2]$[201] to be the inverse of the function $\sin : [-\pi/2, \pi/2] \to [-1, 1]$. Here is the graph of $\sin^{-1}$:



[201] Sometimes called "arcsin".

These two functions have nice derivatives. For $x \in [-1, 1]$,

$$(\cos^{-1})'(x) = \frac{1}{\cos'(\cos^{-1}(x))} = \frac{-1}{\sin(\cos^{-1}(x))}.$$

Now

$$1 = \sin^2(\cos^{-1}(x)) + \cos^2(\cos^{-1}(x)) = \sin^2(\cos^{-1}(x)) + x^2,$$

, so

$$\sin^2(\cos^{-1}(x)) = 1 - x^2 \quad \text{and} \quad \sin(\cos^{-1}(x)) = \sqrt{1-x^2}$$

(we take the positive square root since $\arccos(x) \in [0, \pi]$, on which interval sin is non-negative). It follows that

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

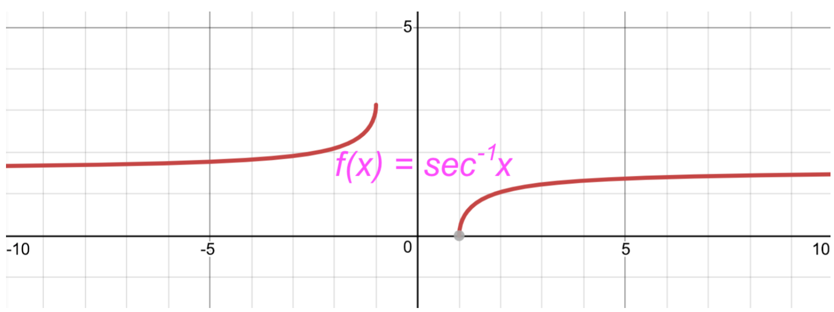
Similarly,

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

Later, when we talk about trigonometric substitutions, it will be useful to know about the inverse of the secant function. It is conventional here to restrict sec to the domain $[0, \pi/2) \cup (\pi/2, \pi]$. As x ranges over $[0, \pi/2)$, $\cos x$ ranges over $[1, \infty)$ (and is increasing), and as x ranges over $(\pi/2, \pi]$, $\cos x$ ranges over $(-\infty, -1]$ (and is also increasing). So

$$\sec^{-1} : (-\infty, -1] \cup [1, \infty) \to [0, \pi/2) \cup (\pi/2, \pi]$$

and has the following graph:

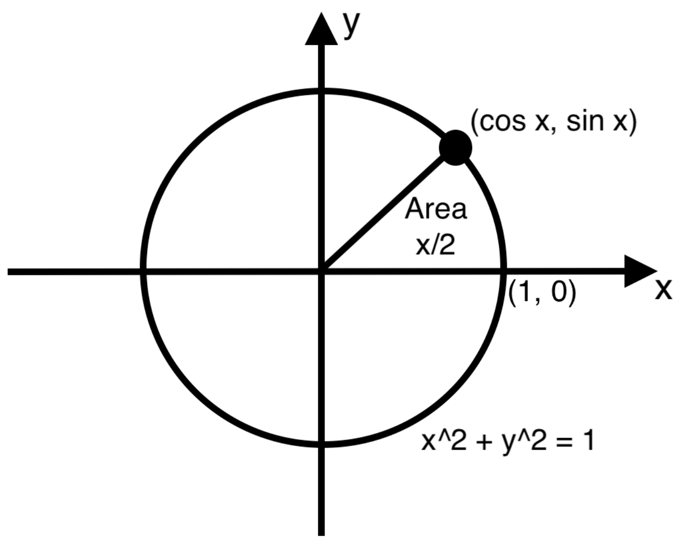


This is probably the first natural instance of a function whose domain is a union of intervals.

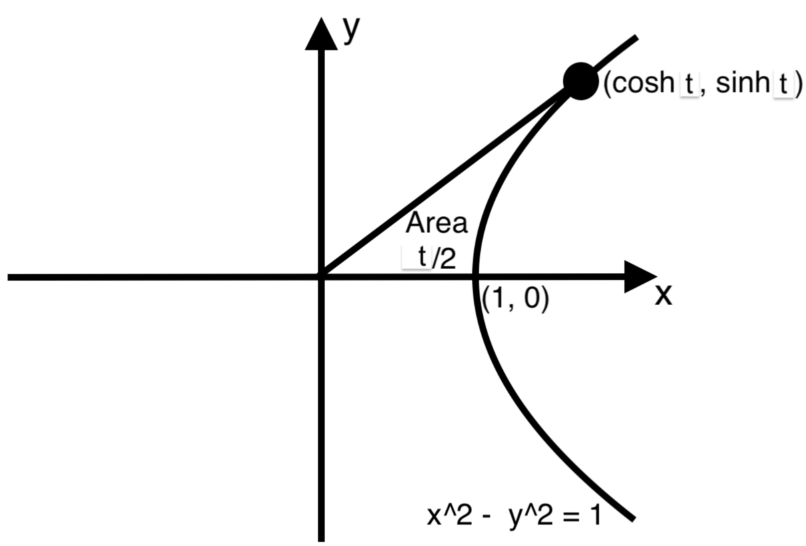
12.5 The hyperbolic trigonometric functions

This section discusses the definitions of, and derivation of the basic properties of, the so-called *hyperbolic* functions.

Just as the trigonometric functions were defined via the following picture:



the hyperbolic functions are defined via the following picture:



That is: let $P = (a, b)$ be a point on the curve $x^2 - y^2 = 1$, with $a \geq 1$ and $b \geq 0$. If the area A bounded by

- the x-axis between $(1, 0)$ and $(0, 0)$,
- the line segment from $(0, 0)$ to P, and
- the curve $x^2 - y^2 = 1$ between P and $(1, 0)$

is $t/2$, then $a = \cosh t$ and $b = \sinh t$. The curve $x^2 - y^2 = 1$ is a hyperbola, hence the adjective "hyperbolic".

It's obvious that this defines the functions cosh, sinh, both on some domain that starts at 0 (and includes 0). It's not entirely obvious just what that domain is — that depends on what is the area of the slanted needle bounded by the line $x = y$ (that the hyperbola is approaching, for large x), the hyperbola, and the the x-axis. If this needle has infinite area, then cosh, sinh have just been defined on $[0, \infty)$, whereas if it has finite area, L say, then cosh, sinh have just been defined on $[0, L)$.

We will discover the answer to the question, "what is the domain on cosh, sinh" in a quite direct way. Unlike with the trigonometric functions, it is possible to come up an explicit expression for $\cosh x$ and $\sinh x$ in terms of functions we have previously defined; specifically, in terms of the exponential function:

$$\cosh t = \frac{e^t + e^{-t}}{2}, \qquad \sinh t = \frac{e^t - e^{-t}}{2}.$$

To see this, first consider the function $f : [0, \infty) \to \mathbb{R}$ given by $f(t) = (e^t + e^{-t})/2$. It is an easy check that this is a is monotone increasing function, with range $[1, \infty)$. This says that for any point $P = (a, b)$ on $x^2 - y^2 = 1$ with $a \geq 1$ and $b \geq 0$, there is a unique $t \in [0, \infty)$ with $a = (e^t + e^{-t})/2$.

Next, it is easy the verify that if $a = (e^t + e^{-t})/2$, then $b = (e^t - e^{-t})/2$. Indeed, given $a^2 - b^2 = 1$ and $a = (e^t + e^{-t})/2$, simple algebra gives that b is one of $\pm(e^t - e^{-t})/2$; and the correct choice to make $b \geq 0$ is easily seen to be $(e^t - e^{-t})/2$. This, together with the observation of the last paragraph, shows that we can parameterize the points of the hyperbola in the first quadrant by $\{(e^t + e^{-t})/2, (e^t - e^{-t})/2 : t \in (0, \infty)\}$. It also says that if we can show that $\cosh t = (e^t + e^{-t})/2$, then we automatically get that $\sinh t = (e^t - e^{-x})/2$.

Let $P = ((e^t + e^{-t})/2, (e^t - e^{-t}/2)$ be a parameterized point on the hyperbola, with $t \geq 0$. The function $A(t)$ that calculates the area of the region A bounded by

- the x-axis between $(1, 0)$ and $(0, 0)$,
- the line segment from $(0, 0)$ to P, and
- the curve $x^2 - y^2 = 1$ between P and $(1, 0)$

is

$$A(t) = \frac{1}{2}\left(\frac{e^t + e^{-t}}{2}\right)\left(\frac{e^t - e^{-t}}{2}\right) - \int_1^{\frac{e^t+e^{-t}}{2}} \sqrt{x^2 - 1}\ dx.$$

This does not look like a very pleasant function to work with! But in fact, it has a very simple re-formulation. Using the fundamental theorem of calculus to differentiate $A(t)$, after a lot of algebra one gets to $A'(t) = 1/2$. Since $A(0) = 0$, it follows that $A(t) = t/2$ for all $t \geq 0$. We conclude that indeed

$$\cosh t = \frac{e^t + e^{-t}}{2}, \qquad \sinh t = \frac{e^t - e^{-t}}{2}, \tag{12}$$

for $t \geq 0$. We extend cosh and sinh to all t by simply taking (12) as the defining relation for all $t \in \mathbb{R}$. This makes $\cosh : \mathbb{R} \to [1, \infty)$ an even function, and $\sinh : \mathbb{R} \to \mathbb{R}$ an odd function.

Here are some basic facts about the hyperbolic functions, including the function tanh defined by $\tanh x = \sinh x / \cosh x$, which can all be verified very easily from (12):

- sinh has domain and range $\mathbb{R}$, and is increasing on its domain. This says that there is a function $\sinh^{-1}$, domain and range $\mathbb{R}$, also increasing, that is the inverse of sinh.

- cosh has domain $\mathbb{R}$ and range $[1, \infty)$. It is not monotone, so not invertible. However, it is increasing on $[0, \infty)$, and on this restricted domain its range is still $[1, \infty)$. This says that there is a function $\cosh^{-1}$, domain $[1, \infty)$ and range $[0, \infty)$, also increasing, that is the inverse of cosh.

- tanh has domain $\mathbb{R}$ and range $(-1, 1)$ (this last is the most non-obvious fact to verify). It is increasing on its domain. This says that there is a function $\tanh^{-1}$, domain $(-1, 1)$ and range $\mathbb{R}$, also increasing, that is the inverse of tanh.

It is worthwhile to look at the graphs of the curves of sinh, cosh, tanh, $\sinh^{-1}$, $\cosh^{-1}$ and $\tanh^{-1}$. The graph of cosh looks like that of a parabola, but it is not (as we will see below, the second derivative of cosh is not zero, but the second derivative of a parabolic function is zero). This graph (cosh) has physical significance — it is the shape formed by a hanging chain, acted on only by the force of gravity.[202]

The hyperbolic functions satisfy many identities that are similar to familiar trigonometric identities. It's easy to verify the following.

- $\cosh^2 - \sinh^2 = 1$.

- $\tanh^2 + 1/\cosh^2 = 1$.

- $\sinh(x + y) = \sinh x \cosh y + \sinh y \cosh x$

- $\cosh(x + y) = \cosh x \cosh y + \sinh y \sinh x$.

- $\sinh' = \cosh$.

- $\cosh' = \sinh$.

- $\tanh' = 1/\cosh^2$.

Just as the inverse trigonometric functions have derivatives that are either rational functions or square roots of rational functions, so too are the derivatives of the inverse hyperbolic functions quite simple. This is one reason why the hyperbolic functions will be important for us: they provide information about the integrals (primitives, antiderivatives) of some very simple functions. Following the approach we took to computing the derivatives of the inverse trigonometric functions, the following are all fairly straightforward to verify:

[202] Google *catenary*. The most famous catenary in the world is upside-down, and is located in St. Louis, Missouri.

- $(\sinh^{-1})'(x) = \frac{1}{\sqrt{x^2+1}}$.
- $(\cosh^{-1})'(x) = \frac{1}{\sqrt{x^2-1}}$, for $x > 1$.
- $(\tanh^{-1})'(x) = \frac{1}{1-x^2}$, for $-1 < x < 1$.

Just as the hyperbolic functions can be explicitly expressed in terms of functions we have defined earlier, so too can the inverse hyperbolic functions. For example, if $y = \cosh x$ (with $x \in [0, \infty)$ and $y \in [1, \infty)$) then $x = \cosh^{-1} y$. So to get an expression for $\cosh^{-1} y$, we can solve

$$y = \frac{e^x + e^{-x}}{2}$$

for x in terms of y. One way to do this is to say that since $y = (e^x + e^{-x})/2$ we have $e^{2x} - 2ye^x + 1 = 0$. This is a quadratic equation in e^x, with solutions

$$y + \sqrt{y^2 - 1} \quad \text{and} y - \sqrt{y^2 - 1}.$$

The expression $y - \sqrt{y^2 - 1}$ is decreasing from 1 on $[1, \infty)$, so taking this solution would give x $(= \log y) \leq 0$. On the other hand the expression $y + \sqrt{y^2 - 1}$ is increasing from 1 on $[1, \infty)$, so this is the right expression to take. We conclude that $x = \log(y + \sqrt{y^2 - 1})$, so that

$$\cosh^{-1}(x) = \log(x + \sqrt{x^2 - 1}).$$

We will use this formula in the next section, when we discuss antiderivatives.

12.6 The length of a curve

How long is a piece of string? If it is stretched straight, we can measure it with a ruler, but if it is curved (and we don't have a possibility to straighten it), it is less clear what to do.

Here's a mathematical formulation of this question:

> Let $f : [a, b] \to \mathbb{R}$ be a bounded function. What is the length, $\ell(f)$, of the graph of f from the points $(a, f(a))$ to $(b, f(b))$?[203][204][205]

[203] This doesn't cover all possible ways in which a piece of string could be curved. For example, it doesn't cover any string that intersects itself (and so can't be modeled as the graph of a *function*). We could get over this by introducing *parameterized curves*, but we won't (yet). The treatment we give here already covers many practically important cases.

[204] On the other hand, it also covers many more possibilities than the arrangement of a piece of string. Presumably any sensible model for the position of a piece of string on the plane would use a *continuous* function; but we are just assuming bounded. As we'll soon see, not a whole lot can be said if we don't add the assumption of continuity.

[205] The notation really should mention a and b, to allow us to talk about lengths of different portions of the graph of the same function. But that would lead to a pretty cumbersome expression (maybe something like $\ell_{[a,b]}(f)$), so we don't bother with the extra information. Usually the interval we are working over is going to be clear from the context.

If f is a linear function $x \mapsto mx + d$ then we can use the Pythagorean formula, since the graph of f is the straight line joining $(a, ma + d)$ to $(b, mb + d)$, to get

$$\ell(f) = \sqrt{(b-a)^2 + (mb - ma)^2} = (b-a)\sqrt{1+m^2}.$$

If f is piecewise linear, we could just compute the lengths of each linear segment, and add them all up. But what if f is nowhere linear?

We have seen how the Darboux integral arises as an answer to a question about understanding area. That suggests an approach to understanding length: let $P = (t_0, t_1, \ldots, t_n)$ be a partition of $[a, b]$ (so $a = t_0 < t_1 < \cdots < t_n = b$). The piecewise linear curve that joins $(t_{i-1}, f(t_{i-1}))$ to $(t_i, f(t_i))$ by a straight line, for each $i = 1, \ldots, n$ is a piecewise linear approximation to the graph of f between $(a, f(a))$ to $(b, f(b))$, and it has length

$$\ell(f, P) = \sum_{i=1}^{n} \sqrt{(t_i - t_{i-1})^2 + (f(t_i) - f(t_{i-1}))^2}.$$

As we add points to P, the above expression gets larger (or at least doesn't get smaller). Indeed, if $P = (t_0, t_1, \ldots, t_k, t_{k+1}, \ldots, t_n)$ and $P' = (t_0, t_1, \ldots, t_k, u, t_{k+1}, \ldots, t_n)$, then where $\ell(f, P)$ has a term that measures the straight-line distance between $(t_k, f(t_k))$ and $(t_{k+1}, f(t_{k+1}))$, $\ell(f, P')$ has *two* terms that measures the straight-line distance between $(t_k, f(t_k))$ and $(t_u, f(t_u))$ *plus* the straight-line distance between $(t_u, f(t_u))$ and $(t_{k+1}, f(t_{k+1}))$. Although we haven't actually proven a "triangle inequality" in two dimensions[206], it is both true and intuitively clear that if A, B, C are three distinct points in the plane, then the straight line distance from A to B is never more than the sum of the straight line distances from A to C and from C to B — in other words, it is never shorter to travel between A and B by going via a third point C. So we get that $\ell(f, P') \geq \ell(f, P)$, and more generally we get that if P, Q are two unrelated partitions of $[a, b]$, then the common refinement partition $P \cup Q$ satisfies that $\ell(f, P \cup Q)$ is at least as large as both $\ell(f, P)$ and $\ell(f, Q)$.[207]

[206] and don't need to, for the purposes of defining $\ell(f)$ — it just helps motivate the eventual definition.

[207] As mentioned in an earlier footnote, we don't actually need the triangle inequality in two dimensions to define length. But it is going to be useful, for some examples. So here's a fairly precise statement, and a proof:

Theorem: Let A, B, C be three points in the plane. Denote by $d(X, Y)$ the straight line distance between points X and Y in the plane. Then

$$d(A, B) \leq d(A, C) + d(C, B),$$

and in fact the inequality is strict ($d(A, B) < d(A, C) + d(C, B)$) unless C lies on the line segment joining A and B (in which case there is equality — $d(A, B) = d(A, C) + d(C, B)$).

Proof: If A and B are the same point, the result is obvious. So we from now on assume that A and B are different points. By translating, rotating and scaling we may put A at $(0, 0)$ and B at $(1, 0)$ (so $d(A, B) = 1$). If C lies along the x-axis (i.e., C is at $(x, 0)$ for some $x \in \mathbb{R}$) then the result is easy — if $x > 1$ then $d(A, C) + d(C, B) = x + (x - 1) > 1 = d(A, B)$; if $x < 0$ then $d(A, C) + d(C, B) = -x + (-x + 1) > 1 = d(A, B)$; while if $0 \leq x \leq 1$ (the only case where C lies on the line segment joining A and B) then $d(A, C) + d(C, B) = x + (1 - x) = 1 = d(A, B)$. So what is left to consider is the case where C is at coordinates

It may be that there is no bound to the possible values that $\ell(f,P)$ can take on as P varies over all partitions[208]. If this happens, then we cannot use this idea of piecewise linear approximations to make sense of the length of the graph. But if there *is* an absolute upper bound on the possible values that $\ell(f,P)$ can take on as P varies over all partitions, then there is a *least* such upper bound, and that number seems like a very good candidate for the length of the curve: it is a number that we can approach arbitrarily closely by a piecewise linear approximation, and there is no larger number with that property.

> **Definition**: Suppose that $f:[a,b]\to\mathbb{R}$ is bounded function. The *length* $\ell(f)$ of the graph of f (from $(a,f(a))$ to $(b,f(b))$) is
>
> $$\begin{aligned}\ell(f) &= \sup\{\ell(f,P) : P \text{ a partition of } [a,b]\}\\ &= \sup\left\{\sum_{i=1}^{n}\sqrt{(t_i-t_{i-1})^2+(f(t_i)-f(t_{i-1}))^2} : P \text{ a partition of } [a,b]\right\},\end{aligned}$$
>
> if this supremum exists. If the supremum does not exist, then the curve graph does not have a length.

Let's check that this definition gives the answer we would expect, for a linear function, say the function $f:[a,b]\to\mathbb{R}$ from earlier (given by $x\mapsto mx+d$). For any partition

---

(x,y) with $y\neq 0$. In this case $d(A,C)=\sqrt{x^2+y^2}$ and $d(C,B)=\sqrt{(x-1)^2+y^2}$, and our goal is to show that

$$\sqrt{x^2+y^2}+\sqrt{(x-1)^2+y^2}>1. \quad (\star)$$

This is equivalent to $\sqrt{(x-1)^2+y^2}\geq 1-\sqrt{x^2+y^2}$. Now we may assume that $\sqrt{x^2+y^2}\leq 1$ (C is on or inside the circle of radius 1 around A), because otherwise $(\star)$ is trivially true. So $\sqrt{(x-1)^2+y^2}\geq 1-\sqrt{x^2+y^2}$ is equivalent to $(\sqrt{(x-1)^2+y^2})^2\geq(1-\sqrt{x^2+y^2})^2$ (both sides are non-negative), which (after squaring out, doing some canceling and rearranging, and dividing both sides by -2) is equivalent to $x<\sqrt{x^2+y^2}$. If $x<0$ this is trivially true (a negative is less than a positive, which $\sqrt{x^2+y^2}$ is, since $y\neq 0$ — notice that $x<0$ puts C outside the circle of radius 1 centered at B, so of course in this case $1=d(A,B)<d(A,C)+d(C,B)$). If $x>0$ it is also true, because (again using $y\neq 0$) $\sqrt{x^2+y^2}>\sqrt{x^2}=x$. So we have established $(\star)$, and finished the proof of the two-dimensional triangle inequality.

[208] It is a good exercise to try to come up with a bounded continuous function on domain $[0,1]$ for which $\ell(f,P)$ can take on arbitrarily large values as P varies over all partitions of $[0,1]$. If you think about it in terms of how the function is built, and don't get overly hung up on giving an exact formula for $f(x)$ at each x, it should be quite easy. Hint: there's a piecewise linear example.

$P = (t_0, \ldots, t_n)$ of $[a, b]$ we have

$$\begin{aligned}
\ell(f, P) &= \sum_{i=1}^{n} \sqrt{(t_i - t_{i-1})^2 + ((mt_i + d) - (mt_{i-1} + d))^2} \\
&= \sum_{i=1}^{n} \sqrt{(t_i - t_{i-1})^2 + m^2(t_i - t_{i-1})^2} \\
&= \sum_{i=1}^{n} (t_i - t_{i-1})\sqrt{1 + m^2} \\
&= \sqrt{1 + m^2} \sum_{i=1}^{n} (t_i - t_{i-1}) \\
&= (b - a)\sqrt{1 + m^2} \quad \text{(via a telescoping sum)}.
\end{aligned}$$

This last expression is a constant (not depending on P), so

$$\ell(f) = \sup\{(b - a)\sqrt{1 + m^2} : P \text{ a partition}\} = (b - a)\sqrt{1 + m^2},$$

exactly as we calculated previously using the Pythagorean formula.

On the other hand, suppose that $g : [a, b] \to \mathbb{R}$ is a bounded function satisfying $g(a) = ma + d$ and $g(b) = mb + d$ (i.e., whose graph starts and finishes at the same points as the linear function f above), but which is *not* the linear function f. Since g is different from f (but agrees with f at a and b) there must be a $c \in (a, b)$ with $f(c) \neq g(c)$. Consider the partition $P = (a, c, b)$. It is intuitively clear (and can be made precise — triangle inequality in two dimensions, again) that $\ell(g, P)$ is strictly greater than $\ell(f)$ — indeed, $\ell(f)$ is the straight-line distance between $(a, ma + d)$ and $(b, mb + d)$, while $\ell(g, P)$ is the sum of the straight line distance between $(a, ma + d)$ and $(c, g(c))$ and the straight line distance between $(c, g(c))$ and $(b, mb + d)$, where $(c, g(c))$ is a point *not* on the straight line between $(a, ma + d)$ and $(b, mb + d)$.[209] Since $\ell(g, P) > \ell(f)$ it follows immediately that $\ell(g) > \ell(f)$ (if $\ell(g)$ exists).

What we have just established is the famous

> **Dictum**: "The shortest distance between two points is a straight line". That is, among all graphs of bounded functions on domain $[a, b]$ that start at $(a, ma + d)$ and end at $(b, mb + d)$, and that have a well defined length, the unique function with the shortest length is the linear function $x \mapsto mx + d$.

There is a close connection between length and the integral, and between the length of a piecewise linear approximation of a graph coming from a partition P, and Darboux sums, coming also from P, of a certain function. We explore that connection now.

The expression $\sum_{i=1}^{n} \sqrt{(t_i - t_{i-1})^2 + (f(t_i) - f(t_{i-1}))^2}$ (i.e., $\ell(f, P)$) doesn't much look like a Darboux sum, but it can be made to look more like one by pulling out a factor of

[209] If you read the earlier long footnote on the two dimensional triangle inequality, you will know that this intuition can be made precise.

$t_i - t_{i-1}$ (a.k.a. Δ_i) from each summand:

$$\ell(f,P) = \sum_{i=1}^{n}(t_i - t_{i-1})\sqrt{1+\left(\frac{f(t_i)-f(t_{i-1})}{t_i - t_{i-1}}\right)^2} = \sum_{i=1}^{n}\Delta_i\sqrt{1+\left(\frac{f(t_i)-f(t_{i-1})}{t_i - t_{i-1}}\right)^2}.$$

Now it looks more like a Darboux sum. It's not quite one yet, because the expression

$$\sqrt{1+\left(\frac{f(t_i)-f(t_{i-1})}{t_i - t_{i-1}}\right)^2}$$

isn't obviously either the infimum or the supremum of a function. But, the above expression does strongly suggest thinking about the function $\sqrt{1+(f')^2}$ (if this exists); after all, for $[t_{i-1}, t_i]$ short,

$$\frac{f(t_i)-f(t_{i-1})}{t_i - t_{i-1}} \approx f'(t_i), f'(t_{i-1}).$$

So: let's assume that f is continuous and differentiable on $[a,b]$ and also that f' is bounded on $[a,b]$. Applying the mean value theorem to the interval $[t_{i-1}, t_i]$, we find that there is $c \in (t_{i-1}, t_i)$ with

$$f'(c) = \frac{f(t_i)-f(t_{i-1})}{t_i - t_{i-1}}$$

so

$$\sqrt{1+(f'(c))^2} = \sqrt{1+\left(\frac{f(t_i)-f(t_{i-1})}{t_i - t_{i-1}}\right)^2}.$$

Since

$$\inf\left\{\sqrt{1+(f'(x))^2} : x \in [t_{i-1}, t_i]\right\} \le \sqrt{1+(f'(c))^2} \le \sup\left\{\sqrt{1+(f'(x))^2} : x \in [t_{i-1}, t_i]\right\}$$

we get (summing over i)

$$\begin{aligned}
\sum_{i=1}^{n}\Delta_i \inf\left\{\sqrt{1+(f'(x))^2} : x \in [t_{i-1}, t_i]\right\} &\le \sum_{i=1}^{n}\Delta_i\sqrt{1+\left(\frac{f(t_i)-f(t_{i-1})}{t_i - t_{i-1}}\right)^2} \\
&= \ell(f,P) \\
&\le \sum_{i=1}^{n}\Delta_i \sup\left\{\sqrt{1+(f'(x))^2} : x \in [t_{i-1}, t_i]\right\}.
\end{aligned}$$

The first and last expressions above are *exactly* Darboux sums — lower and upper Darboux sums, respectively, for the (bounded) function $\sqrt{1+(f')^2}$ with respect to the partition P. In other words: for any partition P of $[a,b]$

$$L(\sqrt{1+(f')^2}, P) \le \ell(f,P) \le U(\sqrt{1+(f')^2}, P). \quad (\star)$$

From $(\star)$, in particular from the second inequality therein, we can read off a very useful fact:

Proposition: If bounded $f : [a,b] \to \mathbb{R}$ is continuous and differentiable, and if also f' is bounded, then $\ell(f)$ exists.

Here is the proof: let P and Q be any two (unrelated) partitions of $[a,b]$. We have

$$\begin{aligned} \ell(f,P) &\leq \ell(f, P \cup Q) \quad \text{(by the earlier triangle inequality observation)} \\ &\leq U(\sqrt{1+(f')^2}, P \cup Q) \quad \text{(by the second inequality in } (\star)) \\ &\leq U(\sqrt{1+(f')^2}, Q) \quad \text{(by one of our earliest observations about Darboux sums).} \end{aligned}$$

So em every upper Darboux sum of $\sqrt{1+(f')^2}$ is at least as large as *every* piecewise linear approximation to the length of the graph of f. It follows that $\sup\{\ell(f,P) : P \text{ a partition}\}$ exists, as claimed (any upper Darboux sum for $\sqrt{1+(f')^2}$ provides an upper bound for the $\ell(f,P)$'s).

The argument above gives more, since it says that

$$\sup\{\ell(f,P) : P \text{ a partition}\} \leq \inf\{U(\sqrt{1+(f')^2}, P) : P \text{ a partition}\}. \quad (\star\star)$$

But also, from the first inequality of $(\star)$ (that $L(\sqrt{1+(f')^2}, P) \leq \ell(f,P)$), and the new-found knowledge that $\sup\{\ell(f,P) : P \text{ a partition}\}$ exists, we immediately get[210]

$$\sup\{L(\sqrt{1+(f')^2}, P) : P \text{ a partition}\} \leq \sup\{\ell(f,P) : P \text{ a partition}\}. \quad (\star\star\star)$$

Combining $(\star\star)$ and $(\star\star\star)$ we get upper and lower bounds on $\ell(f)$:

$$L(\sqrt{1+(f')^2}) \leq \ell(f) \leq U(\sqrt{1+(f')^2}).$$

If we know that $\sqrt{1+(f')^2}$ is not just bounded on $[a,b]$, but is also integrable, then the right and left sides of the above string of inequalities are *equal*, and we get the following theorem (the main point of this section):

Theorem: Suppose that bounded $f : [a,b] \to \mathbb{R}$ is continuous and differentiable, with bounded derivative, and that also $\sqrt{1+(f')^2}$ is integrable on $[a,b]$. Then the length of the graph of f (from $(a, f(a))$ to $(b, f(b))$) exists and is

$$\ell(f) = \int_a^b \sqrt{1+(f')^2}.$$

In particular this formula is valid is f is *continuously differentiable* — continuous and differentiable, with continuous derivative.

[210]This is one of those things that you either see instantly, or don't. If you don't, that's fine, because you can *prove* it! What you have to proof is this: if f, g are two functions on domain A, and $f(x) \leq g(x)$ for all $x \in A$, and $\sup\{g(x) : x \in A\}$ exists, then so also does $\sup\{f(x) : x \in A\}$, and moreover $\sup\{f(x) : x \in A\} \leq \sup\{g(x) : x \in A\}$.

It's hard right now to give many applications of this formula for the length of a curve, because for most functions f (even quite reasonable ones) the function $\sqrt{1+(f')^2}$ is hard to integrate (antidifferentiate). Here is one very relevant example, though. We've defined the number π by saying that $\pi/2$ is the area of half of a unit circle. It is much more usual to see π defined by the relation that 2π is the circumference of a unit circle. It is natural to ask

> do these two definitions actually lead to the *same* π?

Another way to put this question is:

> if we *define* the area of the unit circle $x^2+y^2=1$ to be π, then can we *prove* that circumference of the circle is 2π?

Because we have developed a theory of lengths of graphs, we are now in a position to answer this question. The circumference of the circle $x^2+y^2=1$ is twice the length of the graph of the function $f:[-1,1]\to\mathbb{R}$ given by $f(x)=\sqrt{1-x^2}$. This is a bounded, continuous function, with continuous derivative

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

It follows (after some algebra) that

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

So, by the formula we have just developed, the circumference of the circle $x^2+y^2=1$ is[211]

$$2\int_{-1}^{1}\frac{dx}{\sqrt{1-x^2}}.$$

But we know that the derivative of $\sin^{-1}(x)$ is $1/\sqrt{1-x^2}$. So

$$2\int_{-1}^{1}\frac{dx}{\sqrt{1-x^2}}=2\left[\sin^{-1}(x)\right]_{x=-1}^{1}=2\left(\frac{\pi}{2}-\frac{-\pi}{2}\right)=2\pi.$$

So (thankfully) our definition of π is consistent with (for example) Archimedes' definition.

[211] Not quite. The function $\sqrt{1+(f'(x))^2}$ is not bounded on $[-1,1]$, because of that pesky $1-x^2$ in the denominator. It is, however, bounded on $[-1+\varepsilon_1,1-\varepsilon_2]$ for every $\varepsilon_1,\varepsilon_2>0$. So we can treat the integral we have to compute as an improper integral. Essentially we are saying that the length of a proportion α of the circle approaches π as α approaches one half (from below).