PROOFS AND OTHER IDEAS OF MATHEMATICS

MATH 152, SECTION 55 (VIPUL NAIK)

1. General ideas related to proofs

1.1. **The idea of models and proof.** The main idea behind the concept of proof is to establish something clearly, give a fool-proof, error-free explanation of *why* something is true. That means that all cases must be covered, every step of the argument should be justified, and there shouldn't be any hidden assumptions that aren't true. We'll go over some of these in detail, but the first question you might be wondering is: *why proof?*

So the one thing you have to remember about mathematics is that the world of mathematics is a world of its own creation. It's true that mathematical ideas and formalisms are applied a lot to real-world settings, but often the mathematical ideas go far beyond what we can determine or verify through real-world observation. In other words, for a lot of the things we want to do in the mathematical world, it is hard to be sure of them simply by looking around at the real world. So, in mathematics, it is important to develop a way of being sure of things that depends purely on internal reasoning. In fact, this concept of internal reasoning, or *reasoning within the framework of rules of the system*, is what defines mathematics.

If you think about mathematics as a system of rules-based internal reasoning, your most significant introduction to mathematics isn't when you learn numbers, it is when you learn how to play games, and how to deploy strategies within games.

And it so happens that most systems of rules that we deal with involve the concepts of numbers, many of them involve some geometric ideas, and some of them need the tools of algebra and trigonometry and calculus. But you shouldn't think of mathematics as being the same as algebra or trigonometry or calculus. These are just tools. The main feature of mathematics is that in mathematics, there is a strong place for *internal reasoning* – reasoning within the system to try to determine what is true and what is false.

And, if you look at major developments in a number of academic disciplines, you see that mathematics is seeping into most of them. And I don't just mean that they are getting more quantitative, though that's part of the story. Those who've seen Newton's laws and classical mechanics know that there's a specific model (which, it turns out, isn't exactly how the real world operates) and we can predict how things will behave in that model. Nowadays, many papers in the social sciences also start by creating some artificial model that has a reasonable resemblance to reality, and then try to derive formally what happens in that model. And the main difference between mathematicians and those in other sciences is that for people in the other sciences, they need to justify that their model has some kind of resemblance with, or explanatory power about, the real world. But mathematicians aren't subject to that constraint.

So think of mathematical rigor as something that allows mathematicians to explore things where intuition, or real-world checks and balances, are hard to find.

1.2. **Proof by example isn't; cover your bases, consider all cases.** So one of the things people often do in the real world is when they want to know if something is true they take some example and check it. And you see the media and politicians do that kind of thing everyday. So whenever somebody wants to prove that some thing works, they'll find one person to give a testimonial for it.

But in mathematics, we don't consider a few isolated examples to be proof. And the reason is simple: different cases behave differently, so the examples we choose are probably not representative. That's true in the real world, but it is often even more true in mathematics, where things aren't constrained to be realistic.

So mathematicians try to *cover all cases* in proofs. What does this mean? If you want to prove a statement for all real numbers, it isn't enough to prove it for all rational numbers. After all, there are real numbers that aren't rational. So you need to prove the statement for all rational numbers *and* all irrational numbers.

Now, *how* we choose to break down the problem into cases is up to us. For some problems, the natural way of breaking up the problem may be to first consider rational numbers and then consider irrational numbers. Sometimes, it may be helpful to first consider positive numbers and then consider negative numbers. If you have to prove a statement for all numbers in a finite set, the ultimate break-up would be to check it separately for every element of the finite set.

The thing you should remember is that *if you are breaking things up into cases*, you should *remember to cover all cases*. And one way to remember this is to think of mathematics as just about the smartest adversary you can find in the battlefield. If you don't cover every possible line of attack on your adversary, your adversary will hide in the one place you forgot to cover.

1.3. **Conditional implication.** In mathematics, we often consider statements of the form:

"If A, then B"

Now, these kinds of statements can sometimes be confusing, so let's try to understand what exactly this means. This roughly means that, assuming that A is given to be true, B is true. For instance, "if I don't oversleep, I will attend the calculus lecture on Friday". That is a conditional statement.

There are a lot of subtleties about conditional implications that we need to understand. The first is that "If A, then B" only means that A is *sufficient* for B. It doesn't mean that A is necessary for B. There may be other ways that B could become true, even if A were false. For instance, you may say "If I have enough money, I'll eat lunch". But you may be able to eat lunch even though you don't have enough money – by going to one of U of C's Free Food events.

So "if A, then B" means that if, somehow, one could guarantee A to be true, B would follow – but there may be other ways to guarantee B. In particular, if you prove a statement "if A then B" and then you separately prove that A is true, then you would have proved that B is true.

1.4. **Rough work and fair work.** In many situations where we need to do a proof, there are two parts to doing the proof. The first is the exploratory phase, or the discovery phase, where we need to find a strategy that works for the proof. For instance, in the case of $\epsilon - \delta$ proofs, the exploratory phase involved coming up with a winning strategy for the prover or the skeptic as the case may be. In this exploratory phase, we may do some rough calculations, make some wild guesses, check out our intuition on examples, etc. The exploratory phase may involve *working backwards*, *splitting into cases*, etc. At the end of exploratory phase, we have an overall proof strategy.

At the end of the exploratory phase, we hopefully have a clear proof strategy. The next phase is that of clearly expressing the strategy and showing that it works. When writing this final stategy and the proof, you do not need to cover everything you went through in the exploratory phase. Stick only to that which is more relevant to the final proof strategy. Also, state the strategy right upfront and proceed, to the extent possible, starting from what you know and proceeding towards what you need to show.

1.5. **Opposite statement.** Another concept that I should mention, and that you've had a bit of past experience with, is the *opposite* of a statement. This is related to the question: *how do I prove that A is not true*? Well, in order to prove that, you first need a clear formulation of what it means for A to *not* be true. This new statement is sometimes called the *negation* or *opposite* of A.

Now, some of you may have seen some Boolean algebra or logic, so you might have some idea of the formal process of negating a statement, but even if you haven't, most of the rules are intuitive provided you pause to think and don't just try to rush. Keep your cool, and it's not hard. I'll just mention some important ideas:

(1) Negation turns *and* to *or*, and *or* to *and*. For instance, the negation of the statement $x = 1$ or $x = 2$ is the statement $x \neq 1$ *and* $x \neq 2$.
(2) Negation on a $\forall$ quantifier gives a $\exists$ quantifier and negation on a $\exists$ quantifier gives a $\forall$ quantifier. For instance, the negation of the statement $\forall x \in \mathbb{R}, f(x^2) = f(x)^2$ is the statement $\exists x \in \mathbb{R}, f(x^2) \neq f(x)$. This came up when we looked at $\epsilon - \delta$ proofs.

1.6. **Proof by contradiction.** One of the useful proof techniques is proof by contradiction. This comes up sometimes, and I'll talk more about it when it does, but the way it works is like this: suppose you are trying to prove A. So the first thing you may try to do is prove A straightforward, but that may seem tricky. So what you do is this. You assume that the opposite of A is true. So you write down the opposite of A, and start with that as given. And then, from that, you derive some statement that is plainly *not* true. Since the conclusion isn't true, the statement you started by assuming, namely, the opposite of A, couldn't have been true either. And since the opposite of A is false, A itself must be true.

Some of you may have seen the proof that $\sqrt{2}$ is irrational. That proof is a classic example of proof by contradiction.

2. Specific issues

The material in the previous section is very general and I think most of you would lap it up pretty easily. Most of you seem to have a reasonable understanding of these ideas, but there are some more specific issues that you may have with expressing your proofs. Below are listed some of the specific issues that students in past years have had in the first two advanced homeworks.

2.1. **Making your strategy and specific claims clear upfront.**

This issue has occurred in the past with some of the $\epsilon - \delta$ proofs. If you're the prover, then the stategy involves finding an expression for δ that works in terms of ϵ. If you're the skeptic, the strategy involves finding an ϵ for which no δ works, and then being able to choose a value of x in $(c - \delta, c + \delta) \setminus \{c\}$.

In the exploratory phase, you try to figure out a strategy that works. Then, in the actual proof phase, you show that the strategy works.

When writing up the final proof, please do not show the exploratory phase. Please write the final winning strategy upfront. Then, proceed to translate the general statement about the existence or non-existence of limit into a specific claim based on your strategy. Then, do some algebraic manipulation or case-by-case reasoning to prove that your strategy works.

Some examples:

- For the homework problem $\lim_{x \to 2} x^2 = 4$, state right at the beginning that the winning strategy is $\delta = \min\{1, \epsilon/5\}$. Then, state the specific claim: if $0 < |x - 2| < \min\{1, \epsilon/5\}$, then $|x^2 - 4| < \epsilon$. Now, prove the specific claim.

  Some people do some algebraic manipulation to discover the δ that works. Others are comfortable using the general formula that works for the quadratic. Whichever thing you choose to do, please remember that the less of this exploratory work you show, the clearer your proof is. This is because exploratory work, as a general rule, is mesy, with conditionals much more complicated, steps going forward and backward, etc. So please skip this and write your winning strategy clearly.
- Consider problems where, for instance, we need to select a δ value for a given ϵ value, and the function is defined differently on rationals and irrationals. Here, we need to find a δ_1 that works for rationals, a δ_2 that works for irrationals, and then take $\delta = \min\{\delta_1, \delta_2\}$.

  You should write down the strategy for choosing δ right on top, *make the specific claim*, and split into cases *to prove the specific claim*.

  Some of you split into cases first, proved things in each case, and gave the overall winning strategy at the end. *This is probably the way that you discover things in the exploratory phase, but it's not the prettiest way of presenting a final proof.*

Caveat: There are situations where it is advantageous to show your exploratory phase. For instance, if you were a teacher and were guiding students through a learning process, this exploratory phase might be helpful. If you were trying to break ground with a similar new problem, it might help to revisit the exploratory phase.

However, you should think of showing your exploratory phase as filming the process of the manufacture of sausage, and the fair work proof phase as the phase of enjoying the final sausage.

2.2. **Doing the general case clearly.**

This problem arose with some advanced homework solutions in Homework 1, and a subsequent clarification was made. However, it's worth reiterating here.

In the exploratory phase, we may use some specific numerical examples to check if something is true. Then, we discover that the actual steps work in somewhat greater generality, and we need that greater generality in order to do the whole proof.

When writing down the final proof, jump directly to proving that the actual steps work in somewhat greater generality.

The example from the first advanced homework was: "Show that the function $f(x) := \begin{cases} 1, & x \text{ rational} \\ 0, & x \text{ irrational} \end{cases}$ is periodic but has no period."

One possible discovery approach is as follows:

(1) We notice that the number 1 works in the sense that $f(x+1) = f(x)$ for all $x \in \mathbb{R}$. We prove this by splitting into the cases where x is rational and x is irrational.
(2) After finishing that proof, we notice that, in fact, the proof depended only on the fact that rational + rational = rational and irrational + rational = irrational. Crucially, the only thing we were using about 1 was that it is rational.
(3) We thus conclude that any rational $h > 0$ works in place of 1.
(4) Since there are arbitrarily small positive rational numbers, we concluded that there is no period.

In the final write-up of the proof, we remove steps (1) and (2) and directly proceed with the claim of step (3), with the proof of that claim basically mimicking our original proof of (1).

2.3. **Meta-strategies.** Some of the advanced problems involve constructing a strategy for one game using strategies for other games *as black boxes*. For instance, in problems 1 and 5 of advanced homework 2, you are asked to come up with winning strategies for the prover for $|f|$, $\max\{f, g\}$, and $\min\{f, g\}$, assuming that there exist winning strategies for f and g.

Here, you assume that the winning strategies for f and g are given to you on a platter, but you have to treat them as black boxes. In other words, you assume some statement of the form:

"For every $\epsilon > 0$, there exists a $\delta_1 > 0$ such that if $0 < |x - c| < \delta_1$, then $|f(x) - L| < \epsilon$."

and:

"For every $\epsilon > 0$, there exists a $\delta_2 > 0$ such that if $0 < |x - c| < \delta_2$, then $|g(x) - L| < \epsilon$."

Our "winning strategy" for $H := \max\{f, g\}$, is to choose, for a given $\epsilon > 0$, $\delta = \min\{\delta_1, \delta_2\}$, i.e., the minimum of the δs that work for f and g.

We then make the specific claim: "If $0 < |x - c| < \min\{\delta_1, \delta_2\}$, then $|H(x) - L| < \epsilon$."

After this, we prove the specific claim by splitting into cases for x, based on whether $H(x) = f(x)$ or $H(x) = g(x)$.

Meta-strategies are tricky to understand at first, because the strategies that we are using as black boxes are *unknown knowns* – we can use them, but have to treat them as black boxes.

2.4. **Fixed but arbitrary.** Another note about the $\epsilon-\delta$ proofs. In all these proofs, ϵ is "fixed but arbitrary." What this basically means is that ϵ is fixed, but it is fixed by the skeptic, so we (as the provers) have no control over the choice so we should be prepared for the worst.

INEQUALITY SOLVING

MATH 152, SECTION 55 (VIPUL NAIK)

Difficulty level: Moderate for people who have been inequalities before, high for people who haven't.

Corresponding material in the book: Section 1.3. However, this section does *not* cover inequalities involving rational functions, and only does the polynomial case. The rational functions case is only a mild generalization of the polynomial case, *but it is essential that you understand the rational functions case from the notes.*

Things that students should definitely get: Set notation for writing intervals, how to manipulate and solve inequalities, inequalities for rational functions, inequalities for absolute value.

Executive summary

Words...

(1) When solving inequalities, we can add the same thing to both sides. We can subtract the same thing from both sides. We can add two inequalities with the same direction of inequality.
(2) We can multiply a positive number to both sides and preserve the direction of inequality. If we multiply by a negative number, we reverse the direction of inequality. If we multiply by a number that we know is nonnegative (But we're not sure if it is positive or zero) then the direction of inequality is preserved but an equality case gets introduced. e.g., if $a > b$ and $x \geq 0$, we get $ax \geq ab$.

Actions (think back to examples where you've dealt with these issues)...

(1) If $|x|$ takes values in a certain set A, then the set of possible values for x is the union of A and the negatives of numbers in A.
(2) Generally, when solving inequalities involving absolute value, consider the case that the thing inside the absolute value is negative and the case that it is nonnegative. Solve both cases and then take the union of the solutions.
(3) When solving inequalities involving rational functions, bring everything to one side. So it reduces to trying to find when a given rational function is positive, negative, zero, and undefined.
(4) A rational function is undefined wherever the denominator is 0. It is 0 where the numerator is 0 but the denominator isn't. For positive and negative, start from the far positive end and remember that every time you cross over a linear factor of the numerator or the denominator, the sign changes...
(5) ... except that the sign changes only if the total multiplicity of that factor is odd. If the total multiplicity is even, then the sign doesn't change. (What do I mean? Review the examples we did and try figuring out)...
(6) ... and what if there are quadratic factors in the numerator or denominator, such as $x^2 + 1$, that do not factorize further? These are anyway always positive or always negative, so they don't affect the sign of things ...
(7) ... and what if there are quadratic factors that you don't know how to factorize? Well, use the quadratic formula.

1. Equality versus inequality

In this lecture, we cover inequalities. This is almost like solving equations, except that instead of an equality sign, there is an inequality sign. The main difference is that because we have an inequality sign, the solution sets are usually much bigger – they're not just a few isolated points, they are unions of intervals. Further, the solution to the corresponding *equation* is usually on the *boundary* of the solution set to the inequality.

2. Review of notation

2.1. Review of set notation. So I hope that you've either seen the set notation before, either before coming here or in the last two days. But I'll review some of the basics anyway, very quickly. As you may have heard, a set is a *well-defined* collection of objects, called its *members* or *elements*. If A is a set and x is an object, we say $x \in A$ if x is a member of A, and $x \notin A$ otherwise. We have the symbols $\cup$ for union of sets, $\cap$ for intersection of sets. We have the symbol $\emptyset$ for empty set, symbols $\subset$ and $\subseteq$ for strict and not necessarily strict containment. For reverse containment, we have $\supset$ and $\supseteq$. And, as we saw last time, $\setminus$ represents the set difference.

Now, let's discuss two ways of writing sets. One way is what we might call the *laundry list* or *naive* way: just list everything. I think a more professional word for this is the *roster method*. So, for instance, the set of all one-digit positive integers is $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. That's a complete list of all of them.

The other approach, which is the *set builder* method or constructive method, specifies a qualification. For instance, if $\mathbb{R}$ denotes the set of real numbers (and it always does in this course), then the set:

$$\{x \in \mathbb{R} \mid x > 0 \text{ and } x^2 + x + 1 > 2^x\}$$

is basically a set given by some condition. To determine whether an x is in the set, we check that it satisfies the condition. Is 1 in the set, for instance? Well, $1 > 0$, and $1^2 + 1 + 1 = 3$, which is greater than $2^1 = 2$. Good, so it is in. What about 0.5? Well, $0.5^2 + 0.5 + 1 = 1.75$, which is greater than $2^{0.5} = \sqrt{2}$.

Now the book does not use the $\mid$ separator, it uses the : separator, and that's fine too. So in the book's notation, the above becomes:

$$\{x \in \mathbb{R} : x > 0 \text{ and } x^2 + x + 1 > 2^x\}$$

2.2. Review of interval notation. Certain subsets of $\mathbb{R}$ that come up pretty frequently are the *intervals*, and there is special notation for these. Let's discuss both the notation and the way these intervals are shown on the number line.

The interval (a, b) is the set $\{x \in \mathbb{R} : a < x < b\}$. In other words, it is the set of (real) numbers *strictly between* a and b. It is represented on the number line by creating unfilled circles at a and b and shading or darkening the region in between. This interval is termed the *open interval* between a and b. By the way, one of your homework problems asks you to think of this open interval in another way – in terms of a *center* and a *radius*. That's a very important homework problem not because it is particularly difficult, but because it is critical to many of the $\epsilon - \delta$ definitions of limits.

The interval (a, b) is also denoted $]a, b[$. The latter notation has both advantages and disadvantages. The primary advantage is that while (a, b) may be confused with an *ordered pair* representing a *point in the coordinate plane* with coordinates a and b, the notation $]a, b[$ has no alternative interpretations. However, for these notes and the rest of this course, we'll use the (a, b)-notation. This is used in the book and is also standard in most mathematics courses you will see.

Note, by the way, that the interval (a, b) is empty, and shouldn't be talked about, if $a \geq b$.

The interval $[a, b]$ is the set $\{x \in \mathbb{R} : a \leq x \leq b\}$. This is called the *closed interval* between a and b, and we use filled circles at a and b instead of the unfilled circles used earlier. Okay, here's a question: what is $[a, b] \setminus (a, b)$? In other words, what happens when you remove the open interval from the closed interval? What's the difference? The answer is: the two points a and b. So $[a, b] \setminus (a, b) = \{a, b\}$, and $(a, b) \cup \{a, b\} = [a, b]$.

There are also notions of half-open, half-closed intervals. So what does it mean for an interval to be left-open and right-closed? Well, that's an interval of the form $(a, b] = \{x \in \mathbb{R} : a < x \leq b\}$. And it's represented by an unfilled circle at a and a filled circle at b. In the other notation, it would be $]a, b]$. Similarly, $[a, b) = \{x \in \mathbb{R} : a \leq x < b\}$ is represented by a filled circle at a and an unfilled circle at b. In the other notation, it would be $[a, b[$.

So I think you're getting the general philosophy. The round parentheses () represent openness, or the endpoint excluded, while the square braces [] represent closedness, or the endpoint included. Pictorially, an excluded point is an unfilled circle, and an included endpoint is a filled circle.

To complete the discussion, we need to talk about ∞. Now, ∞ is a big and mind-boggling concept and we're not really going to discuss it here. For our purposes, ∞ (positive infinity) is a placeholder for *no upper limit* and $-\infty$ is a placeholder for *no lower limit*.

So the interval (a, ∞) is the set $\{x \in \mathbb{R} : a < x\}$. and $[a, \infty)$ is the set $\{x \in \mathbb{R} : a \leq x\}$. And similarly, $(-\infty, a)$ is the set $\{x \in \mathbb{R} : x < a\}$ and $(-\infty, a]$ is the set $\{x \in \mathbb{R} : x \leq a\}$. And on the number line, you just use an arrow to indicate that it'll go on forever.

Notice that the parentheses around ∞ are always the round ones, meaning that ∞ is never included. And that makes sense because ∞ isn't real. Nor is $-\infty$. Whether these things exist and what they mean is beyond the scope of our discussion. I just want you to think of them as placeholders.

2.3. **As a union of intervals.** Okay, here's a quick test. Express $(-1, 1) \setminus \{0\}$ as a union of intervals? What is it? [Draw diagram]. You see the picture. It is $(-1, 0) \cup (0, 1)$.

Or, what about $(-2, 2) \setminus (-1, 1)$? [Draw diagram]. That's $(-2, -1] \cup [1, 2)$. Note the way the circles fill and unfill.

3. Inequality solving

3.1. **Some basic rules.** Let's discuss some very basic rules of inequality-solving. The way you solve inequalities is very similar to the way you solve equalities, except this: when you multiply both sides by a negative number, you change the direction of the inequality. And the one thing you shouldn't do is multiply both sides by zero.

Okay, so consider the inequality:

$$x + 4 \leq 5(x - 1)$$

Moving all stuff to one side gives:

$$-4x + 9 \leq 0$$

Now, you multiply both sides by -1, so change the sign of the inequality, and get:

$$4x - 9 \geq 0$$

And then simplify to $x \geq 9/4$. Which is the interval $[9/4, \infty)$ on the number line.

Inequality-solving for polynomial functions. Let's now consider some polynomial functions. The goal is to consider a polynomial function f, and try to determine where it is zero, where it is positive, and where it is negative. For now, we'll focus on polynomial functions that split completely into linear factors.

For instance, consider the polynomial function:

$$f(x) = x(x + 1)(x - 1)$$

This polynomial is a product of three linear factors. How do we figure out the sign of this polynomial function at a point? Its sign is the product of the signs of the three factors. Now, if you have a linear factor $x - a$, where is it positive, where is it zero, and where is it negative? Answer: it is positive for $x > a$, zero for $x = a$, and negative for $x < a$.

So, we see that each linear factor switches sign at the corresponding zero (i.e., the linear factor $x - a$ switches sign at a). If only one linear factor switches sign, and the signs of the other factors remain the same, then the product also switches sign. So, for the above polynomial f, the sign changes could potentially occur at $-1, 0, 1$. Let's view this on a number line.

To the right of 1, x is pretty large, so all the three factors are positive. A product of three positives is a positive, so for $x > 1$, $f(x)$ is positive. At 1, the function becomes zero. Immediately to the left of 1, the factor $x - 1$ becomes negative, but the other two factors are still positive. So, the overall product is negative. And so the story goes till we hit $x = 0$. At $x = 0$, the function again takes the value 0. Then, to the immediate left of 0, both x and $x - 1$ are negative, but $x + 1$ is positive. So, the function is positive, and remains so till we reach $x = -1$. There it becomes 0, and to the left of -1, all factors are negative, hence so is the product.

The upshot: $f(x) = 0$ for $x \in \{-1, 0, 1\}$, $f(x) > 0$ for $x \in (-1, 0) \cup (1, \infty)$, and $f(x) < 0$ for $x \in (-\infty, -1) \cup (0, 1)$.

Let's consider another function:

$$g(x) := x^2(x - 1)$$

The points where interesting things could happen are, in this case, 0 and 1. But 0 is *doubly* interesting, because the linear factor x has multiplicity 2.

So, to the right of 1, all factors are positive, so $g(x) > 0$ for $x > 1$. At 1, g takes the value 0. To the immediate left of 1, $x - 1$ becomes negative but the remaining factors are still positive. So the function becomes negative. Then, at 0, it becomes 0. What happens to the left of 0? Does the function switch sign again? No! And that's because although the x switches sign to negative, there are two of them, so their effects cancel each other. So the overall effect is no sign change, and the function remains negative.

So $g(x)$ is positive for $x \in (1, \infty)$, it is negative for $x \in (-\infty, 0) \cup (0, 1)$, and it is zero for $x \in \{0, 1\}$.

3.2. **Inequality solving for rational functions.** *This is not described explicitly in the book, but is important for some later material, so please go through it carefully. We'll try to briefly go over it in lecture.*

Let's now consider the rational function $P(x)/Q(x)$, where P and Q are polynomials. We want to figure out where this function is positive, zero, and negative.

Now the first thing you should remember about rational functions is that they aren't always globally defined. The points where they aren't defined are the points where the denominator implodes and the expression explodes: the roots of Q. So at these points, there is no inequality satisfied: the function just isn't defined.

The first thing we do is factor both P and Q. So we obtain:

$$\frac{a(x - \alpha_1)(x - \alpha_2) \dots (x - \alpha_m)}{b(x - \beta_1)(x - \beta_2) \dots (x - \beta_n)}$$

So, we already figured out that the values $\beta_1, \beta_2, \dots \beta_n$ are points where this function isn't defined. And by the way, do not cancel before figuring out these points, because we have to take the function and treat it *as is* (remember FORGET?). But once we've excluded all these points, then away from these points, we can cancel, so let's assume from now on that there is no common factor between the numerator and the denominator.

Also, a/b has some sign (positive or negative) so let's ignore that too, because that'll just flip signs for everything. So we're really looking at:

$$\frac{(x - \alpha_1)(x - \alpha_2) \dots (x - \alpha_m)}{(x - \beta_1)(x - \beta_2) \dots (x - \beta_n)}$$

The first thing to observe is that the overall sign is the product of signs of each of the pieces. And the sign of $x - t$ is positive if x is greater than t and negative is x is less than t. So each time x crosses one of the α_is or β_js, one expression flips sign. Thus, the only points where the expression changes sign, or crosses zero, are the α_is and β_js [See caveat to this in a later example].

Okay, now if your head is already swimming with this, it's time to take a simple example. Now if I were trying to impress you with my pedagogy, I would probably start with the simplest example, then gradually build up to more and more complex examples. And that's a very valuable approach that I'll use most of the time. But, often, I'll start by just trying to attack the general problem, make a bit of headway, get stuck, and *then* proceed to a simple example. It's important that you develop a tolerance for both styles because you'll need both styles. There are times when building up gradually from one example to another isn't a luxury you can afford, and there are times where it's the only thing that makes sense.

So let's consider:

$$\frac{x(x + 1)(x + 2)}{x(x - 1)(x - 2)}$$

So what are the points where this function isn't defined? 0, 1 and 2, the three roots of the denominator. Note that at 0, our worry about undefinedness is a little silly because if we were a little smarter we'd cancel

x and get a new function that is defined at 0. But as written, the function is undefined at 0. [In the limit/continuity jargon, the function has a *removable* discontinuity at 0.]

Anyway, away from these three points, the function simplifies to:

$$\frac{(x+1)(x+2)}{(x-1)(x-2)}$$

So, the points where we could have a sign change are $-2, -1, 1, 2$. Let's start from the right. On the far right, everything is positive, because x is greater than all four numbers. So the expression is positive. When you cross 2, $x-2$ becomes negative, the others remain positive. So, negative. Then, when you cross 1, both $x-1$ and $x-2$ become negative, the others remain positive, so positive. And then after -1 it becomes negative and after -2 again positive. So the function is positive at $(-\infty, -2) \cup (-1, 1) \cup (2, \infty)$. But wait. The *original* function wasn't defined at 0, so we need to exclude that. So the original function is positive on $(-\infty, -2) \cup (-1, 0) \cup (0, 1) \cup (2, \infty)$. The function is negative at $(-2, -1) \cup (1, 2)$. The function is zero on $\{-2, -1\}$. It is undefined at $\{0, 1, 2\}$. By the way, there's an important difference between the point 0, where it is undefined only because we wrote the function stupidly, and $\{1, 2\}$, where it is undefined for more fundamental reasons. Namely, at 0, the discontinuity is removable, but at 1 and 2, it cannot be removed.

So we have a pretty good feel. But there's an important thing that we didn't take into account: higher powers.

For instance, consider:

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

So here, the function is not defined at 1. It's positive to the right of 1. What happens as we cross 1? The $(x-1)$ term changes sign, but it is squared, so the sign change doesn't affect the sign of the overall expression. And indeed, there's no sign change. So the expression continues to be positive. Then, at 0, there is a sign change, and the expression becomes negative. At -1, there is another sign change, and the expression becomes positive again. The reason? The exponent in this case is odd.

So, the function is positive on $(-\infty, -1) \cup (0, 1) \cup (1, \infty)$, negative on $(-1, 0)$, zero on $\{-1, 0\}$, and undefined at 1.

To view a consolidated summary on how to handle inequalities involving rational functions, go to the executive summary at the beginning of the document.

4. Inequalities and absolute value

4.1. **Absolute value.** We've talked about the absolute value function, so I'll just say a bit about solving inequalities involving the absolute value function. The main thing to remember is: if the absolute value of a number is in a set A of nonnegative numbers, then that number itself is in the set $A \cup -A$, where $-A$ is the set of negatives of elements of A. So, for instance:

$$|x-2| < 3$$

is equivalent to:

$$-3 < x - 2 < 3$$

So an inequality involving absolute values is really an inequality that involves both an upper bound and a lower bound. So it is two inequalities rolled into one. Now, when you have two inequalities rolled into one, you can solve them separately and *intersect* the solutions. But in this case, it's easy to sort of solve them both together. Just add 2 to both sides:

$$-1 < x < 5$$

so the solution set is the open interval $(-1, 5)$.

I suggest you look at more examples in the book involving absolute values, rational functions and inequalities. There are a lot of tricks. These ideas are very important. They'll come up again and again, particularly in the context of $\epsilon - \delta$ proofs.

4.2. **More ways of thinking about absolute values.** Absolute values and the inequalities related to them could be a little tricky, so here are some further intuitive ways of thinking about the absolute value function that might be useful when thinking about inequalities.

One way of thinking of the absolute value function is as a folding function. What I mean is, you think of the number line as a long thin strip of paper, and to calculate the absolute value, you fold it about 0, so that $-x$ comes on top of x. So *undoing* the absolute value function is like *unfolding*.

Let's take some examples. Suppose you know that the absolute value of x is 3. Then, what can you say about x? You may say that $x \in \{-3, 3\}$. In words, $x = 3$ or $x = -3$. A convenient shorthand is $x = \pm 3$. You may have seen this kind of shorthand in the formula for the roots of a quadratic equation.

Now, suppose you are given that the absolute value of x is either 2 or 3. What are the possible values of x? Think again about unfolding, and you'll see that the possibilities for x are ± 2 and ± 3.

Okay, by the way, what can you say about x if you are given that $|x|$ is either 2 or -1? The thing you have to say is that the -1 case is always false. It has no solution. So the only legitimate case is $|x| = 2$, which gives $x = \pm 2$.

Now we can proceed to thinking about inequalities. Suppose we are given that $2 < |x| < 3$. What does this tell you about the possibilities for x? Unfolding, you see that it's either the case that $2 < x < 3$ (that happens when x is positive) or $-3 < x < -2$ (That happens when x is negative). So the solution set is $(-3, -2) \cup (2, 3)$.

Okay, what about $-1 < |x| < 2$? Well, the first thing you should note is that $-1 < |x|$ is a zero information statement, because it's always true. [SIDENOTE: A statement that is always true is termed a *tautology* and a statement that is always false is termed a *fallacy*.] So we can discard that, and we get $|x| < 2$. So, $|x|$ is in the region $[0, 2)$. Unfolding this, we obtain that the possibilities for x are the interval $(-2, 2)$.

What about $0 < |x| < 2$? Well, you can guess the answer by now: it is $(-2, 0) \cup (0, 2)$.

TRIGONOMETRY REVIEW (PART 1)

MATH 152, SECTION 55 (VIPUL NAIK)

Difficulty level: Easy to moderate, given that you are already familiar with trigonometry.

Covered in class?: Probably not (for the most part). Some small segments may be covered in class or in problem session if it helps with some problems. Please go through this if you experience difficulties while doing trigonometry problems.

Corresponding material in the book: Section 1.6 (part).

Corresponding material in homework problems: Homework 1, advanced homework problem 6.

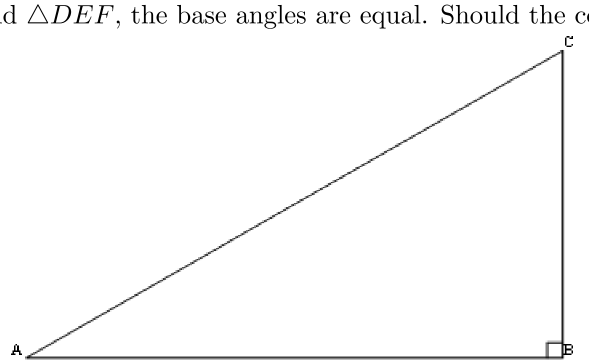
1. Trigonometric functions for acute angles

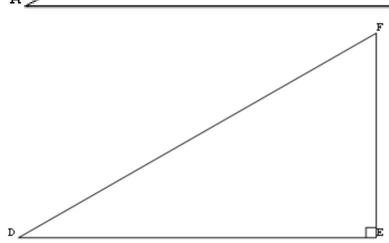
Earlier we talked about the fact that sometimes you know that something is a function (because it sends every element in the domain to a unique output – and satisfies the condition that equal inputs give equal outputs) but you don't have an expression for it. It's like you know somebody is a person but you don't know that person's name. So what do you do? You just make up a name. Well, that's what we're going to do.

So, let's consider an angle, that I'll call θ, and assume that θ is *strictly* between 0 and $\pi/2$ (90°, a right angle). By the way, the word *strictly* when used in mathematics means that the equality case (the *trivial* or *degenerate* case) is excluded. So, in this case, it means $0 < \theta < \pi/2$. The high school term for an angle strictly between 0 and $\pi/2$ is *acute angle*.

So now I define the following function whose domain is the set of acute angles. $f(\theta)$ is the ratio of the height of a right-angled triangle with base angle θ to the hypotenuse of that triangle. In other words, it is the ratio of the opposite side to the angle θ to the hypotenuse.

So, you may say, why is this a function at all? Why does it make sense? There are infinitely many triangles of different sizes with base angle θ. Could different choices of triangle give different values of $f(\theta)$? And if so, doesn't that undermine the claim that f is a function? For instance, in the two triangles $\triangle ABC$ and $\triangle DEF$, the base angles are equal. Should the corresponding side ratios also be equal?



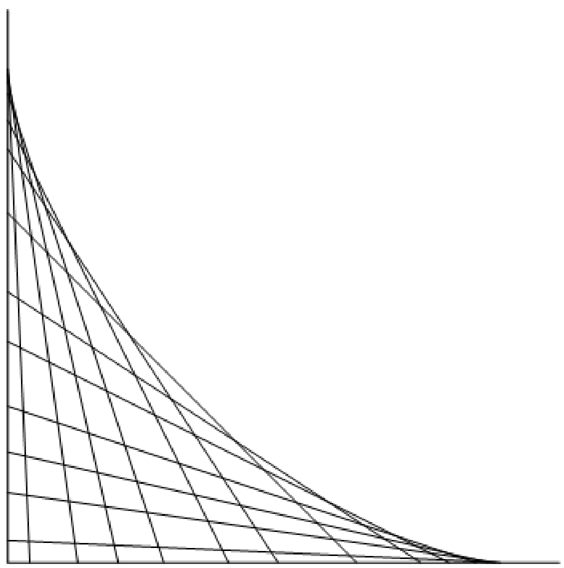


By the way, this first question that I asked is the kind of question you should ask whenever a *function* is defined in a roundabout manner, with some arbitrary choices in between. People often phrase this as *is the function well-defined?* though a more precise formulation is: *is the so-called function a function at all?*

In this case, the answer is *yes*, and the reason is the notion of *similarity* of triangles. For those of you who've taken some high school geometry, you've probably seen this notion. For those who haven't, the idea is that if two triangles have the same angles, they essentially have the same *shape*, even if they have different *sizes*, so the *ratios* of side lengths are the same.

So f is a function. But so what? Do we have an expression for it? Well, yes and no. There's no expression for f as a polynomial or rational function, because it isn't that kind of function. But we can give f its own name, and then we'll be happy. So what's a good name? f? No, f, is too plain and all too common. We need a special name. The name we use is sine, written *sine* in English and sin in mathematics. So $f(\theta)$ is written as $\sin(\theta)$. By the way, when the sin is being taken of a single letter variable or constant, we don't usually put parentheses. So we just say $\sin\theta$.

Okay, so what is the domain of the sin function as we've defined it? It is $(0, \pi/2)$. What is the range? In other words, what are the possible values that $\sin\theta$ can take? Well, think about it this way. Think of a ladder that you have placed with one end touching a vertical wall and the other end on the floor. Now, imagine this ladder sliding down. When the ladder becomes horizontal, the triangle has base angle of zero. When it's almost touched down, the base angle is really small and the opposite side is, too. So $\sin\theta \to 0$ as $\theta \to 0$. That $\to$ here means *tends to* or *approaches* – it's a concept we'll be looking at in more detail when we do limits.



At the other extreme, when the ladder is almost upright, the opposite side is almost equal to 1, so $\sin\theta \to 1$ as $\theta \to \pi/2$. So what we're seeing is that sin is an increasing function starting off from just about the right of zero and ending at just about the left of 1. So the range of this function is $(0, 1)$.

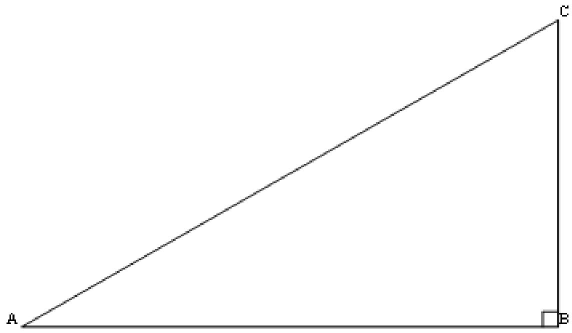
So here's one more point that is worth thinking about. In high school, if you started looking at trigonometric functions before the radian measure was introduced, then you might have seen that angles are denoted differently, e.g., by Greek letters. Why's that? Well, one reason to think of that is that angles aren't ordinary numbers. They are measurements, and denoting an angle by a common letter like x is debasing, because angles come in *degrees*. But after you switch to the radian measure, an angle (in radians) is just any old real number. So we feel free to use x and y to denote angles. We've demystified angles.

Now, let's recall the definitions of cosine. The cosine, denoted cos, is the ratio of the adjacent side to the hypotenuse. So if θ is the base angle of a right triangle, $\cos\theta$ is the ratio of the base to the hypotenuse.

And then there is the tangent function. This is denoted tan, and it is the ratio of the opposite side to the adjacent side. Now, mathematicians can be very creative with naming sometimes but sometimes they just copy names without much deep meaning. So this tangent function doesn't have any deep relation with the concept of *tangent* to a circle or a curve. Yes, they are loosely related, but the relation isn't strong enough to merit the same name. Call that an accident of history.

The other three trigonometric functions are the reciprocals of these. The reciprocal of the sine function is the cosecant function, denoted in mathematical shorthand as csc. The reciprocal of the cosine function is the secant function, denoted in mathematical shorthand as sec. The reciprocal of the tangent function is the cotangent function, denoted in mathematical shorthand as cot.

We summarize the important definitions here.



Next, we define the six trigonometric functions of θ as ratios of side lengths in this triangle:

$$\sin\theta = \frac{\text{Opposite leg}}{\text{Hypotenuse}} = \frac{BC}{AC}$$
$$\cos\theta = \frac{\text{Adjacent leg}}{\text{Hypotenuse}} = \frac{AB}{AC}$$
$$\tan\theta = \frac{\text{Opposite leg}}{\text{Adjacent leg}} = \frac{BC}{AB}$$
$$\cot\theta = \frac{\text{Adjacent leg}}{\text{Opposite leg}} = \frac{AB}{BC}$$
$$\sec\theta = \frac{\text{Hypotenuse}}{\text{Adjacent leg}} = \frac{AC}{AB}$$
$$\csc\theta = \frac{\text{Hypotenuse}}{\text{Opposite leg}} = \frac{AC}{BC}$$

2. Relation between trigonometric functions

The six trigonometric functions are related via three broad classes of relationships. Each of these relationships pairs up the six trigonometric functions into three pairs. We discuss each of these pairings.

2.1. Complementary angle relationships.

The right triangle $\triangle ABC$ has two acute angles. The ratios of side lengths of this triangle give the trigonometric function values for both acute angles. However, a leg that's opposite to one angle becomes adjacent to the other. Thus, the trigonometric functions for $(\pi/2) - \theta$ are related to the trigonometric functions for θ as follows:

$$\begin{aligned}
\sin((\pi/2) - \theta) &= \cos\theta \\
\cos((\pi/2) - \theta) &= \sin\theta \\
\tan((\pi/2) - \theta) &= \cot\theta \\
\cot((\pi/2) - \theta) &= \tan\theta \\
\sec((\pi/2) - \theta) &= \csc\theta \\
\csc((\pi/2) - \theta) &= \sec\theta
\end{aligned}$$

The prefix *co-* indicates a complementary angle relationship. Thus, the functions sine and *co*sine have a complementary angle relationship. The functions tangent and *co*tangent have a complementary angle relationship. The functions secant and *co*secant have a complementary angle relationship.

It is an easy but useful exercise to verify the complementary angle relationships from the definitions of the trigonometric functions.

2.2. **Reciprocal relationships.** Reciprocal relationships between the trigonometric functions are as follows:

(1) $\sin\theta$ and $\csc\theta$ are reciprocals. In other words, $(\sin\theta)(\csc\theta) = 1$ for all acute angles θ.
(2) $\cos\theta$ and $\sec\theta$ are reciprocals. In other words, $(\cos\theta)(\sec\theta) = 1$ for all acute angles θ.
(3) $\tan\theta$ and $\cot\theta$ are reciprocals. In other words, $(\tan\theta)(\cot\theta) = 1$ for all acute angles θ.

It is an easy but useful exercise to verify the complementary angle relationships from the definitions of the trigonometric functions.

2.3. **Square sum and difference relationships.** These are the trickiest and the most important of the relationships. We consider the most important of these first: the square sum relationship between sin and cos.

By the Pythagorean theorem for the right triangle $\triangle ABC$ with the angle at B being the right angle, we have:

$$AB^2 + BC^2 = AC^2$$

Or, in terms of the angle θ:

$$(\text{Adjacent leg})^2 + (\text{Opposite leg})^2 = (\text{Hypotenuse})^2$$

Dividing both sides by AC^2, we obtain:

$$\frac{AB^2}{AC^2} + \frac{BC^2}{AC^2} = \frac{AC^2}{AC^2}$$

Simplifying, we obtain:

$$\left(\frac{AB}{AC}\right)^2 + \left(\frac{BC}{AC}\right)^2 = 1$$

Recall that $\cos\theta = AB/AC$ and $\sin\theta = BC/AC$, so we get:

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

With trigonometric functions, it is a typical convention to write the exponent before the angle, so we write $(\cos\theta)^n$ as $\cos^n\theta$. Using this convention, we can rewrite the above relationship as:

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

Two other square sum and difference relationships of importance are:

$$\begin{aligned} \tan^2\theta + 1 &= \sec^2\theta \\ \cot^2\theta + 1 &= \csc^2\theta \end{aligned}$$

It is a good exercise to prove both of these using the Pythagorean theorem.

2.4. **Everything in terms of** sin **and** cos**.** It is often useful to deal with sin and cos only, so it is helpful to know how to write the other trigonometric functions in terms of sin and cos. The expressions are given below:

It is a good exercise to verify that these expressions are correct using the definitions of the trigonometric functions.

$$\begin{aligned} \tan\theta &= \frac{\sin\theta}{\cos\theta} \\ \cot\theta &= \frac{\cos\theta}{\sin\theta} \\ \sec\theta &= \frac{1}{\cos\theta} \\ \csc\theta &= \frac{1}{\sin\theta} \end{aligned}$$

2.5. **Everything in terms of** sin **or** cos**.** Finally, when it comes to acute angles, we can write all the trigonometric functions in terms of sin alone or cos alone. The key is to use the fact that $\sin^2\theta + \cos^2\theta = 1$. Since both $\sin\theta$ and $\cos\theta$ are positive for an acute angle θ, we can use this to get the expressions:

$$\begin{aligned}
\cos\theta &= \sqrt{1-\sin^2\theta} \\
\sin\theta &= \sqrt{1-\cos^2\theta}
\end{aligned}$$

Once we have this, we can get expressions for all the other trigonometric functions in terms of sin. We can also get expressions for all the other trigonometric functions in terms of cos.

We give here all the expressions in terms of sin. In all these, we just take the previos expressions and replace every occurrence of $\cos\theta$ by $\sqrt{1-\sin^2\theta}$.

$$\begin{aligned}
\cos\theta &= \sqrt{1-\sin^2\theta} \\
\tan\theta &= \frac{\sin\theta}{\sqrt{1-\sin^2\theta}} \\
\cot\theta &= \frac{\sqrt{1-\sin^2\theta}}{\sin\theta} \\
\sec\theta &= \frac{1}{\sqrt{1-\sin^2\theta}} \\
\csc\theta &= \frac{1}{\sin\theta}
\end{aligned}$$

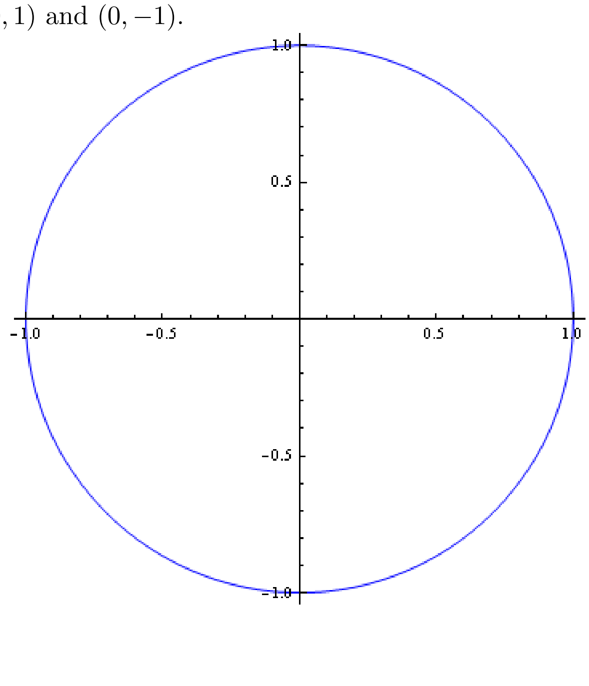
You can work out how the other five trigonometric functions look in terms of $\cos\theta$ by yourself.

Note that these expressions are valid only for acute angles. The problem is that for bigger angles, as we shall soon see, the value of sin or cos can be negative, so even thought $\sin^2\theta + \cos^2\theta = 1$, we cannot write $\sin\theta = \sqrt{1-\cos^2\theta}$ because the $\sqrt{\ }$ symbol always gives a nonnegative output and $\sin\theta$ may well be negative.

3. Unit circle trigonometry

3.1. **The unit circle.** The unit circle centered at the origin is defined as the set of points (x, y) in the coordinate plane that satisfy $x^2 + y^2 = 1$. This is a circle of radius 1 centered at the origin.

The unit circle intersects the x-axis at the points $(1, 0)$ and $(-1, 0)$. It intersects the y-axis at the points $(0, 1)$ and $(0, -1)$.



3.2. **Sine and cosine using the unit circle.** Suppose θ is an angle. We define $\sin\theta$ and $\cos\theta$ using the unit circle as follows. Start on the unit circle at the point $(1, 0)$. Move an angle of θ in the counter clockwise direction on the unit circle. Call the point you finally reach (x_0, y_0). Then, $\cos\theta$ is defined as x_0 and $\sin\theta$ is defined as y_0.

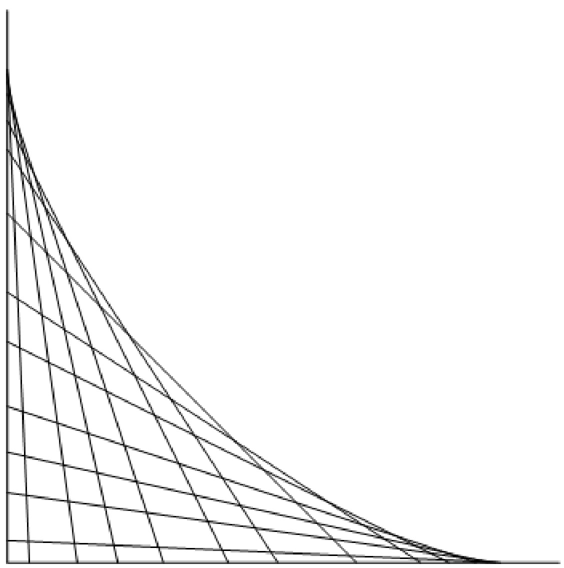
When θ is an acute angle, then the point (x_0, y_0) is in the first quadrant. We see that the definitions of $\cos\theta$ and $\sin\theta$ match the definitions we gave earlier in terms of triangles.

4. Qualitative behavior of sine and cosine

4.1. **For acute angles.** Note that prior to the introduction of unit circle trigonometry, we defined sin and cos only for acute angles. We first discuss the qualitative behavior of these functions for acute angles, but also include the limiting cases of 0 and $\pi/2$. After that, we discuss the behavior for obtuse angles.

For acute angles, we have the following:

(1) sin is a strictly increasing function for acute angles, starting off with $\sin 0 = 0$ and $\sin(\pi/2) = 1$. This can be seen graphically in many ways. For instance, imagine a ladder that is initially vertical along a wall, and gradually slides down. The angle that the foot of the ladder makes with the floor decreases from $\pi/2$ to 0, and the vertical height of the top of the ladder also decreases. [More class discussion on this]



(2) cos is a strictly decreasing function for acute angles, starting off with $\cos 0 = 1$ and $\cos(\pi/2) = 0$. The fact that the behavior of cos is the mirror opposite of that of sin is not surprising – this essentially follows from the complementary angle relationship.

(3) tan is a strictly increasing function for acute angles, starting off with $\tan 0 = 0$. $\tan(\pi/2)$ is not defined, and as an acute angle θ comes closer and closer to being $\pi/2$, $\tan\theta$ approaches ∞. (The precise meaning and explanation of this statement involve familiarity with ideas of limits, which are beyond the current scope of discussion).
(4) cot is a strictly decreasing function for acute angles, and its behavior mirrors that of tan because of the complementary angle relationship. $\cot 0$ is undefined and $\cot(\pi/2) = 0$.

4.2. **For angles between 0 and π.** Using the unit circle trigonometry definition, we can see that:

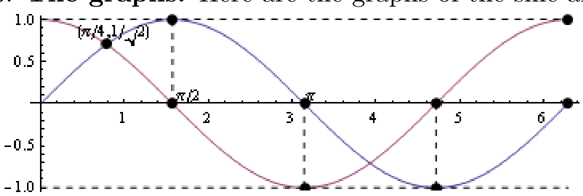
$$\begin{aligned}\sin(\pi - \theta) &= \sin\theta \\ \cos(\pi - \theta) &= -\cos\theta\end{aligned}$$

In particular, sin is positive for all angles θ that are strictly between 0 and π, including both acute and obtuse angles. In terms of unit circle trigonometry, this is because the first half of the unit circle if we go counter-clockwise from $(1, 0)$ has positive y-coordinate. On the other hand, cos is negative for obtuse angles, and this can again be seen from the unit circle trigonometry.

Note that $\tan\theta$ is defined as $\sin\theta/\cos\theta$ wherever $\cos\theta \neq 0$, so we get the formula:

$$\tan(\pi - \theta) = -\tan\theta$$

4.3. **The graphs.** Here are the graphs of the sine and cosine function from 0 to 2π.



We will study the graphs of the other trigonometric functions later.

4.4. **General facts: value taking.**

(1) sin takes its peak value of 1 at numbers of the form $2n\pi + (\pi/2)$, where n is an integer, and its trough values of -1 at numbers of the form $2n\pi - (\pi/2)$, where n is an integer. sin takes the value 0 at multiples of π, i.e., numbers of the form $n\pi$ where n is an integer.
(2) The solutions to $\sin x = \sin\alpha$ come in two families: $x = 2n\pi + \alpha$ and $x = 2n\pi + (\pi - \alpha)$, with n ranging over the integers.
(3) cos takes its maximum value of 1 at multiples of 2π, i.e., numbers of the form $2n\pi$, n ranging over integers. It takes its minimum value of -1 at odd multiples of π, i.e., numbers of the form $(2n+1)\pi$, n ranging over integers. it takes the value 0 at odd multiples of $\pi/2$, i.e., numbers of the form $n\pi + (\pi/2)$, n ranging over integers.
(4) The solutions to $\cos x = \cos\alpha$ come in two families: $x = 2n\pi + \alpha$ and $x = 2n\pi - \alpha$, n ranging over integers. These two families can be combined compactly by writing the general expression as $x = 2n\pi \pm \alpha$.

4.5. **Even, odd, mirror symmetry, half turn symmetry.** Here are some general facts about the sine and cosine functions:

(1) The sine function is an *odd function*, i.e., $\sin(-\theta) = -\sin\theta$. In particular, the graph of the sine function enjoys a half turn symmetry about the origin. In fact, the graph of the sine function enjoys a half turn symmetry about any point of the form $(n\pi, 0)$.
(2) The graph of the sine function enjoys mirror symmetry about any vertical line through a peak or trough, i.e., any line of the form $x = n\pi + (\pi/2)$.
(3) The cosine function is an *even function*, i.e., $\cos(-\theta) = \cos\theta$. In particular, the graph of the cosine function enjoys a mirror symmetry about the origin. In fact, the graph enjoys a mirror symmetry about all vertical lines through a peak or trough, i.e., any line of the form $x = n\pi$.

(4) The graph of the cosine function enjoys a half turn symmetry about any point of the form $(n\pi + (\pi/2), 0)$.
(5) The tangent, cotangent, and cosecant functions are odd functions on their domains of definition. The secant function is even on its domain of definition.

4.6. **Periodicity.** Here are some important facts:

(1) The sine and cosine are periodic functions and they both have period 2π.
(2) The tangent and cotangent are periodic functions and they both have period π.
(3) The secant and cosecant are periodic functions and they both have period 2π.

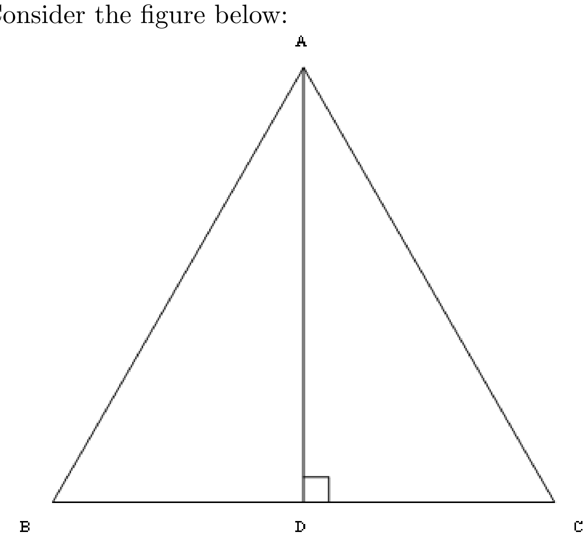
5. Values of trigonometric functions for important angles

5.1. $\pi/4$**.** To determine the values of trigonometric functions for $\pi/4$, we need to examine closely the right isosceles triangle. In this triangle, both legs have the same length, and by the Pythagorean theorem, the length of the hypotenuse is $\sqrt{2}$ times this length. Thus, we obtain the following:

$$\begin{aligned}
\sin(\pi/4) = \cos(\pi/4) &= \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2} && \approx 0.707 \\
\tan(\pi/4) = \cot(\pi/4) &= 1 \\
\sec(\pi/4) = \csc(\pi/4) &= \sqrt{2} && \approx 1.414
\end{aligned}$$

Notice something else too. For the angle $\pi/4$, the values of any two trigonometric functions that are related by the complementary angle relationship are equal. Thus, $\sin(\pi/4) = \cos(\pi/4)$, $\tan(\pi/4) = \cot(\pi/4)$, and $\sec(\pi/4) = \csc(\pi/4)$. Geometrically, this is because we are working with a right isosceles triangle, and there is a symmetry between the two legs. Algebraically, this is because the angle $(\pi/4)$ is its own complement, i.e., $(\pi/4) = (\pi/2) - (\pi/4)$.

5.2. $\pi/6$ **and** $\pi/3$**.** We now consider the triangle where one of the angles is $\pi/6$ and the other angle is $\pi/3$. Consider the figure below:



In the figure, the triangle $\triangle ABC$ is an equilateral triangle and the line AD is an altitude so $AD \perp BC$. Since $\triangle ABC$ is equilateral, all its angles are also equal and hence all the angles are $\pi/3$. Also, some elementary geometry tells us that AD bisects BC so $DC = (1/2)BC$, so $DC/BC = 1/2$.

Now consider the triangle $\triangle ADC$. The angle at C is $\pi/3$ and the angle $\angle CAD$ is $\pi/6$. Thus, we obtain that $\cos(\pi/3) = \sin(\pi/6) = DC/AC$. Since $\triangle ABC$ is equilateral, $AC = BC$, so $DC/AC = DC/BC$, which is $1/2$.

We thus have the first important fact: $\cos(\pi/3) = \sin(\pi/6) = 1/2$. Now, using the relations between trigonometric functions, we can obtain the other values. The full list is given below:

$$
\begin{aligned}
\sin(\pi/6) = \cos(\pi/3) &= \frac{1}{2} && = 0.5 \\
\cos(\pi/6) = \sin(\pi/3) &= \frac{\sqrt{3}}{2} && \approx 0.866 \\
\tan(\pi/6) = \cot(\pi/3) &= \frac{1}{\sqrt{3}} && \approx 0.577 \\
\cot(\pi/6) = \tan(\pi/3) &= \sqrt{3} && \approx 1.732 \\
\sec(\pi/6) = \csc(\pi/3) &= \frac{2}{\sqrt{3}} && \approx 1.154 \\
\csc(\pi/6) = \sec(\pi/3) &= 2
\end{aligned}
$$

You should be able to reconstruct all these values. You can choose either to memorize all of them, or to memorize the first two rows and reconstruct the rest from them on the spot. Alternatively, you can remember the side length configuration of the triangle $\triangle ADC$ and read off the trigonometric function values by looking at that triangle.

FUNCTIONS: A RAPID REVIEW (PART 1)

MATH 152, SECTION 55 (VIPUL NAIK)

Difficulty level: Easy to moderate. Most of these are ideas you should have encountered either implicitly or explicitly in the past.

Covered in class?: This will roughly correspond to material covered on Monday September 26. Most of the trickier aspects of this will be covered in class, but many small points will be omitted due to time constraints. Hence, it is recommended that you read through these notes either before or after lecture.

Corresponding material in the book: Sections 1.5, 1.6. Note that the book covers the same material with somewhat different language and different examples of functions, so you should go through it before or while doing the homework problems.

Corresponding material in homework problems: Homework 1, routine problems 1–6 (all from section 1.5), advanced problem 1.

Things that students should get immediately: The concepts of function, domain, range, expression for function, table of values for a function, graph of a function, the notion of piecewise defined function, polynomials, rational functions, absolute value function, signum function, positive part function.

Things that students should get with effort: How to obtain a piecewise definition for a maximum or minimum of two functions, how to determine the domain and range of a function.

Executive review

This review will probably be reproduced (with minor modifications) in the midterm review sheet. It is meant as a *review summary* of these lecture notes – capturing those aspects of these notes that are important on a second reading, and *ignoring* those things that are significant for first time learning but not so important later on.

For first time reading, skip to the next section.

Words ...

(1) The *domain* of a function is the set of possible inputs. The *range* is the set of possible outputs. When we say $f : A \to B$ is a function, we mean that the domain is A, and the range is a *subset* of B (possibly equal to B, but also possibly a proper subset).
(2) The main fact about functions is that *equal inputs give equal outputs*. We deal here with functions whose domain and range are both subsets of the real numbers.
(3) We typically define a function using an algebraic expression, e.g. $f(x) := 3+\sin x$. When an algebraic expression is given without a specified domain, we take the domain to be the largest possible subset of the real numbers for which the function makes sense.
(4) Functions can be defined piecewise, i.e., one definition on one part of the domain, another definition on another part of the domain. Interesting things happen where the function changes definition.
(5) Functions involving absolute values, max of two functions, min of two functions, and other similar constructions end up having piecewise definitions.

Actions (think back to examples where you've dealt with these issues)...

(1) To find the (maximum possible) domain of a function given using an expression, exclude points where:
   (a) Any denominator is zero.
   (b) Any expression under the square root sign is negative.
   (c) Any expression under the square root sign in the denominator is zero or negative.
(2) To find whether a given number a is in the range of a function f, try solving $f(x) = a$ for x in the domain.

(3) To find the range of a given function f, try solving $f(x) = a$ with a now being an *unknown constant*. Basically, solve for x in terms of a. The set of a for which there exists one or more value of x solving the equation is the range.
(4) To write a function defined as $H(x) := \max\{f(x), g(x)\}$ or $h(x) := \min\{f(x), g(x)\}$ using a piecewise definition, find the points where $f(x) - g(x)$ is zero, find the points where it is positive, and find the points where it is inegative. Accordingly, define h and H on those regions as f or g.
(5) To write a function defined as $h(x) := |f(x)|$ piecewise, split into regions based on the sign of $f(x)$.
(6) To solve an equation for a function with a piecewise definition, solve for each definition within the piece (domain) for which that definition is satisfied.

1. What is a function?

1.1. **Inputs and outputs, or so they say.** We're going to begin by talking about functions. You've probably already seen functions in some form in calculus and precalculus. You may have seen both the general concept of function and lots of specific examples. In this course, we try to be a lot more precise about what a function means. This precision will be very important because functions are used for modeling purposes throughout mathematics and mathematically based disciplines.

A function is something that "takes in" (or *eats* or *gobbles*) an input and "gives out" (or *spits*) an output. Some people think of a function as a *black box* or *machine* into which you feed in input and get output. For instance, you put in money into a cola vending machine and get out a cola. In today's computer age, you might enter an input value onto a computer screen and get an output. We say that a function *maps* the input to the output, so functions are also called *mappings* or *maps*. Some people say that a function *sends* an input to an output. Functions can also be thought of as *rules* or *assignments*.

1.2. **The real bite: equal inputs give equal outputs.** So what's missing from this description? Well, the most important thing about a function is that when you put in one input, you get one output, and the output depends only on the input. In other words, *equal inputs should give equal outputs*. So it doesn't depend on who feeds the input or how the machine is feeling at the time it is fed in. The output depends on the input, and only on the input. This dependence is what we call a *functional* dependence.

So is Google a function? It takes in your query and outputs a bunch of search results. But in another sense, it isn't a function, because Google's results keep changing with time and other factors. What about temperature? Is temperature a function of time? No, because what the temperature is at a given time depends on where you measure it. On the other hand, the temperature at a particular point in space *is* a function of time.

So, in order to make something a function, you have to specify enough input so that the output is determined based on that.

In this course sequence, we will not be looking at functions with weird inputs and weird outputs. For the most part, our inputs will be single real numbers and our outputs will be single real numbers. So, although the concept of function is very general, we will be restricting to very particular kinds of functions to which we can apply tools specifically developed for real numbers.

There are two important concepts related to functions: the *domain* and *range*. We'll talk about these in more detail as we proceed, but here's the rough description: the *domain is the set of all possible (sensible) inputs to the function* and the *range is the set of all possible outputs of the function*.

1.3. **Of circumferences and diameters: an illustrative example.** Let's first consider a "real-world" problem. The wheel of your bicycle has diameter d. You want to find out the circumference of your wheel. In more abstract jargon, you want to find out the circumference of a circle *in terms of* its diameter.

The first question you should ask is: does the circumference *depend only on* the diameter? That's not a silly question, even if in this case it seems intuitively clear to some people. What does my question mean? What it means is that if two circles have the same diameter, is it necessary that they have the same circumference? If that isn't the case, then we don't have a function.

In this case, the answer is *yes*. If the diameter is the same, the circles are congruent – you can translate one over to cover the other. So their circumference should be the same. So yes, we do have a function. But what kind of function is it?

Now, here's the thing. We have this loose reasoning that says that there is a function that takes in an input d and spits out an input. Let's call this function f. The value obtained by applying f to d is denoted $f(d)$. What this really means is f evaluated at d or the value of f at d. [Sidenote: In geometric settings, I will abuse matters a little bit and conflate real numbers with length measurements. The implicit idea everywhere is to choose a unit of measurement, and all inputs are measured in those units. I will perform this abuse on a regular basis. I could be more precise, but you'll find this abuse everywhere so might as well get used to it.] So $f(1)$ denotes the circumference of a circle with diameter 1, and $f(2)$ denotes the circumference of a circle with diameter 2.

So far so good. Happy? Not quite. We've just shown there is a function, but from a computational point of view, we haven't done anything. What we would like to do is have some expression for f that makes it easy to calculate. As it happens, various ancient civilizations (the Greeks and Indians) showed that $f(d) = \pi d$, where π is some number. They also calculated the first few digits of π, $\pi = 3.141592\ldots$. So this finally gives a formula.

So $f(1) = \pi$, $f(2) = 2\pi$, $f(3) = 3\pi$, $f(\pi) = \pi^2$, and so on. What about $f(a+b)$? What's that? Well, to calculate f of something, you do π times that thing, that's what the formula tells you. So $f(a+b)$ is $\pi(a+b)$. What's that? That's $\pi a + \pi b$, by the distributivity laws you learned in primary school.

When you apply f to something, you should make sure that you apply f to *the whole thing*. A common error that students make is to just write $f(a+b) = \pi a + b$. That's wrong, because the *whole expression* $a+b$ has to be multipled by π. So, at the first stage of simplifying a function where the input is itself an expression, *please put parentheses around the input* wherever you write it down.

Hey, but $\pi a = f(a)$ and $\pi b = f(b)$, so we have this really cool fact:

$$f(a+b) = f(a) + f(b) \ \forall \ a, b$$

The $\forall$ symbol above means *for all*. What we have is a rule that holds for all values you plug in for a and b. Is this rule true for any function f, or only for the f that we wrote down? Well, it turns out that this is true for f because f is a *linear function*: it is of the form $f(x) = cx$ for some constant c.

What is $f(-1)$? The formula tells you it is $-\pi$. But hang on. What does it even mean to have a circle of diameter -1? Nothing. It's nonsense. It doesn't make sense. The diameter of a circle cannot be negative, even though the formula makes perfect sense for negative diameters.

Which brings us to the concept of *domain*. The domain of a function is the set of values you can feed in as inputs. What's the domain of f? It is the set of *positive real numbers*. There are two ways to write this set: $\{x \in \mathbb{R} \mid x > 0\}$ and $(0, \infty)$. [Explain both, if many students don't understand.]

So f is a function *from* the positive reals *to* something – where? The set of values that f can possibly take is termed the *range* [SIDENOTE: There is a related concept of *co-domain* that we will discuss later.] For this choice of function f, the range is also the set of positive real numbers.

We write this as follows:

$$f : (0, \infty) \to (0, \infty), \qquad f(x) := \pi x$$

Note that I changed the letter from d to x. That was bad board technique, but it is not mathematically a problem at all. Why? Because that's just a name, and there's nothing in a name. May be d's real name is d, but I prefer to call it by the nickname (or alias) x. The main thing to take care of is that the letter inside the parentheses is the one used on the right side where the input should go.

And by the way, that earlier equation was not quite correct. We should really have:

$$f(a+b) = f(a) + f(b) \ \forall \ a, b \in \operatorname{dom}(f)$$

or:

$$f(a+b) = f(a) + f(b) \ \forall \ a, b \in (0, \infty)$$

[SIDENOTE, may not be covered in class: By the way, what we're using here is the fact that, for this function f, if a and b are in the domain of f, so is $a+b$. Why is that? This basically goes back to the fact that the sum of two positive numbers is positive.]

[SIDENOTE: Local memory, don't give two functions the same letter name in the same context, but feel free to reuse letters in different contexts. Function name letters are just like variables in this respect.]

Let's look at some other functions coming from geometry:

(1) Area of a rectangle as a function of the perimeter? No, sorry. The perimeter does not give enough information to calculate the area of the rectangle. For instance, we can have a long and thin rectangle and a square of the same perimeter but very different areas.
(2) Area of a square as a function of the perimeter? Yes, it is $g(x) = x^2/16$, where g has domain and range the set of positive real numbers.

Before we go into examples of functions, I just want to reiterate the following: *whether something is a function* is a very different question from *whether we have an expression to compute it*. You may be able to successfully prove that something is a function but be completely at a loss to actually compute it.

2. Some important classes of functions

2.1. Constant functions.

2.1. Constant functions. A constant function is a function where the output is the same for all inputs. A constant function can be identified by the constant value of the output. For instance, the *zero* function is the function that sends all its inputs to 0.

2.2. Polynomial functions. The general expression for a polynomial looks like:

$$p(x) = a_0 + a_1 x + \cdots + a_n x^n$$

Here, the a_i are all real numbers, and they're termed the *coefficients* of the polynomial. If $a_n \neq 0$, n is the *degree* of the polynomial. a_0 is termed the *constant term* of the polynomial and a_n is termed the *leading coefficient* of the polynomial.

The coefficients of the polynomial are constants, in the sense that they do not depend on x. However, we haven't specified beforehand the values of these constants. So they're *unknown knowns*.

Here are some concrete examples of polynomials:

- $2x - 5$ is a polynomial of degree 1, with the constant term $a_0 = -5$ and the leading coefficient $a_1 = 2$.
- $x^2 - 7x + 3$ is a polynomial of degree 2 with the constant term $a_0 = 3$, the middle coefficient $a_1 = -7$, and the leading coefficient $a_2 = 1$.
- $x^3 - 2$ is a polynomial of degree 3 with constant term $a_0 = -2$, leading coefficient $a_3 = 1$, and $a_1 = a_2 = 0$.

Polynomials are *globally defined functions*. In other words, they have domain the whole real numbers $\mathbb{R}$. This is just a fancy way of saying that you can evaluate a polynomial at any real number without getting into trouble.

However, even though a function may be globally defined, we may sometimes be interested in restricting it to a smaller domain. For instance, in the circle example, we had the linear function $f(x) = \pi x$. That expression is defined for all real numbers. However, the real-world context from which we were getting the function required us to *restrict the function* to a smaller domain: the set of positive real numbers.

[SIDENOTE, cover in class only if somebody raises the question: What about the range of a polynomial function? That turns out to be a trickier question. We will need to build more machinery before we can answer that question for arbitrary polynomials.]

2.3. Rational functions. Next, we consider rational functions. Here, we have the problem of vanishing denominators. So, the *largest possible domain* for a rational function is the real numbers minus all the points where the denominator vanishes. By the way, the points where a polynomial vanishes are called its *zeros* or *roots*. [SIDENOTE: See "Convention on domains", Page 28, in the book.]

For instance, consider the rational function $T(x) = x/(x^2 + 1)$. What is the largest possible domain for this function? To answer this, first ask: where does the denominator vanish? Now, those of you who've not seen complex numbers may say – nowhere. And those of you who've seen complex numbers will say $\pm i$. Yes, $\pm i$ are roots of the polynomial $x^2 + 1$. But in this course we are dealing with the real world. In all our discussions, whether I say it or not, all numbers that we deal with are real numbers. And $x^2 + 1$ has no real roots. So the denominator does not vanish anywhere and this rational function is globally defined on $\mathbb{R}$.

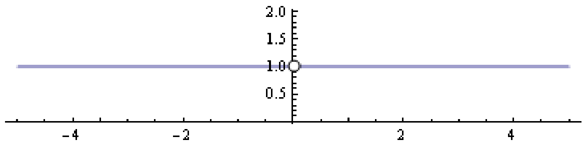
Okay, what about this function called $FORGET$? $FORGET$ is defined by:

$$FORGET(x) = \frac{x}{x}$$

You may be tempted to cancel the x and the x and say that $FORGET(x) = 1$ and so it is always defined. But one of the things about rational functions is that you need to look at the rational function *as it is written*. You cannot cancel something unless it is guaranteed to be nonzero.

So, at the point 0, the function becomes $0/0$, which is undefined. At any other point, the function takes the value 0. So, we find that the domain is the set of nonzero real numbers. How do we express this?

We can use the set difference notation. The domain is written as $\mathbb{R} \setminus \{0\}$. Or, we can think of it as the union of the negative and the positive numbers. In this case, the domain is $(-\infty, 0) \cup (0, \infty)$.



3. Computational tools

3.1. **The domain.** When the domain is not explicitly specified or clear from the situational context, the convention (cf. Page 28, "conventions on domains", subtopic of Section 1.5) is to define the domain as the largest possible subset of $\mathbb{R}$ where the function *as given* can be evaluated. Some of the things you need to check for are:

- The denominator should not vanish. In other words, *we need to exclude from the domain all points where any denominator becomes zero*. For instance, $1/(x(x-1))$ is not defined at the points 0 and 1 because the denominator vanishes at these points.
- When you are taking the square root of some expression as a sub-expression of the function, then the thing under the square root should be nonnegative. For instance, for the function $\sqrt{x} + \sqrt{1-x}$, we should have both $x \geq 0$ and $1 - x \geq 0$.
- When an expression under the square root is in the denominator, then the thing under the square root should be positive.

When we introduce new functions such as the logarithm and exponents with arbitrary bases and exponents, we will introduce further rules for determining the domain of the function based on the domain properties of these functions.

3.2. **The range.** Let's try to translate the statement "a is in the range of f" into a form that is computationally tractable. What does it mean for a to be in the range of f? It means that there exists a value of x such that $a = f(x)$.

For instance, consider the function $f(x) = 1/(x-1)$. How do we determine whether a given $a \in \mathbb{R}$ is in the range of f? Okay, now this might be a little too abstract and symbolic for some people, so I would urge you to do this little trick. Imagine that a is some constant, some number *known to you* but not to me. So in your mind, instead of a, you see a specific number. But since that number is secret, you cannot reveal it to me and you have to call it a.

So you have this number a that's known to you and you need to find x such that $f(x) = a$. Well, let's solve. We have:

$$\begin{aligned} 1/(x-1) &= a \\ \implies 1/a &= x - 1 \\ \implies x &= 1 + (1/a) \end{aligned}$$

The goal is to determine whether there exists a x such that $f(x) = a$. What we've actually done is obtained a formula for x in terms of a. So for those a where this formula makes sense, we actually do have a value of x. What are those a? Well, all nonzero reals. So when $a \neq 0$, we can find a x such that $f(x) = a$. What about when $a = 0$? In this case, it's clear that $f(x) = a$ has no solution.

Okay, so this is the rough idea. But other situations can be a little trickier. In some cases, you may get multiple values of x mapping to a single value of a.

Also see Examples 1 and 2, Page 28-29, and look at Exercises 18-30, Page 30 (all of these are within Section 1.5).

4. Describing a function

4.1. Description by algebraic expression. So far we've discussed functions. Now, we want to discuss ways of describing functions. One way of describing functions is using an expression. We discussed examples of this last time, such as:

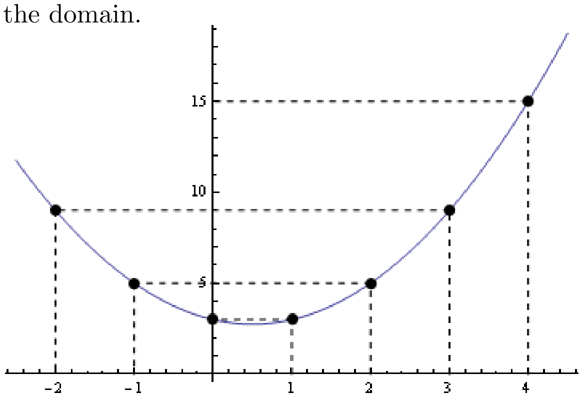
$$f : (0, \infty) \to (0, \infty), \qquad f(x) = \pi x \text{ sends diameter of circle to circumference}$$
$$g : (0, \infty) \to (0, \infty), \qquad g(x) = x^2/16 \text{ sends perimeter of square to area}$$

But, unless you have a deep algebraic understanding of the expression, this doesn't give a very good *feel* for the function. And in many cases, an expression may not exist, or we may not know what it is. So we employ two other tools: tabular listing and graphs.

4.2. Description by tabular listing and graphs. So let's do this tabular listing thing. Let's look at the function $f(x) = x^2 - x + 3$. We want to get a feel for this function. Where is it going up and down? How does it change? Let's try some numbers.

x	$f(x)$
-2	9
-1	5
0	3
1	3
2	5
3	9
4	15

Do you see a picture here? Let's try to draw a graph for this function. What's a graph? Well, it is a picture we draw in the plane that allows us to read the value of the function at any point. We choose a system of coordinate axes. The horizontal axis is conventionally chosen to be the x-axis, with right positive, and the vertical axis is the $f(x)$-axis, with up positive. We then plot the points $(x, f(x))$ for all values of x in the domain.



Now, there are lots of real numbers, and we cannot plot the values for all of them. So what I'm going to do here is a little imprecise. We'll just plot the values at a few numbers (the ones we calculated in the table) and then try to find an easy-to-draw curve that passes through all those points.

So we draw this graph. Now notice that I sort of assumed that the graph moves smoothly, it doesn't have any unexpected kinks, like, it doesn't jump wildly in between the points I plotted. You should take that with a grain of salt. I haven't presented any evidence. To really check that the graph I have drawn represents reality, you need to check a lot of intermediate values.

So, by the way, now that we have drawn the graph, two questions emerge: what's that bottom point for the graph? Or another way of putting it: what is the minimum value of $f(x)$ and at what value of x is it attained? For the case of the quadratic, there is a neat algebraic manipulation trick that can give us the answer. But since you have seen some basic calculus, you are also aware of a general procedure/approach to answering that kind of question.

Scaling issues with graphs. For most of the graphs that we will draw in class, we will use the same scale for both the x-axis and the y-axis. However, when using graphs to study functions in practice, this is not useful. Indeed, for many of the pictures of functions in these lecture notes using Mathematica, the scale used for the two axes is different.

In addition, it is also sometimes helpful, when drawing graphs, to shift the origin. Mathematica, and some other graphing softwares, may do this automatically for many graphs. However, for graphs drawn in class (as well as the Mathematica pictures included here) we will assume that there is no shifting of origin.

4.3. **The vertical line test, domain and range.** Given a picture in the coordinate plane, we have the following:

- The picture represents the graph of a function if every vertical line intersects it at most once.
- The domain of the function is the set of values of x such that the vertical line for that value of x intersects the graph.
- The range of the function is the set of values of y such that the horizontal line for that value of y intersects the graph.

We will return to these points a little later when we study techniques for drawing graphs.

4.4. **Functions defined piecewise.** Now, we're going to consider functions that have explicit expressions, but they have different expressions for different values. In other words, the domain of the function is split into parts and the definition of the function is different for each part. We will say that such functions are *piecewise defined*.

Can you think of an example? Let's think about taxes. Now, in a simple world, the tax you pay is a (non-decreasing) function of your income. The real world is a lot more complicated, with the tax you pay being a function of many other factors. But let's ignore all this. Let's consider the simplest tax system, which is called a *flat tax*. [1]

So here's how a simple version of the flat tax works. There is a *basic exemption* amount, which I'll call B, and a tax rate r for all income earned over and above B. In other words, the first B units of money that you earn don't get taxed, and of the remaining money you earn, a fraction r is taxed. By the way, $0 < r < 1$, and if you write the tax rate as a percentage, you have to divide it by 100 to get r. So, for instance, a tax rate of 10% means that r is 0.1.

[SIDENOTE: So, why did I put *strict inequality* (a $<$ sign instead of a $\leq$ sign and a $>$ sign instead of a $\geq$ sign)? Well, what does $r = 0$ mean? It means that there is effectively no tax at all for any income, which isn't a case of interest here. And what does $r = 1$ mean? It means that all money you earn beyond B belongs to the government, and that doesn't provide people with much incentive to earn. So in fact r should be between 0 and 1. What the optimal value of r should be is a question beyond the scope of this discussion.]

So if T is the tax function, we have:

$$T(I) = r(I - B)$$

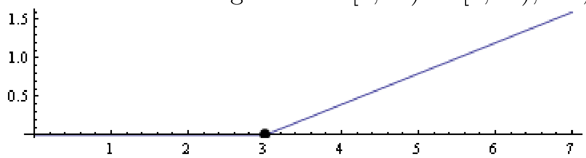
This formula is correct for people who earn as much as or more than B. But what about people who earn less than B? What if, for instance, your income is 0? The formula then says that your tax is $T(0) = -rB$, which means you have a *negative tax*. But that's not the way flat tax systems usually work. So, the real formula is:

$$T(I) = \begin{cases} 0 & \text{if } I < B \\ r(I - B) & \text{if } I \geq B \end{cases}$$

[1] For instance, income tax in the state of Illinois is a flat tax, with a tax rate of 3% or 0.03. Eastern European countries such as Estonia, Latvia, Russia, and Bulgaria have flat or near-flat tax systems.

In other words, $T(I)$ is zero for income up to B, and then rises *linearly* (or proportionally) with income.

So let's draw the graph. In the graph, the x-axis is now the I-axis, or income axis, because the income is the input variable. And the y-axis is the T-axis or the tax axis, because that's the output variable. Note that the tax function goes from $[0, \infty)$ to $[0, \infty)$, i.e., both the income and the tax are nonnegative.



The graph starts off along the horizontal axis (the I-axis) from 0 to B. Then, at B, it takes a turn and goes in a straight line forever. This line points northeast. Now, what can you say about how steep that line can be? It depends on the rate r, but can you say something in general? Sure. Since $r < 1$, the largest angle that line can make with the horizontal axis is $\pi/4$ – that's the angle when $r = 1$. The smaller the r, the smaller the angle.

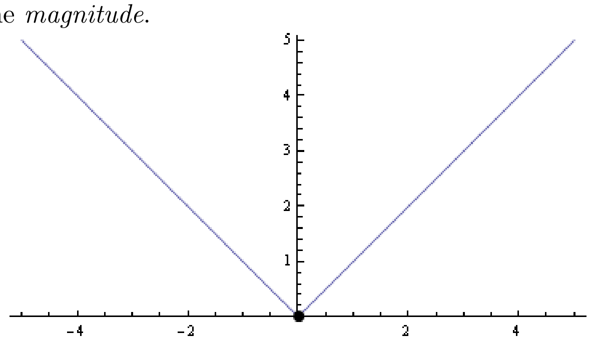
Note that the function takes a turn at the value B, but it does not jump in value. In other words, you can draw the graph without lifting your pencil. So what's happening is that when you cross B, there is a shift in the tax regime, but your tax function doesn't jump suddenly. Since most of you have some idea of what continuous and differentiable means, you can probably make this precise in those terms: the tax function is continuous everywhere (including at the point B where it changes definition) but it is not differentiable at B.

So, just as a fun question, what happens with a *progressive tax function*?[2] By the way, the United States, and most countries, have progressive tax systems. What that would mean is that in addition to the base exemption, there are likely to be other income cutoff values at which the graph takes turns. So you might start off with an almost horizontal line, then turn to a slightly steeper slope, then an even steeper slope and so on. Of course, the tax rate should never exceed 1, so you'll never get steeper than an angle of $\pi/4$ with the horizontal axis.

And what happens with a *regressive tax function*? For instance, the payroll tax in the United States is a regressive tax. Here, the graph becomes *less steep* as the income increases.

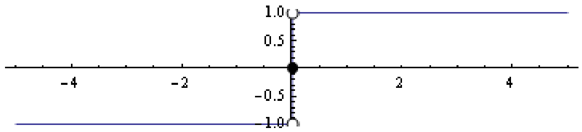
4.5. **Back to mathematics.** Coming back to mathematics, here are some important functions with piecewise definitions:

(1) **Absolute value function**: The function is denoted, not by a letter, but by bars. For a real number x, the absolute value of x, denoted $|x|$, is defined as x if $x \geq 0$ and as $-x$ if $x < 0$. It is also termed the *magnitude*.

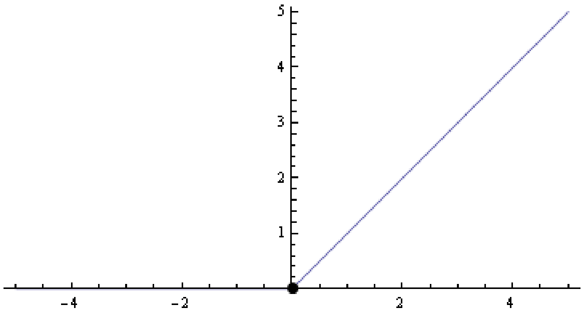


(2) **Signum function**: The sign function or signum function sends all positive numbers to $+1$, 0 to 0, and all negative numbers to -1. For $x \neq 0$, the signum function is thus equal to the function that sends x to $x/|x|$. In an alternate conventions, the signum function is considered undefined at 0, so its domain is $\mathbb{R} \setminus \{0\}$.

[2]The term "progressive" here is used in a strictly mathematical, rather than a political sense, even though self-identified political progressives on average tend to favor more progressive tax systems.



(3) **Positive part function**: This function applied to a real number x is denoted x^+, and is defined as $\max\{0, x\}$. The positive part of x is equal to 0 if $x \leq 0$ and equal to x if $x > 0$.



How do we think of functions defined piecewise? The main thing to remember is that most of the action happens at the place where the function changes definition. Think of it like switching gears or taking a turn. That's what happens literally with the tax function, and the absolute value function and the signum function. And whenever you have piecewise defined functions and you know that the functions on each of the parts are very well-behaved, these turning points are the ones where mischief is most likely to occur.

5. The max and min operators

Among the constructs used to create piecewise functions, that arise naturally, are the max and min operators. For instance, suppose f and g are functions on the real numbers. Then, consider the function:

$$h(x) := \max\{f(x), g(x)\}$$

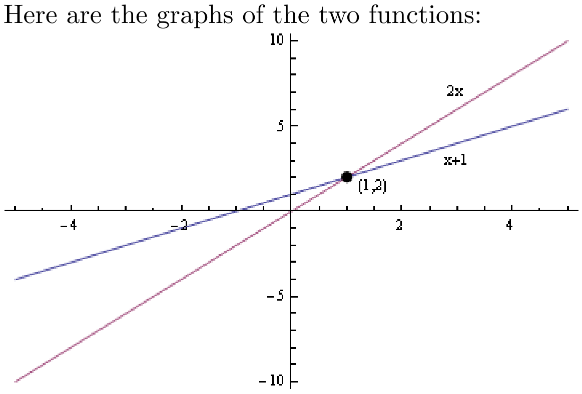
What does this mean? Here is how we evaluate $h(x)$ at a given value of x. We compute both $f(x)$ and $g(x)$. If $f(x) > g(x)$, then we set $h(x) = f(x)$. If $g(x) > f(x)$, then we set $h(x) = g(x)$. If both are equal, we set $h(x)$ to be that equal value. In other words:

$$h(x) := \{ \begin{array}{cc} f(x) & \text{if } f(x) > g(x) \\ g(x) & \text{otherwise} \end{array}$$

For instance, consider the function:

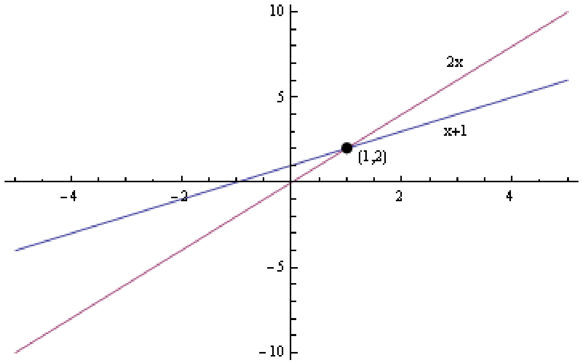
$$h(x) := \max\{x + 1, 2x\}$$

Here are the graphs of the two functions:



At $x = 0$, $x + 1 = 1$ and $2x = 0$, and the maximum of these two values is 1. So $h(0) = 1$.
At $x = 1$, $x + 1 = 2$ and $2x = 2$. Both are equal to 2, so $h(1) = 2$.
At $x = 2$, $x + 1 = 3$ and $2x = 4$. The maximum of these is 4, so $h(2) = 4$.

A similar approach works for calculating the min of two functions.
Here are the graphs of the max and min functions for $x+1$ and $2x$.



Many of the piecewise defined functions that we encounter can naturally be described using the max operator. For instance:

(1) The flat tax function with base exemption B and rate r can be defined as the function $T(I) = \max\{0, r(I - B)\}$.
(2) The absolute value function, that sends x to $|x|$, can be defined as $\max\{x, -x\}$.
(3) The positive part function, that sends x to x^+, can be defined as $\max\{x, 0\}$.

Now, here's a way of thinking of the maximum of two functions. What we need to do is determine, at every point, which one of them is bigger. Think of it as a race. The two functions are constantly racing against each other. At some values of x, one function might come on top, and at other values of x, the other function might come on top.

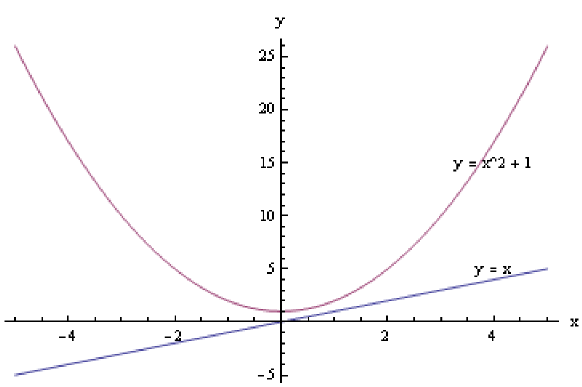
For instance, think of the absolute value function. Let's first plot the graph of the function that sends x to x and the function that sends x to $-x$. The first is a straight line pointing north-east to soth-west, the second is a straight line pointing north-west to south-east.

Now, if you start off at $-\infty$ and move right, the function $-x$ dominates, as is obvious both graphically and algebraically. And it keeps dominating up till the point $x = 0$, where the two functions become equal. After that the function x dominates. Thus, we see that $|x| = -x$ for $x < 0$ and $|x| = x$ for $x \geq 0$.

So the main point is that the places where the functions switch roles are the places where the two functions become equal. [SIDENOTE: Technically, this statement requires both functions to be continuous.] So if we have $h(x) := \max\{f(x), g(x)\}$, then the first thing we should do is find the points where $f(x) = g(x)$. Then, in the intervals between these points, we can try to find out which of the two functions is greater.

For instance, consider the function:

$$f(x) := \max\{x, x^2 + 1\}$$



The first thing you want to know is when those two functions become equal. So you try to solve:

$$x = x^2 + 1$$

which simplifies to:

$$x^2 - x + 1 = 0$$

Now, that function that we just wrote down there has no real roots. You can check this by evaluating the *discriminant* – the $b^2 - 4ac$ term. The discriminant is negative, hence there are no real roots. So this function is never zero.

Thus, one of the two functions, x or $x^2 + 1$, always has the upper hand. Now you can just plug in one value of x and see that $x^2 + 1$ is in fact the upper hand. So, in fact:

$$f(x) = x^2 + 1 \ \forall \ x \in \mathbb{R}$$

By the way, there is another way of showing that $x^2 + 1 > x$ for all $x \in \mathbb{R}$. This is by writing:

$$x^2 - x + 1 = (x - 1/2)^2 + (3/4) \geq 3/4 > 0$$

The secret I used here is *completing the square using the middle term.* This technique will be of importance later when we study integration techniques.

FUNCTIONS: A RAPID REVIEW (PART 2)

MATH 152, SECTION 55 (VIPUL NAIK)

Difficulty level: Easy to moderate. Some of the symmetry concepts (half turn symmetry and mirror symmetry) and some of the proof techniques are likely to be new to students. The rest should be straightforward.

Covered in class?: This will roughly correspond to material covered on Wednesday September 28. Most of the trickier aspects of this will be covered in class, but many small points will be omitted due to time constraints. Hence, it is recommended that you read through these notes either before or after lecture.

Corresponding material in the book: Sections 1.6, 1.7. However, some ideas (including mirror symmetry and half turn symmetry) are not covered in the book. In some other cases, ideas covered later in the book are introduced at this early stage since you are already familiar with calculus.

Corresponding material in homework problems: Homework 1, routine problems 7–8, advanced problems 2–3.

Things that students should get immediately: Definitions of various notions of pointwise combination of functions, scalar multiples of functions, and compositions of functions. Definitions of even, odd, and periodic functions.

Things that students should get with effort: Definitions and graphical interpretations of half turn symmetry and mirror symmetry. Proof techniques related to showing even, odd, and periodic.

Executive summary

For first time reading, skip to the next section.

Words ...

(1) Given two functions f and g, we can define pointwise combinations of f and g: the sum $f + g$, the difference $f - g$, the product $f \cdot g$, and the quotient f/g. For the sum, difference, and product, the domain is the intersection of the domains of f and g. For the quotient, the domain is the intersection of the domain of f and the set of points where g takes a nonzero value.
(2) Given a function f and a real number α, we can consider the scalar multiple αf.
(3) Given two functions f and g, we can try talking of the composite function $f \circ g$. This is defined for those points in the domain of g whose image lies in the domain of f.
(4) One interesting kind of symmetry that we often see in the graph of a function is *mirror symmetry* about a vertical line. This means that the graph of the function equals its reflection about the vertical line. If the vertical line is $x = c$ and the function is f, this is equivalent to asserting that $f(x) = f(2c - x)$ for all x in the domain, or equivalently, $f(c + h) = f(c - h)$ whenever $c + h$ is in the domain. In particular, the domain itself must be symmetric about c.
(5) A special case of mirror symmetry is the case of an *even function*. An even function is a function with mirror symmetry about the y-axis. In other words, $f(x) = f(-x)$ for all x in the domain. (Even also implies that the domain should be symmetric about 0).
(6) Another interesting kind of symmetry that we often see in the graph of a function is *half-turn symmetry* about a point on the graph. This means that the graph equals the figure obtained by rotating it by an angle of π about that point. A point (c, d) is a point of half-turn symmetry if $f(x) + f(2c - x) = 2d$ for all x in the domain. In particular, the domain itself must be symmetric about c. If f is defined at c, then $d = f(c)$.
(7) A special case of half-turn symmetry is an odd function, which is a function having half-turn symmetry about the origin. By definition, the domain of an odd function is symmetric about $\mathbb{R}$. An odd function, if defined at 0, takes the value 0 at 0.

(8) A function f defined on $\mathbb{R}$ is periodic if there exists $h > 0$ such that $f(x+h) = f(x)$ for every $x \in \mathbb{R}$. If there is a smallest $h > 0$ satisfying this, such a h is termed the *period*. Constant functions are periodic but have no period. The sine and cosine functions are periodic with period 2π.

Actions ...

(1) To prove that a function is periodic, try to find a h that *works* for every x. To prove that a function is periodic but has no period, try to show that there are arbitrarily small $h > 0$ that work.
(2) To prove that a function is even or odd, just try proving the corresponding equation for all x. Nothing but algebra.
(3) If a function is defined for the positive or nonnegative reals and you want to extend the definition to negatives to make it even or odd, extend it so that the formula is preserved. So define $f(-x) = f(x)$, for instance, to make it even.

1. Ways of creating new functions from old

1.1. **Pointwise combinations of functions.** Suppose $f, g : \mathbb{R} \to \mathbb{R}$ are functions. By the way, before we proceed, a clarification on notation and terminology. When I say $f : A \to B$ is a function, then the *domain* of the function is A. However, the *range* of the function need not be B. All that notation means is that the range of the function is a *subset* of B. It might be equal to B, but there's no guarantee. And the reason why we allow this kind of latitude is that it makes it a lot easier to write things down if we do not need to calculate the exact range all the time. And, by the way, the set B is termed the *co-domain*.

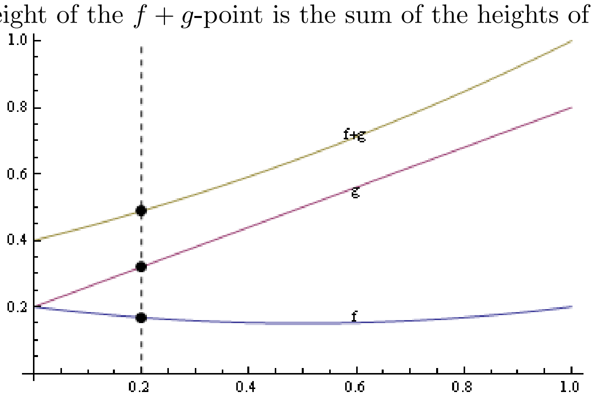
What does $f + g$ mean? So the first thing you might say is: "How can we add functions? I thought we could add only numbers." And the answer is that we don't yet know how to make sense of it, but once we do, it seems intuitive.

So remember, to define $f + g$ as a function, we need to describe where it sends x. So $(f + g)(x)$ is defined as follows:

$$(f + g)(x) := f(x) + g(x)$$

This is the sum of the two functions, and if you stick around in the world of mathematics, you'll hear people say that the sum is defined *pointwise*. What this really means is that to add two functions, what we do is add the *values* of the functions at each *point*.

Here's a picture showing two functions f and g and their sum $f + g$. Note that for each vertical line, the height of the $f + g$-point is the sum of the heights of the f-point and the g-point:



Now, I assumed that the functions are both defined from $\mathbb{R}$ to $\mathbb{R}$. And so, what is the domain of the function $f + g$? Well, it is $\mathbb{R}$ again, because as you can see from the definition, since you can evaluate f and g at a point, you can also evaluate $f + g$ at that point. And, by the way, I use the word *point* where I actually mean *real number* – secretly, I'm thinking of real numbers as points on the number line.

What if f is a function defined on a smaller domain (i.e., a subset of $\mathbb{R}$) and g on another smaller domain (i.e., another subset of $\mathbb{R}$)? In that case, $f + g$ is defined on the *intersection* of dom(f) and dom(g). Why the intersection? Because to evaluate $f + g$ at a point, you need to evaluate f at the point *and* g at the point, and then add those values. And to be able to evaluate *both*, the input should be in the domain of both functions.

We similarly define:

$$(f - g)(x) := f(x) - g(x)$$
$$(f \cdot g)(x) := f(x)g(x)$$
$$(f/g)(x) := f(x)/g(x)$$

For the case of the difference and product, the domain is the intersection of the domains. For the ratio, or quotient, we need to be a little more careful: the domain of the new function is inside the intersection of the domains of f and g, but there's a caveat: we need to exclude points at which $g = 0$.

1.2. **Scalar multiples of functions.** Suppose f is a function and α is a real number. The function αf is defined as:

$$(\alpha f)(x) := \alpha f(x)$$

For instance, $2f$ is the function that sends x to $2f(x)$, while $-f$ is the function sending x to $-f(x)$.

1.3. **Composition of functions.** Suppose $f, g : \mathbb{R} \to \mathbb{R}$ are functions. Then $f \circ g$ is defined as the following function:

$$(f \circ g)(x) := f(g(x))$$

$f \circ g$ is termed the *composition* of the functions f and g. Orally, we say "f composed with g". Note that the function written on the right is the one the we apply *first*, so in the case of function composition, we work from right to left. This can be potentially confusing.

We can also define the composite of two functions when their domains are subsets of $\mathbb{R}$. The domain of the composite $f \circ g$ is that subset of $\text{dom}(g)$ whose image under g lies inside $\text{dom}(f)$. There is a more precise way of expressing this, but it will take us too far afield, so we will skip it.

1.4. **Are there other ways of creating new functions?** Yes, but we will see them later. The most significant of these are differentiation, integration, and taking inverse functions.

1.5. **Why do ways of creating new functions from old matter?** First, of course, ways of creating new functions from old help us create new functions from old. However, just as new food recipes are of little interest to those unenthusiastic about cooking and eating, new function recipes may seem pointless to those unenthusiastic about playing with functions. There is a deeper reason.

The point is that these ways of creating new functions from old are *already in use when we think of and create new functions*. By *explicitly identifying* the various recipes used to create new functions from old, we hope to get a better mental model of functions that *already exist*. Both pointwise combination and composition are implicitly used all the time without our even knowing it. Making them explicit is like writing down an explicit recipe for a dish that we've already been cooking and eating.

We will better be able to understand a new phenomenon for *all functions* when we are able to break the process of such understanding into two steps: (i) understanding the phenomenon for a list of basic building block functions, (ii) understanding how the phenomenon interacts with the recipes for creating new functions from old. For instance:

(1) In order to learn how to differentiate functions, we do two things: (i) learn formulas for differentiating a list of basic functions (e.g., derivatives of power functions, trigonometric functions, etc.) (ii) learn formulas for the derivative of a new function created by a recipe from other functions, in terms of those other functions and their derivatives (e.g., derivatives of sums, differences, scalar multiples, product rule, quotient rule, and chain rule).
(2) In order to learn how to integrate functions, we do two things: (i) learn integration formulas for a list of basic functions (ii) learn procedures for integrating complicated functions in terms of their basic building blocks (unfortunately, the rules for product and composition are not straightforward, making integration a much more messy and also much more interesting business).

(3) To prove that all functions in a particular collection satisfy a property such as continuity or differentiability, it suffices to prove the property for the basic building blocks of the collection, and then to prove that the various ways of building new functions from old within the collection preserve the property.

2. Symmetries of functions

2.1. **Even and odd functions.** Let's first discuss the concept of even function and odd function for globally defined functions, i.e., functions defined for *all* real numbers.

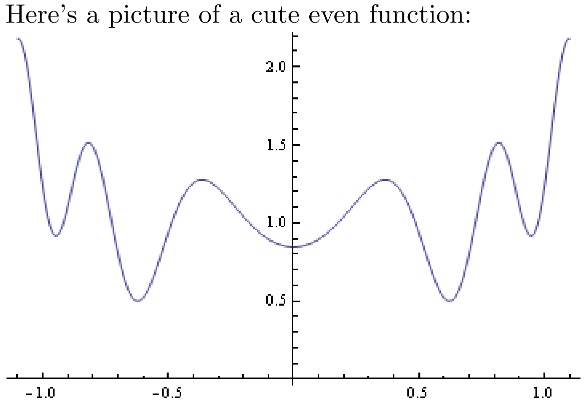
By the way, as I pointed out earlier, when I say $f : \mathbb{R} \to \mathbb{R}$, you should *not* assume that the range is $\mathbb{R}$. I just mean a globally defined function that takes real values.

So we say that f is an *even function* if:

$$f(x) = f(-x)\ \forall\ x \in \mathbb{R}$$

So, what does this mean from the point of view of its graph? Well, it turns out that this is equivalent to saying that the graph is symmetric about the y-axis.

Here's a picture of a cute even function:

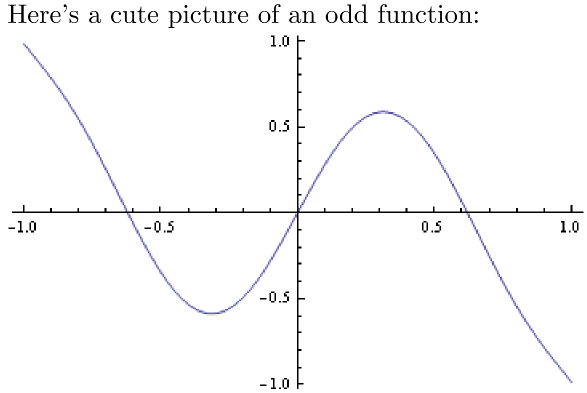


We say that f is an *odd function* if:

$$f(-x) = -f(x)\ \forall\ x \in \mathbb{R}$$

This is equivalent to saying that the graph has a *rotational* symmetry about the origin. If you rotate the graph by π (that's 180°) you get back to the original thing.

Here's a cute picture of an odd function:



The notion of even and odd function also makes sense for functions whose domain is not the whole real numbers, but rather, is a subset of the real numbers. The notion makes sense only when the domain is *symmetric* about 0, i.e., whenever x is in the domain of the function, so is $-x$. Some examples of domains symmetric about 0 are: intervals of the form $[-a, a]$, intervals of the form $(-a, a)$, intervals of the form $(-a, a) \setminus \{0\}$, intervals of the form $[-a, a] \setminus \{0\}$, the set of all integers $\mathbb{Z}$, the set of all rational numbers $\mathbb{Q}$, a union of intervals of the form $(-b, -a) \cup (a, b)$, and many more.

2.2. **Mirror symmetry.** We say that a function f possesses mirror symmetry about the line $x = c$ if the domain of f is symmetric about c and, for all $x \in \mathrm{dom}(f)$, we have:

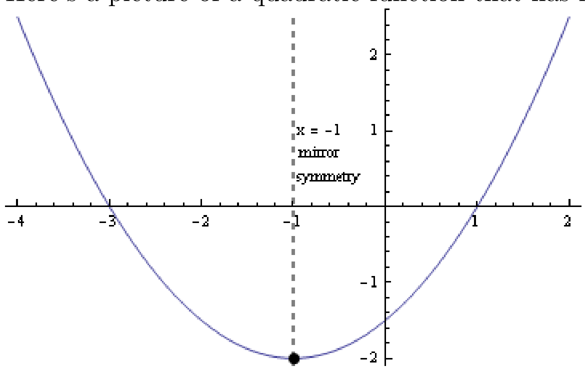
$$f(x) = f(2c - x)$$

Equivalently, for all $h > 0$, $c + h \in \mathrm{dom}(f)$ if and only if $c - h \in \mathrm{dom}(f)$, and if so, then:

$$f(c + h) = f(c - h)$$

Even functions are a special case: they have mirror symmetry about the y-axis.

Here's a picture of a quadratic function that has mirror symmetry about the line $x = -1$.



2.3. **Half turn symmetry.** We say that a function f possesses half turn symmetry about the point (c, d) if the domain of f is symmetric about c and, for all $x \in \mathrm{dom}(f)$, we have:

$$f(x) + f(2c - x) = 2d$$

Equivalently, for all $h > 0$, $c + h \in \mathrm{dom}(f)$ if and only if $c - h \in \mathrm{dom}(f)$, and if so, then:

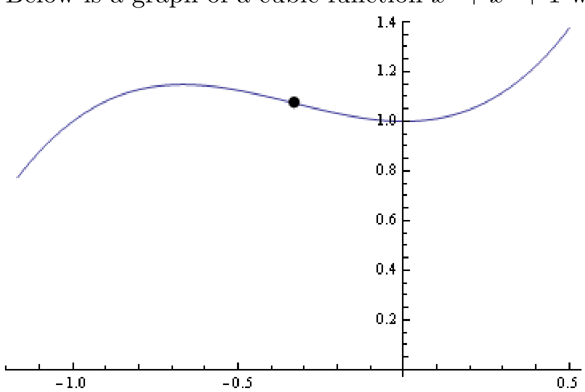
$$f(c + h) + f(c - h) = 2d$$

In other words, the point (c, d) is the midpoint between $(c + h, f(c + h))$ and $(c - h, f(c - h))$.

If $c \in \mathrm{dom}(f)$, then we are forced to have $d = f(c)$.

Odd functions are a special case with the point of half turn symmetry about the origin $(0, 0)$.

Below is a graph of a cubic function $x^3 + x^2 + 1$ with half turn symmetry about the point $(-1/3, 29/27)$.



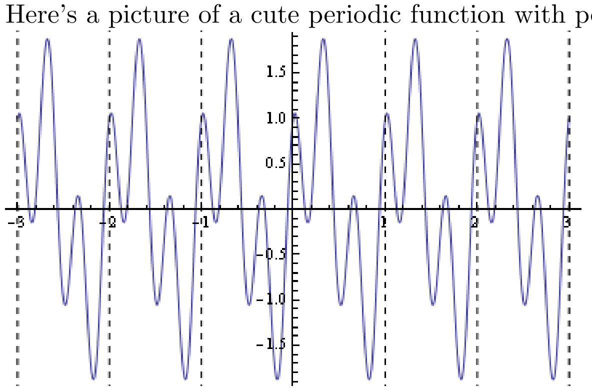
2.4. **Periodic functions.** Suppose $f : \mathbb{R} \to \mathbb{R}$ is a function. We say that f is a *periodic function* if there exists a $h > 0$ such that:

$$f(x + h) = f(x) \; \forall \, x \in \mathbb{R}$$

The *period* (more correctly called the *fundamental period*) of f is the smallest $h > 0$ for which the above holds (for all $x \in \mathbb{R}$).

The trigonometric functions are examples of periodic functions. For instance, sin and cos have period 2π. What about the $\sin^2$ function? Well, 2π works, but it isn't the smallest thing that works. The smallest h that works is π.

Here's a picture of a cute periodic function with period 1:



2.5. **Other notions of symmetry.** There are many other notions of symmetry for functions that we will encounter as we start drawing graphs. The most significant of these is the periodic + linear symmetry, which is observed for functions that can be expressed as a sum of a periodic function and a linear function. More on this later.

2.6. **Why do notions of symmetry matter?** Notions of symmetry are important for a number of reasons, including the following:

(1) For functions which possess symmetry, graphing the function can be a lot easier since the symmetry allows us to fill in the graph at many points based on a small part.
(2) Symmetry allows us to deduce properties about derivatives of the function.
(3) Symmetry allows us to deduce properties about definite integrals. Often, definite integrals can be computed using symmetry properties even though antiderivatives are hard or impossible to express explicitly.

3. Proving and reasoning involving these functions

3.1. **Proof positive: showing something to be even, odd, or periodic.** To show that a function f is even, we start with a *generic* x, compute $f(x)$ and $f(-x)$, and show that both are equal.

To show that a function f is odd, we start with a *generic* x, compute $f(x)$ and $f(-x)$, and show that the results are negatives of each other.

Showing that a function f is periodic is somewhat trickier. f is defined to be periodic if there exists $h > 0$ such that $f(x+h) = f(x)$ for all x in the domain of f. Thus, to show that f is periodic, we first need to find a value of h that works. After we have chosen a specific numerical value of h, we then pick a *generic* x and show that $f(x+h) = f(x)$.

In logic notation, periodicity states that:

$$\exists h > 0 \text{ such that } \ \forall\ x \in \text{dom}(f), f(x+h) = f(x)$$

The $\exists$ stands for an existential quantifier and the $\forall$ stands for a universal quantifier. For existentially quantified variables, we need to come up with a specific value that "works" while for universally quantified variables, we need to show that every value works, which we do by picking a *generic* value.

3.2. **Relation between symmetry and creation of new functions.** Here are some important facts that can be proved using the techniques mentioned in the previous subsection:

(1) The set of even functions is closed under addition, subtraction, scalar multiples, pointwise multiplication, and pointwise division (where defined). All constant functions are even. [Sidenote: In mathematical jargon, we say that even functions form an algebra.]
(2) The set of odd functions is closed under addition, subtraction, and scalar multiples. It is also closed under composition.

(3) A product of two odd functions is even.
(4) A product of an even function and an odd function is odd.
(5) A composite $f \circ g$ where g is even is also even.
(6) If f is even and g is odd, then the composite $f \circ g$ is even.
(7) If f_1 and f_2 are periodic functions with periods h_1 and h_2 such that h_1/h_2 is a rational number, then $f_1 + f_2$, $f_1 - f_2$, and $f_1 \cdot f_2$ are all periodic functions.
(8) If f and g are functions such that g is periodic, so is $f \circ g$.

3.3. **Negative proofs: not even, not odd, not periodic.** To show that a function f is not even, it suffices to find just one counterexample, i.e., to find one value of x such that both x and $-x$ are in the domain of f but $f(-x) \neq f(x)$. A similar technique works for showing that f is not odd.

Let's look at an example Consider the function:

$$f(x) := x^2 - x + 1$$

Claim. f is not an even function.

Proof. If f were an even function, then we would have, for every $x \in \mathbb{R}$, that $f(x) = f(-x)$. Thus, to show that f is not an even function, it suffices to find one value of x at which $f(x) \neq f(-x)$.

In fact, the value $x = 1$ suffices:

$$f(1) = 1, \qquad f(-1) = 3$$

So clearly $f(1) \neq f(-1)$. □

Claim. f is not an odd function.

Proof. If f were an odd function, then we would have, for every $x \in \mathbb{R}$, that $f(x) = -f(-x)$. Thus, to show that f is not an odd function, it suffices to find one value of x at which $f(-x) \neq -f(x)$.

In fact, the value $x = 1$ suffices:

$$f(1) = 1, \qquad f(-1) = 3$$

So clearly $f(-1) \neq -f(1)$. □

Showing that a function is not periodic is trickier. Recall that f being periodic is equivalent to the following:

$$\exists h > 0 \text{ such that } \ \forall\, x \in \operatorname{dom}(f), f(x+h) = f(x)$$

Showing this to be false would entail showing that there is *no value* of h that works in the above. Equivalently, we need to show that every value of h fails. Thus, we want to show the following:

$$\forall h > 0, \exists x \in \operatorname{dom}(f) \text{ such that } f(x+h) \neq f(x)$$

Note that the $\exists$ quantifier gets replaced by a $\forall$ quantifier and the $\forall$ quantifier gets replaced by a $\exists$ quantifier. This is a universal feature of logical negation, and shall be crucial to a clear understanding of $\epsilon - \delta$ proofs that we will encounter soon in this course. Let's now show that the function $f(x) := x^2 - x + 1$ is not a periodic one.

Claim. f is not a periodic function.

The proof technique we use here is what is called *proof by contradiction*. What we do is start out by assuming that f is a periodic function and then do some work and show that we have come up with something that is obviously false.

Proof. Suppose f were a periodic function. By the definition of periodic function, there exists $h > 0$ such that:

$$f(x+h) = f(x) \; \forall \; x \in \mathbb{R}$$

Simplifying this, we obtain that:

$$\begin{aligned}
(x+h)^2 - (x+h) + 1 &= x^2 - x + 1 \\
\implies (x+h)^2 - x^2 + x - (x+h) &= 0 \\
\implies 2xh + h^2 - h &= 0 \\
\implies h(2x+h-1) &= 0 \\
\implies 2x+h-1 &= 0 \text{ (using } h > 0\text{, so } h \neq 0) \\
\implies x &= \frac{1-h}{2}
\end{aligned}$$

Thus, there is *exactly* one value of x, namely $x = (1-h)/2$, such that $f(x+h) = f(x)$. Thus, it is certainly not true that $f(x+h) = f(x)$ for *all* $x \in \mathbb{R}$, and we have the desired contradiction. So, f is not a periodic function. □

3.4. **Extending the domain with even/odd/periodic constraint.** Given a function f defined on the nonnegative reals, there is a unique way of extending the domain of f to all reals to obtain an even function. Similarly, if in addition $f(0) = 0$, there is a unique way of extending the domain of f to all reals to obtain an odd function.

For $x < 0$, the even way of extending *defines* $f(x)$ as equal to $f(-x)$, and the odd way of extending defines $f(x)$ as equal to $-f(-x)$. Graphically, for even functions, the part of the graph of the function for $x < 0$ is obtained by reflecting about the y-axis the part of the graph of the function for $x > 0$. For odd functions, the $x < 0$ part of the graph is obtained from the $x > 0$ part of the graph by performing a half turn about the origin.

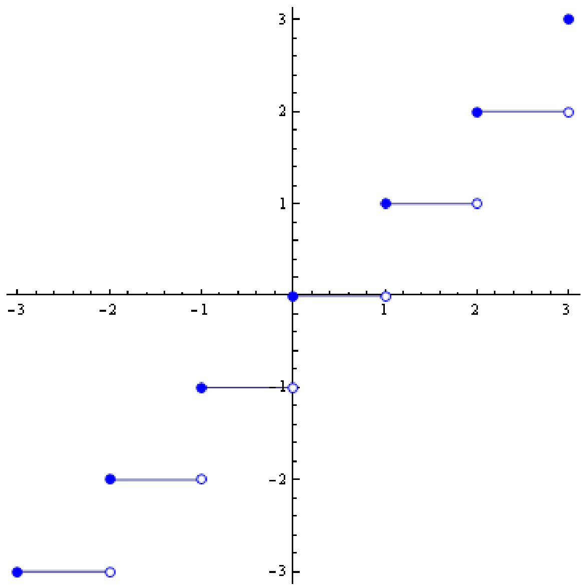
Similarly, given a function defined on a closed interval $[a, b]$ such that $f(a) = f(b)$, we can extend f uniquely to a periodic function for which $h = b - a$ works.

4. More offbeat functions

4.1. **Greatest integer function and fractional part function.** The *greatest integer function*, denoted by [], is defined as follows. For $x \in \mathbb{R}$, the greatest integer function of x, denoted $[x]$, is defined as the greatest integer less than or equal to x. Thus, $[3] = 3$, $[\pi] = 3$, $[0.6] = 0$, $[\sqrt{47}] = 6$, $[-\sqrt{2}] = 2$, $[-7/3] = 3$, and so on.

The greatest integer function is a *piecewise constant function* or *step function* and it has a discontinuity at every integer, with an upward step size of 1. The greatest integer function is also termed the *floor function*.

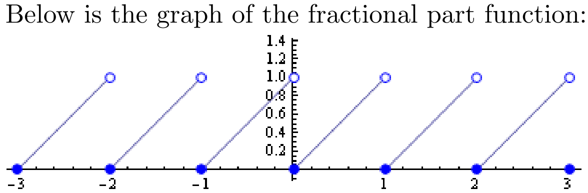
Here is the graph of the greatest integer function:



Closely related to the greatest integer function is the *fractional part function*. The fractional part of x, denoted $\{x\}$, is defined as $x - [x]$. Thus, the fractional part is 3.42 is 0.42, while the fractional part of -0.42 is 0.58.

The fractional part function is piecewise linear, with discontinuities at every integer. Between consecutive integers n and $n + 1$, the function rises linearly from 0 to 1, but just when it is about to reach 1, it slips back down to 0 to start all over again.

Below is the graph of the fractional part function:



4.2. **Functions defined differently for rationals and irrationals.** For the piecewise definitions of functions that we have seen so far, the *pieces* are intervals or unions of intervals, and thus there are points at the boundaries between the pieces where the function can be thought of as *changing* definition. There is a much more messy kind of piecewise definition, where the pieces do not look like intervals or unions of intervals, but are instead scattered across the domain.

One example is where the pieces are taken to be the rational numbers and irrational numbers respectively. both the rational numbers and irrational numbers are dense in the real numbers – in other words, every nonempty open interval in the reals contains both rational and irrational numbers.

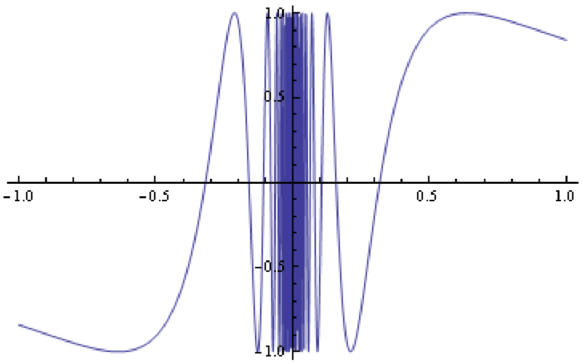
For instance, the *Dirichlet function* is defined as:

$$f(x) := \left\{ \begin{array}{ll} 1 & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational} \end{array} \right.$$

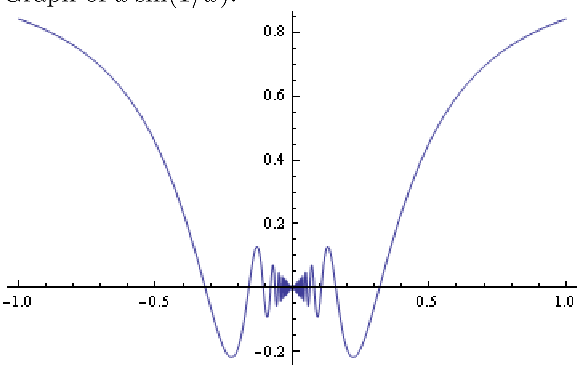
There are some variants of this where there is one constant value (not necessarily 1) for the rational numbers and another constant value (not necessarily 0) for the irrational numbers. There are also other variants that you will see as we explore continuity and differentiability further.

4.3. **The topologist's sine curve.** We will also be looking at the functions $\sin(1/x)$, $x\sin(1/x)$, $x^2\sin(1/x)$, $x^3\sin(1/x)$ throughout the course. Graphs of these functions are given below.

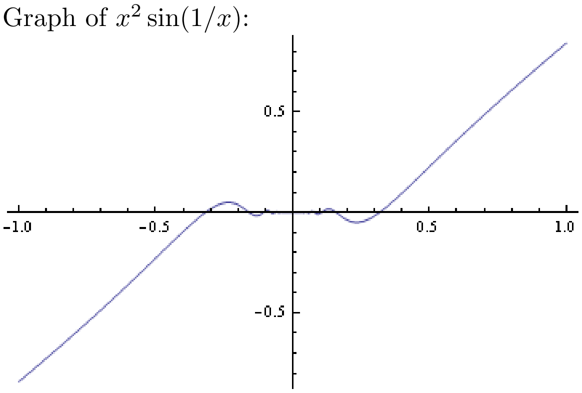
Graph of $\sin(1/x)$:



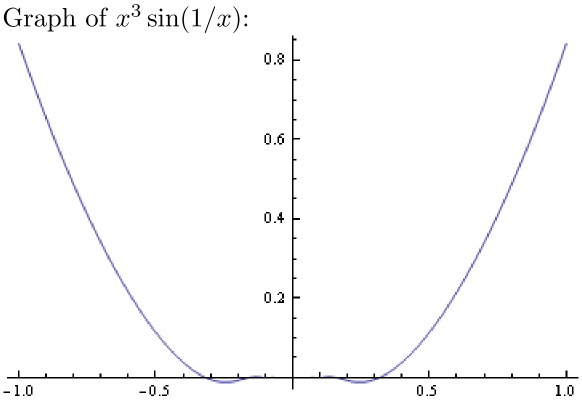
Graph of $x \sin(1/x)$:



Graph of $x^2 \sin(1/x)$:



Graph of $x^3 \sin(1/x)$:



4.4. **Why do we care about weird functions?** The greatest integer function and fractional part function have applications to real world situations, particularly when those real world situations have integer constraints. For instance, you can only buy and sell integer quantities of some commodity.

However, the rational-irrational dichotomy functions and the topologist's sine curve have very few practical applications as functions. Their main utility is to provide *examples that allow us to test the soundness of definitions and notions of continuity and differentiability.* Most of the natural examples of functions are too nice for us to test whether our definitions of continuity and differentiability can rough it out. You can think of them as the equivalent of high school bullies who make you a strong person, as long as you don't cave in to them.

INFORMAL INTRODUCTION TO LIMITS

MATH 152, SECTION 55 (VIPUL NAIK)

Corresponding material in the book: Section 2.1, parts of Sections 2.4.

Difficulty level: Easy to moderate, assuming you have seen some intuitive concept of limits before.

Covered in class?: Probably not. We may go over some small part of this quickly before covering $\epsilon - \delta$ definitions of limits.

Things that students should definitely get: To define limits, you need to get really really close. There are two directions from which to approach a real number: left and right. The notation for limits and one-sided limits. The intuitive meaning of the existence of limits and of continuity,, one-sided continuity, continuity on intervals. The notions of removable and jump discontinuity.

Things that students should hopefully get: The allusion to why taking limits for functions on a plane is harder because of multiple directions of approach.

Executive summary

Words ...

(1) On the real line, there are two directions from which to approach a point: the *left* direction and the *right* direction.
(2) For a function f, $\lim_{x \to c} f(x)$ is read as "the limit as x approaches c of $f(x)$. Equivalently, as x approaches c, $\lim_{x \to c} f(x)$ is the value that $f(x)$ approaches.
(3) $\lim_{x \to c} f(x)$ makes sense only if f is defined *around* c, i.e., both to the immediate left and to the immediate right of c.
(4) We have the notion of the *left hand limit* $\lim_{x \to c^-} f(x)$ and the *right hand limit* $\lim_{x \to c^+} f(x)$. The *limit* $\lim_{x \to c} f(x)$ exists if and only if (both the left hand limit and the right hand limit exist and they are both equal).
(5) f is termed *continuous* at c if c is in the domain of f, the limit of f at c exists, and $f(c)$ equals the limit. f is termed *left continuous* at c if the left hand limit exists and equals $f(c)$. f is termed *right continuous* at c if the right hand limit exists and equals $f(c)$.
(6) f is termed *continuous* on an interval I in its domain if f is continuous at all points in the interior of I, continuous from the right at any left endpoint in I (if I is closed from the left) and continuous fromthe left at any right endpoint in I (if I is closed from the right).
(7) A *removable discontinuity* for f is a discontinuity where a two-sided limit exists but is not equal to the value. A *jump discontinuity* is a discontinuity where both the left hand limit and right hand limit exist but they are not equal.

Actions: See the procedure in the last subsection on computing limits for polynomial and rational functions.

Note: These notes cover only the informal and intuitive concept of limits that you should be familiar with, and do not include the $\epsilon - \delta$ definitions. The $\epsilon - \delta$ definitions are covered in subsequent notes which we will go through very carefully in class.

1. Intuitive conception of limits

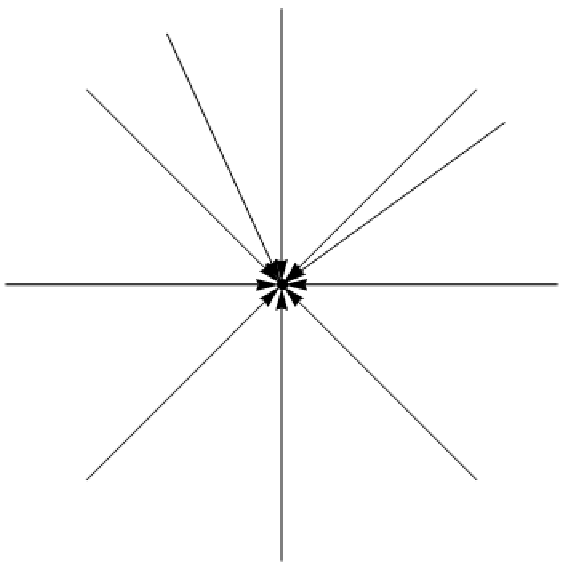
1.1. **The real numbers and two-sidedness of approach.**

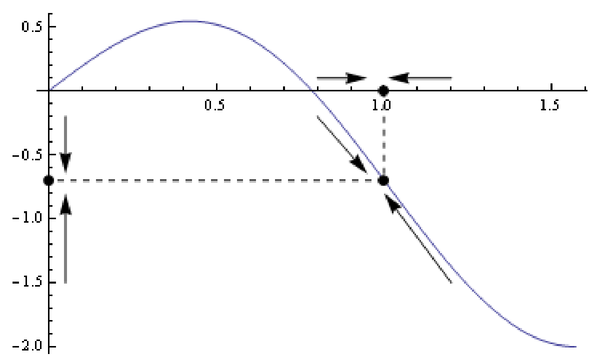
The first thing you need to know is that in order to understand limits, you really need to appreciate the real numbers. There's something particularly interesting about the real numbers: you can get *really really close* to a real number without equaling it. That's not something you can do with more sparse sets such as the integers.

It is this ability to sneak really close to something without being equal to it that allows us to talk of limits. This is something you should keep in mind – we'll get back to it later again when we talk of another kind of limit in 153 – the limit of a sequence. That's a very different but in some ways remarkably similar notion, but we'll come to it in due course.

Now, in the picture, I sneaked up on this number from one side – the left side. But I could sneak up to it from another side – the right side. This two-sidedness makes things pretty interesting. By the way, this is one advantage of dealing with a line – there are only two directions to worry about. Imagine, just imagine, if you were dealing with a plane. Then there would be the left side, the right side, the up side, the down side, this side, that side – too many! Luckily for us, we can postpone all those headaches for multivariable calculus, which is beyond the scope of the 150s. So we focus right now on this simple real line.

Here's the kind of picture you can avoid thinking about for now:





1.2. **Beginning and verbal gymnastics.**

So let's be really abstract. Suppose f is a function from a subset of the reals to a subset of the reals and c is a real number. What we would like to know is the answer to this question: as x gets really close to c, what does $f(x)$ get close to? If $f(x)$ is heading towards a specific destination, that's called its limit, and it has the notation:

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

This is read as "the limit as x approaches c of $f(x)$". An equation such as:

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

can be read in two ways: "the limit as x approaches c of $f(x)$ is b" or "as x approaches c, $f(x)$ approaches b". By the way, some people say *tends to* instead of *approaches*. Some people say *goes to* and those who're living at the point c may say *comes to*.

Now, let's take some examples. Suppose $f(x) = x$. So f is what is called the *identity function*. It is like a mirror that gives back what is put into it. Well, what then is $\lim_{x\to 0} f(x)$? Well, $f(x) = x$, so this is $\lim_{x\to 0} x$. So this reads like a word puzzle: "as x tends to 0, what does x tend to?" Of course, 0. In fact, more generally,

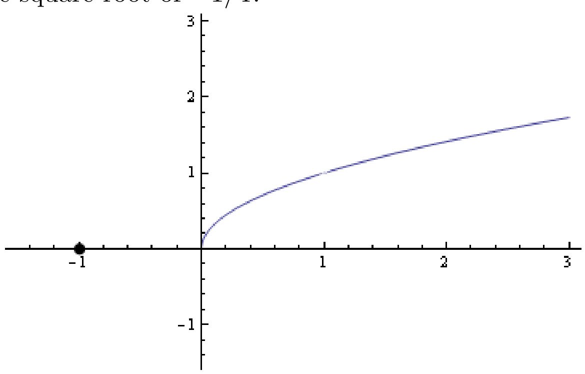
$$\lim_{x\to c} x = c$$

1.3. **Beyond our reach: can't limit to what you can't approach.** Okay, here's the next question: what is $\lim_{x\to -1} \sqrt{x}$. By the way, remember that $\sqrt{x}$ is defined as the nonnegative square root. So what is this limit? In other words, as x approaches, gets really really close to -1, what does $\sqrt{x}$ approach?

Verbal gymnastics doesn't work the same way as it does for the previous limit, so this one requires some serious thought. Okay, to simplify matters, let's first tackle the right side and then the left side.

Okay, let's try the right side. Let's start from far off. What is the square root of 4? It's 2. What is the square root of 3? $\sqrt{3}$ is approximately $1.732\dots$. Square Root of 2 is $1.414\dots$, square root of 1 is 1, square root of 0 is 0. Hmmm. So the square root seems to be decreasing. So you might guess right now that the limit is some negative number.

But to check this guess, you need to go down the negative aisle. And because there's a paucity of integers, we need to use fractional numbers. So let's try some negative number between 0 and -1. Say $-1/4$. What's the square root of $-1/4$?



It doesn't *have a square root*. It's a negative number. In fact, the domain of the square root function is the nonnegative reals, the interval $[0, \infty)$. And that's bad. Which means that as we get even a little close to -1, we cannot evaluate the function from the right side. So the limit from the right side doesn't make sense.

The limit from the left side doesn't make sense either, because the function isn't defined *anywhere* to the left of -1.

What's the message to take from this? It makes sense to talk of the limit of a function at a point, if the function is defined at places very close to the point. If it isn't, it's like, as some people say, putting "lipstick on a pig." You can take the limit of a function that doesn't exist, and it still doesn't exist.

1.4. **One-sided limits.** There are two further notions, that are mirror images of each other: the *left hand limit* and the *right hand limit*. The left hand limit at a is denoted as $\lim_{x\to a^-} f(x)$ and $\lim_{x\to a^+} f(x)$.

The left hand limit of a function is the limit as you approach the domain value from the left side. The right hand limit of a function is the limit as you approach the domain value from the right side. If a function is defined on both the left and the right side of a point, there are five possibilities:

(1) Neither the left hand limit nor the right hand limit exist.
(2) The left hand limit exists but the right hand limit does not exist.
(3) The left hand limit does not exist but the right hand limit exists.
(4) Both the left hand and the right hand limit exist, but they are not equal.
(5) Both the left hand and the right hand limit exist, and they are equal. In this case, we say that the function *has a limit* and the limit is equal to both these values.

Phew! What a wide range of possibilities! But you should be happy that there are only two directions of approach: left and right. If and when you study multivariable calculus, you'll be studying functions on a plane, where you have not two, but *infinitely many* directions. If computing limits from two directions is a headache, computing limits from infinitely many directions is like enduring torture for eternity.

Things are a little different for values that are at extreme ends of the domain. For instance, think about the function $f : [0, 1] \to [0, 1]$ given by $f(x) = \sqrt{1 - x^2}$. Now, at the point -1, a left hand limit doesn't make sense because the function is not defined to the left of -1. So, only the right hand limit does. Similarly at the point 1, the right hand limit doesn't make sense but the left hand limit does.

There's a little confusion about conventions in what I'm going to say, but I'll just stick with the book on this one: if the point c is at the boundary of the domain and so the function isn't defined on one side, the book says that talking of the limit at c, or writing $\lim_{x\to c} f(x)$, is not meaningful. However, we can talk of the one-sided limit from the side that the function is defined. You may see a different convention at other places, but we'll just stick to the book for now. That means that if the point is at the boundary of the domain, you should clearly specify a one-sided limit instead of just taking *the limit*.

2. Continuity

2.1. **The concept of continuity.**

Suppose f is a function and it is defined *around* a point a, i.e., f is defined at the point a and is also defined in some open interval containing a. Then f is continuous at a if the limit of f exists at a and equals $f(a)$. This means that the left hand limit and the right hand limit of f exist at a and are equal to $f(a)$. In symbols:

$$f \text{ continuous at } a \iff \lim_{x\to a^-} f(x) = \lim_{x\to a^+} f(x) = f(a)$$

If f is defined at a and on the *immediate right* of a, then we say that f is *right continuous* or *continuous from the right* at a if the right hand limit of f at a equals $f(a)$. In symbols:

$$f \text{ right continuous at } a \iff \lim_{x\to a^+} f(x) = f(a)$$

If f is defined at a and on the *immediate left* of a, then we say that f is *left continouus* or *continuous from the left* at a if the left hand limit of f at a equals $f(a)$. In symbols:

$$f \text{ left continuous at } a \iff \lim_{x\to a^-} f(x) = f(a)$$

2.2. **Continuity on an interval.**

Suppose I is an interval (open, closed, half-open half-closed, possibly infinite in one or both directions). A function f whose domain contains I is termed *continuous* on I if f is continuous for all *interior* points of I (i.e., all points of I that are not at the boundary of I) and is continuous from the appropriate side at all boundary points. We consider all cases below:

1. If $I = (a, b)$, f needs to be continuous at all points of I.
2. If $I = [a, b]$, f needs to be continuous at all points of (a, b), right continuous at a, and left continuous at b.
3. If $I = [a, b)$, f needs to be continuous at all points of (a, b) and right continuous at a.
4. If $I = (a, b]$, f needs to be continuous at all points of (a, b) and left continuous at b.
5. If $I = (a, \infty)$ or $(-\infty, b)$ or $(-\infty, \infty)$, f needs to be continuous at all points of I.
6. if $I = [a, \infty)$, f needs to be continuous at all points of (a, ∞) and right continuous at a.
7. If $I = (-\infty, b]$, f needs to be continuous at all points of $(-\infty, b)$ and left continuous at b.

3. Plumbing leaks

3.1. **Filling in the hole in the FORGET function.**

I hope you remember the *FORGET* function that we defined a little earlier:

$$FORGET(x) = \frac{x}{x}$$

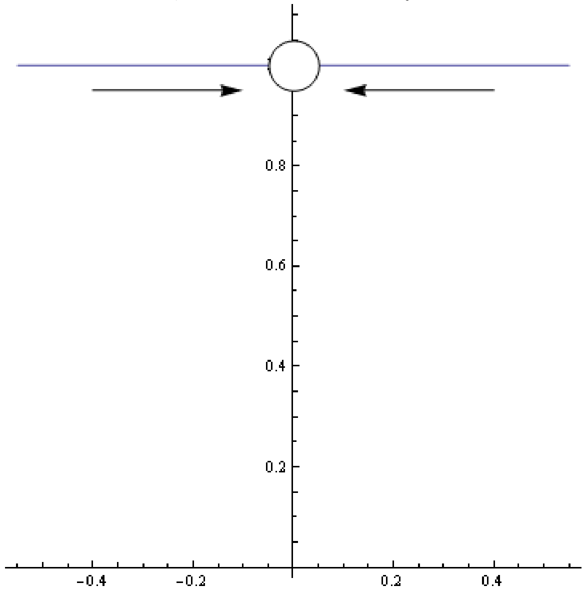
When I defined this function, we discussed that the function is *not* defined at zero. Why? Because at 0, when we plug in, we get a 0 in the numerator and a 0 in the denominator. Zero in the denominator is bad! This expression makes no sense. So forget about evaluating this function at 0.

However, it definitely makes sense to ask whether the *limit* exists at 0:

$$\lim_{x \to 0} FORGET(x)$$

Why does it make sense? Because the function is defined at all points other than 0, it is defined at all points that are close to 0 but not equal to it. It's defined at all points to the left of 0 and at all points to the right of 0. The *only* point where it is not defined is 0. So, it makes sense to evaluate the limit at 0.

It makes sense, but can we actually do this? Well, let's use the graph.



Okay, so we see that the function is 1 to the left of zero, 1 to the right of zero. What should the limit be? If there's any justice in the world, it should be 1. And it is.

Let's see this mathematically:

$$FORGET(x) = \frac{x}{x}, \qquad x \neq 0$$

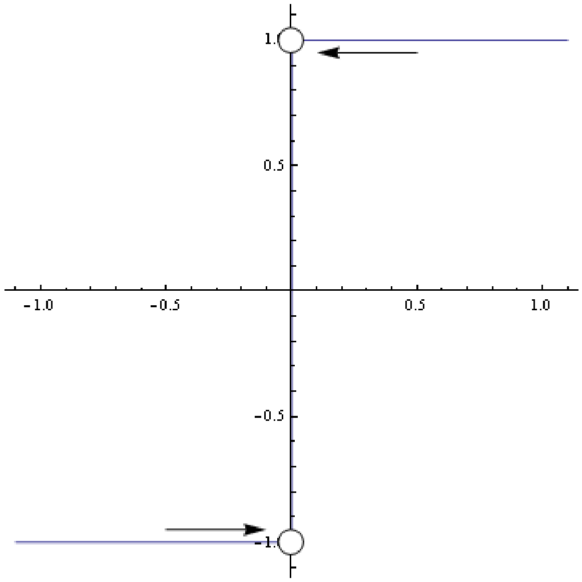
Thus:

$$\lim_{x \to 0} FORGET(x) = \lim_{x \to 0} \frac{x}{x} = \lim_{x \to 0} 1 = 1$$

Now you may say: *why can we cancel now?* The reason why we can cancel x in that step above is that now, since we are *only approaching* 0 and *are not equal to it*, x can be canceled. And that is the beauty of limits. The thing that gives you trouble *at* a point doesn't give you any trouble *near* the point, and because you are sneaking up from nearby rather than evaluating at the point, you evade trouble. It's the roundabout maneuver when a direct assault fails.

3.2. **Looking back at the signum function.** The signum function, denoted sgn, is the function defined as $x/|x|$ when $x \neq 0$. Under some conventions, it is considered undefined at $x = 0$. Under other conventions, its value at 0 is defined to be 0.

The signum function is continuous – in fact, locally constant – at all nonzero points in the domain. At the point 0, it takes the value -1 everywhere on the left and it takes the value 1 everywhere on the right. Thus, the left hand limit is -1 and the right hand limit is 1. Thus, the function jumps from the value -1 to 1 at 0.



The signum function differs from the FORGET function in this important respect: for the FORGET function, we could *fix* or *remove* the discontinuity at 0 by filling in the value 1, because the *limit at* 0 *exists*. However, for the signum function, there is no way of fixing the discontinuity at 0 because the limit does not exist, which happens in turn because the left and right hand limits differ.

3.3. **Kinds of discontinuities.** We now consider some important kinds of *discontinuities* for functions, i.e., situations where a function f is defined aound a point c but is not continuous at c. Here are two important kinds of discontinuities:

(1) *Removable discontinuities* are discontinuities where the limit of f at c exists but is not equal to $f(c)$. There could be two reasons for this: either $f(c)$ is not defined (as for the FORGET function) or it is defined but is not equal to the limit.
(2) *Jump discontinuities* are discontinuities where both the left hand limit and the right hand limit exist and are finite, but they are not equal. Note that the value of the function at c can be changed to make the function continuous from the left at c. It can also be changed to make the function continuous from the right at c. But we cannot choose a value $f(c)$ such that both things happen simultaneously. An example is the signum function.

These are not the only possibilities. There are also infinite discontinuities (where the left hand limit or the right hand limit or both is/are $\pm\infty$) and oscillatory discontinuities. We will return to this topic later.

Aside:Left and right are based on source, not target. The left hand limit is the limit where the approach is *from* the left, but the direction in which the approach happens is *toward* the right. In other words, the choice of hand is based on the direction *from* which approach is being made rather than the direction in which the approach is happening.

A similar convention is followed when specifying the direction of winds. A *northern* wind is a wind that blows *from* north to south. If you took geography in school or read weather forecasts, you should be familiar with this.

[I have a guess as to the reason for choosing this convention, but it's purely speculative, so I won't include it here.]

4. Our nice cocooned world

Living on Planet Earth in an age of affluence, we are used to taking niceties for granted. Of course, if we consider the whole world throughout history, poverty is more the default condition of humans than affluence.

In the same way, the functions you've been dealing with so far are nice and sweet. For all their complications and complexities, they don't throw tantrums. In this course, we'll continue to deal, for the most part, with nice functions, but we'll explore the more rugged terrain every once in a while to appreciate our good fortune and the boundaries of our understanding. For now, let's review how good we have it and how to handle the occasional hiccup.

4.1. **Limits of polynomial and rational functions.** All polynomial functions are continuous, so the limit of a polynomial function at a point equals the value of the polynomial function at that point. In other words, if f is a polynomial function and c is a number, then $\lim_{x \to c} f(x) = f(c)$.

For rational functions, we evaluate the limit of a rational function f at a point c using the following rules:

(1) First, try to evaluate the numerator and the denominator at the point. If the denominator is nonzero at the point, the limit equals the value. If the denominator evaluates to 0 at the point, and the numerator evaluates to something nonzero at the point, then the limit is not defined. If both the numerator and the denominator evaluate to 0 at the point, then *more work is needed.*

(2) If both the numerator and the denominator are 0 at the point c, then there is a factor of $x - c$ in both the numerator and the denominator. Cancel this factor. This is permissible because we are only doing this cancellation *near* the point c, not *at* the point c. Keep doing such cancellations till there are no common factors of $x - c$ in the numerator and the denominator, and go back to step (1).

Here are some examples:

The function $f(x) = (x^2 - 3x + 2)/(x - 3)$ has $\lim_{x \to 1} f(x) = 0/(-2) = 0$.

The function $f(x) = (x^2 - 3x + 2)/(x - 1)$ has $\lim_{x \to 1} f(x) =$?? Well, evaluation gives 0 for both the numerator and the denominator, so we cancel the $(x - 1)$ factor, and we get:

$$\lim_{x \to 1} \frac{x^2 - 3x + 2}{x - 1} = \lim_{x \to 1} \frac{(x-1)(x-2)}{x - 1} = \lim_{x \to 1} x - 2 = 1 - 2 = -1$$

Similarly, if $f(x) = (x - 1)/(x^2 - 3x + 2)$, we have:

$$\lim_{x \to 1} \frac{x - 1}{x^2 - 3x + 2} = \lim_{x \to 1} \frac{1}{x - 2} = \frac{1}{1 - 2} = -1$$

And here's another example:

$$\lim_{x \to 1} \frac{x - 3}{x^2 - 3x + 2}$$

This limit is not defined, because the numerator approaches -2 and the denominator approaches 0.

FORMAL DEFINITION OF LIMIT

MATH 152, SECTION 55 (VIPUL NAIK)

Difficulty level: Hard. Full attention needed.

Covered in class?: Yes. But it is strongly recommended that you read this, as well as the book, preferably *prior* to the lecture.

Corresponding material in the book: Section 2.2.

Corresponding material in homework problems: Homework 2 advanced problems 1–6 and 10.

Things that students should get immediately: The definition of limit is tricky but there is a "method behind the madness". The subtleties in the definition are largely to avoid problems with functions that fluctuate too much.

Things that students should get with effort: The full definition of limit in terms of ϵs and δs. Ideally, write the definition for both generic functions and specific functions, and be able to clearly identify the bounding task that is needed. Also, be able to execute $\epsilon - \delta$ proofs for constant, linear, quadratic functions and for functions that are piecewise of these types.

Things that students should hopefully get: The approach to showing that certain limits do not exist.

Executive summary

Words ...

(1) $\lim_{x \to c} f(x) = L$ if, for every $\epsilon > 0$, there exists $\delta > 0$ such that, for every $x \in \mathbb{R}$ satisfying $0 < |x - c| < \delta$ (in other words, $x \in (c - \delta, c) \cup (c, c + \delta)$, we have $|f(x) - L| < \epsilon$ (in other words, $f(x) \in (L - \epsilon, L + \epsilon)$.
(2) What that means is that however small a trap (namely ϵ) the skeptic demands, the person who wants to claim that the limit does exist can find a δ such that when the x-value is δ-close to c, the $f(x)$-value is ϵ-close to L.
(3) The negation of the statement $\lim_{x \to c} f(x) = L$ is: there exists $\epsilon > 0$ such that for every $\delta > 0$ there exists $x \in \mathbb{R}$ such that $0 < |x - c| < \delta$ but $|f(x) - L| \geq \epsilon$.
(4) The statement $\lim_{x \to c} f(x)$ doesn't exist: for every $L \in \mathbb{R}$, there exists $\epsilon > 0$ such that for every $\delta > 0$ there exists $x \in \mathbb{R}$ such that $0 < |x - c| < \delta$ but $|f(x) - L| \geq \epsilon$.
(5) We can think of $\epsilon - \delta$ limits as a game. The skeptic, who is unconvinced that the limit is L, throws to the prover a value $\epsilon > 0$. The prover must now throw back a $\delta > 0$. Then, the skeptic provides a value of x within a δ-distance of c. If the $f(x)$-value is within an ϵ-distance of L, the prover wins. Otherwise, the skeptic wins. L being the limit means that the prover has a winning strategy, i.e., the prover has a way of picking, for any $\epsilon > 0$, a value of $\delta > 0$ suitable to that ϵ.
(6) The function $f(x) = \sin(1/x)$ is a classy example of a limit not existing. The problem is that, however small we choose a δ around 0, the function takes all values between -1 and 1, and hence refuses to be confined within small ϵ-traps.
(7) We say that f is continuous at c if $\lim_{x \to c} f(x) = f(c)$.

Actions...

(1) If a δ works for a given ϵ, then every smaller δ works too. Also, if a δ works for a given ϵ, the same δ works for any larger ϵ.
(2) Constant functions are continuous, we can choose δ to be anything. In this $\epsilon - \delta$ game, the person trying to prove that the limit does exist wins no matter what ϵ the skeptic throws and no matter what δ is thrown back.
(3) For the function $f(x) = x$, it's continuous, and $\delta = \epsilon$ works.

(4) For a linear function $f(x) = ax + b$ with $a \neq 0$, it's continuous, and $\delta = \epsilon/|a|$ works. That's the largest δ that works.
(5) For a function $f(x) = x^2$ taking the limit at a point p, the limit is p^2 (the function is continuous) and $\delta = \min\{1, \epsilon/(1 + |2p|)\}$ works. It isn't the best, but it works.
(6) For a function $f(x) = ax^2 + bx + c$, taking the limit at a point p, the limit is $f(p)$ (the function is continuous) and $\delta = \min\{1, \epsilon/(|a| + |2ap + b|)\}$ works. It isn't the best, but it works.
(7) If there are two functions f and g and $\lim_{x \to c} f(x) = \lim_{x \to c} g(x) = L$, and h is a function such that $h(x) = f(x)$ or $h(x) = g(x)$ for every x, then $\lim_{x \to c} h(x) = L$. The δ that works for h is the minimum of the δs that work for f and g. This applies to many situations: functions defined differently on the left and right of the point, functions defined differently for the rationals and the irrationals, functions defined as the max or min of two functions.

Pep talk

In this really important lecture, we're going to try and understand the formal definition of limit. This formal definition is really tricky to understand and it is sort of like a "rite of passage", like in some places the boys have to kill a tiger to become men. So it's the same way – this is when all the wishy-washy precalculus stuff ends and proper college calculus begins. And this definition has a lot of subtleties, and I hope you took the time to read the book and try to understand the definition.

1. Rugged terrain

1.1. The topologist's sine curve. First, let's recall the graph of the sine curve. This is a nice function – it is a periodic function with period 2π, and it is not just continuous, it is very smooth, waving smoothly. As good a function as you can get.

We now consider a slightly different function, which is the function:

$$f(x) := \sin(1/x)$$

Can you think of f as a composite of two functions? Indeed, $f = \sin \circ g$, where $g(x) = 1/x$. So what are the points where f is defined? Well, what could be the problem? The first problem could be that the function $g(x) = 1/x$ isn't defined. And that happens at $x = 0$. So 0 is a problem point. Once we've done the $1/x$ part, what next? Well, we need to take sin of that, which isn't a problem, because sin is defined for all real numbers. Thus, the domain of the function is all nonzero real numbers, or $\mathbb{R} \setminus \{0\}$, also written as $(-\infty, 0) \cup (0, \infty)$.

So when I first saw this function, I was like: *ouch*! What does it even mean to be such a function? What would the graph of such a function look like? Well, there are ways to compose graphs pictorially, but we don't want to go into those right now. So we'll just do some *ad hoc* stuff.

First, let's do the positive side. Suppose $x \to +\infty$. What happens to $1/x$? Well, it approaches 0, from the right side. And as you can see from the graph of x, that means it goes to zero. Now in the graph of sin, you see a very quick, almost straight line descent to zero. But when you are seeing this for $1/x$, this is almost painfully slow, so this is how it is going to look – it is going to sort off go more and more horizontal. By the way, in this picture, the x-axis is called a *horizontal asymptote*. We'll talk about asymptotes a little later in the course.

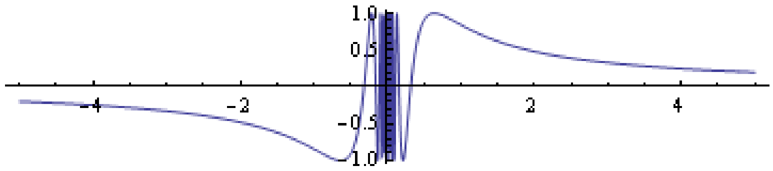
So the graph of $\sin(1/x)$ reaches 1, here, at $x = 2/\pi$. And as you know, $2/\pi$ is around $7/11$, so it is less than 1. So that part from 0 to $\pi/2$ in the sin graph gives rise to this part from $2/\pi$ to ∞. That really small part in the sin graph becomes this really huge part out here.

And all the other oscillations in the sin graph get compressed into the little region between 0 and $2/\pi$. So the graph falls to 0 here at $1/\pi$, then, it falls to -1 at $2/3\pi$, then comes back to zero at $1/2\pi$, and then the oscillations are faster than ever before. And as you get closer and closer to zero from the right, it is almost like it's madly just going up and down between -1 and 1.

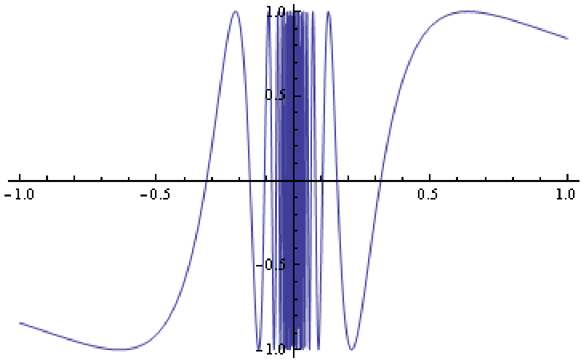
What about the negative side? Well, if you wanted to build your skills, you would do the whole thing again, but we can save some time by making an observation. The sine function is an odd function, and the function sending x to $1/x$ is an odd function, so the composite is also an odd function. Or another way of thinking about this is that $\sin(1/(-x)) = \sin(-1/x) = -\sin(1/x)$. So, what's going to happen is that the graph on the left is just obtained from the graph on the right using two flips, or a half turn about the origin.

Okay, now this is the graph that you need to look at and remember. By the way, some mathematicians have a name for the graph, or more specifically, the $x > 0$ part; they call it the *topologist's sine curve*.

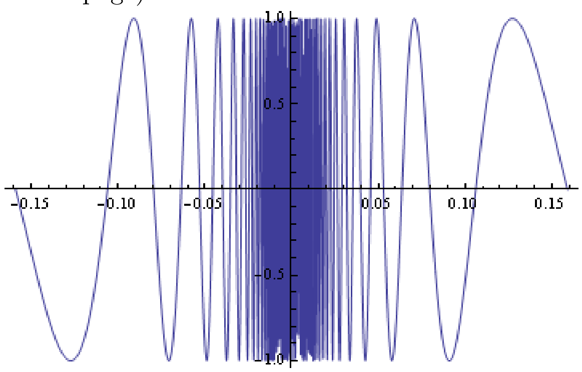
Here's the zoomed out picture:



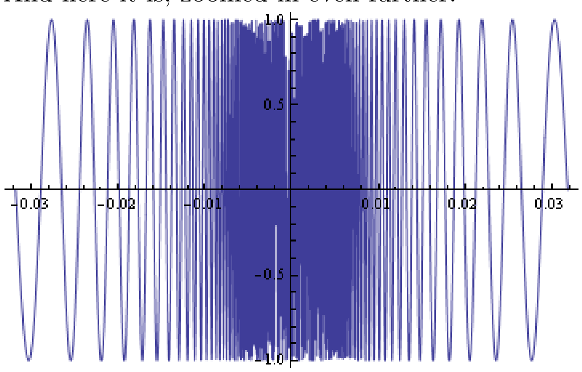
Here's a somewhat intermediate picture where we restrict the domain to $[-1, 1]$.



Here is the picture zoomed in further (note: the x and y axes are scaled differently to make the picture fit in the page):



And here it is, zoomed in even further:



1.2. **One definition of limit.** We first discuss a definition of limit that is wrong but it is wrong in an interesting way. And although it is wrong, it could still be a meaningful definition and is a useful concept in some cases.

Well, let's try to build this wrong definition. We want to interpret the sentence:

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

So one way of thinking of this is: as x gets arbitrarily close to c, $f(x)$ gets arbitrarily close to L. And so here's one guess:

Wrong definition of limit:

For every $\epsilon > 0$ and every $\delta > 0$, there exists x such that $0 < |x - c| < \delta$ and $|f(x) - L| < \epsilon$.

Okay, let's try to interpret this. This is saying that if you pick a really small interval $(c - \delta, c + \delta)$ around c on the x-axis, and a really small interval $(L - \epsilon, L + \epsilon)$ around L on the $f(x)$-axis, then you get this rectangle. And what my definition is saying that you'll have some point of the curve in this rectangle (but is not the point (c, L) itself). So what this definition is saying is that however small a rectangle you make around the point (c, L), you will have some points of the form $(x, f(x))$.

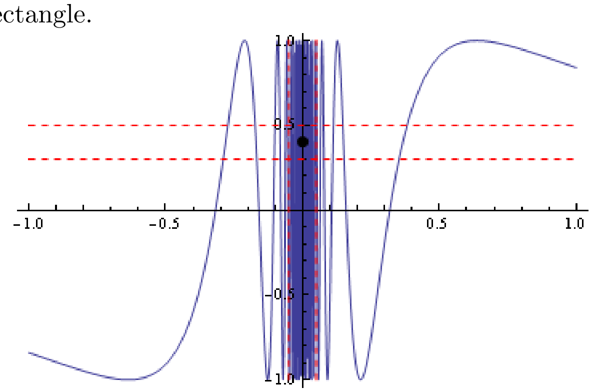
This seems like a reasonable description, because it says that you have these points that are really close. And this is certainly satisfied for most of the functions you have seen. But this is not the correct definition.

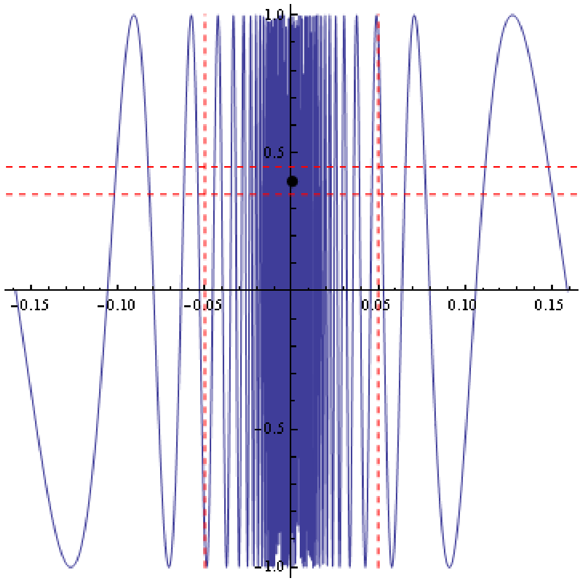
1.3. **Back to the topologist's sine curve.** So to understand what is wrong with this definition, we need to look again at the topologist's sine curve. And we need to look at this curve at the point 0, and ask: what is the limit at 0?

Well, I claim that by the definition I gave, the limit at 0 can be any number between -1 and 1. Why? Well, think about what is happening to the function close to zero on the positive side. It is madly oscillating. This means that however small an interval around 0 you take for the x-value, the $\sin(1/x)$ function takes all possible values between -1 and 1. So, however small a rectangle you draw at any point, you're going to get points that are in there.

So the thing with the function is that its rapid oscillation is creating a problem: by our definition, we get all points in the interval $[-1, 1]$ as the limit. But that's not the way we would like limits to behave – we want the limit to be unique, and it should be a reasonable description of where the function *tends to.* If this definition says that the limit at 0 could be both 0 and 1/2 and $-1/2$ and 1, and this jars your intuition, it is this definition that you need to throw out.

Here's a graphical illustration, where we are trying to study the approach of $\sin(1/x)$ to 0.4 as x approaches 0. Note that however small a rectangle we make around the point, it contains lots of points of the $\sin(1/x)$ graph. However, the $\sin(1/x)$ graph is all over the place, so the function is never *trapped* within a small rectangle.





1.4. **Challenging!** So I know that you're not happy with me – I just tried to shove down your throat a wrong definition and you spent the time trying to understand it and then I told you that this is the wrong definition. So, this was just to give you a flavor that getting the definition is challenging, and a definition that might seem right could be riddled with holes. I know this is exhausting, so I'll skip right to the correct definition and explain why it works. And then after some time, I suggest you come back and compare the correct definition and the wrong definition and try to understand exactly what the difference is.

The way I think about the definition is in terms of a *cage* (or *trap*). And the reason why we need this notion of a cage or trap is precisely to avoid these kinds of oscillations that give rise to multiple limits. So, here is the formal definition:

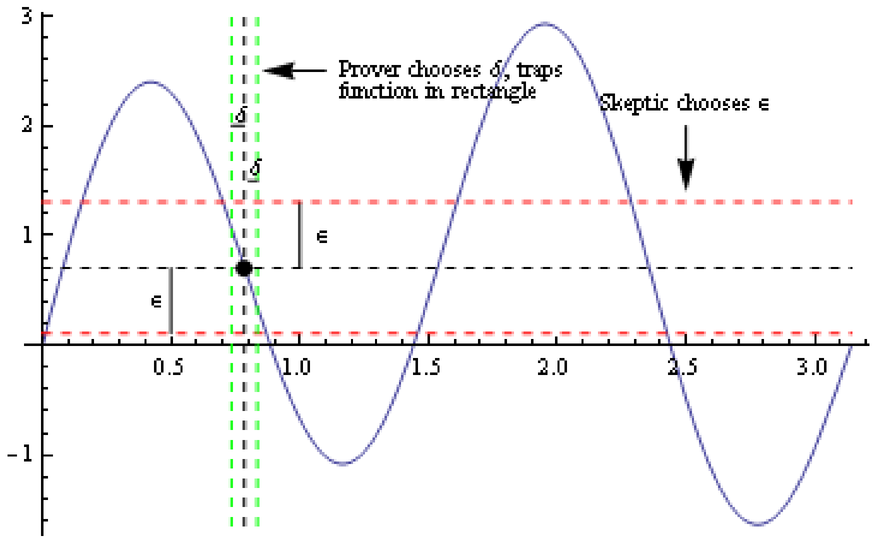
We say that $\lim_{x\to c} f(x) = L$ (as a two-sided limit) if, for every $\epsilon > 0$, there exists $\delta > 0$ such that, for every x such that $0 < |x - c| < \delta$, we have $|f(x) - L| < \epsilon$.

That's quite a mouthful. Let's interpret it graphically. What it is saying is that: "for every ϵ" so we consider this region $(L - \epsilon, L + \epsilon)$, so there are these two horizontal bars at heights $L - \epsilon$ and $L + \epsilon$. Next it says, there exists a δ, so there exist these vertical bars at $c + \delta$ and $c - \delta$. So we have the same rectangle that we had in the earlier definition.

What is different this time is that we not only demand that the graph have a few points in that rectangle, but rather, that it lies completely inside the rectangle. And this is the crucial difference between this definition and the previous definition, that allowed sometimes-in-sometimes-out graphs. Because, here we are saying that the condition holds for *all* x such that $0 < |x - c| < \delta$. Which is the same as saying that the condition holds on the interval $(c - \delta, c + \delta)$ minus the point $\{c\}$ itself. Or, you can think of that set as $(c - \delta, c) \cup (c, c + \delta)$. And we're insisting that the condition hold for all things, not just that there exists a point here or a point there.

The other main difference from the earlier definition is that in this definition, the value of δ depends on ϵ. So here is another way of thinking of this definition that I find useful. Suppose I claim that as x tends to c, $f(x)$ tends to L, and you are skeptical. So you throw me a value $\epsilon > 0$ as a challenge and say – can I trap the function within ϵ? And I say, yeah, sure, because I can find a $\delta > 0$ such that, within the ball of radius δ about c, the value $f(x)$ is trapped in an interval of size ϵ about L. So basically you are challenging me: can I create an ϵ-cage? And for every ϵ that you hand me, I can find a δ that does the job of this cage.

Here's a pictorial illustration:



1.5. **Formal description as a game.** We make the discussion above somewhat more formal by encoding it as a game. Consider the assertion:

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

where specific values are provided for c and L and for the function f. Suppose, further, that f is defined at all points on the immediate left and the immediate right of c (otherwise, the game would be meaningless).

The game has two players, a *prover*, whose goal is to show that the limit statement above is correct, and a *skeptic* (in higher mathematics jargon, this person is called a *verifier*, but I think you'll find *skeptic* a more intuitive term) who is far from convinced and wants to raise the best counter-arguments. The game has three moves:

- The skeptic chooses an $\epsilon > 0$ (the subtext being the interval $(L - \epsilon, L + \epsilon)$), effectively telling the prover: "try to trap the function with ϵ of L if you can."
- The prover chooses a $\delta > 0$ (the subtext being the interval $(c - \delta, c + \delta)$, excluding the point c), effectively telling the skeptic: "here's a trap that works.
- The skeptic chooses a value of x such that $0 < |x - c| < \delta$, (i.e., within the set $(c - \delta, c + \delta) \setminus \{c\}$), challenging the prover at that specific x.

Once these moves are complete: we compute $|f(x) - L|$. If it is less than ϵ, the prover wins. Otherwise, the skeptic wins.

We say that $\lim_{x \to c} f(x) = L$ is *true* if the prover has a *winning strategy*. In other words, *no matter what* choice the skeptic makes for ϵ, the prover has a (smart) choice of δ such that *no matter what* value of x the skeptic chooses in $(c - \delta, c) \cup (c, c + \delta)$, $f(x)$ lies in the interval $(L - \epsilon, L + \epsilon)$.

If, in contrast, the *skeptic* has a winning strategy, then we declare the statement to be false.

The key take-away is that in order to show a limit statement to be true, we need to devise a winning strategy for the prover in the above game. Note that the winning strategy must work against an extremely smart skeptic, not merely against a skeptic who makes a silly choice of ϵ.

1.6. **One-sided limits.** The definition of limit we have given is:

$\lim_{x \to c} f(x) = L$ if, for every $\delta > 0$, there exists $\delta > 0$ such that, for every x such that $0 < |x - c| < \delta$, then $|f(x) - L| < \delta$.

This defintion is fine when the function is defined on both sides. Note the way we are using *both sides* of c, because when we say $0 < |x - c| < \delta$, we are including the δ-interval on the left side $(c - \delta, c)$ and the right side $(c, c + \delta)$.

For the right hand limit, we want to restrict x to the interval $(c, c + \delta)$ on the right side, and for the left hand limit, we want to restrict x to the interval $(c - \delta, c)$.

Here's how we define the *left hand limit*: $\lim_{x \to c^-} f(x) = L$ if, for every $\epsilon > 0$, there exists a $\delta > 0$ such that, for all x satisfying $0 < c - x < \delta$, we have $|f(x) - L| < \epsilon$.

Note that $0 < c - x < \delta$ is the same as saying that $x \in (c - \delta, c)$.

Here's how we define the *right hand limit*: $\lim_{x \to c^+} f(x) = L$ if, for every $\epsilon > 0$, there exists $\delta > 0$ such that, for all x satisfying $0 < x - c < \delta$, we have $|f(x) - L| < \epsilon$.

Note that $0 < x - c < \delta$ is the same as saying that $x \in (c, c + \delta)$.

2. Strategy stockpile

We now discuss strategies for showing that a particular limit exists and has a particular value, and hence, for showing that a function is continuous.

2.1. **What we need to do.** Recall the way of thinking of a limit in terms of a *cage* or a *trap*:

We say that $\lim_{x \to c} f(x) = L$ (as a two-sided limit) if, for every $\epsilon > 0$, there exists $\delta > 0$ such that, for every x such that $0 < |x - c| < \delta$, we have $|f(x) - L| < \epsilon$.

This is quite a mouthful, so let's slow down and try to understand what it means. Let's interpret it graphically. What it is saying is that: "for every ϵ"so we consider this region $(L - \epsilon, L + \epsilon)$, so there are these two horizontal bars at heights $L - \epsilon$ and $L + \epsilon$. Next it says, there exists a δ, so there exist these vertical bars at $c + \delta$ and $c - \delta$. And what we're saying is that if the x-value is trapped between the vertical bars $c - \delta$ and $c + \delta$ (but is not equal to c), the $f(x)$-value is trapped between $L - \epsilon$ and $L + \epsilon$.

The important thing to note here is that the value of δ depends on the value of ϵ. As I said earlier, we can think of this as a game, where I (as the prover) am trying to prove to you that the limit $\lim_{x \to c} f(x) = L$ and you are a skeptic who is trying to catch me out. So you throw ϵs at me, and challenge me to show that I have a δ to trap that ϵ. And if I have a winning strategy, that enables me to find a δ for every ϵ that you throw at me, then yes, the limit is equal to L.

2.2. **The winning strategy for constant functions.** So the real question is: can I obtain a winning strategy? And what would such a strategy be? It would be some procedure, some function, that takes as input a value of ϵ and outputs a value of δ in terms of that ϵ. Now I know that's a mouthful, so let's look at some simple examples.

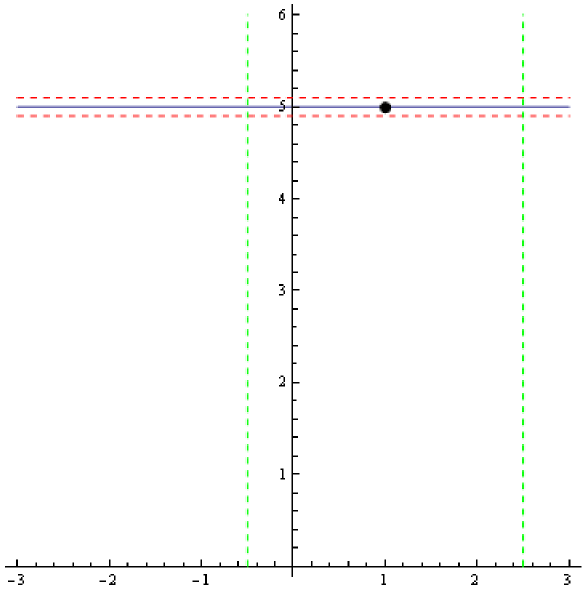
We'll take two kinds of functions for which the limit is particularly easy to compute: *constant functions* and the identity function. Let's first look at a constant function that sends every real number to k. So let me call this f. So f is a function with the property that $f(x) = k$ for all $x \in \mathbb{R}$. So, what can we say about the graph of f? Well, it is this horizontal line at height k. So far, so good.

Let's look at a point c, and try to calculate $\lim_{x \to c} f(x)$. Which is the same as trying to compute $\lim_{x \to c} k$. The guess, from looking at the graph, is that the limit equals k. So how do we show this in terms of the $\epsilon - \delta$ definition?

Okay, lost? No problem. Sometimes, when unwinding mathematical expressions, you (and even I) get lost. That's not the time to give up. Rather, it is the time to refocus and go back to the original definition and work things out again.

So we want to show that $\lim_{x \to c} f(x) = k$. In other words, we want to show that for every $\epsilon > 0$, there exists a $\delta > 0$ such that if $0 < |x - c| < \delta$, then $|f(x) - k| < \epsilon$. By the way, that k that appeared at the end there is because we're claiming the limit is k.

Before we unravel that (and it's a bit of an anti-climax once we do), let's just think of this graphically. We are saying that for every $\epsilon > 0$, so we are thinking of the region between the horizontal bars at heights $k - \epsilon$ and $k + \epsilon$. And then we want to say that there exists a $\delta > 0$ (that we haven't determined) so we are thinking of the vertical bars at $c - \delta$ and $c + \delta$, so we have this rectangle. And we have to choose δ such that that part of the graph lies inside that rectangle.



But the picture makes it clear that we can choose just about any $\delta > 0$! So in the case of the constant function, we can choose just about anything and not run into trouble. So, let's see that from the algebra.

We want to show that there exists a $\delta > 0$ such that if $0 < |x - c| < \delta$, then $|f(x) - k| < \epsilon$. but what's $|f(x) - k|$? it is $|k - k| = 0$, and so is less than ϵ, so the condition is tautologically satisfied for all x. So any δ will do.

Another way of thinking of this is that in this case, I don't need a trap at all – any trap will do because the function is already at the right place all along! Also, remember that the kind of examples that gave us headaches were things where the function changed a lot from point to point, and constant functions change as little as possible – they don't change at all.

Or, thinking of it in terms of the two-person game where you (the skeptic) throw me ϵs and I throw back δs that work, this is a really easy game. Whatever you throw at me, I can throw back anything at you. This is a great game for me – no matter how smart you are or how stupid I am, I always win.

I urge you to go through this example carefully and understand it thoroughly. It's also Example 5 on page 67 of the book.

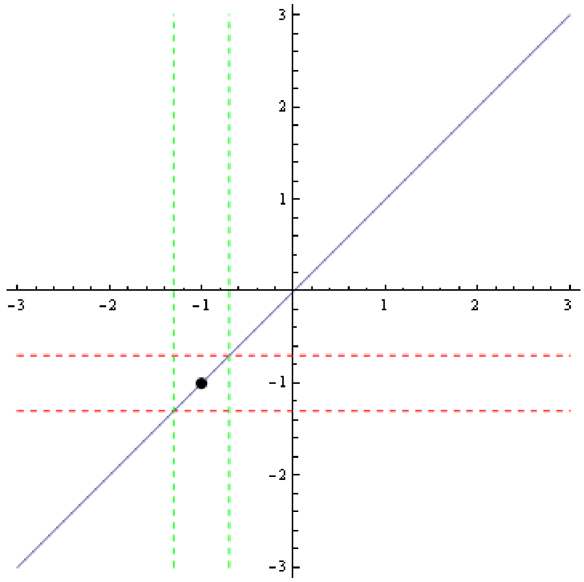
2.3. **The winning strategy for the identity function.** Okay, now when I asked you to calculate the limit of the identity function, you might remember I said it's like a word puzzle: "as x approaches c, what does x approach?" Well, obviously c. So now we want to do the fancy $\epsilon - \delta$ version of that same argument.

Let's unwind the definition. And by the way, the main thing, once you have got the correct definition, is to be able to carefully apply it by interpreting it correctly. That isn't easy but it isn't impossible. In fact, at some stage you should be able to take a definition that you're seeing for the first time and apply it to a given problem well.

So define $g(x) = x$. So we want to show that for every $\epsilon > 0$, there exists a $\delta > 0$ such that whenever $0 < |x - c| < \delta$, then $|g(x) - c| < \epsilon$. And by the way, the second c came because we're claiming that the limit is c. And by the way, since $g(x) = x$, I can rewrite this as follows:

I want to show that for every $\epsilon > 0$, there exists a $\delta > 0$ such that whenever $0 < |x - c| < \delta$, then $|x - c| < \epsilon$. Hmm. So what δ has the property that whenever $0 < |x - c| < \delta$, then $|x - c| < \epsilon$? Again, the answer is almost tautological: set $\delta = \epsilon$.

For the game where the function is the identity function, I can choose the identity function as my winning strategy, in the sense that whatever ϵ you throw at me, I throw back the same value to you for δ. We can also see this graphically:



I urge you to go through this example carefully and understand it thoroughly. It's also Example 5 on page 67 of the book.

2.4. **The winning strategy for linear functions: motivation.** Okay, let's up the ante a little bit, getting more abstract without actually making our function a lot harder. Suppose we have the function:

$$f(x) := ax + b$$

So this f is a *linear function*. And we want to determine the limit $\lim_{x \to c} f(x)$. Here, the numbers a and b are unknown constants, i.e., in a specific problem situation, their values will be known.

The first thing you can do is draw the graph, and this graph is a straight line. The slope of that straight line depends on the value of a. So if a is positive, this is a south-west to north-east line, and if a is negative, this is a north-west to south-east line. And, if $a = 0$, the line is flat. And the parameter b tells you further where the line is placed – different bs give different lines that are all parallel to each other.

Your guess would be that the limit at c is $f(c)$, which is sort of pictorially clear, because a straight line looks very continuous. And you'll notice that this generalizes both the constant function and the identity function: the constant function would be the case $a = 0$, and the identity function would be the case $a = 1$ and $b = 0$. Since we already settled the constant function case, we'll assume $a \neq 0$.

So, what again do I want to show? $\lim_{x \to c} f(x) = f(c)$. And so, I need to find, for every ϵ, a suitable δ such that something holds. What thing?

So, what I want is to find, for every $\epsilon > 0$, a value $\delta > 0$ such that for $0 < |x - c| < \delta$, we have $|f(x) - f(c)| < \epsilon$. We now try to substitute the actual expression for f in that last inequality. So let's do this simplification on the side.

$$f(x) - f(c) = (ax + b) - (ac + b) = a(x - c)$$

So, what we want is the following: for every $\epsilon > 0$, find a value $\delta > 0$ such that for $0 < |x - c| < \delta$, we have $|a(x - c)| < \epsilon$.

Okay, now this looks reasonable, but it isn't tautological as the previous examples were. We need to do some thinking, and there is a real logical impasse here. So let's look at the part:

$$|x - c| < \delta$$

What does this allow us to say about $|a(x - c)|$? Well, we can try multiplying the above inequality by $|a|$, and we get:

$$|a||x - c| < |a|\delta$$

And, note that there isn't any sign change because $|a| > 0$ (as we assumed $a \neq 0$). And, using the absolute value of a product is the product of the absolute values, we get:

$$|a(x-c)| < |a|\delta$$

Okay, so what we *have* is the above, and what we *want* is $|a(x-c)| < \epsilon$. And remember, we have the freedom to choose any δ that we want. So what value of δ do we choose? Well, a little thought should reveal that a simple way of choosing a δ that works is to set $|a|\delta = \epsilon$. That gives $\delta = \epsilon/|a|$. So that's a value of δ that works.

2.5. **The winning strategy: proved succinctly. Problem**: Prove that $\lim_{x \to c} f(x) = f(c)$ where $f(x) := ax + b, a \neq 0$.

Winning strategy for prover-skeptic game: Choose $\delta = \epsilon/|a|$.

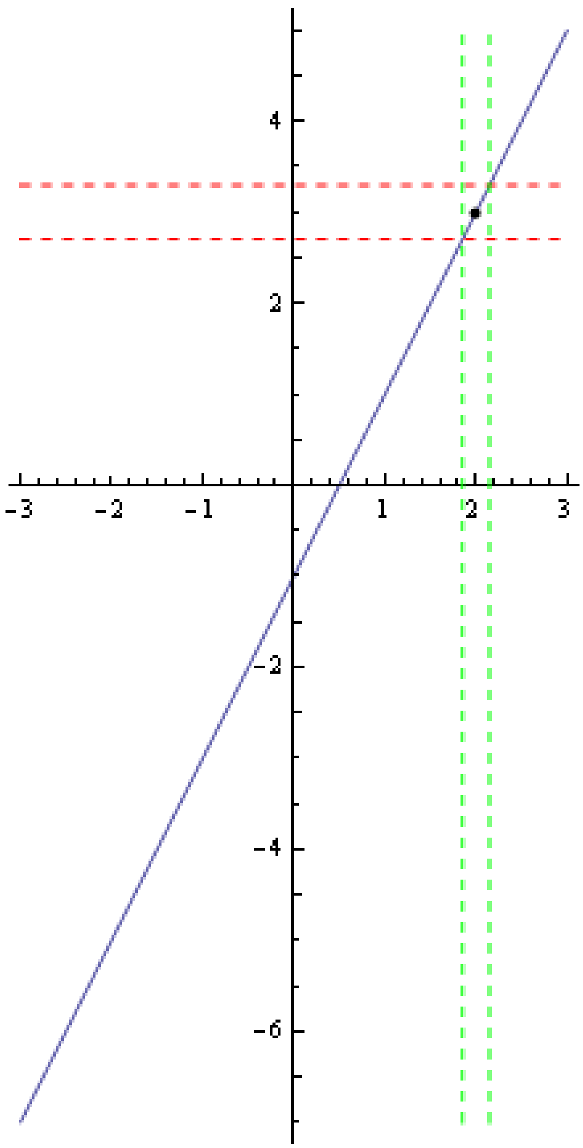
We want to show that, for any $\epsilon > 0$, if $0 < |x - c| < \delta$, then $|f(x) - f(c)| < \epsilon$, where $\delta = \epsilon/|a|$. We do this as follows:

$$\begin{aligned}
&|f(x) - f(c)| \\
=&|(ax + b) - (ac + b)| \\
=&|ax - ac| \\
=&|a(x - c)| \\
=&|a||x - c| \\
<&|a|\delta \qquad \text{using } |x - c| < \delta \\
=&|a|\frac{\epsilon}{|a|} \qquad \text{using } \delta = \epsilon/|a| \\
=&\epsilon
\end{aligned}$$

The chain has one strict inequality and the rest all equalities, so we get, overall, that $|f(x) - f(c)| < \epsilon$.

So that's it. We've proved that for a linear function, the limit always exists at any point, and moreover, the limit equals the value of the function at the point. As you saw in the informal introduction earlier, that's basically saying that linear functions are continuous.

Here's the $\epsilon - \delta$ picture for a linear function:



Note that the larger the value of $|a|$, the smaller the δ needs to be for a given ϵ. This makes sense, because the steeper the slope, the more rapidly the function is changing, and hence, the smaller the trap has to be to catch the function.

2.6. **Smaller δs work too.** One other important thing to note is that we have some latitude in choosing δ. In other words, the winning strategy isn't always unique. In fact, if, for a given ϵ, one value of δ works, then any smaller positive value of δ also works. I urge you to think about this graphically and also in strategic terms.

2.7. **Principles and practice.** So there are two aspects to everything: there's the *theory* and there's the *practice*. And one of the popular quotes on U of C T-shirts and merchandise is "that's all well and good in practice... but how does it work in theory?" So what we did in this and the previous lecture was take a little glimpse at how limits work in theory. And by these attempts to calculate the δ in terms of the ϵ, we are basically trying to bridge the gap between theory and practice.

Going from the theory to the practice requires an understanding of inequalities, which you already have. But it doesn't just require that – it requires a better conceptual grasp of *what we have to determine* and *what is given*. The thing here is that in the proof, we have to *work backwards*, in the sense that we have to first guess what the limit is going to be, then we have to guess how to get a δ that works, and then we can check that it works. And this working backwards can be a little tricky because of the implications stuff.

That is why it is very important that you go home and look at the examples we did in class and the other examples in the book. As far as expectations from you are concerned, you are not expected to be able to do the $\epsilon - \delta$ computations for things other than constant, linear, and quadratic functions, and we've already done the constant and linear cases, and we will now consider general quadratic case. But I suggest you understand the book's examples for a couple of the slightly trickier cases – basically to get an idea of what their goal is and how they're approaching it.

2.8. **General case of quadratic functions: obtaining the formula. NOTE**: On your homework or in the midterm, you cannot assume the $\epsilon - \delta$ formulas we have derived for linear and quadratic functions. However, knowing these formulas will enable you to skip the *rough work* needed to find the δ and move directly to the *fair work* of showing that that δ works.

We look here at the general case of a quadratic function:

$$f(x) := ax^2 + bx + c$$

with $a, b, c \in \mathbb{R}$ and $a \neq 0$. Note that if $a = 0$, we get a linear or constant function, and we already know how to deal with it.

The letter c is used up as a variable, so we should not use c as a limiting point. Let's call the limiting point p instead. So we claim that, for $p \in \mathbb{R}$, we have $\lim_{x \to p} f(x) = f(p)$.

In other words, we need to show that for every $\epsilon > 0$, there exists a value $\delta > 0$ such that for $0 < |x - p| < \delta$, we have $|f(x) - f(p)| < \epsilon$.

So we again do the usual thing: simplify $|f(x) - f(p)|$. We have:

$$f(x) - f(p) = (ax^2 + bx + c) - (ap^2 + bp + c) = a(x^2 - p^2) + b(x - p)$$

Okay, now that's nice, but can we simplify it a little more? Yes, and this is the important idea you should take from this problem: the idea is to factor out the $(x - p)$ factor from the expression. So we get:

$$f(x) - f(p) = (x - p)(a(x + p) + b)$$

Why did we factor out $x - p$? Think of what we need to do. We need to say that if the absolute value of $x - p$ is small, then the absolute value of $f(x) - f(p)$ is small too. And if you go back to the linear example, what we did was to show that $f(x) - f(p)$ is a scalar multiple of $x - p$, i.e., it is $x - p$ times a constant. In the case of the quadratic function, we do not have a constant for the other factor, so it's going to be a little harder, but this is the right direction.

Okay, now what next? We need to find a δ such that if $0 < |x - p| < \delta$, then $|f(x) - f(p)| < \epsilon$. Now if $|x - p| < \delta$, then we have:

$$|f(x) - f(p)| = |x - p||a(x + p) + b| < \delta|a(x + p) + b|$$

Okay, and now we're stuck, because unlike the linear case, the second part isn't a constant, and it involves an x. But can we bound it by a constant? Well, let's think about this intuitively. We are really interested in situations where $x - p$ is really small, so x is really close to p. So $x + p = (x - p) + 2p$, with the $x - p$ part being small. So:

$$|a(x + p) + b| = |a(x - p) + 2ap + b| \leq |a(x - p)| + |2ap + b| < |a|\delta + |2ap + b|$$

So this is progress, and what exactly is the nature of the progress? Well, what we've found is that that expression which isn't constant is still bounded from above by some constant plus $|a|\delta$. So, we get:

$$|f(x) - f(p)| < \delta(|a|\delta + |2ap + b|)$$

And now the original question: given a value $\epsilon > 0$, how do we find a value $\delta > 0$ such that that right side in terms of δ is not more than ϵ? Well, you can try solving a quadratic inequality, but that's a pain, so I'll show you a simpler approach.

First, notice that we're really interested in small values of δ. So let's assume $\delta \leq 1$. Then, we get:

$$|f(x) - f(p)| < \delta(|a| + |2ap + b|)$$

And now, if we have $\delta \le \epsilon/(|a| + |2ap + b|)$, we are in good shape.

So, a value of δ that works is $\min\{1, \epsilon/(|a| + |2ap + b|)\}$.

What we've done above is obtained a value of δ. You should now be able to retrace the steps to confirm that this value of δ works.

By the way, this generalizes Example 6 (Section 2.2, Page 68) of the book. In that example, $a = 1, b = c = 0$, and $p = 3$. So if there are too many symbols in this example and you want a simpler example with fewer symbols, you should look at that example in the book and then, armed with that, come back to master this one.

Please go through both these very thoroughly; this is very important to understand.

2.9. **General case of quadratic function: how to present the solution.** We illustrate how the solution would be *presented* for the general case of a quadratic function.

Problem: Prove that $\lim_{x \to p} f(x) = f(p)$ where $f(x) := ax^2 + bx + c, a \neq 0$.

Winning strategy for prover-skeptic game: Choose $\delta = \min\{1, \frac{\epsilon}{|a|+|2ap+b|}\}$.

Proof: We need to show that, for any $\epsilon > 0$, if $0 < |x - p| < \delta$, then $|f(x) - f(p)| < \epsilon$, where $\delta = \min\{1, \frac{\epsilon}{|a|+|2ap+b|}\}$.

We have:

$$\begin{aligned}
&|f(x) - f(p)| \\
=&|(ax^2 + bx + c) - (ap^2 + bp + c)| \\
=&|a(x^2 - p^2) + b(x - p)| \\
=&|x - p||a(x + p) + b| \\
=&|x - p||a(x - p) - 2ap + b| \\
\le&|x - p|(|a||x - p| + |2ap + b|) \qquad \text{by triangle inequality} \\
<&\delta(|a|\delta + |2ap + b|) \qquad \text{using } |x - p| < \delta \\
\le&\delta(|a| + |2ap + b|) \qquad \text{using } \delta \le 1 \text{ for the inner } \delta \\
\le&\frac{\epsilon}{|a| + |2ap + b|}(|a| + |2ap + b|) \qquad \text{using } \delta \le \epsilon/(|a| + |2ap + b|) \text{ for the outer } \delta \\
=&\epsilon
\end{aligned}$$

Each step of the process involves one of the three signs $=, <, \le$, with one of the steps involving strict inequality. Thus, overall, we obtain that $|f(x) - f(p)| < \epsilon$.

Qualitatively, the process can be described as follows:

- Factor the quadratic $|f(x) - f(p)|$ with one of the factors being $|x - p|$.
- Rewrite the other factor as a constant times $x - p$ plus another constant.
- Now split using the triangle inequality.
- Use $|x - p| < \delta$ at both places.
- For the inner factor of δ, use $\delta \le 1$.
- For the outer factor, use $\delta \le \epsilon/(...)$.

2.10. **Concrete case of quadratic.** We consider a concrete example of a quadratic with actual numerical values of a, b, c, and p, and walk through what the general steps just described would look like in the concrete example.

Consider the limit proof:

$$\lim_{x \to 5}(2x^2 + 3x + 17) = 82$$

Winning strategy: Take $\delta = \min\{1, \epsilon/25\}$. We obtain 25 using the formula $|a| + |2ap + b|$, where $a = 2$, $p = 5$, and $b = 3$, so we got $|2| + |2 \cdot 2 \cdot 5 + 3| = |2| + |23| = 25$.

Proof: We want to show that, for any $\epsilon > 0$, if $0 < |x-5| < \delta$, then $|(2x^2+3x+17)-(2(5)^2+3(5)+17)| < \epsilon$, where $\delta = \min\{1, \epsilon/25\}$.

We consider:

$$
\begin{aligned}
& |2x^2 + 3x + 17 - (2(5)^2 + 3(5) + 17)| \\
=& |2x^2 + 3x - 65| \\
=& |x-5||2x+13| \\
=& |x-5||2(x-5)+23| \\
\le& |x-5|(2|x-5|+23) \qquad \text{by triangle inequality} \\
<& \delta(2\delta+23) \qquad \text{using } |x-5| < \delta \\
\le& \delta(2+23) \qquad \text{using } \delta \le 1 \\
=& 25\delta \\
\le& 25\frac{\epsilon}{25} \qquad \text{using } \delta \le \epsilon/25 \\
=& \epsilon
\end{aligned}
$$

More examples will be done in the relevant review sessions.

3. More $\epsilon - \delta$ limit computations

3.1. **Function with left and right definition.** One kind of situation, that we have already seen, is a situation where a function given to us has one definition on the left side of a point and another definition on the right side of the point. Your intuition would tell you that if the left hand limit and the right hand limit exist separately *and are equal*, then the limit exists on the whole as well, and equals both these values.

How can we make this intuition precise in terms of the $\epsilon - \delta$ game? The idea is to think of it as two games in parallel: the left hand limit game, where I (the prover) have to throw back a δ at you that works for x approaching c from the left side, and an analogous right hand limit game. The fact that the left hand limit and right hand limit are both equal to L tells me that I have winning strategies for both games. I now need to combine them into a winning strategy for the two-sided limit game. How do I do this?

Basically, I need to choose δ small enough that it's good enough on both the left and the right. The idea is simple: pick δ as the *minimum* of the δs that I picked on the left and on the right. This will work for *both* the left side *and* the right side.

See, for instance, the example of the absolute value function (given on Page 67 of the book). At the point 0, the absolute value function has the definition $-x$ on the left and x on the right.

So what we first need to do is figure out the left strategy and the right strategy separately. For the function $f(x) = -x$, what is the δ that works for a given ϵ? Well, you can do the calculations again, but since we already did the general case of a linear function, we can just plug that in and get that $\delta = \epsilon$ works. Similarly, on the right side, we again get that $\delta = \epsilon$ works. So in this case, the same strategy works on both sides, so we can use $\delta = \epsilon$ as the winning strategy for the two-sided limit.

Okay, let's take a more interesting example. Suppose $f(x) = \{ \begin{array}{ll} 3x, & x < 0 \\ -5x, & x \geq 0 \end{array}$. It is an obvious guess that the limit at 0 equals 0. Since both sides are linear functions, we know how to get the strategies for the left side and the right side separately. The strategy for the left side is $\delta = \epsilon/3$ and the strategy for the right side is $\delta = \epsilon/5$. So, what is our overall strategy?

The overall strategy should pick a δ small enough that it works for both sides. In this case, the smaller of the numbers $\epsilon/3$ and $\epsilon/5$ is $\epsilon/5$ (remember, $\epsilon > 0$). So, we get a winning strategy by choosing $\delta = \epsilon/5$.

REMINDER: On your homework or in the midterm, you cannot assume the $\epsilon - \delta$ formulas we have derived for linear and quadratic functions. However, knowing these formulas will enable you to skip the *rough work* needed to find the δ and move directly to the *fair work* of showing that that δ works.

Let's now do the general case, with a full proof:

Claim. Suppose $a_1, a_2, b_1, b_2 \in \mathbb{R}$ with $a_1 \neq 0$ and $a_2 \neq 0$. Suppose $f(x) = \{ \begin{array}{ll} a_1x + b_1, & x < c \\ a_2x + b_2, & x > c \end{array}$. Then, if $L = a_1c + b_1 = a_2c + b_2$, we have $\lim_{x \to c} f(x) = L$. For any $\epsilon > 0$, a δ that works is $\epsilon / \max\{|a_1|, |a_2|\} = \min\{\epsilon/|a_1|, \epsilon/|a_2|\}$ (note that δ is positive).

Proof. To prove the claim, we need to show that for any x such that $0 < |x - c| < \delta$, we have $|f(x) - L| < \epsilon$.

We split this into two cases: the case where $x < c$ and the case where $x > c$.

Case that $x < c$: In this case, we have $f(x) = a_1x + b_1$. Thus, we get:

$$|f(x) - L| = |a_1x + b_1 - (a_1c + b_1)| = |a_1(x - c)| = |a_1||x - c| < |a_1|\delta \leq |a_1|(\epsilon/|a_1|) = \epsilon$$

Thus, $|f(x) - L| < \epsilon$.

Case that $x > c$: In this case, we have $f(x) = a_2x + b_2$. Thus, we get:

$$|f(x) - L| = |a_2x + b_2 - (a_2c + b_2)| = |a_2(x - c)| = |a_2||x - c| < |a_2|\delta \leq |a_2|(\epsilon/|a_2|) = \epsilon$$

□

3.2. **Weirder functions, mundane ideas.** So far, we have considered the case where the function has one definition on the left and another definition on the right, and how we can combine the winning strategies. Basically, we combine the winning strategies by picking the smaller of the two δs. This allows us to use our strategies for constant, linear, and quadratic functions to tackle functions that are piecewise constant, linear, and quadratic.

The strategy extends to some more weirdly defined functions, where there are multiple definitions, but they are not clearly separated in left-right terms. For instance, consider the function:

$$f(x) = \{ \begin{array}{ll} x^2, & x \text{ rational} \\ x, & x \text{ irrational} \end{array}$$

What we have done is split the domain of definition into two subsets, but this time the two subsets aren't nicely left and right of some point, they are both scattered all over the place. However, trying to prove limit problems for such functions follows *essentially the same strategy*. For instance, if, for the above function, we are trying to prove that $\lim_{x \to 1} f(x) = 1$, then what we do is to find a winning strategy for the x^2 function, a winning strategy for the x function, and take the smaller of the δs. And, instead of splitting into cases based on left and right, we split into cases based on – you guessed it – *rational* and *irrational*.

Let's try to formally prove that $\lim_{x \to 1} f(x) = 1$. The first thing we need to do is rough work to figure out the δs that work for the individual functions x^2 and x. But we've already done this. Recall that for x^2, we can choose $\delta = \min\{1, \epsilon/3\}$, and for x, we can choose $\delta = \epsilon$. So the overall δ that we need to choose is $\min\{1, \epsilon/3, \epsilon\}$. We can simplify that to $\min\{1, \epsilon/3\}$.

To complete the proof, we first consider the case where x is rational and then consider the case that x is irrational.

Proof. Case that x is rational: We want to show that if $0 < |x - 1| < \min\{1, \epsilon/3\}$, then $|x^2 - 1| < \epsilon$. Let's do this:

$$|x^2 - 1| = |x - 1||x + 1| < (\epsilon/3)|x + 1| \leq (\epsilon/3)(|x - 1| + 2) < \epsilon/3(1 + 2) = \epsilon$$

In the second step, we used that $|x - 1| < \epsilon/3$, and in the fourth step, we used $|x - 1| < 1$.

Case that x is irrationa: We want to show that if $0 < |x - 1| < \min\{1, \epsilon/3\}$, then —$x - 1| < \epsilon$. Let's do this:

$$|x - 1| < \min\{1, \epsilon/3\} \leq \epsilon/3 < \epsilon$$

□

3.3. **Max or min of two functions.**

Suppose f and g are two functions and we define $h(x) = \min\{f(x), g(x)\}$. Suppose, further, that at some point c, we have $\lim_{x\to c} f(x) = L$ and $\lim_{x\to c} g(x) = L$. Then, we'll also have $\lim_{x\to c} h(x) = L$.

The reason? Well, the minimum of two functions is a special situation where, at every point, we are picking one of the functions. And we've just discussed that, whenever you get a function by always picking one of two functions, and both of them are approaching the same limit, the new function that you're picking also approaches the same limit. The reason for that is that we can determine the winning strategies for both the $\epsilon - \delta$ games and then choose the smaller of the δs that we have for both functions.

So, this general strategy works for the minimum. It also works for the maximum.

3.4. **ADDENDUM: A harder problem.**

Suppose $f : \mathbb{R} \to \mathbb{R}$ is a function with the property that, for every $x \in \mathbb{R}$, we have $(f(x))^2 - 3xf(x) + 2x^2 = 0$. We want to show that $\lim_{x\to 0} f(x) = 0$.

Here's how we do this problem. First, note that the given statement is that:

$$\begin{aligned} (f(x))^2 - 3xf(x) + 2x^2 &= 0 \\ \implies (f(x) - x)(f(x) - 2x) &= 0 \end{aligned}$$

This means that for every value of x, we either have $f(x) = x$ or we have $f(x) = 2x$. Unlike the previous cases where the domain was split based on left-right or rationa-irrational, we do not know exactly how the domain splits up into these two definitions. But we do know that everywhere in the domain, the function behaves like one of these linear functions. And that itself is enough.

Thus, to show that the limit at 0 equals 0, we use the same old trick: we find $\epsilon - \delta$ winning strategies for the functions x and $2x$, and then we combine these winning strategies by picking the smaller of the δs and show that it works.

So I began by calling this a hard problem but we have somehow overcome the hard part and you should now consider it an easy problem.

4. Skeptic's victory: showing that something is not the limit

4.1. **Negating the existence of limit.**

Recall the definition of limit: We say that $\lim_{x\to c} f(x) = L$ if f is defined in a neighborhood of c (except possibly at c) and the following holds:

For every $\epsilon > 0$, there exists $\delta > 0$ such that, for all $x \in \mathbb{R}$ satisfying $0 < |x-c| < \delta$, we have $|f(x)-L| < \epsilon$.

Now, what does it mean to say that it is *not true* that $\lim_{x\to c} f(x) = L$? It means:

There exists $\epsilon > 0$ such that for all $\delta > 0$, there exists $x \in \mathbb{R}$ such that $0 < |x - c| < \delta$ but $|f(x) - L| \geq \epsilon$.

Note that the quantifiers change roles: *for all* becomes *there exists*, and *there exists* becomes *for all*.

In our game interpretation, this means that the skeptic has a winning strategy. In other words, the skeptic has a strategic choice of $\epsilon > 0$ such that whatever $\delta > 0$ the prover tries to use, the prover fails to trap the function within ϵ of the claimed limit.

4.2. **What does it mean to say that no limit exists?**

What does it mean to say that there is no L for which $\lim_{x\to c} f(x) = L$? In other words, what does it mean to say that the limit *does not exist*? It means that:

For all $L \in \mathbb{R}$, there exists $\epsilon > 0$ such that for all $\delta > 0$, there exists $x \in \mathbb{R}$ such that $0 < |x - c| < \delta$ but $|f(x) - L| \geq \epsilon$.

4.3. **An example of showing that no limit exists.**

Let's think back to the picture of the function $f(x) := \sin(1/x)$. This example was the example we used to realize that we need a certain kind of definition of limit. We noticed that, as $x \to 0$, the value of $\sin(1/x)$ didn't really approach anything because it was oscillating rapidly between -1 and 1. This led us to define limits in terms of traps; hence the $\epsilon-\delta$ definition.

We have to now come back full circle and try to explain *why*, as $x \to 0$, there does not exist a limit for $\sin(1/x)$. The way to think about this is that the function takes both the values 1 and -1 arbitrarily close to $x = 0$. We want to somehow make this be in contradiction with the fact that we can set ϵ-traps for arbitrarily small ϵ. Basically, there are two things going on:

(1) Suppose $\lim_{x\to 0}\sin(1/x) = L$. Suppose $\epsilon = 0.1$. Then, there exists a value of δ such that, if $0 < |x - 0| < \delta$, then $|\sin(1/x) - L| < 0.1$. In other words, we have $\sin(1/x) \in (L - 0.1, L + 0.1)$. Now, the interval $(L - 0.1, L + 0.1)$ has width 0.2, and hence, it cannot contain both the numbers 1 and -1. Thus, what we have is that there exists a $\delta > 0$ such that for $0 < |x| < \delta$, the function $\sin(1/x)$ does not take both the values 1 and -1.
(2) On the other hand, we have that for every $\delta > 0$, the function $\sin(1/x)$ takes the values $+1$ and -1 for $0 < |x| < \delta$. In fact, it takes *all* values in $[-1, 1]$. To see this, note that for $0 < |x| < \delta$, the set of possible values for $1/x$ is $(-\infty, -1/\delta) \cup (1/\delta, \infty)$. You can now see from the graph of the sine function that $\sin(1/x)$ takes all values in $[-1, 1]$.

Something similar works for the *Dirichlet function*, except that instead of using -1 and 1, we use 1 and 0. And, the main point that needs to be made for the Dirichlet function is that any open interval, no matter how small, contains infinitely many rational numbers and infinitely many irrational numbers.

5. Stability of limit

5.1. **Statement of the problem.** Let f be a function defined on some open interval $(c - p, c + p)$. Now change the value of f at a finite number of points $x_1, x_2, \dots, x_n$ and call the resulting function g. Show that if $\lim_{x\to c} f(x) = L$, then $\lim_{x\to c} g(x) = L$.

5.2. **Meaning of the problem.** Here's what the problem means. We start with some function f. For simplicity you can just imagine f to be a continuous function – a wiggly wavy curve where the x-value varies in the interval $(c-p, c+p)$. Now, we choose a few points at random and just move the value of the function. So, if say $f(5) = 7$, and we don't like that, we just move the value to 9, creating a *hole* at the point 7 here and filling in the poitn at 9. And we do this for finitely many points.

The new function that we get after we do all these moves, we choose to call g. Note that although the function f that we started with was continuous, g isn't. Of course, it isn't necessray that the function that we start with is continuous either. We could juts start with some function that wiggles and waves and jumps and then move a few values here and there and get a new function that wiggles and waves and jumps.

5.3. **What the question asks for, and the intuitive reason it is true.** The question says that if the limit $\lim_{x\to c} f(x)$ exists and is equal to L, then the limit $\lim_{x\to c} g(x)$ *also* exists and is equal to L. In other words, what this is saying is that the notion of limit is *stable* under changes of value at only finitely many points.

Now, the first thing you should notice is that if c itself is one of the x_is (i.e., one of the points where we are changing the value of the function) that should make no difference to the limit. Because, if you recall, the definition of the limit specifically excludes behavior *at the point* where we are taking the limit. The definition says:

For every $\epsilon > 0$, there exists $\delta > 0$ such that if $0 < |x - c| < \delta$, then $|f(x) - L| < \epsilon$. Note the $0 <$ part. This basically says that we aren't really imposing any condition when $x = c$. Specifically, remember that the set of x satisfying $0 < |x-c| < \delta$ is the set $(c-\delta, c+\delta) \setminus \{c\}$, which is the same as the set $(c-\delta, c) \cup (c, c+\delta)$.

So changing the value of the function at c doesn't affect anything. And once we've acknowledged this, we'll just assume that if c is equal to one of the x_is, we can just throw out that value of x_i. In other words, we'll just assume that we retain only those x_is that are not equal to c.

Now, all the *other* x_is – the points where we change the value of the function – are far away from c. What do I mean by that? I mean that we can choose a *small open interval* about c that excludes all the other x_is. (It may be helpful to think of the x_is as *bad points* that should be excluded from all decent society for being traitors to their function.)

And the main idea of limit is that it is intensely local – it only matters what is happening really really close to the point.[1] So, if all those other points are far away, the value of the function at those points shouldn't affect the value of the limit.

[1]There is a deep mathematical concept related to this called a *germ*, which means the essence of a function really really close to a point, but you don't have to bother about that right now.

5.4. **A reality check: why finitely many?** What is the significance of the fact that we are allowed to change the value of the function at only finitely many points? The thing is that if we are allowed to change the value of the function at infinitely many points, then we can affect the limits at some points. The reason is that when we have infinitely many points, it may not be possible to avoid all of them.

For example, suppose we have a function defined as the constant 5 on $(-1, 1)$ and we are interested in the limit at 0. Suppose now that we change the value of the function to the constant 7 at the points $1/2, 1/3, \ldots, 1/n, \ldots$. There are two things you should note: first, we have changed the value of the function at infinitely many points, and second, the value of the function at *most* points is still 5. So we haven't really changed the function all that much. However, even this small change is enough to disrupt the limit at 0. Because now, no matter how small an interval I choose about 0, that interval will contain some of those *bad points* – those points where the definition changed.

5.5. **Formalization of the proof in terms of ϵ and δ.** *Read this only after you feel you have understood the ideas at an intuitive level.*

We begin by throwing out any x_i that equals c, because the value of f or g at c is clearly of no relevance to the limits of the functions at c.

We need to show that, if $\lim_{x \to c} f(x) = L$, then, for every $\epsilon > 0$, there exists a $\delta > 0$ such that if $0 < |x - c| < \delta$, then $|g(x) - L| < \epsilon$.

We begin by noting that since $\lim_{x \to c} f(x) = L$, plugging in the $\epsilon - \delta$ definition shows that there exists $\delta_1 > 0$ such that if $0 < |x - c| < \delta_1$, then $|f(x) - L| < \epsilon$.

Now, let $\delta = \min\{\delta_1, |c - x_1|, |c - x_2|, \ldots, |c - x_n|\}$. The intuition here is to pick δ small enough so that none of the x_i are in the interval $(c - \delta, c + \delta)$.

Note that $\delta_1 > 0$, and since c is not equal to any of the x_is, $|c - x_i| > 0$. Thus, $\delta > 0$. We claim that this δ works.

To see this, we note two things.

First, if $0 < |x - c| < \delta$, we also have $0 < |x - c| < \delta_1$, and this gives:

$$|f(x) - L| < \epsilon \tag{1}$$

Second, if $0 < |x - c| < \delta$, we cannot have x equal to any of the x_is (why? think about this. This is the only part that I haven't justified in the proof). Hence, we get:

$$f(x) = g(x) \tag{2}$$

Combining (1) and (2), we get $|g(x) - L| < \epsilon$ for $0 < |x - c| < \delta$, completing the proof.

THEOREMS ON LIMITS AND CONTINUITY

MATH 152, SECTION 55 (VIPUL NAIK)

Difficulty level: Moderate to hard. There are a couple of proofs here that are hard to understand; however, you are not responsible for the proofs of these theorems this quarter. The statements of the theorems will be easy if you have seen them before, and somewhat hard if you have not.

Corresponding material in the book: Sections 2.3, 2.4, 2.5, 2.6.

Things that students should definitely get: The uniqueness theorem for limits. The statements of the theorems for limits and continuity in relation to the pointwise combinations of functions and composition. The fact that all the results for pointwise combination hold for one-sided limits and one-sided continuity. The statement and simple applications of the pinching theorem, intermediate-value theorem, and extreme-value theorem.

Things that students should hopefully get: The fact that the limit theorems have both a conditional existence component and a formula component. The way that the triangle inequality is used critically to prove the uniqueness theorem and the theorem on limits of sums. The importance of the continuity assumption for the intermediate-value theorem and of that as well as the closed interval assumption for the extreme-value theorem. How to think about counterexamples and weird functions in a way that builds intuition about the significance of the hypotheses to the theorems.

Executive summary

Limit theorems + quick/intuitive calculation of limits. Words...

(1) If the limits for two functions exist at a particular point, the limit of the sum exists and equals the sum of the limits. Similarly for product and difference.
(2) For quotient, we need to add the caveat that the limit of the denominator is nonzero.
(3) If $\lim_{x\to c} f(x) = L \neq 0$ and $\lim_{x\to c} g(x) = 0$, then $\lim_{x\to c}(f(x)/g(x))$ is undefined.
(4) If $\lim_{x\to c} f(x) = \lim_{x\to c} g(x) = 0$, then we cannot say anything offhand about $\lim_{x\to c}(f(x)/g(x))$.
(5) Everything we said (or implied) can be reformulated for one-sided limits.

Continuity theorems. Words ...

(1) If f and g are functions that are both continuous at a point c, then the function $f + g$ is also continuous at c. Similarly, $f - g$ and $f \cdot g$ are continuous at c. Also, if $g(c) \neq 0$, then f/g is continuous at c.
(2) If f and g are both continuous in an interval, then $f + g$, $f - g$ and $f \cdot g$ are also continuous on the interval. Similarly for f/g provided g is not zero anywhere on the interval.
(3) The composition theorem for continuous functions states that if g is continuous at c and f is continuous at $g(c)$, then $f \circ g$ is continuous at c. The corresponding composition theorem for limits is *not true but almost true*: if $\lim_{x\to c} g(x) = L$ and $\lim_{x\to L} f(x) = M$, then $\lim_{x\to c} f(g(x)) = M$.
(4) The one-sided analogues of the theorems for sum, difference, product, quotient work, but the one-sided analogue of the theorem for composition is not in general true.
(5) Each of these theorems at points has a suitable analogue/corollary for continuity (and, with the exception of composition, for one-sided conitnuity) on intervals.

Three important theorems. Words ...

(1) The pinching theorem states that if $f(x) \leq g(x) \leq h(x)$, and $\lim_{x\to c} f(x) = \lim_{x\to c} h(x) = L$, then $\lim_{x\to c} g(x) = L$. A one-sided version of the pinching theorem also holds.
(2) The intermediate-value theorem states that if f is a continuous function, and $a < b$, and p is between $f(a)$ and $f(b)$, there exists $c \in [a, b]$ such that $f(c) = p$. Note that we need f to be defined and continuous on the entire closed interval $[a, b]$.

(3) The extreme-value theorem states that on a closed bounded interval $[a, b]$, a continuous function attains its maximum and minimum.

Actions ...

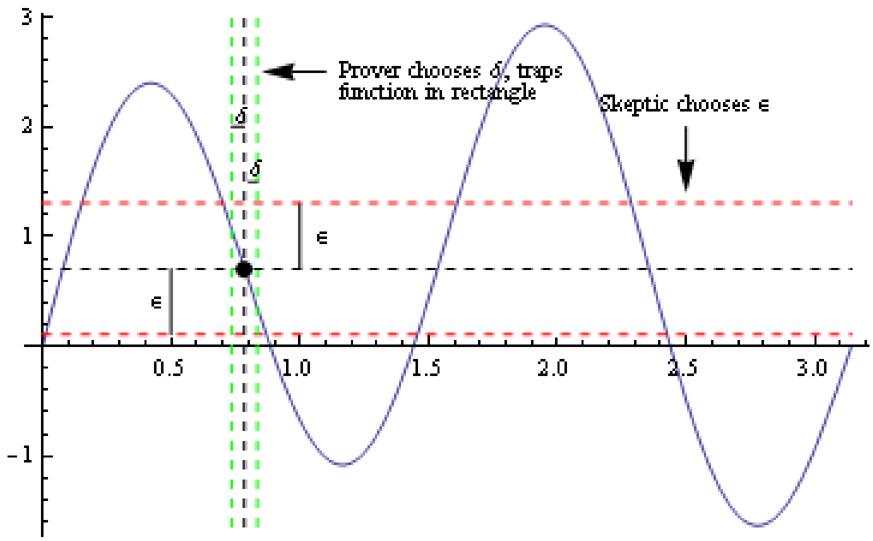
(1) When trying to calculate a limit that's tricky, you might want to bound it from both sides by things whose limits you know and are equal. For instance, the function $x \sin(1/x)$ taking the limit at 0, or the function that's x on rationals and 0 on irrationals, again taking the limit at 0.
(2) We can use the intermediate-value theorem to show that a given equation has a solution in an interval by calculating the values of the expression at endpoints of the interval and showing that they have opposite signs.

1. Limit theorems

This section discusses some important limit theorems.

Recall the definition. We say that $\lim_{x \to c} f(x) = L$ (as a two-sided limit) if, for every $\epsilon > 0$, there exists $\delta > 0$ such that, for every x such that $0 < |x - c| < \delta$, we have $|f(x) - L| < \epsilon$.

Now, that's quite a mouthful. Let's interpret it graphically. What it is saying is that: "for every ϵ"so we consider this region $(L - \epsilon, L + \epsilon)$, so there are these two horizontal bars at heights $L - \epsilon$ and $L + \epsilon$. Next it says, there exists a δ, so there exist these vertical bars at $c + \delta$ and $c - \delta$. And what we're saying is that if the x-value is trapped between the vertical bars $c - \delta$ and $c + \delta$ (but is not equal to c), the $f(x)$-value is trapped between $L - \epsilon$ and $L + \epsilon$.



The important thing to note here is that the value of δ depends on the value of ϵ. As I said last time, we can think of this as a game, where I am trying to prove to you that the limit $\lim_{x \to c} f(x) = L$ and you are a skeptic who is trying to catch me out. So you throw ϵs at me, and challenge me to show that I have a δ to trap that ϵ. And if I have a winning strategy, that enables me to find a δ for every ϵ that you throw at me, then yes, the limit is equal to L.

1.1. **What are limit theorems?** And why do we need them?

In the absence of limit theorems, we have two alternatives:

- Use our "intuition" – this is problematic, because while intuition works great for nice functions such as polynomials, it tsarts failing us as soon as we get to weirder functions.
- Use "first principles,", i.e., the $\epsilon - \delta$ definition of limits every time – this is very tedious even for experienced mathematicians.

Limit theorems provide a sort of middle ground that avoids the pitfalls at either end. Basically, what these theorems do is, show, using the $\epsilon - \delta$ definition of limits, that certain "intuitive" facts about limits are always true. Then, we can use these theorems guilt-free without having to wade through a mess of ϵs and δs.

So think of proving a theorem as an investment. It's like putting money in a savings account. You put money once, and you keep getting the interest from it. But the first thing you need to do is put in the hard work of earning and saving the money. And proving the theorems is like doing that hardwork.

1.2. **Review of some inequalities involving the absolute value.** So, before we plunge into the proofs, I want to review some facts about the absolute value function that are very important. I'll stick to two facts. The first is what is called the *triangle inequality*, and it goes like this:

$$|a+b| \leq |a| + |b| \ \forall \ a, b \in \mathbb{R}$$

Now, equality holds if either a or b is zero, or if they both have the same sign. Equality does *not* hold if a and b are of opposite signs. So, you may wonder, why the name *triangle inequality*? And to understand this, we need to think about triangles on the real line.

But before we get into that, first, let's recall what the triangle inequality in geometry states. It says that the sum of two sides of a triangle is greater than the third side. That's basically a manifestation of the fact that *straightest is shortest* – the straight line path between two points is shorter than a path that involves two straight lines.

Now, the inequality sign is strict, because in geometry, we don't use the word *triangle* if all the three vertices are collinear. By the way, if all the three vertices are collinear, we call the triangle a *degenerate* triangle. Let's say we included degenerate triangles. Then, the equality case could occur. In fact, it'll occur in precisely the case where the single side is between the two more extreme points.

So, here's how this relates to the statement involving absolute values. Consider the degenerate triangle with vertices the points 0, a, and $a+b$ on the number line. What are the side lengths? Well, the length of the side from 0 to a is $|a|$, the length of the side from a to $a+b$ is $|b|$, and the length of the side from 0 to $a+b$ is $|a+b|$. And so the result we have is:

$$|a+b| \leq |a| + |b|$$

which is our triangle inequality.

Verbally, what this is saying is that if you travel a distance of a and then travel a distance of b along the real line, you cannot be more than $a+b$ away from where you started. The farthest you can get is if both your two pieces of travel were in the same direction.

The other result, which we've already used a few times, is that the absolute value of the product of two real numbers equals the product of their absolute values.

$$|ab| = |a||b| \ \forall \ a, b \in \mathbb{R}$$

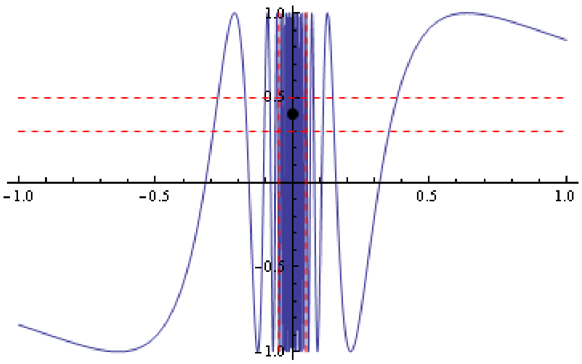
1.3. **Uniqueness theorem.** [**Note**: We will probably not go over this proof in class in full detail, and you are not expected to know this proof for any of your tests. However, I strongly suggest that you at least try to understand this proof temporarily. The ideas involved here are extremely useful for understanding some of the more advanced limits stuff that we will see in 153.]

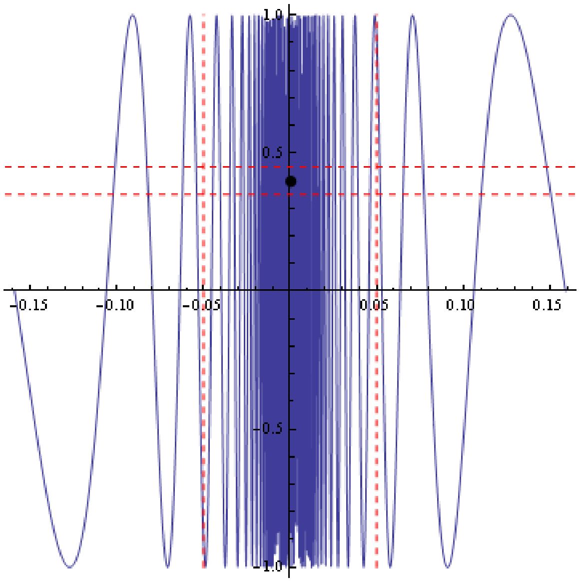
The first result that we establish is a theorem on the *uniqueness* of limits. And the intuition behind this goes back to our original discussion where I gave you a wrong definition of limit and found that the problem with that definition was that it gave rise to multiple limits. And we tweaked it and got the correct definition.

So the first thing we need to establish is that if the limit exists, then it is unique.

Okay, can you give an intuitive reason why that should be true?

Well, think back to our discussion on traps. And think back to the wrong definition of limit, and why $\sin(1/x)$ was problematic. The reason was that it was jumping a lot, and our right definition of limits, by creating traps, avoided that.





So you can think of what we are doing here as a proof by contradiction. We will show that if there are two real numbers $L \neq M$, it cannot be the case that *both* L and M satisfy the $\epsilon - \delta$ definition for $\lim_{x \to c} f(x)$.

Now, this is an example where we want to show that the limit *cannot* be something. So this is an example of a situation where you are trying to prove the opposite of a statement. We already wrote down what it means to say that as $x \to c$, $f(x)$ does not approach L. Let's recall it: there exists $\epsilon > 0$ such that for every $\delta > 0$, there exists x satisfying $0 < |x - c| < \delta$ and $|f(x) - L| \geq \epsilon$.

Okay, now let's take a step back and see what we're really trying to achieve. Simply put, think about it as two games. There's the L-game, which is the game where I'm trying to prove the limit is L. And there's the M-game, which is where I'm trying to prove the limit is M. And what is true is that no matter what, you, the skeptic, have a winning strategy for at least one of the games.

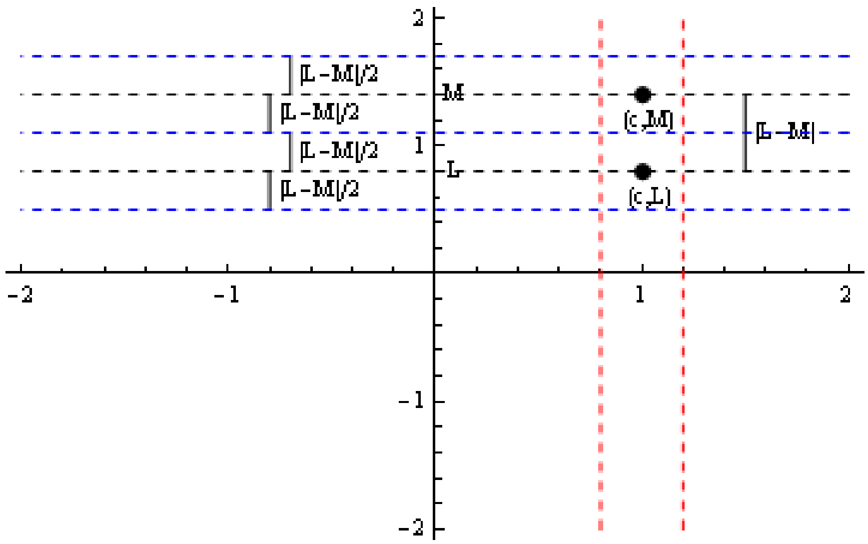
So what you do is to try to trap me with my own trap, literally speaking. Because what I'm claiming is that for x sufficiently close to c, the function is trapped really close to L, but it's also trapped really close to M. So what you need to do is call the bluff on me, by forcing me to get too close to both L and M for comfort. Basically, what you want to call me out on is the assertion that a function can be trapped in two places at the same time. And, to cut a long story short, the ϵ that you stump me with is:

$$\epsilon = \frac{|L - M|}{2}$$

So, you stump me with this ϵ in the L-game and in the M-game. And suppose I throw back δ_1 at you in the L-game and δ_2 at you in the M-game. So I'm claiming that when the x-value comes within δ_1 of c, the function value is trapped with ϵ of L, and when the x-value comes within δ_2 of c, the function value is trapped within ϵ of M.

So let $\delta = \min\{\delta_1, \delta_2\}$. Then, what we have is that:

if $0 < |x - c| < \delta$, then $|f(x) - L| < \epsilon$, and $|f(x) - M| < \epsilon$.



So for x in this small region around c, $f(x)$ is within ϵ of L and within ϵ of M. But the picture makes clear that this is not possible, because the ϵ-disks around L and M don't meet. The formalism behind this is the triangle inequality:

$$|L - M| = |(L - f(x)) + (f(x) - M)| \leq |L - f(x)| + |f(x) - M| < 2\epsilon = |L - M|$$

So, we get $|L - M| < |L - M|$, a contradiction.

1.4. **Limits for sum, difference, product, ratio.** Okay, now we'll state the results for the limits of sums, differences, scalar multiples, products, and ratios. None of the proofs are in the syllabus, but I've sketched the proof for sums. The book has proofs for all the limits.

Suppose f and g are two functions defined in a neighborhood of the point c. Then, if $\lim_{x \to c} f(x)$ and $\lim_{x \to c} g(x)$ are well-defined, we have the following:

(1) $\lim_{x \to c}(f(x) + g(x))$ is defined, and equals the sum of the values $\lim_{x \to c} f(x)$ and $\lim_{x \to c} g(x)$.
(2) $\lim_{x \to c}(f(x) - g(x))$ is defined, and equals $\lim_{x \to c} f(x) - \lim_{x \to c} g(x)$.
(3) $\lim_{x \to c} f(x)g(x)$ is defined, and equals the product $\lim_{x \to c} f(x) \lim_{x \to c} g(x)$.

The scalar multiples result basically states that if $\lim_{x \to c} f(x)$ exists, and $\alpha \in \mathbb{R}$, $\lim_{x \to c} \alpha f(x) = \alpha \lim_{x \to c} f(x)$.

By the way, a week ago, we defined the notions of sum, difference, and product, of functions. So with that notation, we can rewrite the results as:

$$\begin{aligned} \lim_{x \to c}(f+g)(x) &= \lim_{x \to c} f(x) + \lim_{x \to c} g(x) \\ \lim_{x \to c}(f-g)(x) &= \lim_{x \to c} f(x) - \lim_{x \to c} g(x) \\ \lim_{x \to c}(f \cdot g)(x) &= \lim_{x \to c} f(x) \lim_{x \to c} g(x) \end{aligned}$$

A couple of important additional points. The first is that the results I mentioned state a little more than what is captured in the formulas. The subtlety arises because every limit that we write need not exist.

What the sum result says is that *if* the two limits $\lim_{x \to c} f(x)$ and $\lim_{x \to c} g(x)$ *both exist*, the the limit for $f + g$ exists *and* is given by the formula. So, the result is a *conditional existence result plus a formula.* Note that it may very well be the case that the limit for $f + g$ exists but the individual limits – those for f and g, do not exist. For instance, if $f(x) = 1/x$ and $g(x) = -1/x$, then f and g do not have limits at 0, but $f + g$ does have a limit at 0.

Similarly for the results about difference, product, and scalar multiples.

1.5. **Result for the ratio.** For the ratio (also called the quotient), we have the result that if $\lim_{x\to c} f(x)$ and $\lim_{x\to c} g(x)$ exist, *and* if $\lim_{x\to c} g(x) \neq 0$, then we have:

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

1.6. **Indeterminate form $\to 0/\to 0$ limits.** In addition to the previous limit theorems, there are some important facts that should feel familiar if you have done limit computations.

Suppose $\lim_{x\to c} f(x) = \lim_{x\to c} g(x) = 0$. Then the quotient limit $\lim_{x\to c} f(x)/g(x)$ has what is called a $0/0$ form – more precisely, it is of the form $(\to 0)/(\to 0)$. This is an example of an *indeterminate form*. An indeterminate form is a form of a limit that does not allow us to conclude anything specific about the value of the limit. A limit of the form $(\to 0)/(\to 0)$ may or may not exist. Further, it could "not exist" in practically all the possible values that limits happen to "not exist" (infinite limits, oscillatory limits, etc.). $(\to 0)/(\to 0)$ may be finite and nonzero, it may be zero, it may be going to $+\infty$, to $-\infty$, different infinities from different sides, oscillatory between finite bounds, oscillatory between infinite bounds, etc.

An indeterminate form does *not* mean that we can throw up our hands. To the contrary, indeterminate form means that the limit *needs more work*. This extra work typically involves understanding more about the nature of the *specific functions* f and g near the point of approach. For rational functions, we try to cancel common factors between the numerator and denominator. There are more general approaches such as trigonometric limits, l'Hopital's rule and power series, which we will see later in the course.

Intuitively, what matters is: does the numerator go to 0 more quickly, does the denominator go to 0 more quickly, or do they both go to 0 at roughly the same rate. We will explore this theme in mind-numbing theme later in 152 and even more in 153.

1.7. **Lonely denominator blow-ups: undefined limit.** If $\lim_{x\to c} f(x) = L \neq 0$ and $\lim_{x\to c} g(x) = 0$, then $\lim_{x\to c} f(x)/g(x)$ is not defined. In other words, the limit *does not exist*. When we later study infinity as a limit, we will consider in more detail whether the (one-sided) limit exists as an infinity. Note that if a limit is infinite, we still say that the limit "does not exist."

1.8. **One-sided and two-sided limits.** So far, in all the situations where we have been saying that *the limit exists*, we mean that the *two-sided* limit exists. Recall that we have that $\lim_{x\to c} f(x) = L$ if f is defined in an open interval about c (except possibly at c) and if, for every $\epsilon > 0$, there exists $\delta > 0$ such that for every x such that $0 < |x - c| < \delta$, we have $|f(x) - L| < \epsilon$.

Let's also recall what it means to say that the *left-hand limit exists*. We say that $\lim_{x\to c^-} f(x) = b$ if f is defined to the immediate left of c, and if, for every $\epsilon > 0$, there exists $\delta > 0$ such that for every x such that $0 < c - x < \delta$, we have $|f(x) - b| < \epsilon$.

Basically, what's happening is that now, we need only a one-sided trap for δ, i.e., a trap of the form $(c - \delta, c)$ rather than a trap of the form $c - \delta, c + \delta)$.

Similarly, we say that $\lim_{x\to c^+} = b$ if, for every $\epsilon > 0$, there exists $\delta > 0$ such that for every x such that $0 < x - c < \delta$, we have $|f(x) - b| < \epsilon$.

Now, it turns out that all the results we proved *so far* about limits also hold for one-sided limits from either side. So, for instance, we have the following for left-hand limits.

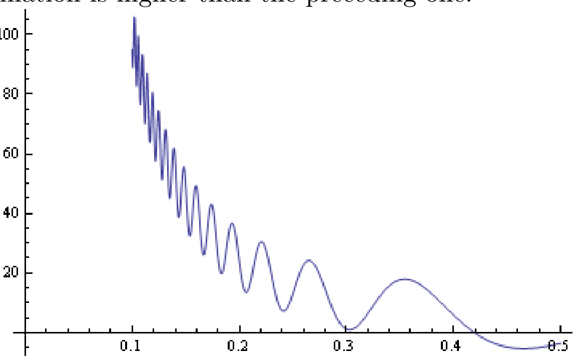
(1) If the left-hand limit $\lim_{x\to c^-} f(x)$ exists, it is unique.
(2) If $\lim_{x\to c^-} f(x)$ and $\lim_{x\to c^-} g(x)$ exist, then $\lim_{x\to c^-}(f(x) + g(x))$ exists and equals the sum of the individual limits.
(3) If $\lim_{x\to c^-} f(x)$ and $\lim_{x\to c^-} g(x)$ exist, then $\lim_{x\to c^-}(f(x) + g(x))$ exists and equals the difference of the individual limits.
(4) If $\lim_{x\to c^-} f(x)$ and $\lim_{x\to c^-} g(x)$ exist, then $\lim_{x\to c^-}(f(x)g(x))$ exists and equals the product of the individual limits.
(5) If $\lim_{x\to c^-} f(x)$ exists and α is a real number, then $\lim_{x\to c^-} \alpha f(x) = \alpha \lim_{x\to c^-} f(x)$.

Analogous results hold for right-hand limits.

1.9. **Infinity as a limit.** In the discussion so far, when we said that the limit *exists*, we meant that it *exists* and is *finite*. And that's the way it'll continue to be. Nonetheless, since the book makes some mentions of infinity, and since you may have seen these concepts when going through intuitive introductions to limits, let me briefly describe the role of infinity. We will master these definitions formally a little later in the course. For now, the coverage is (relatively) informal.

When we're thinking in terms of limits, then we say that the limit is ∞, or $+\infty$, when the number is getting larger and larger and not bouncing back to very small values. This doesn't mean it can't oscillate – it can. But rather, for every number N you pick, there exists a $\delta > 0$ such that, if $0 < |x - c| < \delta$, then $f(x) > N$. In other words, it eventually gets above any barrier *and stays above*. And if that's the case, we say that the limit is $+\infty$.

For instance, the picture below is of a function that approaches $+\infty$ as $x \to 0$ – because of page-fitting considerations, the function has been plotted only till $x = 0.1$ – but the function oscillates. However, each oscillation is higher than the preceding one.



Similarly, there is the notion of the limit being $-\infty$.

Now, what happens with functions, and the classic example is the function $1/x$, but you'll see this for other functions, is that the left-hand limit approaches infinity from one direction and the right-hand limit approaches infinity from the other direction. So in this case, for the function $f(x) := 1/x$, the left-hand limit is $-\infty$ and the right-hand limit is $+\infty$. By the way, neither limit *exists* in the sense that we will use the word, because neither limit is finite. But a two-sided limit doesn't even exist as an infinity, because the two sides are approaching two different infinities.

On the other hand, for the function $g(x) := 1/x^2$, both the left-hand limit and the right-hand limit are $+\infty$.

And generalizing from these examples, we can see some general rules emerging:

(1) If $\lim_{x \to c} f(x)$ is a positive number and $\lim_{x \to c} g(x) = 0$, but the approach is *from the positive direction*, then $f(x)/g(x) \to \infty$ as $x \to c$. Analogous observations apply to negative numbers.
(2) If $\lim_{x \to c} f(x)$ is a positive number and $\lim_{x \to c} g(x) = 0$, but the approach is *from the positive direction*, then $f(x)/g(x) \to \infty$ as $x \to c$. Analogous observations apply to negative numbers.

We'll talk more about infinities as limits, and consider more complicated cases, a little later in the course.

1.10. **Proof that limit of sums is sum of limits.** [**Note**: We will probably not go over this proof in class in full detail, and you are not expected to know this proof for any of your tests. However, I strongly suggest that you at least try to understand this proof temporarily. The ideas involved here are extremely useful for understanding some of the more advanced limits stuff that we will see in 153.]

We'll now discuss how to prove the statement that, as an English phrase, would read: "the sum of the limits is the limit of the sums". In other words, we are trying to prove that if $\lim_{x \to c} f(x) = L$ and if $\lim_{x \to c} g(x) = M$, then $\lim_{x \to c} f(x) + g(x) = L + M$.

So the way to think about it is that f comes close to L and g comes close to M, so doesn't $f + g$ comes close to $L + M$? Yes, it does. But to make that precise, what we need to do is to loosen the definition of what it means to come close.

You've probably heard of *rounding errors*. For instance, you may say that 1.4 rounds off to 1 and 2.3 rounds off to 2. But when you add the two numbers, you get 3.7, and 3.7 rounds off to 4, rather than $1 + 2 = 3$.

So, the upshot is that just because a is close to a' and b is close to b', doesn't necessarily mean that $a+b$ is just as close to $a'+b'$. However, even if it isn't *as close*, it is still close. The point is that when you add things, the *margins of error add*.

So the way we use this idea is to make sure that our margins of error for the functions f and g are both so small that when you add them up, you still get a small margin of error.

Let's now flesh out the proof details. We need to show that if $\lim_{x \to c} f(x)$ exists and $\lim_{x \to c} g(x)$ exists, then $\lim_{x \to c}(f(x)+g(x))$ exists. Let's call $L = \lim_{x \to c} f(x)$ and $M = \lim_{x \to c} g(x)$.

Let's discuss this. What we need to do is to show that, for every $\epsilon > 0$, we need to find a $\delta > 0$ such that if $0 < |x-c| < \delta$, then $|(f(x)+g(x)) - (L+M)| < \epsilon$. Here's how we do this.

Since f is continuous, consider the value $\epsilon/2$. There exists a value $\delta_1 > 0$ such that, if $0 < |x-c| < \delta_1$, we have $|f(x) - L| < \epsilon/2$. What we're doing here is using $\epsilon/2$ as the value of ϵ for f.

Similarly, for g, we use $\epsilon/2$ again. So, there exists a value $\delta_2 > 0$ such that, if $0 < |x-c| < \delta_2$, we have $|g(x) - M| < \epsilon/2$.

Now consider $\delta = \min\{\delta_1, \delta_2\}$. Then, if $0 < |x-c| < \delta$, we have $0 < |x-c| < \delta_1$ and $0 < |x-c| < \delta_2$, so:

$$|f(x) - L| < \epsilon/2 \tag{*}$$
$$|g(x) - M| < \epsilon/2 \tag{**}$$

We thus get, using the triangle inequality and (*) and (**):

$$|(f(x)+g(x)) - (L+M)| \le |f(x) - L| + |g(x) - M| < \epsilon/2 + \epsilon/2 = \epsilon$$

2. Continuity theorems

We now proceed to a discussion of theorems on continuity. Most of these are analogous to corresponding theorems about limits.

2.1. Recall the definition.

It may be worthwhile recalling the definitions of limit and continuity side by side. Let's do that.

We say that $\lim_{x \to c} f(x) = L$ if f is defined in an open interval about c (except possibly at c itself) and, for every $\epsilon > 0$, there exists $\delta > 0$ such that, for every x satisfying $0 < |x-c| < \delta$, we have $|f(x) - L| < \epsilon$.

We say that f is continuous at c if f is defined in an open interval about c (including at the point c) and, for every $\epsilon > 0$, there exists $\delta > 0$ such that, for every x satisfying $|x-c| < \delta$, we have $|f(x) - f(c)| < \epsilon$.

Some differences: for the definition of limit, we do not require the function to be defined *at* the point c, but for the definition of continuity, we do. The L that we use for the definition of continuity is the value $f(c)$. Also, we can drop the $0 <$ part in the definition of continuity.

2.2. Theorems about continuity.

The definition we gave above was for a function being continuous *at a point*, and we can use that definition, along with the limit theorems, to prove that continuity at a point is preserved by sums, scalar multiplies, differences, products, and, if the denominator function is not zero, ratios. Explicitly (Theorem 2.4.2, Page 84):

If f and g are functions that are both continuous at a point $c \in \mathbb{R}$, then:

(1) $f+g$ is continuous at c.
(2) $f-g$ is continuous at c.
(3) $f \cdot g$ is continuous at c.
(4) If $g(c) \neq 0$, then f/g is continuous at c.

Further, if f is continuous at c, then αf is continuous at c for any real number α.

Why are these statements true? Essentially, they are all immediate corollaries of the corresponding statements for limits. For instance, if you take for granted the theorem that $\lim_{x \to c}(f+g)(x) = \lim_{x \to c} f(x) + \lim_{x \to c} g(x)$, then given that f and g are both continuous at c, you can substitute the values and get $\lim_{x \to c}(f+g)(x) = f(c) + g(c) = (f+g)(c)$. The same proof idea works for differences and scalar multiples.

2.3. **Composition of continuous functions.** We'll next state a result, which is Theorem 2.4.4 of the book, about the composition of continuous functions. You don't need to know the proof of this statement.

The results is that if f and g are functions such that g is continuous at c and f is continuous at $g(c)$, then the composite function $f \circ g$ is continuous at c. And remember the way the composition works – the function written on the right is the one that is applied first. So here's this element c, and we first apply g, and we get to $g(c)$. Then we apply f, and we get to $f(g(c)) = (f \circ g)(c)$.

So I hope you see why it is important to require that f is continuous at $g(c)$ and g is continuous at c. In particular, it does not matter whether f is continuous at c, because the input that is fed into f is not c but $g(c)$.

2.4. **One-sided continuity.** In a previous lecture, we defined one-sided continuity. Left-continuity means that the left-hand limit exists and equals the value of the function at the point. Right-continuity means that the right-hand limit exists and equals the value of the function at the point.

And, as you might have guessed, the results we stated about sum, difference, scalar multiples, product, and quotient of functions being continuous extends to one-sided continuity. It turns out, interestingly, that the results about composition do *not* necessarily hold in the one-sided sense. In fact, one of your quiz questions on Friday of the first week was exactly about this issue. The main reason is that, for $f \circ g$ to have the required one-sided continuity at c, we need that $g(x)$ approach $g(c)$ from the correct direction.

Some weaker versions can be salvaged: for instance, if f and g are both left-continuous functions, and g is an increasing function, then $f \circ g$ is also left-continuous. There are other weaker versions too, which we will not get into here.

2.5. **Continuity theorems also hold on intervals.** So far, we have stated the continuity theorems at individual points. From these, we can deduce continuity theorems on intervals. Specifically, if I is an interval (of whatever sort), and f and g are continuous functions on I, then:

(1) $f + g$ is continuous on I.
(2) $f - g$ is continuous on I.
(3) $f \cdot g$ is continuous on I.
(4) f/g is continuous at those points of I where it is defined, i.e., where g takes nonzero values.

The interval version of the result for composition is a little trickier, because to deduce the continuity of $f \circ g$, we need f to be continuous, not on the original domain of g, but on the range of g which feeds into the domain of f. Here is one formulation.

Suppose I and J are intervals in $\mathbb{R}$, f is a continuous function on J, and g is a continuous function on I such that the range of g is contained in J. Then $f \circ g$ is a continuous function on I.

2.6. **Most functions you've seen are continuous.** So this is the time to pause and review and think: "which of the functions that we have seen are continuous, and do we have the tools to make sure?"

Well, with the tedious $\epsilon - \delta$ definition of limits, we actually proved that the constant functions are continuous, and that the function $f(x) = x$ is continuous. And we did this basically by showing that the limit at any point equals the value at the point. And now, we know that we can multiply things together, multiply by scalars, and add. And if you think for a moment, you'll see that that shows that polynomial functions are continuous.

For instance, the polynomial $2x^3 - 3x^2 + x + 1$ is the sum of the functions that send x to $2x^3$, $-3x^2$, x, and 1 respectively. Each of these functions itself is a product of multiple copies of the function sending x to itself, multiplied by some scalar. And each step of the construction/deconstruction of a polynomial preserves continuity. So we see why/how polynomial functions are continuous.

Rational functions, which are functions obtained as the ratio of two polynomials, are not necessarily globally defined, because the denominator may blow up at some point. However, the ratio results that we have established show that, at all the places where the denominator does not blow up, the rational function is continuous. Hence, with rational functions, we are in the position that *wherever the function is defined, it is continuous.*

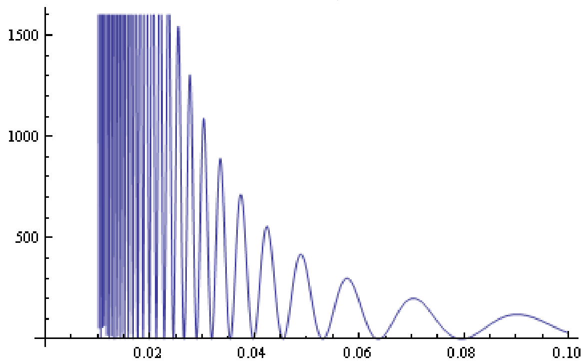
Another thing you should know is that the trigonometric functions sin and cos are continuous. And, since the trigonometric function tan is defined as the ratio of these, tan is again continuous at all the points where it is defined – the points where it is not continuous are the points where it is not defined, which are the points where cos takes the value 0.

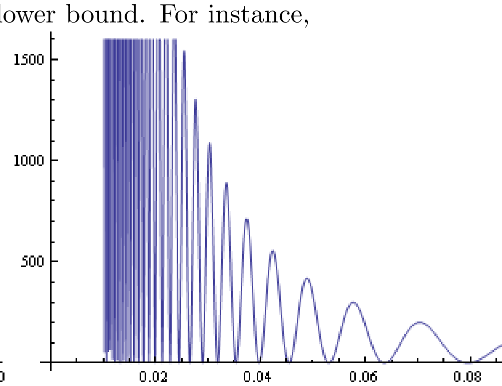
2.7. **Partying wild with freaky functions.** There are two directions of approach on the real line: left and right. And hence we can talk of the left-hand limit and the right-hand limit. And it is important that the real line has two directions.

So you're happy, because there are only two directions of approach, and the function comes in nicely form both sides – but that's not really the case. We saw the example of $\sin(1/x)$, frolicking back and forth cheerfully and indifferent to the coming crash at 0. And then in the homework you've been plagued by things like the Dirichlet function that is defined separately on the rationals and irrationals. And what these things show is that even the functions defined on what appears to be a straight line can show incredible diversity and jumping up and down.

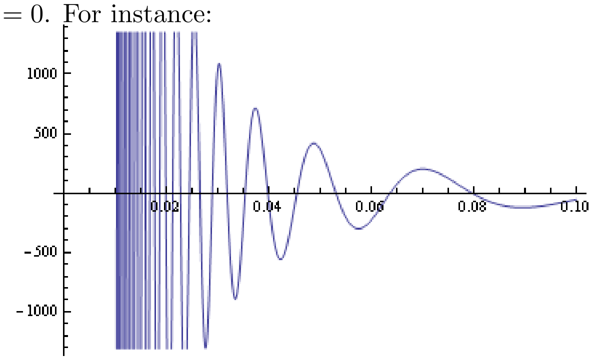
So let's think a bit about what can happen when a function is defined around a point but a one-sided limit is not defined. Here are some possibilities:

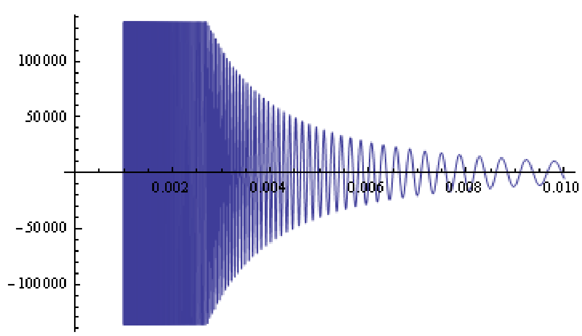
(1) As x approaches c, the function *oscillates* or jumps around, so it doesn't settle down, but it is still bounded. Some examples of this are the $\sin(1/x)$ function and the Dirichlet function.
(2) As x approaches c, the function heads for $+\infty$. This means that whatever height you set, the function eventually crosses that height and stays above. For instance, $1/x^2$, for $c = 0$, from either side. Or $1/x$ from the right side.
(3) As x approaches c, the function heads for $-\infty$. For instance, $-1/x^2$, for $c = 0$. Or $1/x$ from the left side.
(4) As x approaches c, the function oscillates between $+\infty$ and some finite lower bound. For instance, $\sin^2(1/x)/x^2$, for $c = 0$.





(5) As x approaches c, the function oscillates between $+\infty$ and $-\infty$. For instance, $\sin(1/x)/x^2$, for $c = 0$. For instance:





This is just scratching the surface. And to complicate matters further, we could have different behavior from the left and from the right. So you see that there's realy a wild party going on here.

3. Three important theorems

3.1. **The pinching theorem.** This theorem is also called the *squeeze theorem* or the *sandwich theorem.* Here's what it says:

If f, g, h are functions defined in a neighborhood of c, with the property that close to c, we have $f(x) \leq g(x) \leq h(x)$ for all x, and if $\lim_{x \to c} f(x) = \lim_{x \to c} h(x) = L$ for all x, then we also have $\lim_{x \to c} g(x) = L$.

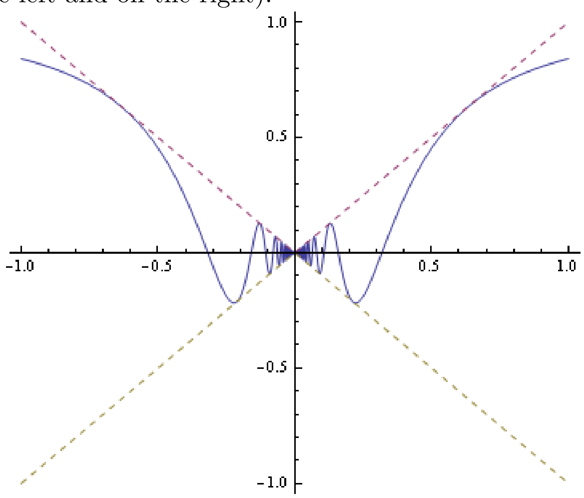
Analogous results hold for the left-hand limits and right-hand limits.

Basically, what this says is that if a function is trapped between two functions, both of which are approaching a particular value, then the function trapped in between also approaches that same value.

Here are some applications of this theorem.

(1) Recall the function in Homework 2: a funcion g defined as $g(x) = x$ for rational values of x and $g(x) = 0$ for irrational values of x. We want to show that $\lim_{x \to 0} g(x) = 0$. We first argue this for the right-hand limit. On the right, if we define $f(x) := 0$ and $h(x) := x$, then $f(x) \leq g(x) \leq h(x)$, and both f and h approach 0 at 0. Hence, g also approaches 0 at 0. On the left, we have $h(x) \leq g(x) \leq f(x)$, and again, since both f and h approach 0, so does g.

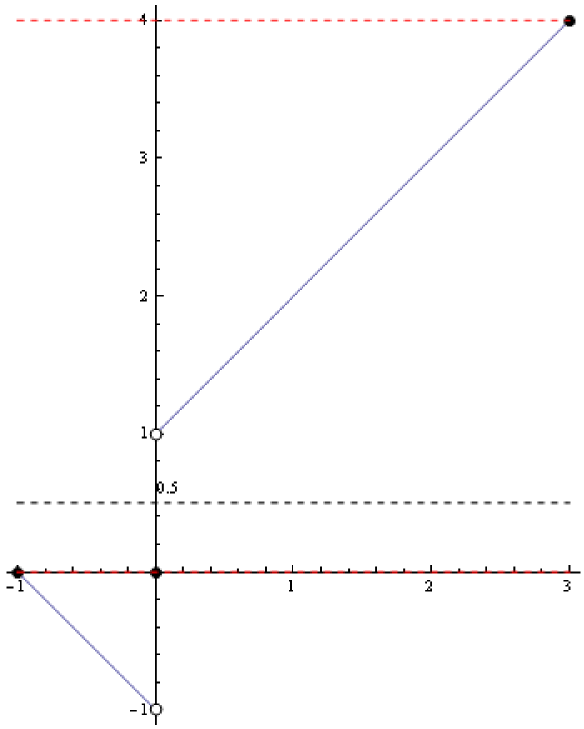
(2) Another example is the function $g(x) := x \sin(1/x)$. This is different from the $\sin(1/x)$ example that we saw earlier, because with this new function, the coefficient x causes a *damping* in the amplitude of the oscillations. To show that $\lim_{x \to 0} g(x) = 0$, we can use the pinching theorem, by squeezing g between the functions $f(x) = x$ and $h(x) = -x$ (again, the pinching will occur in different ways on the left and on the right).



Intermediate-value theorem. The intermediate-value theorem says that if f is a continuous function on a closed interval $[a, b]$, with $f(a) = c$ and $f(b) = d$, f takes every possible value between c and d. If $c < d$, this would mean that the range of f contains $[c, d]$. If $c > d$, this would mean that the range of f contains $[d, c]$.

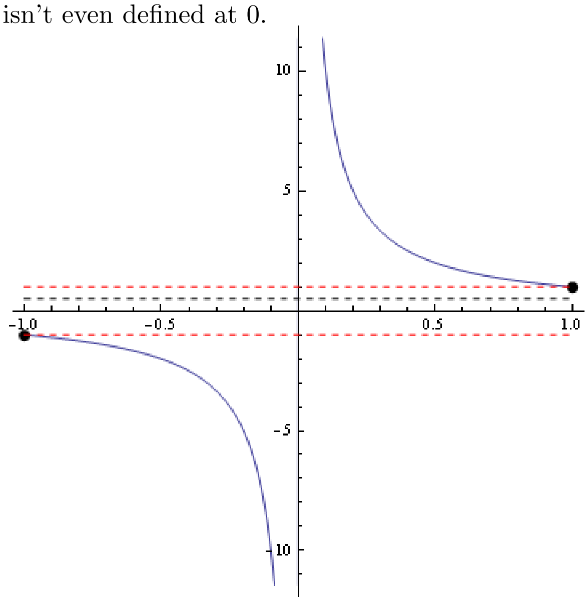
Note that f may take other values as well – for instance, it could go really high somewhere in between and then come back down, but what this theorem tells us is that it takes *at least* all the values between c and d.

The important thing here is that the function needs to be *continuous*. And the significance of continuity is that the function cannot suddenly *jump* from one place to the other – it has to go through all the intermediate steps. The graph below, for instance, plots the function $f(x) := (x + 1) \operatorname{sgn}(x)$ on the interval $[-1, 3]$, where we define sgn as the signum functions, which is -1 on negative numbers, 0 at 0, and 1 on positive numbers. Note that $f(-1) = 0$ and $f(3) = 4$, but f does not take the value $1/2$ (which is between $f(-1)$ and $f(3)$) anywhere on its domain.



Here's an important caveat. If the function isn't continuous, we may get wrong conclusions by applying the intermediate-value theorem.

For instance, consider the function $f(x) = 1/x$. Then, $f(-1) = -1$ and $f(1) = 1$. Hence, a naive application of the intermediate-value theorem would suggest that there exists $x \in [-1, 1]$ such that $f(x) = 0$. But this is nonsense – $1/x$ can never be equal to 0. So, something went wrong in our application of the intermediate-value theorem. What went wrong? Just the fact that $1/x$ is not continuous on $[-1, 1]$ – in fact, it isn't even defined at 0.



Another wrong application would be to say that since $f(-1) = -1$ and $f(1) = 1$, there exists $x \in [-1, 1]$ such that $f(x) = 1/2$. This is wrong, because the only x for which $f(x) = 1/2$ is $x = 2$, which is not in the interval $[-1, 1]$. Again, the reason we went astray is that the function $f(x) = 1/x$ is not continuous on $[-1, 1]$.

3.2. **Justifying the existence of the square root function.** One way in which the intermediate-value theorem gets used, and that is to justify the existence of inverse functions. We'll discuss this in greater generality probably toward the end of the course.

Suppose we needed to justify the existence of $\sqrt{3}$. In other words, we needed to show the existence of a positive number x such that $x^2 = 3$. Here's how the intermediate-value theorem can be used. Consider the function $f(x) := x^2$. This function is everywhere continuous. Now, $f(1) = 1$ and $f(2) = 4$. Hence, by the intermediate-value theorem there exists some $x \in [1, 2]$ such that $f(x) = 3$.

3.3. **"Solving" equations using the intermediate-value theorem.** When you looked at an equation, your first urge was probably to solve it. And you solved it by manipulating stuff, applying the formula for the root of a quadratic, etc., etc. You had a toolkit of methods to solve equations, and you tried to faithfully apply this toolkit.

But there are times when you cannot find precise expressions for the solutions of equations. For instance, if I write a polynomial equation of degree 5, and ask you to solve it, may be you try a few values and then give up, but there's no general formula you can plug in to find the solutions (in fact, mathematicians have actually *proved* that general formulas do not exist – a proof you will probably not see unless you choose to major in mathematics). So, that's bad news. And similarly, if I write $\cos x = x$ and ask you to solve that, you have no mathematical way of finding a solution.

However, even if we cannot find exact solutions, we may be able to determine whether solutions exist, and even narrow down the interval in which they exist. And one tool in doing this is the intermediate-value theorem.

So consider the equation $\cos x = x$. The first thing you do is to take the difference, which in this case is $\cos x - x$. This is continuous, so the intermediate-value theorem applies, so to show that this difference is zero somewhere, it is enough to show that there's some place where it's positive and some place where it's negative. Well, let's draw the graphs of the functions x and $\cos x$ to get a bit if the intuition.

So looking at the graphs, you see that it's likely that a solution exists somewhere between 0 and $\pi/2$, because that's where the function x overtakes the function $\cos x$. Let's try to see this using the intermediate-value theorem. At 0, we have $\cos x - x = 1$, and at $\pi/2$, we have $\cos x - x = -\pi/2$. So, the function $\cos x - x$ goes from a positive value 1 to a negative value $-\pi/2$, which means that at some point in between, the function must be zero, and that's the point where $\cos x = x$.

Now, we can actually narrow down the value where $\cos x = x$ a little further, by trying to evaluate $f(x)$ for other values of x. For instance, we see, by evaluation, that $f(\pi/3) = 1/2 - \pi/3 < 0$, so the intermediate-value theorem tells us that $f(x) = 0$ for some $x \in [0, \pi/3]$. Next, we try $f(\pi/4) = 1/\sqrt{2} - \pi/4 < 0$, so, in fact, $f(x) = 0$ for some $x \in [0, \pi/4]$. And we can narrow things down still further by checking that $f(\pi/6) > 0$, so in fact, there is a solution to $f(x) = 0$ for $x \in [\pi/6, \pi/4]$.

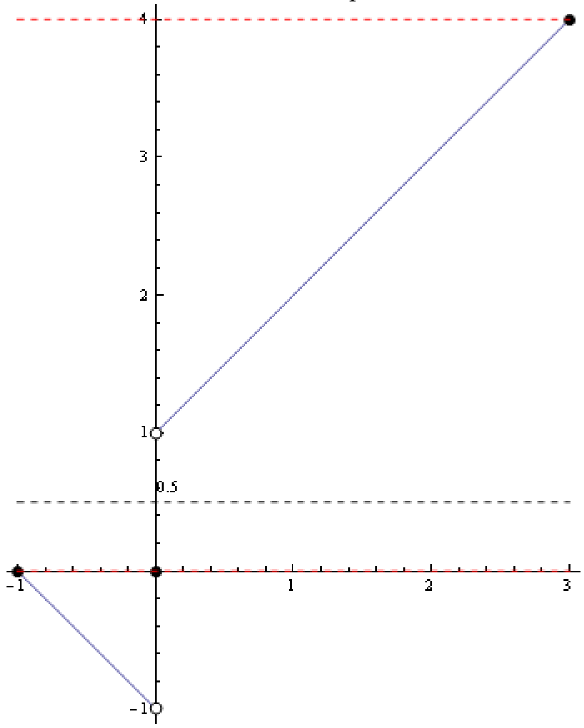
So, we see that we can use the intermediate-value theorem, along with evaluating the function at mutliple points, to narrow down pretty well where a function is zero. If you're more interested in these techniques, you should read more about the *bisection method* – it is discussed on Page 102 under the header *Project 2.6*.

3.4. **Extreme-value theorem.** Another theorem, that you should know, though you will not have the opportunity to apply it much right now, is the extreme-value theorem. This states that if f is a continuous real-valued function on a closed interval $[a, b]$, then f attains its maximum and its minimum, and of course, by the intermediate-value theorem, all the values in betweem. So if M is the maximum and m is the minimum of f, then the range of f is an interval of the form $[m, M]$.

Okay, there are many parts to this theorem, so let's understand it part-by-part. What does it mean to say that the function *attains its maximum*? Basically, we mean that there is some value M in the range of f that is the largest value in the range of f. And similarly, *attains its minimum* happens when there is some value in the range of f that is the smallest value in the range of f.

So, may be you're thinking that every function shoul attain its maximum and minimum. That's not true. In fact, a function on an interval stretching to infinity, or a function defined on an open interval, need not attain a maximum or minimum. For instance, the function $1/x$ on the interval $(0, \infty)$ doesn't attain either a maximum or minimum – it tends to (but doesn't reach) ∞ on one side and tends to (but doesn't reach) 0 on the other side.

Also, a discontinuous function defined on a closed interval need not attain a maximum and minimum. For instance, the function $(x+1)\operatorname{sgn}(x)$ on $[-1,3]$, which we discussed a little while ago, does not attain a minimum value because of the open circle at the low point:



So okay, we are somehow using something about continuity – may be it isn't clear what, but something, and we're also using something about closed intervals, to say that there is a maximum and minimum. What about the next part of the statement, which says that the range is precisely the stuff that's in between? Well, everything in the range has to be between the maximum and minimum – that's the definition of maximum and minimum. But why does everything between the maximum and minimum have to be in the range? Well, you can think of that as the intermediate-value theorem in action again.

COMPOSITION THEOREM FOR LIMITS

MATH 152, SECTION 55 (VIPUL NAIK)

There is a composition theorem for continuous functions: if g is continuous at c and f is continuous at $g(c)$, then $f \circ g$ is continuous at c.

We might suspect an analogous composition theorem for limits: if $\lim_{x \to c} g(x) = L$ and $\lim_{v \to L} f(v) = M$, then $\lim_{x \to c} f(g(x)) = M$.

However, the composition theorem for limits as stated above is not strictly true.

The reason is very subtle, though it is in some sense similar to why composites do not hold in the one-sided sense. It all has to do with the $0 <$ in the limit definition, or, the fact that when we say $\lim_{x \to c} g(x)$, we exclude the behavior at the point c.

Let's consider this more carefully. If $\lim_{x \to c} g(x) = L$, then that means that as $x \to c$, then $g(x) \to L$. Next, we know that as $v \to L$, then $f(v) \to M$. So, why doesn't it follow that as $x \to c$, $f(g(x)) \to M$?

In words:

$x \to c$ implies $g(x) \to L$

and:

$v \to L$ implies $f(v) \to M$

Plugging $v = g(x)$, why doesn't the result follow?

The issue is that the $g(x) \to L$ on the right half of the first line differs from the $v \to L$ on the left half of the second line. The former conclusion is valid, for instance, if $g(x)$ is a constant function L. On the other hand the $v \to L$ specifically includes only a straightforward approach from the left or the right.

Thus, we can construct a counterexample to the "composition theorem for limits" with the following features: (i) g is constant around c with the value L and (ii) f is defined at c but has a removable discontinuity at c.

INTRODUCTION TO DERIVATIVES

MATH 152, SECTION 55 (VIPUL NAIK)

Corresponding material in the book: Sections 3.1, 3.2, 3.3, 3.5.

Difficulty level: Moderate if you remember derivatives to at least the AP Calculus level. Otherwise, hard. Some things are likely to be new.

Covered in class?: Yes, but we'll go *very quickly* over most of the stuff that you would have seen at the AP level, and will focus much more on the conceptual interpretation and the algebraic-verbal-graphical-numerical nexus.

What students should definitely get: The definition of derivative as a limit. The fact that the left-hand derivative equals the left-hand limit and the right-hand derivative equals the right-hand limit. Also, the derivative is the slope of the tangent to the graph of the function. Graphical interpretation of tangents and normals. Finding equations of tangents and normals. Leibniz and prime notation. The sum rule, difference rule, product rule, and quotient rule. The chain rule for composition. Differentiation polynomials and rational functions.

What students should hopefully get: The idea of derivative as an *instantaneous rate of change*, the notion of *difference quotient* and *slope of chord*, and the fact that the definitions of tangent that we use for circles don't apply for general curves. How to find tangents and normals to curves from points outside them. Subtle points regarding tangent lines.

Executive summary

0.1. **Derivatives: basics.** Words ...

(1) For a function f, we define the *difference quotient* between w and x as the quotient $(f(w)-f(x))/(w-x)$. It is also the slope of the line joining $(x, f(x))$ and $(w, f(w))$. This line is called a *secant line*. The segment of the line between the points x and w is sometimes termed a *chord*.

(2) The limit of the difference quotient is defined as the *derivative*. This is the slope of the *tangent line* through that point. In other words, we define $f'(x) := \lim_{w \to x} \frac{f(w)-f(x)}{w-x}$. This can also be defined as $\lim_{h \to 0} \frac{f(x+h)-f(x)}{h}$.

(3) If the derivative of f at a point x exists, the function is termed *differentiable* at x.

(4) If the derivative at a point exists, then the tangent line to the graph of the function exists and its slope equals the derivative. The tangent line is horizontal if the derivative is zero. Note that if the derivative exists, then the tangent line cannot be vertical.

(5) Here are some misconceptions about tangent lines: (i) that the tangent line is the line perpendicular to the radius (this makes sense only for circles) (ii) that the tangent line does not intersect the curve at any other point (this is true for some curves but not for others) (iii) that any line other than the tangent line intersects the curve at at least one more point (this is always false – the vertical line through the point does not intersect the curve elsewhere, but is not the tangent line if the function is differentiable).

(6) In the Leibniz notation, if y is functionally dependent on x, then $\Delta y/\Delta x$ is the difference quotient – it is the quotient of the difference between the y-values corresponding to x-values. The limit of this, which is the derivative, is dy/dx.

(7) The left-hand derivative of f is defined as the left-hand limit for the derivative expression. It is $\lim_{h \to 0^-} \frac{f(x+h)-f(x)}{h}$. The right-hand derivative is $\lim_{h \to 0^+} \frac{f(x+h)-f(x)}{h}$.

(8) Higher derivatives are obtained by differentiating again and again. The *second* derivative is the derivative of the derivative. The n^{th} derivative is the function obtained by differentiating n times. In prime notation, the second derivative is denoted f'', the third derivative f''', and the n^{th} derivative

for large n as $f^{(n)}$. In the Leibniz notation, the n^{th} derivative of y with respect to x is denoted $d^n y/dx^n$.

(9) Derivative of sum equals sum of derivatives. Derivative of difference is difference of derivatives. Scalar multiples can be pulled out.
(10) We have the *product rule* for differentiating products: $(f \cdot g)' = f' \cdot g + f \cdot g'$.
(11) We have the *quotient rule* for differentiating quotients: $(f/g)' = (g \cdot f' - f \cdot g')/g^2$.
(12) The derivative of x^n with respect to x is nx^{n-1}.
(13) The derivative of sin is cos and the derivative of cos is $-\sin$.
(14) The chain rule says that $(f \circ g)' = ((f' \circ g) \cdot g'$

Actions ...

(1) We can differentiate any polynomial function of x, or a sum of powers (possibly negative powers or fractional powers), by differentiating each power with respect to x.
(2) We can differentiate any rational function using the quotient rule and our knowledge of how to differentiate polynomials.
(3) We can find the equation of the tangent line at a point by first finding the derivative, which is the slope, and then finding the point's coordinates (which requires evaluating the function) and then using the point-slope form.
(4) Suppose g and h are everywhere differentiable functions. Suppose f is a function that is g to the left of a point a and h to the right of the point a, and suppose $f(a) = g(a) = h(a)$. Then, the left-hand derivative of f at a is $g'(a)$ and the right-hand derivative of f at a is $h'(a)$.
(5) The k^{th} derivative of a polynomial of degree n is a polynomial of degree $n - k$, if $k \le n$, and is zero if $k > n$.
(6) We can often use the sum rule, product rule, etc. to find the values of derivatives of functions constructed from other functions simply using the values of the functions and their derivatives at specific points. For instance, $(f \cdot g)'$ at a specific point c can be determined by knowing $f(c)$, $g(c)$, $f'(c)$, and $g'(c)$.
(7) Given a function f with some unknown constants in it (so a function that is not completely known) we can use information about the value of the function and its derivatives at specific points to determine those constant parameters.

0.2. **Tangents and normals: geometry.** Words...

(1) The normal line to a curve at a point is the line perpendicular to the tangent line. Since the tangent line is the best linear approximation to the curve at the point, the normal line can be thought of as the line best approximating the perpendicular line to the curve.
(2) The angle of intersection between two curves at a point of intersection is defined as the angle between the tangent lines to the curves at that point. If the slopes of the tangent lines are m_1 and m_2, the angle is $\pi/2$ if $m_1 m_2 = -1$. Otherwise, it is the angle α such that $\tan\alpha = |m_1 - m_2|/(|1 + m_1 m_2|)$.
(3) If the angle between two curves at a point of intersection is $\pi/2$, they are termed *orthogonal* at that point. If the curves are orthogonal at all points of intersection, they are termed *orthogonal curves.*
(4) If the angle between two curves at a point of intersection is 0, that means they have the same tangent line. In this case, we say that the curves *touch* each other or are *tangent* to each other.

Actions...

(1) The equation of the normal line to the graph of a function f at the point $(x_0, f(x_0))$ is $f'(x_0)(y - f(x_0)) + (x - x_0) = 0$. The slope is $-1/f'(x_0)$.
(2) To find the angle(s) of intersection between two curves, we first find the point(s) of intersection, then compute the value of derivative (or slope of tangent line) to both curves, and then finally plug that in the formula for the angle of intersection.
(3) It is also possible to find all tangents to a given curve, or all normals to a given curve, that pass through a given point *not* on the curve. To do this, we set up the generic expression for a tangent line or normal line to the curve, and then plug into that generic expression the specific coordinates of the point, and solve. For instance, the generic equation for the tangent line to the graph of a

function f is $y - f(x_1) = f'(x_1)(x - x_1)$ where $(x_1, f(x_1))$ is the point of tangency. Plugging in the point (x, y) that we know the curve passes through, we can solve for x_1.

(4) In many cases, it is possible to determine geometrically the number of tangents/normals passing through a point outside the curve. Also, in some cases, the algebraic equations may not be directly solvable, but we may be able to determine the number and approximate location of the solutions.

0.3. **Deeper perspectives on derivatives.** Words... (these points were all seen in the quiz on Chapter 3)

(1) A continuous function that is everywhere differentiable need not be everywhere continuously differentiable.
(2) If f and g are functions that are both continuously differentiable (i.e., they are differentiable and their derivatives are continuous functions), then $f + g$, $f - g$, $f \cdot g$, and $f \circ g$ are all continuously differentiable.
(3) If f and g are functions that are both k times differentiable (i.e., the k^{th} derivatives of the functions f and g exist), then $f + g$, $f - g$, $f \cdot g$, and $f \circ g$ are also k times differentiable.
(4) If f and g are functions that are both k times continuously differentiable (i.e., the k^{th} derivatives of both functions exist and are continuous) then $f + g$, $f - g$, and $f \cdot g$, and $f \circ g$ are also k times continuously differentiable.
(5) If f is k times differentiable, for $k \geq 2$, then it is $k - 1$ times continuously differentiable, i.e., the $(k-1)^{th}$ derivative of f is a continuous function.
(6) If a function is *infinitely differentiable*, i.e., it has k^{th} derivatives for all k, then its k^{th} derivatives are continuous functions for all k.

1. Bare-bones introduction

1.1. **Down and dirty with derivatives.** The order in which I'll do things is to quickly introduce to you the mathematical formalism for derivatives, which is really quite simple, and give you a quick idea of the conceptual significance, and we'll return to the conceptual significance from time to time, as is necessary.

The derivative of a function at a point measures the *instantaneous rate of change* of the function at the point. In other words, it measures how quickly the function is changing. Or, it measures the velocity of the function. These are vague ideas, so let's take a closer look.

Let's say I am riding a bicycle on a road, and I encounter a sign saying "speed limit: 30 mph". Well, that sign basically says that the speed limit should be 30 mph, but what speed is it refering to? We know that speed is distance covered divided by time taken, but if I start out from home in the morning, get stuck in a traffic jam on my way, and then, once the road clears, ride at top speed to get to my destination, I might have take five hours to just travel five miles. So my speed is 1 mile per hour. But may be I *still* went over the speed limit, because may be it was the case that at the point where I crossed the speed limit sign, I was going really really fast. So, what's really relevant isn't my *average* speed from the moment I began my ride to the moment I ended, but the speed at the particular *instant* that I crossed that sign. But how do we measure the speed at a particular instant?

Well, one thing I could do is may be measure the time it takes me to get from the lamp post just before that sign to the lamp post just after that sign. So if that distance is a and the time it took me to travel that distance is b, then the speed is a/b. And that might be a more relevant indicator since it is the speed in a small interval of time around where I saw that sign. But it still does not tell the entire story, because it might have happened that I was going pretty fast between the lamp posts, but I slowed down for a fraction of a second while crossing that sign. So may be I still didn't technically break the law because I was slow enough *at the instant* that I was crossing the sign.

What we're really trying to measure here is the change in my position, or the distance I travel, divided by the time, but we're trying to measure it for a smaller and smaller interval of time around that crucial time point when I cross the signpost. Smaller and smaller and smaller and smaller... this suggests that a limit lurks, and indeed it does.

Let's try to formalize this. Suppose the function giving my position in terms of time is f. So, for a time t, my position is given by $f(t)$. And let's say that the point in time when I crossed the sign post was the point c. That means that at time $t = c$, I crossed the signpost, which is located at the point $f(c)$.

Now, let's say I want to calculate my speed immediately after I cross the sign post. So I pick some t that's just a little larger than c, and I look at the speed from time c to time t. What I get is:

$$\frac{f(t) - f(c)}{t - c}$$

This gives the *average* speed from time c to time t. How do we narrow this down to $t = c$? If we try to plug in $t = c$ in the formula, we get something of the form $0/0$, which is not defined. So, what do we do?

This is where we need the crucial idea – we need the idea of a *limit*. So, the instantaneous speed just *after* crossing the signpost is defined as:

$$\lim_{t \to c^+} \frac{f(t) - f(c)}{t - c}$$

This is the limit of the *average speed* over the time interval $[c, t]$, as t approaches c from the right side.

Similarly, the instantaneous speed just before crossing the signpost is:

$$\lim_{t \to c^-} \frac{f(c) - f(t)}{c - t} = \lim_{t \to c^-} \frac{f(t) - f(c)}{t - c}$$

Notice that both these are taking limits of the same expression, except that one is from the right and one is from the left. If both limits exist and are equal, then this is the instantaneous speed. And once we've calculated the instantaneous speed, we can figure out whether it is greater than what the signpost capped speed at.

1.2. **Understanding derivatives – formal definition.** The derivative of a function f at a point x is defined as the following limit, if it exists:

$$f'(x) := \lim_{w \to x} \frac{f(w) - f(x)}{w - x}$$

if this limit exists! If the limit exists, then we say that the function f is *differentiable* at the point x and the value of the derivative is $f'(x)$.

The *left-hand derivative* is defined as the *left-hand limit* for the above expression:

$$\lim_{w \to x^-} \frac{f(w) - f(x)}{w - x}$$

The *right-hand derivative* is defined as the *right-hand limit* for the above expression:

$$\lim_{w \to x^+} \frac{f(w) - f(x)}{w - x}$$

The expression whose limit we are trying to take in each of these cases is sometimes called a *difference quotient* – it is the quotient between the difference in the function values and the difference in the input values to the function.

So, in words:

The derivative of f at x = The limit, as w approaches x, of the difference quotient of f between w and x

Now, what I've told you so far is about all that you need to know – the basics of the definition of derivative. But the typical way of thinking about derivatives uses a slightly different formulation of the definition, so I'll just talk about that. That's related to the idea of substituting variables.

1.3. **More conventional way of writing derivatives.** Recall that $\lim_{w \to x} g(w) = \lim_{h \to 0} g(x + h)$. This is one of those rules for substitution we use to evaluate some trigonometric limits. And we typically use this rule when writing down the definition of derivatives. So, with the $h \to 0$ convention, we define:

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

What this is saying is that what we're trying to measure is the quotient of the difference in the values of f for a small change in the value of the input for f, near x. And then we're taking the limit as the increment gets smaller and smaller. So, with this notation, the left-hand derivative becomes the above limit with h approaching 0 from the negative side and the right-hand derivative becomes the above limit with h approaching 0 from the positive side. In symbols:

$$\text{LHD of } f \text{ at } x = \lim_{h \to 0^-} \frac{f(x+h) - f(x)}{h}$$

and:

$$\text{RHD of } f \text{ at } x = \lim_{h \to 0^+} \frac{f(x+h) - f(x)}{h}$$

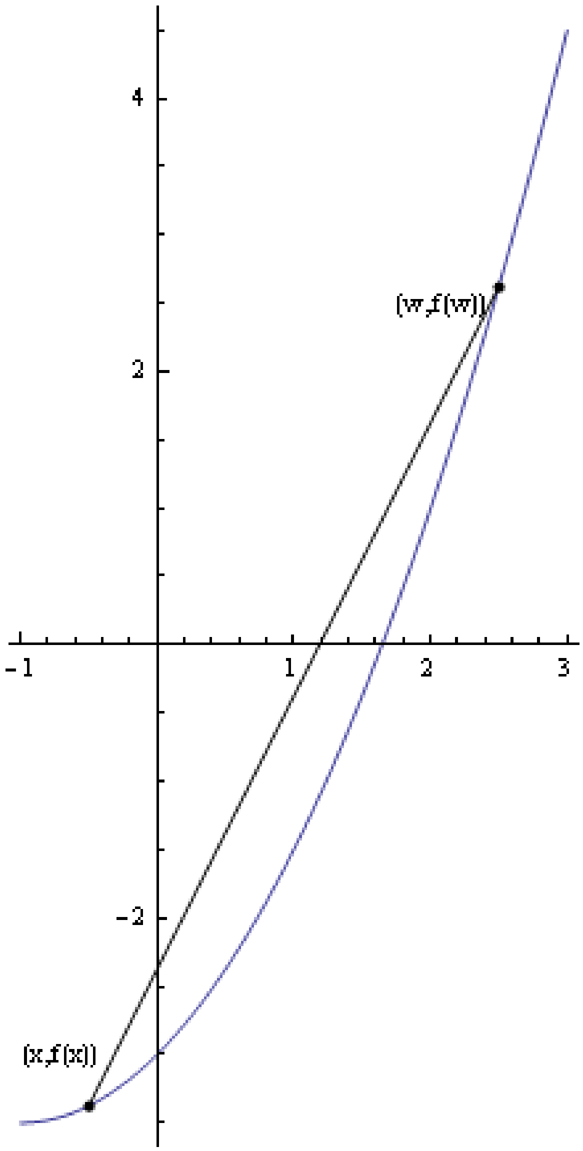
1.4. **Preview of graphical interpretation of derivatives.** Consider the graph of the function f, so we have a graph of $y = f(x)$. Then, say we're looking at two points x and w, and we're interested in the difference quotient:

$$\frac{f(w) - f(x)}{w - x}$$

What is this measuring? Well, the point $(x, f(x))$ represents the point on the graph with its first coordinate x, and the point $(w, f(w))$ represents the point on the graph with its first coordinate w. So, what we're doing is taking these two points, and taking the quotient of the difference in their y-coordinates by the difference in their x-coordinates. And, now this is the time to revive some of your coordinate geometry intuition, because this is essentially the *slope* of the line joining the points $(x, f(x))$ and $(w, f(w))$. Basically, it is the *rise* over the *run* – the vertical change divided by the horizontal change.

So, one way of remembering this is that:

$$\text{Difference quotient of } f \text{ between points } x \text{ and } w = \text{Slope of line joining } (x, f(x)) \text{ and } (w, f(w))$$

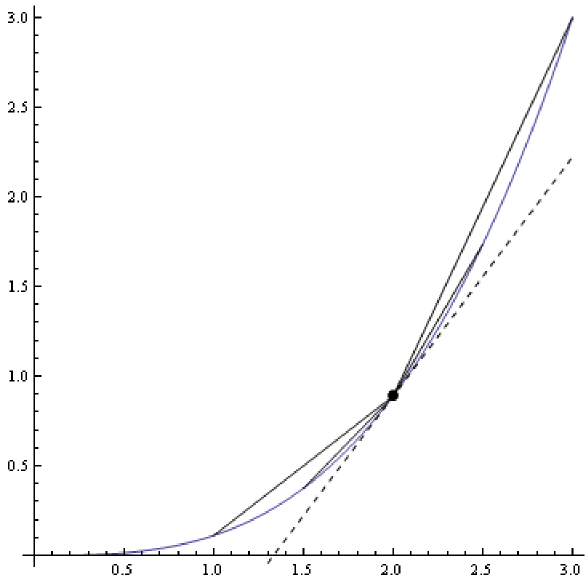


Also, note that the difference quotient is symmetric in x and w, in the sense that if we interchange the roles of the two points, the difference quotient is unaffected. So is the slope of the line joining the two points.

Now, let's try to use this understanding of the difference quotient as the slope of the line joining two points to understand what the derivative means geometrically. So, here's some terminology – this line joining two points is termed a *secant line*. The *line segment* between the two points is sometimes called a *chord* – just like you might have seen chords and secants for circles, we're now using the terminology for more general kinds of curves.

So, what we're trying to do is understand: as the point w gets closer and close to x, what happens to the slope of the secant line? Well, the pictures make it clear that what's happening is that the secant line is coming closer and closer to a line that *just touches* the function at the point $(x, f(x))$ – what is called the *tangent line*. So, we get:

$$\text{Derivative of } f \text{ at } x = \text{ Slope of tangent line to graph of } f \text{ at } (x, f(x))$$



1.5. **More on the tangent line.** I'll have a lot to say on the tangent line and derivatives – our treatment of these issues has barely scratched the surface so far. But before proceeding too much, here's just one very important point I want to make.

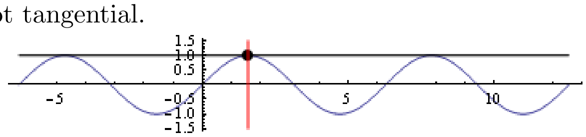
Some of you have done geometry and you've seen the concept of tangent lines to circles. You'll recall that the tangent line at a point on a circle can be defined in a number of ways: (i) as the line perpendicular to the line joining the point to the center of the circle (the latter is also called the radial line), and (ii) as the unique line that intersects the circle at exactly that one point.

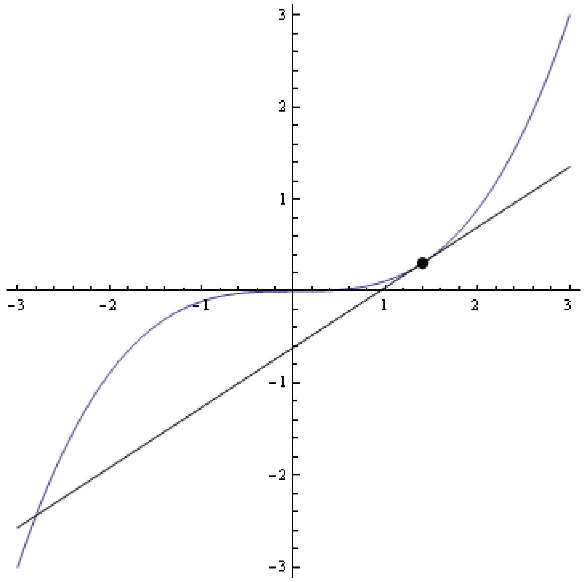
Now, it's obvious that you need to discard definition (i) when thinking about tangent lines to general curves, such as graphs of functions. That's because a curve in general has no center. It doesn't even have an inside or an outside. So definition (i) doesn't make sense.

But definition (ii) makes sense, right? Well, it makes sense, but it is wrong. For general curves, it is wrong. It's wrong because, first, the tangent line may well intersect the curve at other points, and second, there may be many non-tangent lines that also intersect the curve at just that one point.

For instance, consider the function $f(x) := \sin x$. We know the graph of this function. Now, our intuition of the tangent line should say that, at the point $x = \pi/2$, the tangent line should be the limit of chords between $\pi/2$ and very close points, and those chords are becoming almost horizontal, so the tangent line should be horizontal. But – wow! The tangent line intersects the graph of the function at *infinitely* many other points. So banish from your mind the idea that the tangent line doesn't intersect the curve anywhere else.

And, there are other lines that are very much *not* tangent lines that intersect the curve at exactly that one point. For instance, you can make the vertical line through the point $(\pi/2, 1)$, so that's the line $x = \pi/2$, and that intersects the curve at exactly one point – but it's far from the tangent line! And you can make a lot of other lines – ones that are close to vertical, that intersect the curve at exactly that one point, but are not tangential.





So, thinking of the tangent line as the line that intersects the curve at exactly that one point is a flawed way of thinking about tangent lines. So what's the right way? Well, one way is to think of it as a limit of secant lines, which is what we discussed. But if you want a way of thinking that isn't really limits-based, think of it as the line *best line approximating the curve near the point*. It's the answer to the question: if we wanted to treat this function as a linear function near the point, what would that linear function look like? Or, you can go back to the Latin roots and note that tangent comes from *touch*, so the feeling of just touching, or just grazing, is the feeling you should have.

1.6. **Actual computations of derivatives.** Suppose we have $f(x) := x^2$. We want to calculate $f'(2)$. How do we do this?

Let's use the first definition, in terms of w, which gives:

$$f'(2) = \lim_{w \to 2} \frac{w^2 - 4}{w - 2} = \lim_{w \to 2} \frac{(w+2)(w-2)}{w-2} = \lim_{w \to 2} (w+2) = 4$$

So, there, we have it, $f'(2) = 4$.

Now this is the time to pause and think one level more abstractly. We have a recipe that allows us to calculate $f'(x)$ for any specific value of x. But it seems painful to have to do the limit calculation for each x. So if I now want to find $f'(1)$, I have to do that calculation again replacing the 2 by 1. So, in order to save time, let's try to find $f'(x)$ for a *general* value of x.

$$f'(x) = \lim_{w \to x} \frac{w^2 - x^2}{w - x} = \lim_{w \to x} \frac{(w+x)(w-x)}{w-x} = \lim_{w \to x} (w+x) = 2x$$

Just to illustrate, let me use the $h \to 0$ formulation:

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

So good, both of these give the same answer, so we now have a *general* formula that say that for $f(x) = x^2$, $f'(x) = 2x$.

Note that this says that the function f is differentiable *everywhere*, which is a pretty strong statement, and the derivative is *itself* a continuous function, namely the linear function $2x$.

Let's see what this means graphically. The graph of $f(x) := x^2$ looks like a parabola. It passes through $(0, 0)$. What the formula tells us is that $f'(0) = 0$. Which means that the tangent at 0 is horizontal, which it surely is. The formula tells us that $f'(1) = 2$, so the tangent line at 1 has its y-coordinate rising twice as fast as its x-coordinate, which is again sort of suggested by the graph. $f'(2) = 4$, $f'(3) = 6$, and so on. Which means that the slope of the tangent line increases as x gets larger – again, suggested by the graph

– it's getting steeper and steeper. And for $x < 0$, $f'(x)$ is negative, which is again pictorially clear because the slope of the tangent line is negative.

Let's do one more thing: write down the equation of the tangent line at $x = 3$. Remember that if a line has slope m and passes through the point (x_0, y_0), then the equation of the line is given by:

$$y - y_0 = m(x - x_0)$$

This is called the *point-slope form*.

So what do we know about the point $x = 3$? For this point, the value of the function is 9, so one point that we for sure know is on the tangent line is the point $(3, 9)$. And the tangent line has slope $2(3) = 6$. So, plugging into the point-slope form, we get:

$$y - 9 = 6(x - 3)$$

And that's it – that's the equation of the line. You can rearrange stuff and rewrite this to get:

$$6x - y = 9$$

1.7. **Differentiable implies continuous.** Important fact: if f is differentiable at the point x, then f is continuous at the point x. Why? Below is a short proof.

Consider, for $h \neq 0$:

$$f(x + h) - f(x) = \frac{f(x + h) - f(x)}{h} \cdot h$$

Now, taking the limit as $h \to 0$, and using the fact that f is differentiable at x, we get:

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

The first limit on the right is $f'(x)$, and the second limit is 0, so the product is 0, and we get:

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

Since $f(x)$ is a constant (independent of h) it can be pulled out of the limit, and rearranging, we get:

$$\lim_{h \to 0} f(x + h) = f(x)$$

Which is precisely what it means to say that f is continuous at x.

What does this mean intuitively? Well, intuitively, for f to be differentiable at x means that the rate of change at x is finite. But if the function is jumping, or suddenly changing value, or behaving in any of those odd ways that we saw could occur at discontinuities, then the instantaneous rate of change wouldn't make sense. It's like if you apparate/teleport, you cannot really measure the *speed* at which you traveled. So, we can hope for a function to be differentiable at a point only if it's continuous at the point.

Aside: Differentiability and the $\epsilon - \delta$ definition. (Note: This is potentially confusing, so ignore it if it confuses you). The derivative of f at $x = p$ measures the rate of change of f at c. Suppose the derivative value is $\nu = f'(p)$. Then, this means that, close to c, the graph looks like a straight line with slope ν and passing through $(p, f(p))$.

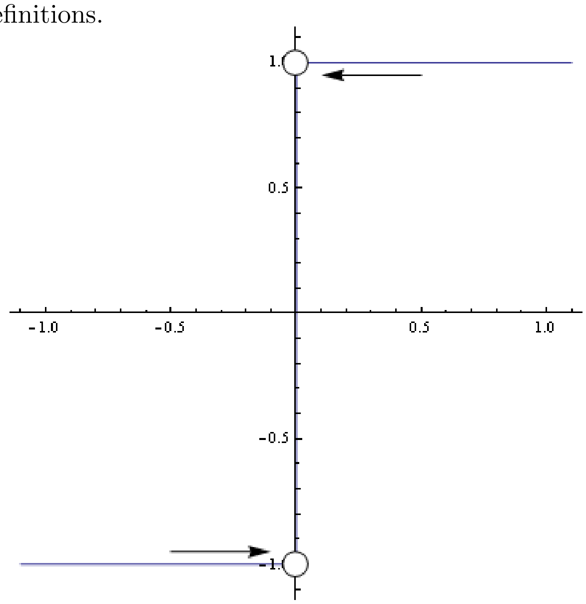
If it were exactly the straight line, then the strategy for an $\epsilon - \delta$ proof would be to pick $\delta = \epsilon/|\nu|$. In practice, since the graph is not exactly a straight line, we need to pick a slightly different (usually smaller) δ to work. Recall that for f a quadratic function $ax^2 + bx + c$, we chose $\delta = \min\{1, \epsilon/(|a| + |2ap + b|)\}$. In this case, $\nu = 2ap + b$. However, to make the proof go through, we need to pad an extra $|a|$ in the denominator, and that extra padding is because the function isn't quite linear.

1.8. **Derivatives of linear and constant functions.** You can just check this – we'll deal more with this in the next section:

(1) The derivative of a constant function is zero everywhere.
(2) The derivative of a linear function $f(x) := ax + b$ is everywhere equal to a.

1.9. **Handling piecewise definitions.** Let's think about functions defined piecewise. In other words, let us think about functions that change definition, i.e., that are governed by different definitions on different parts of the domain. How do we calculate the derivative of such a function?

The first rule to remember is that the derivative of a function at a point depends only on what happens at the point and very close to the point. So, if the point is contained in an open interval where one definition rules, then we can just differentiate the function using that definition. On the other hand, if the point is surrounded, very closely, by points where the function takes different definitions, we need to handle all these definitions.

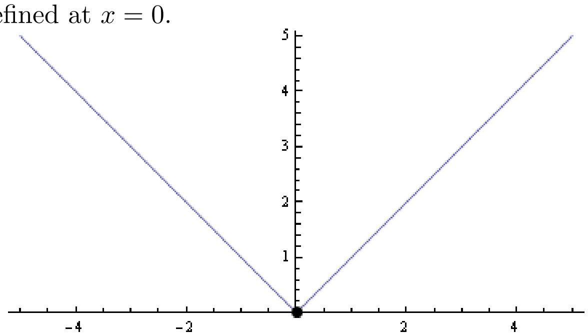


Let's begin with the signum function. Define $f(x) = x/|x|$ for $x \neq 0$, and define $f(0) = 0$. So f is constant at -1 for $x < 0$, $f(0) = 0$, and f is constant at 1 for $x > 0$.

So,what can we say about the derivative of f? Well, think about what's happening for $x = -3$, say. Around $x = -3$, the function is constant at -1. Yes, it's going to change its value *far in the future*, but that doesn't affect the derivative. The derivative is a *local measurement*. So we simply need to calculate the derivative of the function that's constant-valued to -1, for $x = -3$. You can apply the definition and check that the derivative is 0, which makes sense, because constant functions are unchanging.

Similarly, at the point $x = 4$, the function is constant at 1 around the point, so the derivative is against 0.

What about the point $x = 0$? Here, you can see that the function isn't continuous, because it's jumping, so the derivative is not defined. So the derivative of f is the function that is 0 for all $x \neq 0$, and it isn't defined at $x = 0$.



Let's look at another function, this time $g(x) := |x|$. For $x < 0$, this function is $g(x) = -x$, and for $x > 0$, this function is $g(x) = x$. The crucial thing to note here is that g is continuous at 0, so *both definitions* – the definition $g(x) = x$ and the definition $g(x) = -x$ apply at $x = 0$.

For $x < 0$, for instance, $x = -3$, we can treat the function as $g(x) = -x$, and differentiate this function, What we'll get is that $g'(x) = -1$ for $x < 0$. And for $x > 0$, we can use $g(x) = x$, so the derivative, which we can again calculate, is $g'(x) = 1$ for all $x > 0$.

But what about $x = 0$? The function is continuous at $x = 0$, so we cannot do the cop-out that we did last time. So we need to think in terms of left and right: calculate the left-hand derivative and the right-hand derivative.

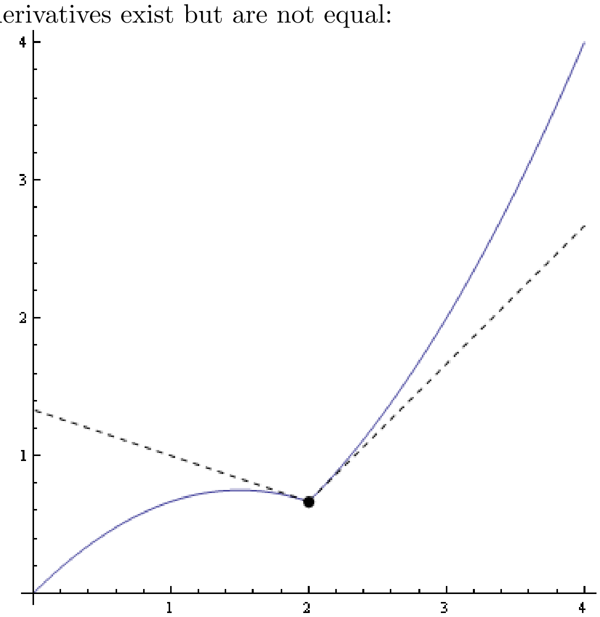
Now, the important idea here is that since the definition $g(x) = -x$ *also applies* at $x = 0$, the formula for g' that we calculated for $-x$ also applies to the *left-hand derivative* at 0. So the left-hand derivative at $x = 0$ is -1, which again you can see from the graph. And the right-hand derivative at $x = 0$ is $+1$, because the $g(x) = x$ definition also applies at $x = 0$.

So, the function g' is -1 for x negative, and $+1$ for x positive, but it isn't defined for $x = 0$. However, the left-hand derivative at 0 is defined – it's -1. And so is the right-hand derivative – it's $+1$.

So, the upshot is: Suppose a function f has a definition f_1 to the left of c and a definition f_2 to the right of c, and both definitions give everywhere differentiable functions on all real numbers:

(1) f is differentiable at all points other than c.
(2) If $f_1(c) = f_2(c)$ and these agree with $f(c)$, then f is continuous at c.
(3) If f is continuous from the left at c, the left-hand derivative at c equals the value of the derivative f_1' evaluated at c.
(4) If f is continuous from the right at c, the right-hand derivative at c equals the value of the derivative f_2' evaluated at c.
(5) In particular, if f is continuous at c, *and* $f_1'(c) = f_2'(c)$, then f is differentiable at c.

Here is an example of a picture where the function is continuous but changes direction, so the one-sided derivatives exist but are not equal:



1.10. **Three additional examples.**

$$f(x) := \begin{cases} x^2, & x \le 0 \\ x^3, & x > 0 \end{cases}$$

Here $f_1(x) = x^2$ and $f_2(x) = x^3$. Both piece functions are differentiable everywhere. Note that f is continuous at 0, since $f_1(0) = f_2(0) = 0$. The left hand derivative at 0 is $f_1'(x) = 2x$ evaluated at 0, giving 0. The right hand derivative at 0 is $f_2'(x) = 3x^2$ evaluated at 0, giving 0. Since both one-sided derivatives agree, f is differentiable at 0 and $f'(0) = 0$.

Now consider the example:

$$f(x) := \{ \begin{array}{rl} x^2+1, & x \leq 0 \\ x^3, & x > 0 \end{array}$$

This function is not continuous at 0 because the two functions x^2+1 and x^3 do not agree at 0. Note that since 0 is included in the left definition, f is left continuous at 0, and $f(0) = 1$. We can also calculate the left hand derivative: it is $2x$ evaluated at 0, which is $2 \cdot 0 = 0$.

However, since the function is not continuous from the right at 0, we *cannot* calculate the right hand derivative by plugging in the derivative of x^3 at 0. In fact, we should not feel the need to do so, because *the function is not right continuous at* 0*, so there cannot be a right hand derivative.*

Finally, consider the example:

$$f(x) := \{ \begin{array}{rl} x, & x \leq 0 \\ x^3, & x > 0 \end{array}$$

This function is continuous at 0, and $f(0) = 0$. The left hand derivative at 0 is 1 and the right hand derivative is $3(0)^2 = 0$. Thus, the function is *not differentiable* at 0, though it has one-sided derivatives.

1.11. **Existence of tangent and existence of derivative.** Our discussion basically pointed to the fact that, for a function f defined around a point x, if the derivative of f at x exists, then there exists a tangent to the graph of f at the point $(x, f(x))$ and the slope of that tangent line is $f'(x)$. There are a few subtleties related to this:

(1) If both the left-hand derivative and the right-hand derivative at a point are defined, but are not equal, then the left-hand derivative is the tangent to the graph on the *left* side and the right-hand derivative is the tangent to the graph on the *right* side. There is no single tangent to the whole graph at the point.
(2) In some cases, the tangent to the curve exists and is vertical. If the tangent exists and is vertical, then the derivative does not exist. In fact, a vertical tangent is the *only* situation where the tangent exists but the derivative does not exist.

An example of this is the function $f(x) = x^{1/3}$ at the point $x = 0$. See Page 110 of the book for a more detailed discussion of this. We will get back to this function (and the general notion of vertical tangent) in far more gory detail in the near future.

2. Rules for computing derivatives

2.1. **Quick recap.** Before proceeding further, let's recall the definition of the derivative.

If f is a function defined around a point x, we define:

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

The left-hand derivative is defined as:

$$f_l'(x) := \lim_{h \to 0^-} \frac{f(x+h) - f(x)}{h} = \lim_{w \to x^-} \frac{f(w) - f(x)}{w - x}$$

The right-hand derivative is defined as:

$$f_r'(x) := \lim_{h \to 0^+} \frac{f(x+h) - f(x)}{h} = \lim_{w \to x^+} \frac{f(w) - f(x)}{w - x}$$

(The subscripts l and r are not standard notation, but are used for simplicity here).

2.2. **Differentiating constant and linear functions.** Suppose a function f is constant with constant value k. We want to show that the derivative of f is zero everywhere. Let's calculate the derivative of f.

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

So the derivative of a constant function is zero. And this makes sense because the derivative is the *rate of change* of the function. And a constant function is unchanging, so its derivative is zero.

Let's look at the derivative of a linear function $f(x) := ax + b$. Let's calculate the derivative:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0} \frac{a(x+h) + b - (ax+b)}{h} = \lim_{h \to 0} \frac{ah}{h} = a$$

So, for the function $f(x) = ax + b$, the derivative is a *constant* function whose value is equal to a. And this makes sense, because the graph of $f(x) = ax + b$ is a straight line with slope a. And now remember that the derivative of a function equals the slope of the tangent line. But now, what's the tangent line to a straight line at a point? It is the line itself. So, the tangent line is the line itself, and its slope is a, which is what our calculations also show.

Remember that the tangent line is the *best linear approximation* to the curve, so if the curve is already a straight line, it coincides with the tangent line to it through any point.

2.3. **The sum rule for derivatives.** Suppose f and g are functions, both defined around a point x. The sum rule for derivatives states that if both f and g are differentiable at the point x, then $f + g$ is also differentiable at the point x, and $(f+g)'(x) = f'(x) + g'(x)$. In words, *the sum of the derivatives equals the derivative of the sum.* This is the first part of Theorem 3.2.3. The proof is given below. It involves very simple manipulation. (You're not expected to know this proof, but this should be sort of manipulation that you eventually are comfortable reading and understanding).

So, what we have is:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{1}$$

and:

$$g'(x) = \lim_{h \to 0} \frac{g(x+h) - g(x)}{h} \tag{2}$$

Now, we also have:

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

Simplifying further, we get:

$$(f+g)'(x) = \lim_{h \to 0} \frac{(f(x+h) - f(x)) + (g(x+h) - g(x))}{h} = \lim_{h \to 0} \left[\left(\frac{f(x+h) - f(x)}{h} \right) + \left(\frac{g(x+h) - g(x)}{h} \right) \right]$$

Now, using the fact that the limit of the sum equals the sum of the limits, we can split the limit on the right to get:

$$(f+g)'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} + \lim_{h \to 0} \frac{g(x+h) - g(x)}{h} \tag{3}$$

Combining (1), (2), and (3), we obtain that $(f+g)'(x) = f'(x) + g'(x)$.

On a similar note, if α is a constant, then the derivative of αf is α times the derivative of f. In other words:

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

2.4. **Global application of these rules.** The rule for sums and scalar multiples apply at each point. In other words, if both f and g are differentiable *at a point* x, then $f+g$ is differentiable at x and the derivative of $f+g$ is the sum of the derivatives of f and g at the point x.

An easy corollary of this is that if f and g are everywhere differentiable functions, then $f+g$ is also everywhere differentiable, and $(f+g)' = f' + g'$. In other words, what I've just done is applied the previous result *at every point*. And, so, the function $(f+g)'$ is the sum of the functions f' and g'.

In words, *the derivative of the sum is the sum of the derivatives.*

2.5. **Rule for difference.** In a similar way that we handled sums and scalar multiples, we can handle differences, so we get $(f-g)' = f' - g'$.

In words, *the derivative of the difference is the difference of the derivatives.*

2.6. **Rule for product.** People less seasoned in calculus than you may now expect the rule for products to be: *the derivative of the product is the product of the derivatives.* But this is *wrong*.

The rule for the product is that, if f and g are differentiable at a point x, then the product $f \cdot g$ is also differentiable at the point x, and:

$$(f \cdot g)'(x) = f(x) \cdot g'(x) + f'(x) \cdot g(x)$$

We won't bother with a proof of the product rule, but it's an important result that you should know. In particular, again, when f and g are defined everywhere, we have:

$$(f \cdot g)' = f' \cdot g + f \cdot g'$$

2.7. **Rule for reciprocals and quotients.** If f is differentiable at a point x, and $f(x) \neq 0$, then the function $(1/f)$ is also differentiable at the point x:

$$\left(\frac{1}{f}\right)'(x) = \frac{-f'(x)}{(f(x))^2}$$

If, at a point x, f and g are both differentiable, and $g(x) \neq 0$, then:

$$\left(\frac{f}{g}\right)'(x) = \frac{g(x)f'(x) - f(x)g'(x)}{(g(x))^2}$$

2.8. **Formula for differentiating a power.** Here's the formula for differentiating a power function: if $f(x) = x^n$, then $f'(x) = nx^{n-1}$. You can actually prove this formula for positive integer values of n using induction and the product rule (I've got a proof in the addendum at the end of this section). So, for instance, the derivative of x^2 is $2x$. The derivative of x^5 is $5x^4$. Basically, what you do is to take the thing in the exponent and pull it down as a coefficient and subtract 1 from it.

Okay, now there are plenty of things I want you to note here. First, note that the n in the exponent really should be a *constant*. So, for instance, we cannot use this rule to differentiate the function x^x, because in this case, the thing in the exponent is itself dependent on x.

The second thing you should note is that you are already familiar with some cases of this formula. For instance, consider the constant function $x^0 = 1$. The derivative of this is the function $0x^{0-1} = 0x^{-1} = 0$. Also, for the function $x^1 = x$, the derivative is 1, which is again something we saw. And for the function x^2, the derivative is $2x^1 = 2x$, which is what we saw last time.

The third thing you should think about is the scope and applicability of this formula. I just said that when n is a positive integer, we can show that for $f(x) = x^n$, we have $f'(x) = nx^{n-1}$. But in fact, this formula works not just for positive integers, but for all integers, and not just for all integers, even for non-integer exponents:

(1) If $f(x) = x^n, x \neq 0$, where n is a negative integer, $f'(x) = nx^{n-1}$.
(2) If $f(x) = x^r$, where r is some real number, and $x > 0$, we have $f'(x) = rx^{r-1}$. We will study later what x^r means for arbitrary real r and positive real x.
(3) The above formula also applies for $x < 0$ where r is a rational number with denominator an odd integer (that's necessary to make sense of x^r).

2.9. **Computing the derivative of a polynomial.** To differentiate a polynomial, what we do is use the rule for differentiating powers, and the rules for sums and scalar multiples.

For instance, to differentiate the polynomial $f(x) := x^3 + 2x^2 + 5x + 7$, what we do is differentiate each of the pieces. The derivative of the piece x^3 is $3x^2$. The derivative of the piece $2x^2$ is $4x$, the derivative of the piece $5x$ is 5, and the derivative of the piece 7 is 0. So, the derivative is $f'(x) = 3x^2 + 4x + 5$.

Now, the way you do this is you write the polynomial, and with a bit of practice, which you did in high school, you can differentiate each individual piece mentally. And so you keep writing the derivatives one by one.

In particular, polynomial functions are everywhere differentiable, and the derivative of a polynomial function is another polynomial. Moreover, the degree of the derivative is one less than the degree of the original polynomial.

2.10. **Computing the derivative of a rational function.** What if we have a rational function $Q(x) = f(x)/g(x)$?

Essentially, we use the formula for differentiating a quotient, which gives us:

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

Since both f and g are polynomials, we know how to differentiate them, so we are effectively done.

For instance, consider $Q(x) = x^2/(x^3 - 1)$. In this case:

$$Q'(x) = \frac{(x^3 - 1)(2x) - x^2(3x^2)}{(x^3 - 1)^2} = \frac{-2x - x^4}{(x^3 - 1)^2}$$

Note another important thing – since the denominator of Q' is the square of the denominator of Q, the points where Q' is not defined are the *same* as the point where Q is not defined. Thus, a rational function is differentiable wherever it is defined.

Addendum: Proof by induction that the derivative of x^n is nx^{n-1}. We prove this for $n \geq 1$.

Base case for induction: Here, $f(x) = x^1 = x$. We have:

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

Thus, for $n = 1$, we have $f'(x) = nx^{n-1}$ where $f(x) = x^n$.

Induction step: Suppose the result is true for $n = k$. In other words, suppose it is true that for $f(x) := x^k$, we have $f'(x) = kx^{k-1}$. We want to show that for $g(x) := x^{k+1}$, we have $g'(x) = (k + 1)x^k$.

Using the product rule for $g(x) = xf(x)$, we have:

$$g'(x) = 1f(x) + xf'(x)$$

Substitutng the expression for $f'(x)$ from the assumption that the statement is true for k, we get:

$$g'(x) = 1(x^k) + x(kx^{k-1}) = x^k = kx^k = (k + 1)x^k$$

which is what we need to prove. This completes the induction step, and hence completes the proof.

3. Graphical interpretation of derivatives

3.1. **Some review of coordinate geometry.** So far, our attention to coordinate geometry has been minimal: we discussed how to get the *equation* of the tangent line to the graph of a function at a point using the *point-slope form*, wherein we determine the point possibly through function evaluation and we determine the slope by computing the derivative of the function at the point.

Now, recall that the *slope* of a line equals $\tan \theta$, where θ is the angle that the line makes with the x-axis, measured counter-clockwise from the x-axis. What are the kind of lines whose slope is not defined? These are the vertical lines, in which case $\theta = \pi/2$. Notice that since the slope is not defined, we see that if a function is differentiable at the point, then the tangent line cannot be vertical.

The slope of a line equals $\tan\theta$, where θ is the angle that that line makes with the x-axis. And for some of you, thinking of that angle might be geometrically more useful than thinking of the slope as a number. We will study this in more detail a little later when we study the graphing of functions more intensely.

Caveat: Axes need to be scaled the same way for geometry to work right. When you use a graphing software such as Mathematica, or a graphing calculator for functions, you'll often find that the software or calculator automatically chooses different scalings for the two axes, so as to make the picture of the graph fit in nicely. If the axes are scaled differently, the geometry described here does not work. In particular, the derivative is no longer equal to $\tan\theta$ where θ is the angle with the horizontal; instead, we have to scale the derivative by the appropriate ratio between the scaling of the axes. Similarly, the angles of intersection between curves change and the notion of orthogonality is messed up.

However, the notion of tangency of curves discussed below does *not* get messed up.

3.2. **Perpendicular lines, angles between lines.** Two lines are said to be *orthogonal* or *perpendicular* if the angle between them is $\pi/2$. This means that if one of them makes an angle θ_1 with the x-axis and the other makes an angle θ_2 with the x-axis (both angles measured counter-clockwise from the x-axis) then $|\theta_1 - \theta_2| = \pi/2$. How do we characterize this in terms of slopes?

Well, we need to take a short detour here and calculate the formula for $\tan(A-B)$. Recall that $\sin(A-B) = \sin A\cos B - \cos A\sin B$ and $\cos(A-B) = \cos A\cos B + \sin A\sin B$. When we take the quotient of these, we get:

$$\tan(A-B) = \frac{\sin A\cos B - \cos A\sin B}{\cos A\cos B + \sin A\sin B}$$

If both $\tan A$ and $\tan B$ are defined, this simplifies to:

$$\tan(A-B) = \frac{\tan A - \tan B}{1 + \tan A\tan B}$$

Now, what would happen if $|A-B|$ is $\pi/2$? In this case, $\tan(A-B)$ should be undefined, which means that the denominator on the right side should be 0, which means that $\tan A\tan B = -1$.

Translating back to the language of slopes, we obtain that two lines (neither of which is vertical or horizontal) are perpendicular if the product of their slopes is -1.

More generally, given two lines with slopes m_1 and m_2, we have the following formula for $\tan\alpha$ where α is the angle of intersection:

$$\tan\alpha = \left|\frac{m_1 - m_2}{1 + m_1 m_2}\right|$$

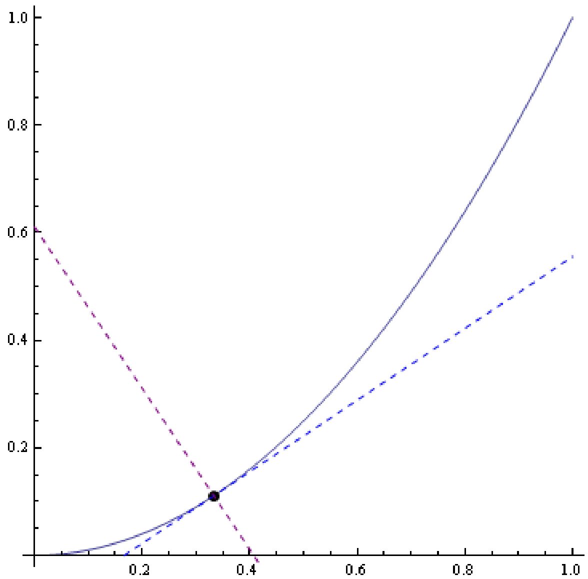
The absolute value means that we are not bothered about the direction (or *orientation* or *sense*) in which we are measuring the angle (if we care about the direction, we have to specify angle from which line to which line, measured clockwise or counter-clockwise).

3.3. **Normal lines.** Given a point on a curve, the *normal line* to the curve at that point is the line perpendicular to the tangent line to the curve at that point. This normal is the same normal you may have seen in *normal force* and other related ideas in classical mechanics.

Now, we know how to calculate the equation of the normal line, if we know the value of the derivative. let me spell this out with two cases:

(1) If the derivative of a function at $x = a$ is 0, then the tangent line is $y = f(a)$ and the normal line is $x = a$.
(2) If the derivative of a function at $x = a$ is $m \neq 0$, then the tangent line is $y - f(a) = m(x-a)$ and the normal line is $y - f(a) = (-1/m)(x-a)$. We got the slope of the normal line using the fact that the product of slopes of two perpendicular lines is -1.

Here is a picture of the usual case where the derivative is nonzero:



3.4. **Angle of intersection of curves.** Consider the graphs of the functions $f(x) = x$ and $g(x) = x^2$. These curves intersect at two points: $(0,0)$ and $(1,1)$. We would like to find the angles of intersection between the curves at these two points.

The *angle of intersection* between two curves at a given point is the angle between the tangent lines to the curve at that point. In particular, if α is this angle of intersection, then for $\alpha \neq \pi/2$, we have:

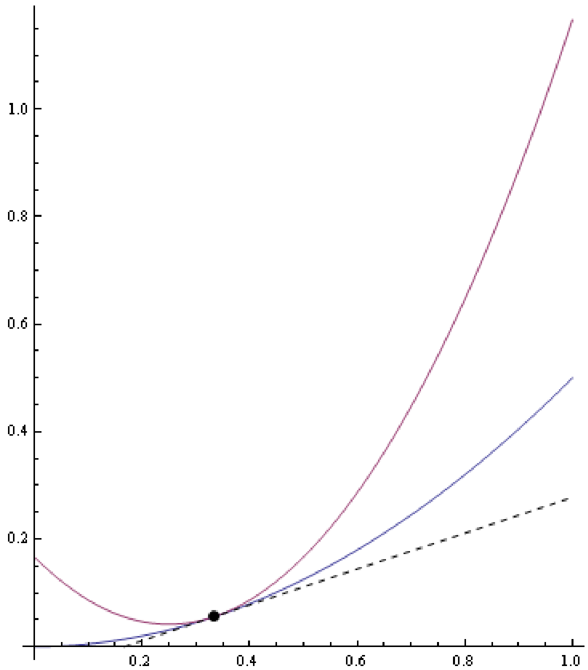
$$\tan\alpha = \left|\frac{f'(a) - g'(a)}{1 + f'(a)g'(a)}\right|$$

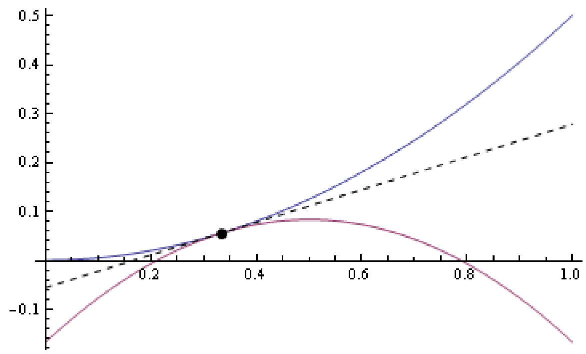
If $f'(a)g'(a) = -1$, then $\alpha = \pi/2$.

So let's calculate $\tan\alpha$ at these two points. At the point $(0,0)$, we have $f'(0) = 1$, $g'(0) = 0$. Plugging intot the formula, we obtain $\tan\alpha = 1$, so $\alpha = \pi/4$.

What about the point $(1,1)$? At this point, $f'(1) = 1$ and $g'(1) = 2$. Thus, we obtain that $\tan\alpha = 1/3$. So, α is the angle whose tangent is $1/3$. We don't know any particular angle α satisfying this condition, but by the intermediate-value theorem, we can see that α is somewhere between 0 and $\pi/6$. In the 153 course, we will look at inverse trigonometric functions, and when we do that, we will write α as $\arctan(1/3)$. For now, you can just leave your answer as $\tan\alpha = 1/3$.

3.5. **Curves that are tangent and orthogonal at specific points.** Here are some pictures of tangent pairs of curves:



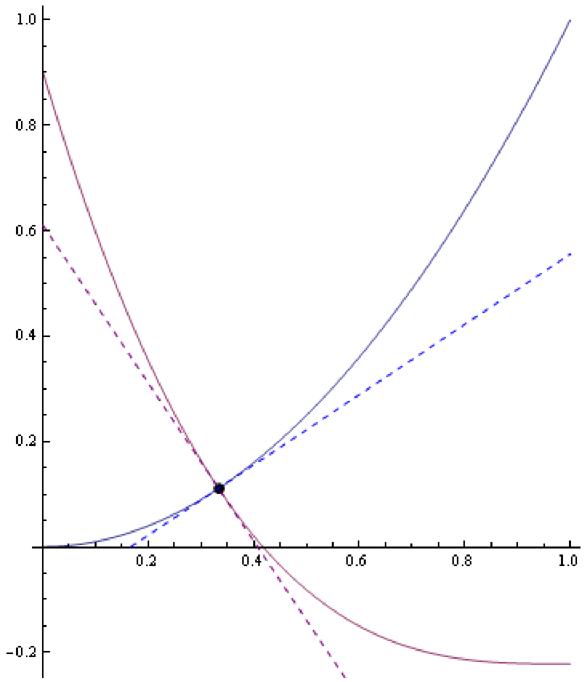


We say that two curves are *tangent* at a point if the angle of intersection between the curves at that point is 0. In other words, the curves share a tangent line. You can think of it as the two curves *touching* each other. [Draw pictures to explain]. Note that it could happen that they are kissing each other outward, as in this picture, or one is crossing the other, as in this picture.

If two curves are tangent at the point, that means that, *to the first order of approximation*, the curves behave very similarly close to that point. In other words, the best *linear* approximation to both curves cloose to that point is the same.

We say that two curves are *perpendicular* or *orthogonal* at a particular point of intersection if their tangent lines are perpendicular. If two curves are orthogonal at *every point of intersection*, we say that they are *orthogonal curves*. You'll see more on orthogonal curves in a homework problem.

Here is a graphical illustation of orthogonal curves:



3.6. **Geometric addendum: finding tangents and normals from other points.** Given the graph of a function $y = f(x)$, we can ask, given a point (x_0, y_0) in the plane (not necessarily on the graph) what tangents to the graph pass through this point. Similarly, we can ask what normals to the graph pass through this point.

Here's the general approach to these questions.

For the tangent line, we assume, say, that the tangent line is tangent at some point (x_1, y_1). The equation of the tangent line is then:

$$y - y_1 = f'(x_1)(x - x_1)$$

Since $y_1 = f(x_1)$, we get:

$$y - f(x_1) = f'(x_1)(x - x_1)$$

Further, we know that the tangent line passes through (x_0, y_0), so we plugging in, we get:

$$y_0 - f(x_1) = f'(x_1)(x_0 - x_1)$$

We now have an equation in the variable x_1. We solve for x_1. Note that when f is some specific function, both $f(x_1)$ and $f'(x_1)$ are expressions in x_1 that we can simplify when solving the equation.

In some cases, we may not be able to solve the equation precisely, but can guarantee the existence of a solution and find a narrow interval containing this solution by using the intermediate value theorem.

A similar approach works for normal lines. We'll revisit this in the near future.

4. Leibniz notation

4.1. **The d/dx notation.** Recall what we have done so far: we defined a notation of derivative, and then introduced a notation, the *prime notation*, for the derivative. So, if f is a function, the *derivative function* is denoted f'. And this is the function you obtain by differentiating f.

We now talk of a somewhat different notation that has its own advantages. And to understand the motivation behind this definition, we need to go back to how we think of derivative as the *rate of change*.

Suppose we have a function f. And, for simplicity, let's denote $y = f(x)$. So, what we want to do is study how, as the value of x changes, the value of $f(x)$, which is y, changes. So if for a point x_1 we have $f(x_1) = y_1$ and for a point x_2 we have $f(x_2) = y_2$, we want to measure the difference between the y-values (which is

$y_2 - y_1$) and compare it with the difference between the x-values. And the quotient of the difference in the y-values and the difference in the x-values is what we called the *difference quotient*:

$$\text{Difference quotient } = \frac{y_2 - y_1}{x_2 - x_1}$$

And, this is also the slope of the line joining (x_1, y_1) and (x_2, y_2).

Now, there's a slightly different way of writing this, which is typically used when the values of x are fairly close, and that is using the letter Δ. And that says:

$$\text{Difference quotient } = \frac{\Delta y}{\Delta x}$$

Here Δy means "change in" y and Δx means "change in" x. So, the difference quotient is the ratio of the change in y to the change in x.

The *limiting* value of this quotient, as the x-values converge and the y-values converge, is called dy/dx. This is read "dee y dee x" or "dee y by dee x" and is also called the derivative of y with respect to x.

So, if $y = f(x)$, the function f' can also be written as dy/dx.

4.2. **Derivative as a function and derivative as a point.** The function f' can be evaluated at any point; so for instance, we write $f'(3)$ to evaluate the derivative at 3. With the dy/dx notation, things are a little different. To evaluate dy/dx at a particular point, we write something like:

$$\frac{dy}{dx}|_{x=3}$$

The bar followed by the $x = 3$ means "evaluated at 3". In particular, it is *not* correct to write $dy/d3$. That doesn't make sense at all.

4.3. **Dependent and independent variable.** The Leibniz d/dx notation has a number of advantages over the prime notation. The first advantage is that instead of thinking in terms of functions, we now think in terms of two variables – x and y, and the relation between them. The fact that y can be expressed as a function of x becomes less relevant.

This is important because calculus is meant to address questions like: "When you change x, what happens to y?" The explicit functional form of y in terms of x is only of secondary interest – what matters is that we have these two variables measuring the two quantities x and y and we want to determine how changes in the variable x influence changes in the variable y. We sometimes say that x is the *independent variable* and y is the *dependent variable* – because y depends on x via some (may be known, may be unknown) functional dependence.

4.4. d/dx **as an operator and using it.** The great thing about the d/dx notation is that you don't need to introduce separate letters to name your function. For instance, we can write:

$$\frac{d}{dx}(x^2 + 3x + 4)$$

No need to define $f(x) := x^2 + 3x + 4$ and ask for $f'(x)$ – we can directly write this stuff down.

This not only makes it easier to write down the first step, it also makes it easier to write and apply the rules for differentiation.

(1) The sum rule becomes $d(u + v)/dx = du/dx + dv/dx$.
(2) The product rules becomes $d(uv)/dx = u(dv/dx) + v(du/dx)$.
(3) The scalar multiplication rule becomes $d(\alpha u)/dx = \alpha(du/dx)$.
(4) The difference rule becomes $d(u - v)/dx = du/dx - dv/dx$.
(5) The quotient rule becomes $d((u/v))/dx = (v(du/dx) - u(dv/dx))/v^2$.

The great thing about this notation is that we can write down partially calculated derivatives in intermediate steps, without naming new functions each time we break up the original function. For instance:

$$\frac{d}{dx}\left(\frac{x^3 + \sqrt{x} + 2}{x^2 + 1/(x + 1)}\right)$$

We can write down the first step:

$$\left[(x^2+1/(x+1))\frac{d}{dx}(x^3+\sqrt{x}+2)-(x^3+\sqrt{x}+2)\frac{d}{dx}(x^2+1/(x+1))\right]/(x^2+1/(x+1))^2$$

and then simplify the individual parts. If using the function notation, we would have to give names to both the numerator and denominator functions, but here we don't have to.

You should think of d/dx as an operator – the *differentiation operator* – that you can apply to expressions for functions.

5. Higher derivatives

5.1. **Higher derivatives and the multiple prime notation.** So far, we defined the derivative of a function at the point as the *rate of change* of the function at that point. The *second derivative* is the *rate of change of the rate of change*. In other words, the second derivative measures the rate at which the rate of change is changing. Or, it measures the rate at which the graph of the function is *turning* at the point, because a change in the derivative means a change in the direction of the graph.

So, this is just to remind you that higher derivatives are useful. Now, let's discuss the notation for higher derivatives.

In the prime notation, the second derivative of f is denoted f''. In other words, $f''(x)$ is the derivative of the function f' evaluated at x. The third derivative is denoted f''', the fourth derivative is denoted f''''. To simplify notation, we have the shorthand $f^{(n)}$ for the n^{th} derivative, where by n^{th} derivative we mean the function obtained after applying the differentiation operator n times. So the first derivative f' can also be written as $f^{(1)}$, the second derivative f'' can also be written as $f^{(2)}$, and so on. Typically, though, for derivatives up to the third derivative, we put the primes instead of the parenthesis (n) notation.

Well, how do we compute these higher derivatives? Differentiate one step at a time. So, for instance, if $f(x) = x^3 - 2x^2 + 3$, then $f'(x) = 3x^2 - 4x$, so $f''(x) = 6x - 4$, and $f'''(x) = 6$. The fourth derivative of f is zero. And all *higher* derivatives are zero.

Similarly, if $f(x) = 1/x$, then $f'(x) = -1/x^2$, $f''(x) = 2/x^3$, $f'''(x) = -6/x^4$, and so on and so forth.

The first derivative measures the rate of change of the function, or the slope of the tangent line. The second derivative measures the rate at which the tangent line is turning, or the speed with which the graph is turning. The third derivative measures the rate at which this rate of turning itself is changing.

So, for the function $f(x) = x^2$, the first derivative is $f'(x) = 2x$, and the second derivative is 2. So, the first derivative is an increasing function, and the second derivative is constant, so the graph of the function is turning at a *constant rate*.

(The detailed discussion of the role of derivatives in terms of whether a function is increasing or decreasing will be carried out later – for now, you should focus on the computational aspects of derivatives).

5.2. **Higher derivatives in the Leibniz notation.** The Leibniz notation for higher derivatives is a bit awkward if you haven't seen it before.

Recall that the first derivative of y with respect to x is denoted dy/dx. Note that what I just called *first derivative* is what I have so far been calling *derivative* – when I just say derivative without an ordinal qualifier, I mean first derivative. So the first derivative is dy/dx. How would we write the second derivative?

Well, the way of thinking about it is that the first derivative is obtained by applying the d/dx operator to y, so the second derivative is obtained by applying the d/dx operator to dy/dx. So the second derivative is:

$$\frac{d}{dx}\left(\frac{dy}{dx}\right)$$

And that's perfectly correct, but it is long to write, so we can write this in shorthand as:

$$\frac{d^2y}{(dx)^2}$$

Basically, we are (only notationally, not mathematically) multiplying the ds on top and multiplying the dx's down below. There's a further simplification we do with the notation – we omit the parentheses in the denominator, to get:

$$\frac{d^2y}{dx^2}$$

This is typically read out as "dee two y dee x two" though some people read it as "dee square y dee x square".

And more generally, the n^{th} derivative is given by the shorthand:

$$\frac{d^ny}{dx^n}$$

Note that the dx^n in the denominator should be thought of as $(dx)^n$ – *not* as $d(x^n)$, even though you omit parentheses.

This is typically read out as "dee n y dee x n".

So for instance:

$$\frac{d^2}{dx^2}\left(x^3+x+1\right) = \frac{d}{dx}\left(\frac{d(x^3+x+1)}{dx}\right) = \frac{d}{dx}\left(3x^2+1\right) = 6x$$

6. Chain rule

6.1. **The chain rule: statement and application to polynomials.** Now, the rules we have seen so far allow us to basically differentiate any rational function any number of times without ever using the limit definition – simply by applying the formulas.

Okay, how would you differentiate $h(x) := (x^2+1)^5$? Well, in order to apply the formulas, you need to expand this out first, then use the termwise differentiation strategy. But taking the fifth power of x^2+1 is a lot of work. So, we want a strategy to differentiate this without expanding.

This strategy is called the chain rule.

The chain rule states that if y is a function of v and v is a function of x, then:

$$\frac{dy}{dx} = \frac{dy}{dv}\frac{dv}{dx}$$

The intuition here may be that we can cancel the dvs – however, that's not a rigorous reason since these are not really ratios but limits. But that's definitely one way to remember the result.

In this case, we have $y = (x^2+1)^5$. What v can we choose? Well, let's try $v = x^2+1$. Then $y = v^5$. So we have:

$$\frac{dy}{dx} = \frac{d(v^5)}{dv}\frac{d(x^2+1)}{dx}$$

Now, the first term on the right is $5v^4$ and the second term on the right is $2x$, so the answer is:

$$\frac{dy}{dx} = (5v^4)(2x)$$

And $v = x^2+1$, so plugging that back in, we get:

$$\frac{dy}{dx} = 5(x^2+1)^4(2x) = 10x(x^2+1)^4$$

6.2. **Introspection: function composition.** What we really did was we had the function $h(x) = (x^2+1)^5$, and we decomposed h into two parts, the function g that sends x to x^2+1, and the function f that sends v to v^5. What we did was to write $h = f \circ g$, for two functions f and g that we can handle easily in terms of differentiation. Then, we used the chain rule to differentiate h, using what we know about differentiating f and g.

In functional notation, what the chain rule says equationally is that:

$$(f \circ g)'(x) = f'(g(x))g'(x)$$

Going back to the y, v terminology, we have $v = g(x)$, and $y = f(v)$. And in that notation, $dy/dv = f'(v) = f'(g(x))$, while $dv/du = g'(x)$. Which is precisely what we have here.

Notice that we apply f' not to x but to the value $g(x)$, which is what we called v, i.e., the value you get after applying one function but before applying the other one. But g' we apply to x.

Another way of writing this is:

$$\frac{d}{dx}\left[f(g(x))\right] = f'(g(x))g'(x)$$

This is a mix of the Leibniz notation and the prime notation.

6.3. **Precise statement of the chain rule.** What we said above is the correct equational expression for the chain rule, but let's just make the precise statement now with the assumptions.

If f, g are functions such that g is differentiable at x and f is differentiable at $g(x)$, then $f \circ g$ is differentiable at x and:

$$(f \circ g)'(x) = f'(g(x))g'(x)$$

Again, this statement is *at a point*, but if the differentiability assumptions hold globally, then the expression above holds globally as well, in which case we get:

$$(f \circ g)' = (f' \circ g) \cdot g'$$

The $\cdot$ there denotes the pointwise product of functions.

7. Additional facts and subtleties

Most of these are things you will discover with experience as you do homework problems, but they're mentioned here just for handy reference. This is something you might want to read more carefully when you review these notes at a later stage.

7.1. **One-sided versions.** The situation with one-sided versions of the results for derivatives is very similar to that with limits and continuity. For all pointwise combination results, one-sided versions hold. Conceptually, each result for derivatives depends (in its proof) on the corresponding result for limits. Since the result on limits has a one-sided version, so does the corresponding result on derivatives. In words:

(1) The left-hand derivative of the sum is the sum of the left-hand derivatives.
(2) The right-hand derivative of the sum is the sum of the right-hand derivatives.
(3) The left-hand derivative of a scalar multiple is the same scalar multiple of the left-hand derivative.
(4) The right-hand derivative of a scalar multiple is the same scalar multiple of the right-hand derivative.
(5) The analogue of the product rule for the left-hand derivative can be obtained if we replace the derivative in all three places in the product rule with left-hand derivative. Similarly for right-hand derivative.
(6) Similar to the above for quotient rule.

But – *you saw it coming* – the naive one-sided analogue of the rule for composites fails, for the same reason as the one-sided analogue of the composition results for limits and continuity fail. We need the additional condition that the direction of approach of the intermediate expression is the same as that of the original domain variable.

7.2. **The derivative as a function: is it continuous, differentiable?** Suppose f is a function. For simplicity, we'll assume the domain of f to be a (possibly infinite) open interval I in $\mathbb{R}$. We're taking an open interval to avoid one-sided issues at boundary points. We say that f is *differentiable* on its domain if f' exists everywhere on I. If f is differentiable, what can we say about the properties of f'?

Your first instinct may be to say that if f' is defined on an open interval, then it should be continuous on that interval. Indeed, in all the simple examples one can think of, the existence of the derivative *on an open interval* implies continuity of the derivative. However, this is not true as a general principle. Some points of note:

(1) It is possible for the derivative to not be a continuous function. An example is the function $g(x) := \{ \begin{matrix} x^2\sin(1/x), & x \neq 0 \\ 0, & x = 0 \end{matrix}$. This function is differentiable everywhere, but the derivative at 0 is *not* the limit of the derivative near zero.

(2) However, the derivative of a continuous function, if defined everywhere on an open interval, satisfies the intermediate value property. This is a fairly hard and not very intuitive theorem called Darboux's theorem, and you might see it if you take up the 203-204-205 analysis sequence. In this respect, it behaves in a manner very similar to a continuous function. In particular, any discontinuities of the derivative must be of the oscillatory kind on both the left and right side. In particular, *if* a one-sided limit *exists* for the derivative, it equals the value of the derivative.

The proof of (2) is well beyond the scope of this course. You don't even need to know the precise statement, and I'm including it here just in order to place the examples you've seen in context.

A fun discussion of the fact that derivatives need not be continuous but satisfy the intermediate value property and the implications of this fact can be found here:

`http://www.thebigquestions.com/2010/09/16/speed-math/`

7.3. **Higher differentiability.** We say that a function f on an open interval I is k times differentiable on I if the k^{th} derivative of f exists at all points of I. We say that f is k times continuously differentiable on I if the k^{th} derivative of f exists *and* is continuous on I.

Recall that differentiable implies continuous. Thus, if a function is twice differentiable, i.e., the first derivative is differentiable function, this implies that the first derivative is a continuous function. We thus see a chain of implications:

$$\text{Continuous} \Leftarrow \text{Differentiable} \Leftarrow \text{Continuously differentiable} \Leftarrow \text{Twice differentiable} \Leftarrow \text{Twice continuously differentiable} \leftarrow \dots$$

In general:

$$k \text{ times differentiable} \Leftarrow k \text{ times continuously differentiable} \Leftarrow k+1 \text{ times differentiable}$$

We say that a function f is *infinitely differentiable* if it is k times differentiable for all k. The above implications show us that this is equivalent to saying that f is k times continuously differentiable for all k.

A k times continuously differentiable function is sometimes also called a C^k-function. (When $k = 0$, we get continuous functions, so continuous functions are sometimes called C^0-functions). An infinitely differentiable function is sometimes also called a C^∞-function. We will not use the term in this course, though we will revisit it in 153 when studying power series, and you'll probably see it if you do more mathematics courses.

All these containments are strict. Examples along the lines of the $x^n \sin(1/x)$ constructions can be used to show this.

7.4. **Carrying out higher differentiation.** Suppose functions f and g are both k times differentiable. Are there rules to find the k^{th} derivatives of $f+g$, $f-g$, $f \cdot g$, etc. directly in terms of the k^{th} derivatives of f and g respectively? For the sum, difference, and scalar multiples, the rules are simple:

$$\begin{aligned}(f+g)^{(k)} &= f^{(k)} + g^{(k)} \\ (f-g)^{(k)} &= f^{(k)} - g^{(k)} \\ (\alpha f)^{(k)} &= \alpha f^{(k)}\end{aligned}$$

Later on, we'll see that this bunch of rules can be expressed more briefly by saying that the operation of differentiating k times is a linear operator.

For products, the rule is more complicated. In fact, the general rule is somewhat like the binomial theorem. The situation with composites is also tricky. We will revisit both products and composites a little later. For now, all we care about are existence facts:

- If f and g are k times differentiable on an open I, so are $f+g$, $f-g$, and $f \cdot g$. If g is not zero anywhere on I, then f/g is also k times differentiable on I.
- Ditto to the above, replacing "k times differentiable" by "k times continuously differentiable."
- If f and g are functions such that g is k times differentiable on an open interval I and f is k times differentiable on an open interval J containing the range of g, then $f \circ g$ is k times differentiable on I.
- Ditto to the above, replacing "k times differentable" by "k times continuously differentiable".

7.5. **Families of functions closed under differentiation.** Suppose $\mathcal{F}$ is a collection of functions that is closed under addition, subtraction and scalar multiplication. We say in this case that $\mathcal{F}$ is a *vector space* of functions. If, in addition, $\mathcal{F}$ contains constant functions and is closed under multiplication, we say that $\mathcal{F}$ is an *algebra* of functions. (You aren't responsible for learning this terminology, but it really helps make clear what we're going to be talking about shortly).

The vector space generated by a bunch of functions $\mathcal{B}$ is basically the set of all functions we can get starting from $\mathcal{B}$ by the processes of addition and scalar multiplication. If $\mathcal{B}$ generates a vector space $\mathcal{F}$ of functions, then we say that $\mathcal{B}$ is a generating set for $\mathcal{F}$.

For instance, if we consider all the functions $1, x, x^2, \dots, x^n, \dots$, these generate the vector space of all polynomials: we can get to all polynomials by the processes of addition and scalar multiplication starting with these functions.

The algebra generated by a bunch $\mathcal{B}$ of functions (which we assume includes constant functions) is the collection of functions $\mathcal{A}$ that we obtain by starting with the functions in $\mathcal{B}$ and the processes of addition, subtraction, multiplication, and scalar multiplication.

For instance, the algebra generated by the identity function (the function $xmapstox$) is the algebra of all polynomial functions.

The point of all this is as follows:

(1) Suppose $\mathcal{B}$ is a bunch of functions and $\mathcal{F}$ is the vector space generated by $\mathcal{B}$. Then, if every function in $\mathcal{B}$ is differentiable and the derivative of the function is in $\mathcal{F}$, then every function in $\mathcal{F}$ is differentiable and has derivative in $\mathcal{F}$. As a corollary, every function in $\mathcal{F}$ is infinitely differentiable and all its derivatives lie in $\mathcal{F}$.

(2) Suppose $\mathcal{B}$ is a bunch of functions and $\mathcal{A}$ is the algebra generated by $\mathcal{B}$. Then, if every function in $\mathcal{B}$ is differentiable and the derivative of the function is in $\mathcal{A}$, then every function in $\mathcal{A}$ is differentiable and the derivative of the function is in $\mathcal{A}$. As a corollary, every function in $\mathcal{A}$ is infinitely differentiable and all its derivatives lie in $\mathcal{A}$.

Let's illustrate point (2) (which is in some sense the more powerful statement) with the example of polynomial functions. The single function $f(x) := x$ generates the algebra of all polynomial functions. The derivative of f is the function 1, which is also in the algebra of all polynomial functions. What point (2) is saying is that just this simple fact allows us to see that the derivative of *any* polynomial function is a polynomial function, and that polynomial functions are infinitely differentiable.

We'll see a similar trigonometric example in the near future. We'll also explore this way of thinking more as the occasion arises.

Another way of thinking of this is that each time we obtain a formula to differentiate a bunch of functions, we have a technique to differentiate *all* functions in the algebra generated by that bunch of functions. While this may seem unremarkable, the analogous statement is *not true at all* for other kinds of operators such as indefinite integration.

TRIGONOMETRIC LIMITS AND DERIVATIVES

MATH 152, SECTION 55 (VIPUL NAIK)

Corresponding material in the book: Section 3.6.

Difficulty level: Easy to moderate, particularly if you remember corresponding stuff from AP level calculus.

Covered in class?: Yes, but not necessarily all examples.

What students should definitely get: The key trigonometric limits. The key differentiation formulas for trigonometric functions.

What students should eventually get: Techniques for computing limits and derivatives involving composites of trigonometric functions with each other and with polynomial and rational functions.

Executive summary

Words ...

(1) The following three important limits form the foundation of trigonometric limits: $\lim_{x\to 0}(\sin x)/x = 1$, $\lim_{x\to 0}(\tan x)/x = 1$, and $\lim_{x\to 0}(1-\cos x)/x^2 = 1/2$.
(2) The derivative of sin is cos, the derivative of cos is $-\sin$. The derivative of tan is $\sec^2$, the derivative of cot is $-\csc^2$, the derivative of sec is $\sec\cdot\tan$, and the derivative of csc is $-\csc\cdot\cot$.
(3) The second derivative of any function of the form $x \mapsto a\sin x + b\cos x$ is the negative of that function, and the fourth derivative is the original function.

Actions ...

(1) Substitution is one trick that we use for trigonometric limits: we translate $\lim_{x\to c}$ to $\lim_{h\to 0}$ where $x = c + h$.
(2) Multiplicative splitting, chaining, and stripping are some further tricks that we often use.
(3) For derivatives of functions that involve composites of trigonometric and polynomial functions, we *have* to use the chain rule as well as rules for sums, differences, products, and quotients when simplifying expressions.

1. Some critical trigonometric limits

Sinning by degrees is costly. For all applications of trigonometry to limits and calculus, *all angles are expressed in radians.* The radian measurement is the natural measurement for an angle.

1.1. The sinc function.

There are many other minor matters related to trigonometric functions that we need to address, but for now, let's get back and focus on one very important function – the so-called sinc function. This function is defined as $(\sin x)/x$ for $x \neq 0$.

Now, the function isn't defined at $x = 0$, and it isn't immediately clear what the limit is. Because when we just try to substitute the value at 0, we get a $0/0$ form. So it seems we need some angel to come and help us out.

Anyway, here's what the angel tells us:

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

1.2. Demystifying our angel.

For an acute angle θ, $\sin\theta$ is the vertical height (the y-coordinate of the point we get after rotating counter-clockwise by an angle of θ from $(1,0)$ along the unit circle, while θ is the arc length. You can see, from this picture, that the arc length θ is greater than $\sin\theta$. That's because $\sin\theta$ falls straight down while the arc moves both horizontally and vertically. However, and this is the crucial point – as θ gets smaller and smaller, you see that the arc length and the vertical line seem to get closer and closer. And this suggests that, perhaps, as θ tends to zero, $\sin\theta/\theta$ tends to 1.

Now, this is hardly a proof, because two numbers getting really close does not necessarily mean that their ratio tends to 1. But it is suggestive. So, with this suggestivity, let's believe that the limit, as x tends to 0, of the fraction $\sin x/x$ is 1.

1.3. **The substitution idea.** We know that $\lim_{x\to 0}(\sin x)/x = 1$. More generally, it is true that if f is continuous at c and $f(c) = 0$, then as $x \to c$, we have:

$$\lim_{x\to c} \frac{\sin(f(x))}{f(x)} = 1$$

For instance:

$$\lim_{x\to 0} \frac{\sin(x^2)}{x^2} = 1$$

Similarly:

$$\lim_{x\to 0} \frac{\sin(2x)}{2x} = 1$$

and:

$$\lim_{x\to 0} \frac{\sin(x/2)}{x/2} = 1$$

On the other hand, if we consider the limit:

$$\lim_{x\to 0} \frac{\sin(x + (\pi/3))}{x + (\pi/3)}$$

This limit is not 1 – the inner expression does *not* go to 0 as x goes to 0.

1.4. **Chaining limit computations.** Let's consider the computation:

$$\lim_{x\to 0} \frac{\sin(\sin x)}{x}$$

This is a special case of a more general limit computation that you have seen in Question 2 of the October 4 quiz. Let's first do this specific example, and then return to how it relates to that question.

For the specific example, we note that the limit in question involves a composite function. For such problems, we typically *chain* the limit by multiplying and dividing by the inner function. We get:

$$\lim_{x\to 0} \frac{\sin(\sin x)}{\sin x} \frac{\sin x}{x}$$

We now use that the limit of products is the product of limits, and obtain:

$$\lim_{x\to 0} \frac{\sin(\sin x)}{\sin x} \lim_{x\to 0} \frac{\sin x}{x}$$

The second limit is clearly 1. The first limit is also 1, because it is of the form $[\sin(f(x))]/f(x)$ where $f(x) \to 0$.

Now, the quiz question was that if $\lim_{x\to 0} g(x)/x = A \neq 0$ with g continuous, then what is $\lim_{x\to 0} g(g(x))/x$? The same chaining idea applies:

$$\lim_{x\to 0} \frac{g(g(x))}{x} = \lim_{x\to 0} \frac{g(g(x))}{g(x)} \frac{g(x)}{x}$$

We again split the limit multiplicatively, and argue that both component limits are A. For one of them, we have to argue that as $x \to 0$, $g(x) \to 0$ – an argument that we make in a somewhat indirect fashion. Go back to the quiz solution for more.

1.5. **Easier chainings.** Here are some easier examples:

$$\lim_{x \to 0} \frac{\sin(mx)}{x} = m$$

Here, we chain via mx.

$$\lim_{x \to 0} \frac{\sin(mx^n)}{x^n} = m$$

where n is positive.

1.6. **The $(1 - \cos x)/x^2$ limit.** We now show another fundamentally important trigonometric limit:

$$\lim_{x \to 0} \frac{1 - \cos x}{x^2} = \frac{1}{2}$$

We first show how to obtain this limit by multiplying both numerator and denominator by $1 + \cos x$. We get:

$$\lim_{x \to 0} \frac{1 - \cos^2 x}{x^2(1 + \cos x)}$$

The $1 + \cos x$ in the denominator pulls out by evaluation, and we get:

$$\frac{1}{2} \lim_{x \to 0} \frac{1 - \cos^2 x}{x^2}$$

We now use $1 - \cos^2 x = \sin^2 x$ and get:

$$\frac{1}{2} \lim_{x \to 0} \frac{\sin^2 x}{x^2}$$

This becomes:

$$\frac{1}{2} \lim_{x \to 0} \left(\frac{\sin x}{x} \right)^2$$

The limit of the square is the square of the limit (basically, a special case of the fact that the limit of the product is the product of the limit, so the inner limit is 1, and we get a 1/2.

We can also obtain this limit using a double angle formula as described below:

The idea here is to use the identity we saw last time, which was that $\cos 2A = 1 - 2\sin^2 A$. So $1 - \cos 2A = 2\sin^2 A$. Here $A = x/2$, so we get:

$$\lim_{x \to 0} \frac{2\sin^2(x/2)}{x^2}$$

We can pull the 2 out, and we get $\sin^2(x/2)/x^2$ inside. Now, the thing with calculating these limits is that we only know how to calculate the limit of the form $\sin\theta/\theta$, and here, $\theta = x/2$. So, we rewrite the denominator as $4(x/2)^2$, and we pull out the 4, so we get:

$$\frac{1}{2} \lim_{x \to 0} \left(\frac{\sin(x/2)}{(x/2)} \right)^2$$

Now, using the *limit of product equals product of limits* meme, we get:

$$\frac{1}{2} \lim_{x \to 0} \frac{\sin(x/2)}{(x/2)} \lim_{x \to 0} \frac{\sin(x/2)}{(x/2)}$$

Now, in both cases, we have $x \to 0$, so $x/2 \to 0$, so both limits are 1, and hence, our final answer is 1/2.

The $\tan x/x$ **limit.** Let's now calculate the limit:

$$\lim_{x\to 0} \frac{\tan x}{x}$$

What we do is to write $\tan x = \sin x/\cos x$:

$$\lim_{x\to 0} \frac{\sin x}{x\cos x}$$

We split this as a product:

$$\lim_{x\to 0} \frac{1}{\cos x} \lim_{x\to 0} \frac{\sin x}{x}$$

Both limits are 1, so the overall limit is 1.

1.7. **Corollaries.** We can now state some easy corollaries of the above results:

$$\lim_{x\to 0} \frac{1-\cos(mx)}{x^2} = \frac{m^2}{2}$$

We obtain by chaining via $(mx)^2$.
Similarly:

$$\lim_{x\to 0} \frac{\tan(mx)}{x} = m$$

1.8. **Substitution that involves translation.** There is a *substitution of variables* trick that we can use for computing limits. When we were computing limits for rational functions, we never really needed that trick, primarily because we knew how to handle rational functions anyway. But this trick comes in useful for trigonometric functions.

The trick is:

$$\lim_{x\to c} f(x) = \lim_{h\to 0} f(c+h)$$

Similarly:

$$\lim_{x\to c^+} f(x) = \lim_{h\to 0^+} f(c+h)$$

and:

$$\lim_{x\to c^-} f(x) = \lim_{h\to 0^-} f(c+h) = \lim_{h\to 0^+} f(c-h)$$

Why is this useful for trigonometric functions? Because for trigonometric functions, the *only* nontrivial limit that we know is the one I just told you: $\lim_{x\to 0} \frac{\sin x}{x} = 1$. So, we need to basically use this for any nontrivial limit that we need to compute. For instance, consider the limit:

$$\lim_{x\to \pi/4} \frac{\sin x - (1/\sqrt{2})}{x - \pi/4}$$

Now, we want to change the thing that's limiting to h, so we rewrite this as:

$$\lim_{h\to 0} \frac{\sin(\pi/4 + h) - (1/\sqrt{2})}{h}$$

We simplify the numerator using the $\sin(A+B)$ formula, which we know is $\sin A\cos B + \cos A \sin B$, and we get:

$$\lim_{h\to 0} \frac{(1/\sqrt{2})(\sin(h) + \cos(h) - 1)}{h}$$

We take out the $1/\sqrt{2}$ factor and now try to split the inner limit additively:

$$\frac{1}{\sqrt{2}}\left(\lim_{h\to 0}\frac{\sin h}{h} - \lim_{h\to 0}\frac{1-\cos h}{h}\right)$$

The first limit is 1. For the second limit, it can be written as $\lim_{h\to 0} h \cdot \lim_{h\to 0}(1-\cos h)/h^2 = 0(1/2) = 0$. So the second limit is 0. Overall, the limit is $1/\sqrt{2}$.

We'll see this same calculation a little later when we try to calculate the derivative of the sin function.

2. Stripping: a sneak peek

We will cover this somewhat delicate operation a little later in 153, when students are more mature (in terms of having seen more kinds of functions) and can handle the issue with the requisite care. But some of you may be equipped to use this approach to compute trigonometric limits more intuitively and rapidly now. [NOTE: To my knowledge, this particular intuitive approach to stripping is not found in any high school or college calculus text and can be considered my own invention, though most people who compute limits on a regular basis do this all the time.]

To motivate stripping, let us look at a fancy example:

$$\lim_{x\to 0}\frac{\sin(\tan(\sin x))}{x}$$

This is a composite of three functions, so if we want to chain it, we will chain it as follows:

$$\lim_{x\to 0}\frac{\sin(\tan(\sin x))}{\tan(\sin x)}\frac{\tan(\sin x)}{\sin x}\frac{\sin x}{x}$$

We now split the limit as a product, and we get:

$$\lim_{x\to 0}\frac{\sin(\tan(\sin x))}{\tan(\sin x)}\lim_{x\to 0}\frac{\tan(\sin x)}{\sin x}\lim_{x\to 0}\frac{\sin x}{x}$$

Now, we argue that each of the inner limits is 1. The final limit is clearly 1. The middle limit is 1 because the inner function $\sin x$ goes to 0. The left most limit is 1 because the inner function $\tan(\sin x)$ goes to 0. Thus, the product is $1 \times 1 \times 1$ which is 1.

If you are convinced, you can further convince yourself that the same principle applies to a much more convoluted composite:

$$\lim_{x\to 0}\frac{\sin(\sin(\tan(\sin(\tan(\tan x)))))}{x}$$

However, *writing that thing out takes loads of time*. Wouldn't it be nice if we could just strip off those sins and tans? In fact, we can do that.

The key stripping rule is this: *in a multiplicative situation* (i.e. there is no addition or subtraction happening), if we see something like $\sin(f(x))$ or $\tan(f(x))$, and $f(x) \to 0$ in the relevant limit, then we can strip off the sin or tan. In this sense, both sin and tan are *strippable* functions. A function g is strippable if $\lim_{x\to 0} g(x)/x = 1$.

The reason we can strip off the sin from $\sin(f(x))$ is that we can multiply and divide by $f(x)$, just as we did in the above examples.

Stripping can be viewed as a special case of the l'Hopital rule as well, but it's a much quicker shortcut in the cases where it works.

Thus, in the above examples, we could just have stripped off the sins and tans all the way through.

Here's another example:

$$\lim_{x\to 0}\frac{\sin(2\tan(3x))}{x}$$

As $x \to 0$, $3x \to 0$, so $2\tan 3x \to 0$. Thus, we can strip off the outer sin. We can then strip off the inner tan as well, since its input $3x$ goes to 0. We are thus left with:

$$\lim_{x\to 0}\frac{2(3x)}{x}$$

Cancel the x and get a 6. We could also do this problem by chaining or the l'Hopital rule, but stripping is quicker and perhaps more intuitive.

Here's yet another example:

$$\lim_{x \to 0} \frac{\sin(x \sin(\sin x))}{x^2}$$

As $x \to 0$, $x \sin(\sin x) \to 0$, so we can strip off the outermost sin and get:

$$\lim_{x \to 0} \frac{x \sin(\sin x)}{x^2}$$

We cancel a factor of x and get:

$$\lim_{x \to 0} \frac{\sin(\sin x)}{x}$$

Two quick sin strips and we get x/x, which becomes 1.

Yet another example:

$$\lim_{x \to 0} \frac{\sin(ax)\tan(bx)}{x}$$

where a and b are constants. Since this is a multiplicative situation, and $ax \to 0$ and $bx \to 0$, we can strip the sin and tan, and get:

$$\lim_{x \to 0} \frac{(ax)(bx)}{x}$$

This limit becomes 0, because there is a x^2 in the numerator and a x in the denominator, and cancellation of one factor still leaves a x in the numerator.

Here is yet another example:

$$\lim_{x \to 0} \frac{\sin^2(ax)}{\sin^2(bx)}$$

where a, b are nonzero constants. We can pull the square out of the whole expression, strip the sins in both numerator and denominator, and end up with a^2/b^2.

When you can't strip. The kind of situations where we are not allowed to strip are where the expression $\sin(f(x))$ is not just multiplied but is being added to or subtracted from something else. For instance, in order to calculate the limit:

$$\lim_{x \to 0} \frac{x - \sin x}{x^3}$$

stripping off the sin would be a sin. The intuition behind what is wrong with this will have to wait till next quarter.

What about $1 - \cos(f(x))$? If we see something like $1 - \cos(f(x))$ with $f(x) \to 0$ in the limit, then we know that it "looks like" $f(x)^2/2$ and we can replace it by $f(x)^2/2$. This can be thought of as a sophisticated version of stripping. For instance:

$$\lim_{x \to 0} \frac{(1 - \cos(5x^2))}{x^4}$$

Here $f(x) = 5x^2$, so the numerator is like $(5x^2)^2/2$, and the limit just becomes $25/2$.

3. Differentiating the trigonometric functions

3.1. Differentiation formulas. The differentiation formulas are as follows:

$$\begin{aligned}
\sin' &= \cos \\
\cos' &= -\sin \\
\tan' &= \sec^2 \\
\sec' &= \sec \cdot \tan \\
\cot' &= -\csc^2 \\
\csc' &= -\csc \cdot \cot
\end{aligned}$$

We can obtain all these just from the differentiation formulas for sin and cos, using the quotient rule:

$$\frac{d}{dx}[\sin x] = \cos x$$

and:

$$\frac{d}{dx}[\cos x] = \sin x$$

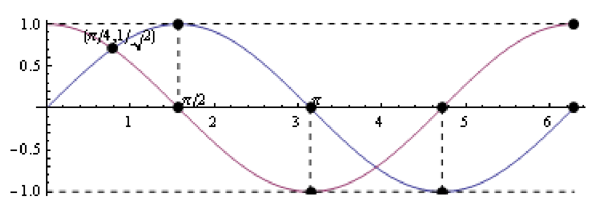
In fact, the differentiation formula for cos can be obtained from that of sin using the chain rule as follows:

$$\frac{d}{dx}[\cos x] = \frac{d}{dx}[\sin((\pi/2) - x)] = -\cos((\pi/2) - x) = -\sin x$$

So, in order to obtain all these formulas, all we need to do is obtain the differentiation formula for sin. The derivation of this formula is given below. Essentially, it uses the fact that $\lim_{h \to 0}(\sin h)/h = 1$:

$$\begin{aligned}
&\lim_{h \to 0} \frac{\sin(x+h) - \sin x}{h} \\
=& \lim_{h \to 0} \frac{\sin x \cos h + \cos x \sin h - \sin x}{h} \\
=& \sin x \lim_{h \to 0} \frac{\cos h - 1}{h} + \cos x \lim_{h \to 0} \frac{\sin h}{h} \\
=& \sin x \cdot 0 + \cos x \cdot 1 \\
=& \cos x
\end{aligned}$$

The fact that $(\cos h - 1)/h \to 0$ can be deduced by multipled by $1 + \cos h$ and simplifying. Basically, $\cos h - 1$ goes to 0 at the rate of h^2, and the denominator is only h.



3.2. Graphical interpretation of these derivatives.

The fascinating thing about the sin function is that its derivative looks just like the function – except it's shifted over by $\pi/2$. Let's see what this means graphically.

At the point $x = 0$, the derivative is $\cos 0 = 1$. This means that the tangent is the line $y = x$. And indeed, close to the point 0, the sine graph looks a lot like the line $y = x$. However, there's an added subtlety that you may not be able to appreciate right now but will be able to later. The graph of the sine function on the right of the point 0 falls slightly on the lower side of the $y = x$ line, and the graph to the left of 0 falls slightly on the upper side of the line. In other words, the curve actually passes from one side of the tangent line at 0 to the other side.

This is a very unusual situation, because most of the time that you think of tangent lines, the graph of the function *close to the point* lies entirely on one side of the tangent line. But here, the graph is *crossing* the tangent line. These kinds of points are called *points of inflection*, and believe it or not, for a twice differentiable function, a point of inflection is a point where the *second* derivative is zero (though there could be points where the second derivative is zero that are not points of inflection – funny things could happen with the *third* derivative). We'll talk more about inflections later.

Anyway, so at zero, the derivative is 1. What happens as x increases from 0 to $\pi/2$? The cosine function keeps decreasing form 1 to 0. But since the cosine function is positive, the tangent to the sine function is still upward-sloping, so the sine function is increasing, but the slope of the tangent line is falling. Finally at $x = \pi/2$, the sine function reaches its peak of 1, and its derivative, the cosine function, becomes 0.

So here's the beautiful things you see, and for which you now have a powerful explanation based on the derivative:

(1) The points where the sine function attains its maxima and minima are precisely the points where the cosine function is 0. Namely, these points are odd multiples of $\pi/2$. The points where the cosine function attains its maxima and minima are precisely the points where the sine function is 0. Namely, these points are the multiples of π.
(2) The regions where the sine function is increasing (respectively, decreasing) are the same as the regions where the cosine function is positive (respectively, negative). The regions where the cosine function is increasing (respectively, decreasing) are the same as the regions where the sine function is negative (respectively, positive).

3.3. **We've done it!** What I've given you is the complete toolkit with which you can calcuate any derivative that involves a mix of trigonometric, polynomial, and rational functions. Let's consider an example: $\sin(x^2)$.

How do you think of this function? To evaluate this function at a particular value of x, what you first do is calculate x^2, and then apply the sin function to that. So, you're doing two operations in sequence, whereby, you're feeding the output of one operation as the input of the other. So what you're doing is essentially *function composition*.

Which brings us back to the chain rule. Remember the way the chain rule works? We set $v = x^2$, first differentiate the function with respect to v, and then multiply by dv/dx. So, we get that the derivative is $(\cos(x^2)) \cdot 2x$.

For composites of trigonometric functions and polynomial, the chain rule is *indispensable* – if you don't want to use the chain rule, there's no way of calculating the derivative without *actually calculating it using the limit definition*. In the case of polynomials and rational functions, the chain rule was a convenience. Now it's a necessity.

3.4. **Periodicity of the sequence of derivative functions.** The derivative of sin is cos. The derivative of cos is $-\sin$. The derivative of $-\sin$ is $-\cos$. The derivative of $-\cos$ is sin.

In other words, the sin function equals its *fourth* derivative. If we consider the sequence of derivative functions of sin, this *sequence* has a *period* of four. (We'll formally define sequence and period of a sequence in 153, but you know what this means). Thus, for instance, the 101^{th} derivative of sin is the same as the first derivative (because the remainder on dividing 101 by 4 is 1) and is thus cos.

[Aside: This is significant in many ways. For instance, when we study antidifferentiation (indefinite integration) we'll notice that the above basically tells us that we can keep taking antiderivatives of sin or cos and still remain sin or cos (up to a plus or minus). This is significant when, for instance, we study integration by parts. There, we will choose sin or cos as the *part to repeatedly integrate* precisely because repeated integration does not increase the complexity of the function.]

3.5. **Second derivative same with a minus sign.** Also note that for the sin function, the second derivative is the *negative* of the function. This is also true for the cos function. It is also true of all *linear combinations* of sin and cos, i.e., all functions of the form $x \mapsto a \sin x + b \cos x$ where $a, b \in \mathbb{R}$.[1] For any function f of this form, $f''(x) = -f(x)$.

What's so special about this fact? *This is the reason why trigonometric functions, specifically linear combinations of* sin *and* cos*, arise in nature.* What happens is that some basic physical/biological/chemical/ecological law or constraint forces the solution function we have to satisfy the equation $f''(x) = -\omega^2 f(x)$ or something

[1]Formally, this is the vector space of functions generated by sin and cos.

like that. Now, functions of the above form (with appropriate scaling) pop up naturally. More on this when we study differential equations.

DERIVATIVE AS RATE OF CHANGE, IMPLICIT DERIVATIVES: ROUGH APPROXIMATION OF LECTURE

MATH 152, SECTION 55 (VIPUL NAIK)

Corresponding material in the book: Sections 3.4 and 3.7

Difficulty level: Easy to moderate, since most of these should be familiar to you and there are no new subtleties being added here.

What students should definitely get: The notion of derivative as a rate of change, handling word problems that ask for rates of change. The main idea and procedure of implicit differentiation.

What students should hopefully get: The distinction between conceptual and computational, the significance of implicit differentiation, understanding the relative rates concept and its intuitive relationship with the chain rule.

Executive summary

Derivative as rate of change. Words...

(1) The derivative of A with respect to B is the rate of change of A with respect to B. Thus, to determine rates of change of various quantities, we can use the techniques of differentiation.
(2) If there are three linked quantities that are changing together (e.g., different measures for a circle such as radius, diameter, circumference, area) then we can use the chain rule.

Most of the actions in this case are not more than a direct application of the words.

Implicit differentiation. Words...

(1) Suppose there is a curve in the plane, whose equation cannot be manipulated easily to write one variable in terms of the other. We can use the technique of implicit differentiation to determine the derivative, and hence the slope of the tangent line, at different points to the curve.
(2) For a curve where neither variable is expressible as a function of the other, the notion of derivative still makes sense as long as *locally*, we can get y as a function of x. For instance, for the circle $x^2 + y^2 = 1$, y is not a function of x, but if we restrict attention to the part of the circle above the x-axis, then on this restricted region, y is a function of x.
(3) In some cases, even when one variable is expressible as a function of the other, implicit differentiation is easier to handle as it may involve fewer messy squareroot symbols.

Actions ...

(1) To determine the derivative using implicit differentiation, write down the equations of both curves, differentiate both sides with respect to x, and simplify using all the differentiation rules, to get everything in terms of x, y, and dy/dx. Isolate the dy/dx term in terms of x and y, and compute it at whatever point is needed.
(2) This procedure can be iterated to compute higher order derivatives at specific points on the curve where the curve locally looks like a function.

1. Conceptual versus computational

Back in the first lecture, I defined the concept of function. A function is some kind of machine that takes an input and gives an output. And the important thing about functions is that *equal inputs give equal outputs.*

The interesting thing about functions is that this way of thinking about functions is a sort of *black box, hands-off* approach. If you think of the function as this box machine which sucks in an input from one side and spits out the output on the other side, we don't really care *how the black box works.* It doesn't matter what is happening inside, as long as we are guaranteed that equal inputs give equal outputs.

With this abstract concept of function, we defined the notion of limit, which was the $\epsilon - \delta$ definition, and this definition didn't really depend on how you compute f. Then we defined the notion of derivative, which is a particular kind of limit, namely, the limit of the different quotient. And in all this, how to *compute* things wasn't the focus. And simply thinking of things conceptually, we got a lot of insights. We understood what limits mean and we understood what derivatives mean, and we saw the qualitative significance.

Complementing this conceptual understanding of the concepts of functions, limits, continuity, derivatives, and differentiation, there is the computational aspect. The computational aspect tells us how, for functions with specific functional forms or expressions, we can calculate limits and derivatives. And in order to do this, we use general theorems (limits for sums, differences, ...; derivatives for sums, differences ...) and specific tricks and formulas.

What you should remember, though, is that *just because you cannot compute something, doesn't mean that it cannot be understood qualitatively.* So, if you encounter a function and there's no formula to differentiate it, that's not the same as saying that it isn't differentiable. Computation is one tool among many to get a conceptual understanding of ideas.

This is really important because a lot of the places where you'll see these mathematical ideas applied are cases where the functions involved are inherently *unknown* or *unknowable* – there aren't explicit expressions for them. Still, we want to talk about the broad qualitative properties – is the function continuous? Is it differentiable? Is it twice differentiable? Is it increasing or decreasing, is it oscillating? Often, we can answer these qualitative questions without having explicit expressions for the functions.

2. Derivative as a rate of change

Recall that if f is a function, the derivative f' is the *rate of change* of the output of f relative to the input. Or, if we are thinking of two quantities x and y, where y is functionally dependent on x, then the rate of change of y with respect to x is dy/dx. That is the limit of the *difference quotient* $\Delta y/\Delta x$.

This means that if we want to ask the question: *if the rate of change of x is this much, what is the rate of change of y*, we should think of derivatives.

For instance, we know that the area of a circle of radius r is πr^2. We may ask the question: what is the rate of change of the area with respect to the radius? This is the derivative of πr^2 with respect to r, and that turns out to be $2\pi r$.

For instance, if $r = 5$, the rate of change of the area with respect to the radius is 10π.

Now, suppose the radius is changing at the rate of $5m/hr$. That means that every hour, the radius increases by $5m$. What is the rate of increase of the area with respect to time, when the radius is $100m$. Well, here we have three quantities, the area A, the radius r, and the time t. r is a function of t, and $dr/dt = 5m/hr$ and $dA/dr = 2\pi r$. So by the chain rule, we have $dA/dt = (dA/dr)(dr/dt = (2\pi r)(5m/hr)$. And since $r = 100m$, we get $1000\pi m^2/hr$.

3. Implicit differentiation

3.1. **Introduction.** So far, when trying to differentiate one quantity with respect to another quantity, what we do is to write one as a function of the other, and then differentiate that function. This is all very good when we have an explicit expression for the function. Sometimes, however, we do not really have a functional expression for one quantity in terms of the other, but we do know of a *relation* between the two quantities.

Let's think of this a little differently. One importance of differentiation is that it allows us to find tangent lines to curves that arise as the graph of a function. This has some geometric significance, if we are trying to understand the geometry of a curve that arises as the graph of a function. But what about the curves that don't arise from explicit functions? Or, where we don't have explicit functional expressions?

For instance, let's look at the circle of radius 1 centered at the origin. This is given by the equation $x^2 + y^2 = 1$. Note that in this case, y is *not* a function of x, because for many values of x, there are two values of y. For instance, for $x = 0$, we have $y = 1$ and $y = -1$. For $x = 1/2$, we have $y = \sqrt{3}/2$ and $y = -\sqrt{3}/2$. So, y is not a function of x.

However, *locally* y is still a function of x, in the following sense. If you just restrict yourself to the part above the x-axis, then you do get y as a function of x. This is the function $y := \sqrt{1 - x^2}$ for $-1 \leq x \leq 1$. If we restrict ourselves to the part below the x-axis, we consider the function $y := -\sqrt{1 - x^2}$ for $-1 \leq x \leq 1$.

Now, how do we calculate dy/dx? Well, it depends on whether we are interested in the part above the x-axis or in the part below the x-axis. For the part above the x-axis, we have the function $\sqrt{1-x^2}$, and we get that the derivative is:

$$\frac{d(\sqrt{1-x^2})}{dx} = \frac{d(\sqrt{1-x^2})}{d(1-x^2)}\frac{d(1-x^2)}{dx} = \frac{1}{2\sqrt{1-x^2}} \cdot (2x) = \frac{-x}{\sqrt{1-x^2}}$$

If we are interested in the lower side, we get $x/\sqrt{1-x^2}$.

Now, in this case, we have to split into two cases, and do a painful calculation involving differentiating a square root via the chain rule.

Here's another way of handling this differentiation, that does not involve a messy square root.

We start with the original expression:

$$x^2 + y^2 = 1$$

This is an *identity*, which means that it's true for every point on the curve. When we have an equation that is identically true, it is legitimate to differentiate both sides and still get an identity. Differentiating both sides with respect to x, we get:

$$\frac{d(x^2)}{dx} + \frac{d(y^2)}{dx} = 0$$

Simplifying and using the chain rule, we get:

$$2x + 2y\frac{dy}{dx} = 0$$

We thus get:

$$\frac{dy}{dx} = \frac{-x}{y}$$

Notice that with this method, we get $-x/y$, which works in *both* cases. When $y = \sqrt{1-x^2}$, we get $-x/\sqrt{1-x^2}$, and when $y = -\sqrt{1-x^2}$, we get $x/\sqrt{1-x^2}$. The method that we used is called *implicit differentiation*.

So the idea of implicit differentiation is that, instead of writing $y = f(x)$ and then differentiating both sides, we differentiate the messy mixed-up expression on both sides with respect to x. Next, we use the various rules (sum rule, difference rule, product rule, quotient rule) to keep splitting things up into smaller and smaller pieces, and in the final analysis, we get everything in terms of x, y, and dy/dx. Then, we try to separate dy/dx completely to one side.

Let's look at another example:

$$\sin(x+y) = xy$$

So, what we do is differentiate both sides:

$$\frac{d(\sin(x+y))}{dx} = \frac{d(xy)}{dx}$$

Now, how would we handle something like $\sin(x+y)$? It is something in terms of $x+y$, so we use the chain rule on the left side, thinking of $v = x+y$ as the intermediate function:

$$\frac{d(\sin(x+y))}{d(x+y)}\frac{d(x+y)}{dx} = x\frac{dy}{dx} + y\frac{dx}{dx}$$

This simplifies to:

$$\cos(x+y)\left[1 + \frac{dy}{dx}\right] = x\frac{dy}{dx} + y$$

Opening up the parentheses, we get:

$$\cos(x+y) + \cos(x+y)\frac{dy}{dx} = x\frac{dy}{dx} + y$$

Now, we move stuff together to one side, to get:

$$(\cos(x+y) - x)\frac{dy}{dx} = y - \cos(x+y)$$

And we now isolate dy/dx:

$$\frac{dy}{dx} = \frac{y - \cos(x+y)}{\cos(x+y) - x}$$

3.2. **Implicit differentiation: understood better.** So, in implicit differentiation, what we're doing is, instead of thinking of an explicit functional form, we are using a relation that is true for every point in the curve, then *differentiating both sides*. Next, we keep trying to simplify the expression we have using the various rules until we land up with something that just involves x, y, and dy/dx. Till this point, it's usually smooth sailing. Now, it may be the case that we can *isolate* dy/dx and hence get an expression for it in terms of x and y. If that's the case, then we're in good shape.

Note the following key difference: when y is an explicit function of x, then the expression we get for dy/dx only involves x and does not have the letter y appearing in it. However, in the implicit case, the expression we get for dy/dx involves both x and y together.

3.3. **Higher derivatives using implicit differentiation.** We can also use implicit differentiation to compute second derivatives and higher derivatives. Here's what we do. First, we get the expression for dy/dx. In other words, we write:

$$\frac{dy}{dx} = \text{Some expression in terms of } x \text{ and } y$$

We now differentiate both sides with respect to x. Again, this differentiation is valid because the above relation holds as an identity, and not just as an isolated point.

The left side becomes d^2y/dx^2. For the right side, we again use the same idea: we split as much as possible using the sum rule, product rule, etc. For the expressions that purely involve x, we differentiate the usual way. For the expressions that purely involve y, we differentiate with respect to y and multiply by dy/dx. The upshot is that we get:

$$\frac{d^2y}{dx^2} = \text{Some expression in terms of } x,\ y, \text{ and } \frac{dy}{dx}$$

Now, we plug back the earlier expression for dy/dx in terms of x and y into this expression, and get an expression for d^2y/dx^2.

ROLLE'S, MEAN-VALUE, INCREASE/DECREASE, EXTREME VALUES

MATH 152, SECTION 55 (VIPUL NAIK)

Corresponding material in the book: Sections 4.1-4.4.

Difficulty level: Moderate to hard. While most of these are ideas you have probably seen at the AP level or equivalent, our treatment of the topics will be somewhat more thorough. Also, this is extemely important as preparation for the process of graphing a function, which in turn is very important as a general tool for understanding all kinds of functions.

What students should definitely get: The statements of Rolle's theorem and the mean value theorem. The relationship between the signs of one-sided derivatives and whether the function value at a point is greater or less than the function value to its immediate left or right. The notions of local maximum, local minimum, point of increase, point of decrease. The definition of critical point. The first derivative test and second derivative test. The procedure for determining absolute maxima and minima.

What students should hopefully get: The distinction between being positive and being nonnegative; similarly, the distinction between being negative and being nonpositive. In particular, the fact that even when difference quotients are strictly positive, the derivative obtained as the limit may be zero. The conceptual distinction between local extreme values (a local condition) and absolute extreme values.

Executive summary

Words...

(1) If a function f is continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b), and $f(a) = f(b) = 0$, then there exists $c \in (a, b)$ such that $f'(c) = 0$. This is called *Rolle's theorem* and is a consequence of the extreme-value theorem.
(2) If a function f is continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b), then there exists $c \in (a, b)$ such that $f'(c)$ is the difference quotient $(f(b) - f(a))/(b - a)$. This result is called the *mean-value theorem*. Geometrically, it says that for any chord, there is a parallel tangent. Another way of thinking about it is that every difference quotient is equal to a derivative at some intermediate point.
(3) If f is a function and c is a point such that $f(c) \geq f(x)$ for x to the immediate left of c, we say that c is a local maximum from the left. In this case, the left-hand derivative of f at c, if it exists, is greater than or equal to zero. This is because the difference quotient is greater than or equal to zero. Local maximum from the right implies that the right-hand derivative (if it exists) is ≤ 0, local minimum from the left implies that the left-hand derivative (if it exists) is ≤ 0, and local minimum from the right implies that the right-hand derivative (if it exists) is ≥ 0. Even in the case of *strict* local maxima and minima, we still need to retain the equality sign on the derivative because it occurs as a *limit* and a limit of positive numbers can still be zero.
(4) If c is a point where f attains a local maximum (i.e., $f(c) \geq f(x)$ for all x close enough to c on both sides), then $f'(c)$, if it exists, is equal to zero. Similarly for local minimum.
(5) A *critical point* for a function is a point where either the function is not differentiable or the derivative is zero. All local maxima and local minima must occur at critical points.
(6) If $f'(x) > 0$ for all x in the open interval (a, b), f is increasing on (a, b). Further, if f is one-sided continuous at the endpoint a and/or the endpoint b, then f is increasing on the interval including that endpoint. Similarly, $f'(x) < 0$ implies f decreasing.
(7) If $f'(x) > 0$ everywhere except possibly at some isolated points (so that they don't cluster around any point) where f is still continuous, then f is increasing everywhere.
(8) If $f'(x) = 0$ on an open interval, f is constant on that interval, and it takes the same constant value at an endpoint where it's continuous from the appropriate side.

(9) If f and g are two functions that are both continuous on an interval I and have the same derivative on the interior of I, then $f - g$ is a constant function.

(10) There is a *first derivative test* which provides a sufficient (though not necessary) condition for a local extreme value: it says that if the first derivative is nonnegative (respectively positive) on the immediate left of a critical point, that gives a strict local maximum (respectively local maximum) from the left. If the first derivative is negative on the immediate left, we get a strict local minimum from the left. If the first derivative is positive on the immediate right, we get a strict local minimum from the right, and if it is negative on the immediate right, we get a strict local maximum from the right.

    The first derivative test is similar to the corresponding "one-sided derivative" test, but is somewhat stronger for a variety of situations because in many cases, one-sided derivatives are zero, which is inconclusive, whereas the first derivative test fails us more rarely.

(11) The second derivative test states that if f has a critical point c where it is twice differentiable, then $f''(c) > 0$ implies that f has a local minimum at c, and $f''(c) < 0$ implies that f has a local maximum at c.

(12) There are also higher derivative tests that work for critical points c where $f'(c) = 0$. These work as follows: we look for the smallest k such that $f^{(k)}(c) \neq 0$. If this k is even, then f has a local extreme value at c, and the nature (max versus min) depends on the sign of $f^{(k)}(c)$ (max if negative, min if positive). If k is odd, then we have what we'll see soon is a point of inflection.

(13) To determine absolute maxima/minima, the candidates are: points of discontinuity, boundary points of domain (whether included in domain or outside the domain; if the latter, then limiting), critical points (derivative zero or undefined), and limiting cases at $\pm\infty$.

(14) To determine absolute maxima and absolute minima, find all candidates (discontinuity, endpoints, limiting cases, boundary points), evaluate at each, and compare. Note that any absolute maximum must arise as a local or endpoint maximum. However, instead of first determining which critical points give local maxima by the derivative tests, we can straightaway compute values everywhere and compare, if our interest is solely in finding the absolute maximum and minimum.

Actions... (think of examples that you've done)

(1) Rolle's theorem, along with the more sophisticated formulations involving increasing/decreasing, tell us that there is an intimate relationship between the zeros of a function and the zeros of its derivative. Specifically, between any two zeros of the function, there is a zero of its derivative. Thus, if a function has r zeros, the derivative has at least $r - 1$ zeros, with at least one zero between any two consecutive zeros of f.

(2) The more sophisticated version tells us that between any two zeros of a differentiable function, the function must attain a local maximum or local minimum. So, if the function is increasing everywhere or decreasing everywhere, there is at most one zero.

(3) The mean-value theorem allows us to use bounds on the derivative of a function to bound the overall variation, or change, in the function. This is because if the derivative cannot exceed some value, then the difference quotient also cannot exceed that value, which means that the function cannot change too quickly on average.

(4) To determine regions where a function is increasing and decreasing, we find the derivative and determine regions where the derivative is positive, zero, and negative.

(5) To determine all the local maxima and local minima of a function, find all the critical points. To find the critical points, solve $f' = 0$ and also consider, as possible candidates, all the points where the function changes definition. *Although a point where the function changes definition need not be a critical point, it is a very likely candidate.*

1. Rolle's theorem and mean-value theorem

1.1. **Rolle's theorem.** Rolle's theorem states that if f is a function defined on a closed interval $[a, b]$ such that the following three conditions hold: (i) f is continuous on $[a, b]$ (ii) f is differentiable on the open interval (a, b) (iii) $f(a) = f(b) = 0$, then there exists $c \in (a, b)$ such that $f'(c) = 0$. (It turns out that the condition that both $f(a)$ and $f(b)$ be equal to *zero* is not necessary – we can weaken it to simply requiring that $f(a)$ equal $f(b)$. The version stated in the book requires them both to be zero).

Now, I'll give you a rough sketch of the proof of Rolle's theorem. One possibility is that f is a constant function, in which case $f'(c) = 0$ for all $c \in (a, b)$. If f is nonconstant, then by the extreme-value theorem, f is either bigger than zero somewhere or smaller than zero somewhere. Assume the former – a similar proof applies for the latter assumption. In this case, f attains a maximum at some point in (a, b). At this point, if we try to calculate the left-hand derivative, we see that the left-hand derivative is greater than or equal to zero. And if we try to calculate the right-hand derivative, we see that the right-hand derivative is less than or equal to zero. Because the function is differentiable at the point, both the left-hand derivative and the right-hand derivative must be equal, which means that they must both be equal to zero.

Now, the crucial point here is understanding *why* the derivative of a function should be zero at a point where it is maximum. And this is very important for some of the stuff we'll be seeing in the near future. So let's understand more clearly what's happening.

At a point c where f attains a maximum, two things are happening. First, $f(c) \geq f(x)$ for $x < c$. This forces that the difference quotient that we form between c and x for any $x < c$ is nonnegative. Hence, the left-hand derivative is the limit of some expression that is nonnegative, so the left-hand derivative itself is nonnegative.

What happens to the right-hand derivative? Well, in this case, $f(x) - f(c)$ is zero or negative, but $x - c$ is positive, so the difference quotient is nonpositive, so the limit, which is the right-hand derivative, is nonpositive. So we have a situation where the left-hand derivative is nonnegative (i.e., positive or zero) and the right-hand derivative is nonpositive (i.e., negative or zero).

Now, for the function to be differentiable, the left-hand derivative and right-hand derivative must be equal, so the derivative must be equal to zero.

Here's another way of thinking about this. Up until the point where the maximum is achieved, the function must be, at least roughly speaking, going up. (This is not correct strictly speaking, but is useful at least in simple cases). And then, immediately after that point, the function must be, at least roughly speaking, going down. So, at that point, it changes from a *going up* to a *going down* function, hence at the point it is oing neither up nor down, so the derivative is zero.

Why is differentiability so important? Well, think of what might happen for a function that has one-sided derivatives but isn't differentiable. In that case, it could have a *sharp peak* – it increases in a straight line, and then takes a turn and starts decreasing in a straight line.

Also, note that differentiability *at the endpoints* is not necessary. So, Rolle's theorem applies for instance to the function $f(x) := \sqrt{1 - x^2}$ on the interval $[-1, 1]$, even though that function is not differentiable *at* the two points -1 and 1.

1.2. **Mean-value theorem.** Here's what the mean-value theorem states. It states that if f is a function that is continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b), then there exists $c \in (a, b)$ such that:

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

In other words, given any two points such that the function is continuous on the closed interval between those two points and differentiable on the open interval, then there is a point in the open interval at which the derivative at the interior point equals the difference quotient between the two endpoints.

In other words, there is a point on the graph in between these two points such that the tangent line at the point is *parallel* to the secant line, or chord, joining the points $(a, f(a))$ and $(b, f(b))$.

Note that this result has a somewhat similar flavor to the intermediate-value theorem, but it is a different result.

Please see the book for a description of how the mean-value theorem can be derived from Rolle's theorem.

Aside: The mean-value theorem can be used to prove Darboux's theorem. Recall that the derivative of a differentiable function on an interval need not be continuous on that same interval. However, it comes very close to being continuous if it is defined everywhere on the interval. Specifically, the derivative satisfies the intermediate value property, and hence all its discontinuities must be of the oscillatory kind.

This result is called Darboux's theorem. Although the result is not part of the syllabus, it can be deduced with a little bit of work from the mean-value theorem.

In fact, there are many similar results about derivatives that would become easy if we assumed that the derivative is continuous, but are true even in general, and the proofs of most of these results relies on the mean-value theorem. Since we're not focused on proving theorems, we will not be talking a lot about the mean-value theorem explicitly, but you should keep in mind that it is at the back of a lot of what we do.

2. Local increase and decrease behavior

We will now try to understand, very clearly, the relationship between the *sign of the derivative* and the *behavior of the function near a point.*

2.1. **Larger than stuff on the left.**

Suppose c is a point and $a < c$ such that $f(x) \leq f(c)$ for all $x \in (a, c)$. In other words, c is a *local maximum from the left.* What do I mean by that? I mean that $f(c)$ is larger than or equal to f of the stuff on the *immediate* left of it. That doesn't mean that $f(c)$ is a maximum over the entire domain of f – it just means it is greater than or equal to stuff on the immediate left.

Now, we claim that, if the left-hand derivative of f at c exists, then it is greater than or equal to 0. How do we work that out? The left-hand derivative is the limit of the difference quotient:

$$\frac{f(x) - f(c)}{x - c}$$

where $x \to c^-$. Note that for x close enough to c, (i.e., $a < x < c$), the numerator is negative or zero, and the denominator is negative, so the difference quotient is zero or positive. Thus, the limit of this, if it exists, is zero or positive.

There are three other cases. Let's just summarize the four cases:

(1) If c is a point that is a local maximum from the left for f, then the left-hand derivative of f at c, if it exists, is zero or positive.
(2) If c is a point that is a local maximum from the right for f, then the right-hand derivative of f at c, if it exists, is zero or negative.
(3) If c is a point that is a local minimum from the left for f, then the left-hand derivative of f at c, if it exists, is zero or negative.
(4) If c is a point that is a local minimum from the right for f, then the right-hand derivative of f at c, if it exists, is zero or positive.

2.2. **Strict maxima and minima.**

We said that for a function f, a point c is a *local maximum from the left* if there exists $a < c$ such that $f(x) \leq f(c)$ for all $x \in (a, c)$. Now, this definition also includes the possibility that the function is constant just before c.

A related notion is that of a *strict local maximum from the left*, which means that there exists $a < c$ such that $f(x) < f(c)$ for all $x \in (a, c)$. In other words, $f(c)$ is *strictly bigger* than $f(x)$ for x to the immediate left of c.

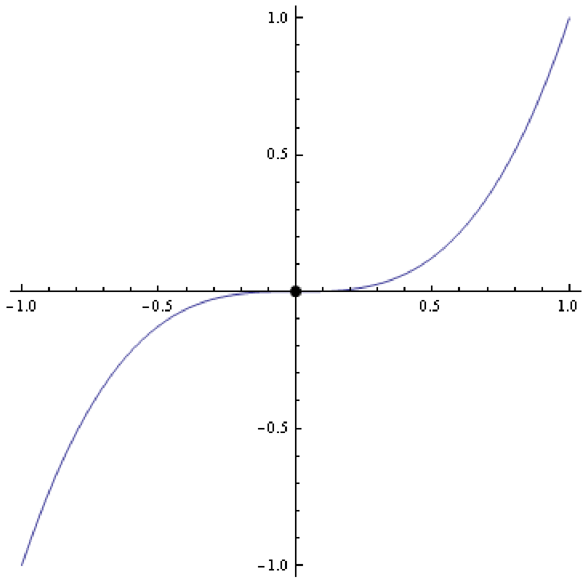
Similarly, we can define the notions of strict local maximum from the right, strict local minimum from the left, and strict local minimum from the right.

2.3. **Does strict maximum/minimum from the left/right tell us more?**

Recall that if c is a point that is a local maximum from the left for f, then the left-hand derivative of f at c, if it exists, is greater than or equal to zero. What if c is a point that is a strict local maximum from the left for f? Can we say something more about the left-hand derivative of f at c?

The first thing you might intuitively expect is that that left-hand derivative of f at c should now not just be greater than or equal to zero, it should be strictly greater than zero. But you would be wrong.

It *is* true that if c is a strict local maximum from the left for f, then the difference quotients, as $x \to c^-$, are all positive. However, the *limit* of these difference quotients could still be zero. Another way of thinking about this is that even if the function is increasing up to the point c, it may happen that the rate of increase is leveling off to 0. An example is the function x^3 at the point 0: 0 is a strict local maximum from the left, but the derivative at 0 is 0. Here's a picture:



Later, we will understand this situation more carefully and it will turn out that we are dealing (in this case) with what is called a *point of inflection*.

2.4. **Minimum, maximum from both sides.** So we have some sign information about the derivative closely related to how the function at the point compares with the value of the function at nearby points. Maximum from the left means left-hand derivative is nonnegative, maximum from the right means right-hand derivative is nonpositive, minimum from the left means left-hand derivative is nonpositive, minimum from the right means right-hand derivative is nonnegative.

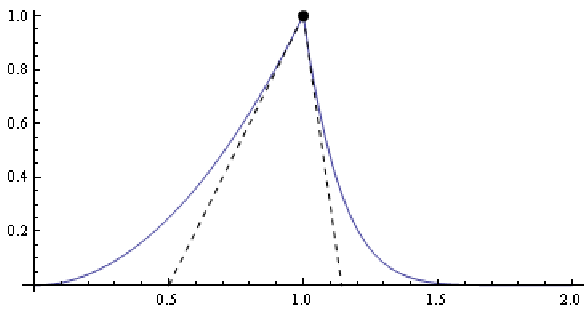
So, let's piece these together:

(1) A *local maximum* for the function f is a point c such that $f(c)$ is the maximum possible value for $f(x)$ in an open interval containing c. Thus, a point of local maximum for f is a point that is both a local maximum from the left and a local maximum from the right. A *strict local maximum* for the function f is a point c such that $f(c)$ is strictly greater than $f(x)$ for all x in some open interval containing c.
(2) A *local minimum* for the function f is a point c such that $f(c)$ is the minimum possible value for $f(x)$ in an open interval containing c. Thus, a point of local minimum for f is a point that is both a local minimum from the left and a local minimum from the right. A *strict local minimum* for the function f is a point c such that $f(c)$ is strictly smaller than $f(x)$ for all x in some open interval containing c.

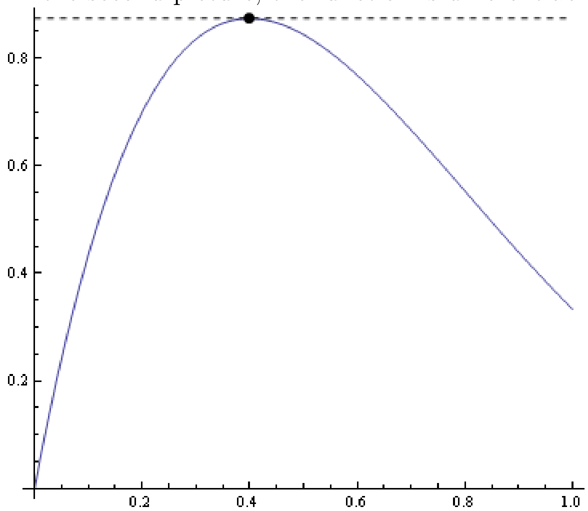
What can we say about local maxima and local minima? We can say the following:

(1) At a local maximum, the left-hand derivative (if it exists) is greater than or equal to zero, and the right-hand derivative (if it exists) is less than or equal to zero. Thus, *if* the derivative exists at a point of local maximum, it *equals zero*. The same applies to strict local maxima.
(2) At a local minimum, the left-hand derivative (if it exists) is less than or equal to zero, and the right-hand derivative (if it exists) is greater than or equal to zero. Thus, *if* the derivative exists at a point of local minimum, it *equals zero*. The same applies to strict local minima.

Below are two pictures depicting points of local maximum. In the first picture, the left-hand derivative is positive, the right-hand derivative is negative, and the function is not differentiable at the point of local maximum.



In the second picture, the function is differentiable, and the derivative is zero.



2.5. **Maximum from the left, minimum from the right.** Suppose c is a point such that it is a local maximum from the left for f and is a local minimum from the right for f. This means that $f(c)$ is greater than or equal to $f(x)$ for x to the immediate left of c, and $f(c)$ is less than or equal to $f(x)$ for x to the immediate right of c. In this case, we say that f is *non-decreasing* at the point c.

In other words, f at c is bigger than or equal to what it is on the left and smaller than or equal to what it is on the right. Well, in this case, the left-hand derivative is greater than or equal to zero and the right-hand derivative is greater than or equal to zero. Thus, if $f'(c)$ exists, we have $f'(c) \geq 0$.

Now consider the case where c is a point that is a local minimum from the left for f and is a local maximum from the right for f. This means that $f(c)$ is less than or equal to $f(x)$ for x to the immediate left of c and greater than or equal to $f(x)$ for x to the immediate right of c. In this case, we say that f is *non-increasing* at the point c.

In other words, f at c is smaller than what it is on the right and larger than what it is on the left. Well, in this case, the left-hand derivative is less than or equal to zero and the right-hand derivative is less than or equal to zero. Thus, if $f'(c)$ exists, we have $f'(c) \leq 0$.

2.6. **Introducing strictness.** We said that f is *non-decreasing* at the point c if $f(c) \geq f(x)$ for x just to the left of c and $f(c) \leq f(x)$ for x just to the right of c. We now consider the *strict* version of this concept. We say that f is *increasing* at the point c if there is an open interval (a, b) containing c such that, for $x \in (a, b)$, $f(x) < f(c)$ if $x < c$ and $f(x) > f(c)$ if $x > c$. In other words, c is a strict local maximum from the left and a strict local minimum from the right.

Well, what can we say about the derivative at a point where the function is increasing, rather than just non-decreasing? We already know that $f'(c)$, if it exists, is greater than or equal to zero, but we might hope to say that the derivative $f'(c)$ is strictly greater than zero. Unfortunately, that is not true.

In other words, a function could be increasing at the point c, in the sense that it is strictly increasing, but still have derivative 0. For instance, consider the function $f(x) := x^3$. This is increasing everywhere, but at the point zero, its derivative is zero.

How can a function be increasing at a point even though its derivative is zero? Well, what happens is that the derivative was positive before the point, is positive just after the point, and becomes zero just momentarily. Alternatively, if you think in terms of the derivative as a limit of difference quotients, all the difference quotients are positive, but the limit is still zero because they get smaller and smaller in magnitude as you come closer and closer to the point. Another way of thinking of this is that you reduce your car's speed to zero for the split second that you cross the STOP line, so as to cimply with the letter of the law without actually stopping for any interval of time.

Similarly, we can define the notion of a function f being *decreasing* at a point c. This means that $f(c) < f(x)$ for x to the immediate left of c and $f(c) > f(x)$ for x to the immediate right of c. As in the previous case, we can deduce that $f'(c)$, if it exists, is less than or equal to zero, but it could very well happen that $f'(c) = 0$. An example is $f(x) := -x^3$, at the point $x = 0$.

2.7. **Increasing functions and sign of derivative.** Here's what we did. We first did separate analyses for what we can conclude about the left-hand derivative and the right-hand derivative of a function based on how the value of the function at the point compares with the value of the function at points to its immediate left. We used this to come to some conclusions about the nature of the derivative of a function (if it exists) at points of local maxima, local minima, and points where the function is nondecreasing and nonincreasing. Let's now discuss a converse result.

So far, we have used information about the nature of changes of the function to deduce information about the sign of the derivative. Now, we want to go the other way around: use information about the sign of the derivative to deduce information about the behavior of the function. And this is particularly useful because now that we have a huge toolkit, we can differentiate practically any function that we can write down. This means that even for functions that we have no idea how to visualize, we can formally differentiate them and work with the derivative. Thus, if we can relate information about the derivative to information about the function, we are in good shape.

Remember what we said: if a function is increasing, it is nondecreasing, and if it is nondecreasing, then the derivative is greater than or equal to zero. Now, a converse for this would mean some condition on the derivative telling us whether the function is increasing.

Unfortunately, the derivative being zero is very inconclusive. The function could be constant, it could be a local maximum, it could be a local minimum, it could be increasing, or it could be decreasing. However, it turns out that if the derivative is *strictly* positive, then we can conclude that the function is increasing.

Specifically, we have the following chain of implications for a function f defined around a point c and differentiable at c:

$f'(c) > 0 \implies f$ is increasing at $c \implies f$ is nondecreasing at $c \implies f'(c) \geq 0$

And each of these implications is strict, in the sense that you cannot proceed backwards with any of them, becausethere are counterexamples to each possible reverse implication.

Similarly, for a function f defined around a point c and differentiable at c:

$f'(c) < 0 \implies f$ is decreasing at $c \implies f$ is nonincreasing at $c \implies f'(c) \leq 0$

2.8. **Increasing and decreasing functions.** A function f is said to be increasing an an interval I (which may be open, closed, half-open, half-closed, or stretching to infinity) if for any $x_1 < x_2$, with both x_1 and x_2 in I, we have $f(x_1) < f(x_2)$. In other words, the larger the input, the larger the output.

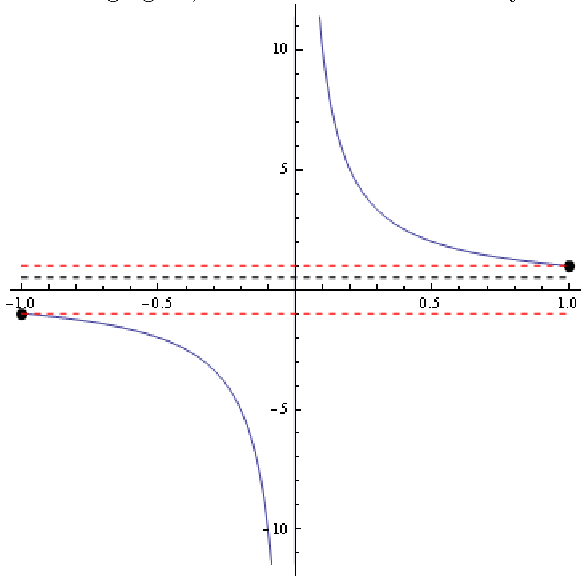
A little while ago, we talked of the notion of a function that is increasing at a point, and that was basically something similar, except that there one of the comparison points was fixed and the other one was restricted to somewhere close by. For a function to be increasing on an interval means that it is increasing at every point in the interior of the interval. If the interval has endpoints, then the function attains a strict local minimum at the left endpoint and a strict local maximum at the right endpoint.

Similarly, we say that f is *decreasing* on an interval I if, for any $x_1, x_2 \in I$, with $x_1 < x_2$, we have $f(x_1) > f(x_2)$. In other words, the larger the input, the smaller the output.

When I do not specify the interval and simply say that a function is increasing (respectively, decreasing), I mean that the function is increasing (respectively, decreasing) over its entire domain. For functions whose domain is the set of all real numbers, this means that the function is increasing (respectively, decreasing) over the set of all real numbers.

An example of an increasing function is a function $f(x) := ax + b$ with $a > 0$. An example of a decreasing function is a function $f(x) := ax + b$ with $a < 0$.

By the way, here's an interesting and weird example. Consider the function $f(x) := 1/x$. This function is not defined at $x = 0$. So, its domain is a union of two disjoint open intervals: the interval $(-\infty, 0)$ and the interval $(0, \infty)$. Now, we see that on each of these intervals, the function is decreasing. In fact, on the interval $(-\infty, 0)$, the function starts out from something close to 0 and then becomes more and more negative, approaching $-\infty$ as x tends to zero from the left. And then, on the interval $(0, \infty)$, the function is decreasing again, down from $+\infty$ all the way to zero.



But, taken together, is the function decreasing? No, and the reason is that at the point 0, where the function is undefined, it is undergoing this *huge* shift – from $-\infty$ to ∞. This fact – that points where the function is undefined can be points where it jumps from $-\infty$ to $+\infty$ or $+\infty$ to $-\infty$ – is a fact that keeps coming up. If you remember, this same fact haunted us when we were trying to apply the intermediate-value theorem to the function $1/x$ on an interval containing 0.

2.9. **The derivative sign condition for increasing/decreasing.** We first state the result for open intervals, where it is fairly straightforward. Suppose f is a function defined on an open interval (a, b). Suppose, further, that f is continuous and differentiable on (a, b), and for every point $x \in (a, b)$, $f'(x) > 0$. Then, f is an increasing function on (a, b).

A similar statement for decreasing: If f is a function defined on an interval (a, b). Suppose, further, that f is continuous and differentiable on (a, b), and for every point $x \in (a, b)$, $f'(x) < 0$. Then, f is a decreasing function on (a, b).

The result also holds for open intervals that stretch to ∞ or $-\infty$.

Note that it is important that f should be defined for all values in the interval (a, b), that it should be continuous on the interval, and that it should be differentiable on the interval. Here are some counterexamples:

(1) Consider the function $f(x) := 1/x$, defined and differentiable for $x \neq 0$. Its derivative is $f'(x) := -1/x^2$, which is negative wherever defined. Hence, f is decreasing on any open interval not containing 0. However, it is *not* decreasing on any open interval containing 0.
(2) Consider the function $f(x) := \tan x$. The derivative of the function is $f'(x) := \sec^2 x$. Note that f is defined for all x that are not odd multiples of $\pi/2$, and the same holds for f'. Also, note that $f'(x) > 0$ wherever defined, because $|\sec x| \geq 1$ wherever defined. Thus, the tan function is increasing on any interval not containing an odd multiple of $\pi/2$. But at each odd multiple of $\pi/2$, it slips from $+\infty$ to $-\infty$.

Let us now look at the version for a closed interval.

Suppose f is a function defined on a closed interval $[a, b]$, which is continuous on $[a, b]$ and differentiable on (a, b). Then, if $f'(x) > 0$ for $x \in (a, b)$, then f is increasing on all of $[a, b]$. Similarly, if $f'(x) < 0$ for $x \in (a, b)$, then f is decreasing on all of $[a, b]$.

In other words, we do *not* need to impose conditions on one-sided derivatives at the endpoints in order to guarantee that the function is increasing on the entire closed interval.

Finally, if $f'(x) = 0$ on the interval (a, b), then f is constant on $[a, b]$.

Some other versions:

(1) The result also applies to half-closed, half-open intervals. So, it may happen that a function f is continuous on $[a, b)$, differentiable on (a, b), and $f'(x) > 0$ for $x \in (a, b)$. In this case, f is increasing on $[a, b)$.
(2) The result also applies to intervals that stretch to infinity in either or both directions.

2.10. **Finding where a function is increasing and decreasing.** Let's consider a function f that, for simplicity, is continuously differentiable on its domain. So, f' is a continuous function. We now note that, in order to find out where f is increasing and decreasing, we need to find out where f' is positive, negative and zero.

Here's an example, Consider the function $f(x) := x^3 - 3x^2 - 9x + 7$. Where is f increasing and where is it decreasing? In order to find out, we need to differentiate f. The function $f'(x)$ is equal to $3x^2 - 6x - 9 = 3(x - 3)(x + 1)$. By the usual methods, we know that f' is positive on $(-\infty, -1) \cup (3, \infty)$, negative on $(-1, 3)$, and zero at -1 and 3. Thus, the function f is increasing on the intervals $(-\infty, -1]$ and $[3, \infty)$ and decreasing on the interval $[-1, 3]$.

Note that it is *not* correct to conclude from the above that f is increasing on the set $(-\infty, -1] \cup [3, \infty)$, although it is increasing on each of the intervals $(-\infty, -1]$ and $[3, \infty)$ *separately*. This is because the two pieces $(-\infty, -1]$ and $[3, \infty)$ are in some sense independent of each other. In general, the positive derivative implies increasing conclusions hold on intervals because they are what mathematicians call *connected sets*, and not for disjoint unions of intervals. In the case of this specific function, we note that $f(-1) = 12$ and $f(3) = -20$, so the value of the function at the point 3 is smaller than it is at -1. Thus, it is not correct to think of the function as being increasing on the union of the two intervals.

Similarly, if f is a rational function $x^2/(x^3 - 1)$, then we get $f'(x) = (-2x - x^4)/(x^3 - 1)^2$. Now, in order to find out where this is positive and where this is negative, we need to factor the numerator and the denominator. The factorization is:

$$\frac{-x(x + 2^{1/3})(x^2 - 2^{1/3}x + 2^{2/3})}{(x - 1)^2(x^2 + x + 1)^2}$$

The zeros of the numerator are 0 and $-2^{1/3}$ and the zero of the denominator is 1. The quadratic factors in both the numerator and the denominator are always positive. Also note that there is a minus sign on the outside.

Hence, f' is negative on $(1, \infty)$, $(0, 1)$, and $(-\infty, -2^{1/3})$, positive on $(-2^{1/3}, 0)$, zero on 0 and $-2^{1/3}$, and undefined at 1. Thus, f is decreasing on $[0, 1)$, $(1, \infty)$, and $(-\infty, -2^{1/3}]$, increasing on $[-2^{1/3}, 0]$.

Now, let's combine this with the information we have about f itself. Note that f is undefined at 1, it is positive on $(1, \infty)$, it is zero at 0, and it is negative on $(-\infty, 0) \cup (0, 1)$. How do we combine this with information about what's happening with the derivative?

On the interval $(-\infty, -2^{1/3})$, f is negative *and* decreasing. What's happening as $x \to -\infty$? f tends to zero (we'll see why a little later). So, as x goes from $-\infty$ to $-2^{1/3}$, f goes down from 0 to $-2^{2/3}/3$. Then, as x goes from $-2^{1/3}$ to 0, f is still negative but starts going up from $-2^{2/3}/3$ and reaches 0. In the interval from 0 to 1, f goes back in the negative direction, from 0 down to $-\infty$. Then, in the interval $(1, \infty)$, f goes emerges from $+\infty$ and goes down to 0 as $x \to +\infty$.

So we see that information about the sign of the derivative helps us get a better picture of how the function behaves, and allows us to better draw the graph of the function – something that we will try to do more of a short while from now.

Point-value distinction. We use the term *point of local maximum* or *point of local minimum* (or simply *local maximum* or *local minimum*) for the point in the domain, and the term *local maximum value* for the value of the function at the point.

3. Determining local extreme values

3.1. **Local extreme values and critical points.** If f is a function and c is a point in the interior of the domain of f, then f is said to have a *local maximum* at c if $f(x) \leq f(c)$ for all x sufficiently close to c. Here, *sufficiently close* means that there exists $a < c$ and $b > c$ such that the statement holds for all $x \in (a, b)$.

Similarly, we have the concept of *local minimum* at c.

The points in the domain at which local maxima and local minima occur are termed the *points of local extrema* and the values of the function at these points are termed the *local extreme values*.

As we discussed last time, if f is differentiable at a point c of local maximum or local minimum, the derivative of f at c is zero. This suggests that we define a notion.

An interior point c in the domain of a function f is termed a *critical point* if either $f'(c) = 0$ or $f'(c)$ does not exist. Thus, all the local extreme values occur at critical points – because at a local maximum or minimum, either the derivative does not exist, or the derivative equals zero.

Note that not all critical points are points of local maxima and minima. For instance, for the function $f(x) := x^3$, the point $x = 0$ is a critical point, but the function does not attain a local maximum or local minimum at that point. However, critical points give us a small set of points that we need to check against. Once we have this small set, we can use other methods to determine what precisely is happening at these points.

3.2. **First-derivative test.** The first-derivative test basically tries to determine whether something is a local maximum by looking, not just at the value of the derivative *at* the point, but also the value of the derivative *close* to the point.

Basically, we want to combine the idea of *increasing on the left, decreasing on the right* to show that something is a local maximum, and similarly, we combine the idea of *decreasing on the left, increasing on the right* to show that something is a local minimum.

The first-derivative test says that if c is a critical point for f and f is continuous at c (Note that f need not be differentiable at c). if there is a positive number δ such that:

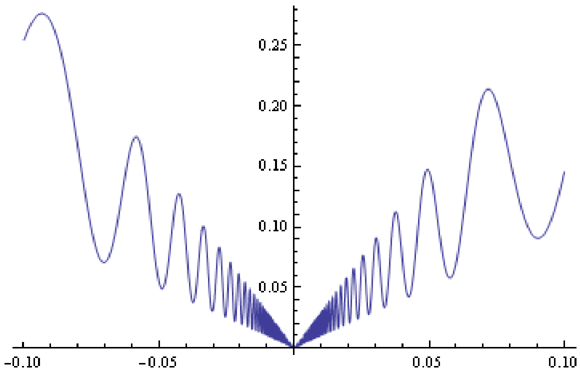
(1) $f'(x) > 0$ for all $x \in (c - \delta, c)$ and $f'(x) < 0$ for all $x \in (c, c + \delta)$, then $f(c)$ is a local maximum, i.e., c is a point of local maximum for f.
(2) $f'(x) < 0$ for all $x \in (c - \delta, c)$ and $f'(x) > 0$ for all $x \in (c, c + \delta)$, then $f(c)$ is a local minimum, i.e., c is a point of local minimum for f.
(3) $f'(x)$ keeps constant sign on $(c - \delta, c) \cup (c, c + \delta)$, then c is not a point of local maximum/minimum for f.

Thus, for the function $f(x) := x^2/(x^3 - 1)$, there is a local minimum at $-2^{1/3}$ and a local maximum at 0.

Recall that for the function $f(x) := x^3$, the derivative at zero is zero, so it is a critical point but it is not a point of local extremum, because the derivative is positive everywhere else.

3.3. **What are we essentially doing with the first-derivative test?** Why does the first-derivative test work? Essentially it is an application of the results on increasing and decreasing functions for closed intervals. What we're doing is using the information about the derivative from the left to conclude that the point is a strict local maximum from the left, because the function is increasing up to the point, and is a strict local maximum from the right, because the function is decreasing down from the point.

3.4. **The first-derivative test is sufficient but not necessary.** For most of the function that you'll see, the first-derivative test will give you a good way of figuring out whether a given critical point is a local maximum or local minimum. There are, however, situations where the first-derivative test fails to work. These are situations where the derivative changes sign infinitely often, close to the critical point, so does not have a constant sign near the critical point. For instance, consider the function $f(x) := |x|(2 + \sin(1/x))$. This attains a local minimum at the point $x = 0$, which is a critical point. However, the derivative of the function oscillates between the positive and negative sign close to zero and doesn't settle into a single sign on either side of zero.



3.5. **Second-derivative test.** One problem with the first-derivative test is that it requires us to make two local sign computations over *intervals*, rather than *at points*. Discussed here is a variant of the first-derivative test, called the second-derivative test, that is sometimes easier to use.

Suppose c is a critical point in the interior of the domain of a function f, and f is twice differentiable at c. Then, if $f''(c) > 0$, c is a point of local minimum, whereas if $f''(c) < 0$, then c is a point of local minimum.

The way this works is as follows: if $f''(c) > 0$, that means that f' is (strictly) increasing at c. This means that f' is negative to the immediate left of c and is positive to the immediate right of c. Thus, f attains a local minimum at c.

Note that the second-derivative test works for critical points where the function is twice-differentiable. In particular, it does not work for the kind of sharp peak points where the function suddenly changes direction. On the other hand, since the second-derivative test involves evaluation of the second derivative at only one point, it may be easier to apply in certain situations than the first-derivative test, which requires reasoning about the sign of a function over an interval.

4. Finding maxima and minima: a global perspective

4.1. **Endpoint maxima and minima.** An *endpoint maximum* is something like a local maximum, except that it occurs at the endpoint of the domain, so the value of the function at the point needs to be compared only with the values of the function at points sufficiently close to it on one side (the side that the domain is in). Similarly, an *endpoint minimum* is like a local minimum, except that it occurs at the endpoint of the domain, so the value of the function at the point needs to be compared only with the values of the function at points sufficiently close to it on one side.

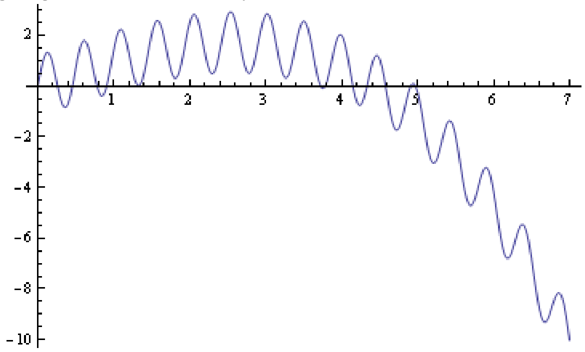
If the endpoint is a left endpoint, then being an endpoint maximum (respectively, minimum) means being a local maximum (respectively, minimum) from the right. If the endpoint is a right endpoint, then being an endpoint maximum (respectively, minimum) means being a local maximum (respectively, minimum) from the left.

4.2. **Absolute maxima and minima.** We say that a function f has an absolute maximum at a point d in the domain if $f(d) \geq f(x)$ for all x in the domain. We say that f has an absolute minimum at a point d in the domain if $f(d) \leq f(x)$ for all x in the domain. The corresponding value $f(d)$ is termed the absolute maximum (respectively, minimum) of f on its domain.

Notice the following very important fact about absolute maxima and minima, which distinguishes them from local maxima and minima. If an absolute maximum value exists, then the value is unique, even though it may be attained at multiple points on the domain. Similarly, if an absolute minimum value exists, then the value is unique, even though it may be attained at multiple points of the domain. Further, assuming the function to be continuous through the domain, and assuming the domain to be connected (i.e., not fragmented into intervals) the range of the function is the interval between the absolute minimum value and the absolute maximum value. This follows from the intermediate value theorem.

For instance, for the cos function, absolute maxima occur at multiples of 2π and absolute minima occur at odd multiples of π. The absolute maximum value is 1 and the absolute minimum value is -1.

Just for fun, here's a picture of a function having lots of local maxima and minima, but all at different levels. Note that some of the local maximum values are less than some of the local minimum values. This highlights the extremely local nature of local maxima/minima.



4.3. **Where and when do absolute maxima/minima exist?** Recall the *extreme value theorem* from some time ago. It said that for a continuous function on a closed interval, the function attains its maximum and minimum. This was basically asserting the existence of absolute maxima and minima for a continuous function on a closed interval.

Notice that any point of absolute maximum (respectively, minimum) is either an endpoint or is a point of local maximum (respectively, minimum). We further know that any point of local maximum or minimum is a critical point. Thus, in order to find all the absolute maxima and minima, a good first step is to find critical points and endpoints.

Another thing needs to be noted. For some funny functions, it turns out that there is no maximum or minimum. This could happen for two reasons: first, the function approaches $+\infty$ or $-\infty$, i.e., it gets arbitrarily large in one direction, somewhere. Second, it might happen that the function approaches some maximum value but does not attain it on the domain. For instance, the function $f(x) = x$ on the interval $(0, 1)$ does not attain a maximum or minimum, since these occur at the endpoints, which by design are not included in the domain.

Thus, the absolute maxima and minima, *if they occur*, occur at critical points and endpoints. But we need to further tackle the question of existence. In order to deal with this issue clearly, we need to face up to something we have avoided so far: limits to infinity.

4.4. **Limits at infinity and to infinity.** We need to tackle two questions: first, what do we mean by trying to evaluate $\lim_{x\to\infty} f(x)$ and $\lim_{x\to-\infty} f(x)$, and second, what do we mean by saying $\lim_{x\to c} f(x) = \infty$ and $\lim_{x\to c} f(x) = -\infty$.

For both, the basic idea is that for something to approach $+\infty$ means that it eventually crosses every arbitrarily large number and does not come back down, while approacing $-\infty$ means that it eventually crosses below every arbitrarily small negative number and does not come back. Formally, we say that:

(1) $\lim_{x\to c} f(x) = +\infty$ if, for every $N > 0$, there exists $\delta > 0$ such that if $0 < |x - c| < \delta$, $f(x) > N$.
(2) $\lim_{x\to c} f(x) = -\infty$ if, for every $N > 0$, there exists $\delta > 0$ such that if $0 < |x - c| < \delta$, $f(x) < -N$.
(3) $\lim_{x\to\infty} f(x) = L$ if, for every $\epsilon > 0$, there exists $N > 0$ such that for $x > N$, $|f(x) - L| < \epsilon$.
(4) $\lim_{x\to-\infty} f(x) = L$ if, for every $\epsilon > 0$, there exists $N > 0$ such that for $x < -N$, $|f(x) - L| < \epsilon$.
(5) $lim_{x\to\infty} f(x) = \infty$ if, for every $N > 0$, there exists $M > 0$ such that if $x > M$, then $f(x) > N$.
(6) $\lim_{x\to\infty} f(x) = -\infty$ if, for every $N > 0$, there exists $M > 0$ such that if $x > M$, then $f(x) < -N$.
(7) $\lim_{x\to-\infty} f(x) = \infty$ if, for every $N > 0$, there exists $M > 0$ such that if $x < -M$, then $f(x) > N$.
(8) $\lim_{x\to-\infty} f(x) = -\infty$ if, for every $N > 0$, there exists $M > 0$ such that if $x < -M$, then $f(x) < -N$.

We will consider limits to infinity in much more detail in 153.

4.5. **Rules of thumb for figuring out limits at infinity.** Here are some rules. Note that each rule also applies to one-sided limits, and often has to be applied in a one-sided sense because the signs of infinity being approached from the two sides may be different:

(1) If the numerator approaches a positive number and the denominator approaches zero from the positive side, then the quotient approaches $+\infty$.

(2) If the numerator approaches a negative number and the denominator approaches zero from the positive side, the quotient approaches $-\infty$.
(3) If the numerator approaches a positive number and the denominator approaches zero from the negative side, the quotient approaches $-\infty$.
(4) If the numerator approaches a negative number and the denominator approaches zero from the negative side, the quotient approaches $+\infty$.

For instance, for the function $f(x) := 1/x^2$, the numerator approaches (in fact, equals) a positive number, and the denominator approaches zero from the positive side, so the limit at 0 is $+\infty$.

For the function $g(x) := 1/x$, as $x \to 0^-$, the numerator is positive and the denominator approaches zero from the negative side, so the quotient approaches $-\infty$. As $x \to 0^+$, the numerator is positive and the denominator approaches zero from the positive side, so the quotient approaches $+\infty$.

Let's use these ideas to revisit our example of the rational function $x^2/(x^3 - 1)$. Recall that here the only point where the function is not defined is $x = 1$. Here, the numerator approaches a positive number (1). As $x \to 1^-$, the denominator approaches 0 from the left, so the quotient approaches $-\infty$, and as $x \to 1^+$, the denominator approaches 0 from the right side, so the quotient approaches $+\infty$. Thus, the left-hand limit is $-\infty$ and the right-hand limit is $+\infty$. Note that, when the limits are $\pm\infty$, we are still allowed to say, and should say, that the limits do not exist. Infinite does not exist.

Next, we look at rules of thumb that guide us when $x \to \infty$. Here are some of these rules:

(1) For a polynomial p of degree one or higher, $p(x) \to \infty$ as $x \to \infty$ if the leading coefficient of p is positive, and $p(x) \to -\infty$ as $x \to \infty$ if the leading coefficient of p is negative.
(2) For a polynomial p of degree one or higher, $p(x) \to \infty$ as $x \to -\infty$ if the leading coefficient of p is positive and the degree of p is even, and $p(x) \to -\infty$ as $x \to -\infty$ if the leading coefficient of p is positive and the degree of p is odd. When the leading coefficient of p is negative, the signs get reversed.
(3) For a rational function, the limits as $x \to \pm\infty$ can be computed by simply looking at the limit of the quotient of the leading terms in the numerator and the denominator.
(4) For a rational function, if the degree of the denominator is greater than the degree of the numerator, the value of the rational function approaches 0 as the input goes to ∞ and as the input goes to $-\infty$. In other words, for such a rational function r, $\lim_{x\to\infty} r(x) = \lim_{x\to-\infty} r(x) = 0$.
(5) For a rational function, if the degree of the denominator is smaller than the degree of the numerator, the limit of the rational function, as $x \to \infty$, is the infinity with the same sign as the quotient of the leading coefficients. As $x \to -\infty$, it is the infinity with the same sign as the product of (the quotient of the leading coefficients) and (-1 to the power of the difference of degrees).
(6) For a rational function, if the numerator and the denominator have equal degrees, then the limit as $x \to \infty$, and the limit as $x \to -\infty$, are both equal to the quotient of the leading coefficients.

4.6. **Strategy for computing absolute maxima and minima.** Here's all the candidates we have to deal with:

(1) All endpoints in the domain, and the function values at those endpoints.
(2) All critical points in the domain, and the function values at those critical points.
(3) For points not in the domain but in the boundary of the domain, as well as at $\pm\infty$ (if the domain stretches to $+\infty$ and/or $-\infty$), we try to compute the limits.

Here's what we get, comparing the values:

(1) If, for any of the points where the function isn't defined, or at $\pm\infty$, the limit is $+\infty$, there isn't any absolute maximum. If, for any of the points where the function isn't defined, or at $\pm\infty$, the limit is $-\infty$, there isn't any absolute minimum.
(2) Consider the values of the function at all the critical points, and the limits at $\pm\infty$ and all other points in the boundary of the domain but not in the domain itself. If the maximum of these is attained by one of the critical points, that is the absolute maximum. If the maximum is attained as one of the limits but not at any of the critical points, there is no absolute maximum. Similar remarks apply for minima.

4.7. **Other subtle issues.** Here are some additional issues:

- When there are only finitely many critical points, endpoints, and limit situations, and we need to find the absolute maximum or absolute minimum, we do *not* need to use the derivative tests to figure out which of them are local maxima, which are local minima, and which are neither. We can simply compute the values and compare.
- However, as the picture shown a little earlier indicates, just looking at the values does not immediately tell us whether we have a local maximum, local minimum, or neither. Some lcoal maximum values may be smaller than some local minimum values.
- If there are infinitely many critical points, endpoints, and limit situations, we may need to think a little more clearly about what is happening. It may be helpful to use derivative tests and facts about even, odd, and periodic functions to eliminate or narrow down possibilities.

5. Important fact critical for integration

We noted a little while back that if f is a continuous function on a closed interval $[a, b]$, and its derivative is zero on the open interval (a, b), then f is constant on $[a, b]$.

This fact has an important corollary, which is critical to the whole setup and process of integration:

> If f and g are continuous functions on a closed interval $[a, b]$ and $f'(x) = g'(x)$ for all $x \in (a, b)$, then $f - g$ is a constant function on $[a, b]$.

We will return to this fact and its implications a little later.

6. Piecewise defined functions

We now consider all the above notions for functions defined piecewise on intervals. As usual, we assume that each of the piece functions is nice enough, which in this case means we assume that it is twice continuously differentiable.

For functions defined piecewise, we need to spearately consider all the points where the definition changes. As far as we are concerned, for each point where the definition changes, we have the following possibilities:

- The function is not continuous at this point: In this case, we need to separately consider the left-hand limit and right-hand limit at the point, and the value at that point, and consider all these as candidates for the local extreme values.
- The function is continuous but not differentiable at this point: Then it is a critical point, and the value there might be a candidate for a local extreme value. Whether it is or not depends on the signs of the one-sided derivatives.
- The function is continuously differentiable at the point, and the derivative is zero: Then again, it is a critical point.
- The function is continuously differentiable at the point, and the derivative is nonzeor: Then, it is not a critical point.

We will consider all these in more detail when we study the graphing of functions.

MAX-MIN PROBLEMS

MATH 152, SECTION 55 (VIPUL NAIK)

Corresponding material in the book: Section 4.5

Difficulty level: Moderate to hard. This is material that you have probably seen at the AP level, but it is very important and there will be many additional subtleties that you may have glossed over earlier.

What students should definitely get: The basic procedure for converting a verbal or real-world optimization problem into a mathematical problem seeking absolute maxima and absolute minima, solving that problem, and reinterpreting the solution in real-world terms.

What students should hopefully get: Important facts about area-perimeter optima. The idea that the maximum is determined by the minimum, or most binding, constraint. The intuiton of tangency (as seen in the tapestry problem). The multiple use heuristic. The idea of transforming a function into an equivalent function that is easier to optimize. The procedure for and subtleties in integer optimization. How single-variable optimization fits into the broader optimization context.

Executive summary

Words...

(1) In real-world situations, maximization and minimization problems typically involve multiple variables, multiple constraints on those variables, and some objective function that needs to be maximized or minimized.

(2) The only thing we know to solve such problems is to reduce everything in terms of one variable. This is typically done by *using up* some of the constraints to express the other variables in terms of that variable.

(3) The problem then typically boils down to a maximization/minimization problem of a function in a single variable over an interval. We use the usual techniques for understanding this function, determining the local extreme values, determining the endpoint extreme values, and determining the absolute extreme values.

Actions... (think of examples; also review the notes on max-min problems)

(1) Extremes sometimes occur at endpoints but these endpoints could correspond to degenerate cases. For instance, of all the rectangles with given perimeter, the square has the maximum area, and the minimum occurs in the degenerate case of a rectangle where one side has length zero.

(2) Some constraints on the variables we have are explicitly stated, while others are implicit. Implicit constraints include such things as nonnegativity constraints. *Some of these implicit constraints may be on variables other than the single variable in terms of which we eventually write everything.*

(3) After we have obtained the objective function in terms of one variable, we are in a position to throw out the other variables. However, before doing so, it is *necessary to translate all the constraints into constraints on the one variable that we now have.*

(4) When our intent is to maximize a function, it is sometimes useful to maximize an equivalent function that is easier to visualize or differentiate. For instance, to maximize $\sqrt{f(x)}$ is equivalent to maximizing $f(x)$ if $f(x)$ is nonnegative. With this way of thinking about equivalent functions, we can make sure that the actual function that we differentiate is easy to differentiate. The main criterion is that the two functions should rise and fall together. (Analogous observations apply for minimizing) Remember, however, that to calculate the *value* of the maximum/minimum, you should go back to the original function.

(5) Sometimes, there are other parameters in the maximization/minimization problem that are *unknown constants*, and the final solution is expected to be in terms of those constants. In rare cases, the nature of the function, and hence the nature of maxima and minima, depends on whether those

constants fall in particular intervals. *If you find this to be the case, go back to the original problem and see whether the real-world situation it came from constrains the constants to one of the intervals.*

(6) For some geometrical problems, the maximization/minimization can be done trigonometrically. Here, we make a clever choice of an angle that controls the *shape* of the figure and then use the trigonometric functions of that angle. This could provide alternate insight into maximization.

Smart thoughts for smart people ...

(1) Before getting started on the messy differentiation to find critical points, think about the constraints and the endpoints. Is it obvious that the function will attain a minimum/maximum at one of the endpoints? What are the values of the function at the endpoints? (If no endpoints, take limiting values as you go in one direction of the domain). Is there an intuitive reason to believe that the function attains its optimal value somewhere *in between* rather than at an endpoint? Is there some kind of trade-off to be made? Are there some things that can be said qualitatively about where the trade-off is likely to occur?
(2) Feel free to convert your function to an equivalent function such that the two functions rise and fall together. This reduces the burden of messy expressions.
(3) It is useful to remember the fact that the function $x^p(1-x)^q$ attains a local maximum at $p/(p+q)$. That's because this function appears in disguise all the time (e.g., maximizing area of rectangle with given perimeter, etc.)
(4) A useful idea is that when dividing a resource into two competing uses, and one use is hands-down better than the other, the *best* use happens when the entire resource is devoted to the better use. However, the *worst* may well happen somewhere in between, because divided resources often perform even worse than resources devoted wholeheartedly to a bad use. This is seen in perimeter allocation to boundaries with the objective function being the total area, and area allocation to surfaces with the objective function being the total volume.
(5) When we want to *maximize* something subject to a collection of many constraints, the most relevant constraint is the *minimum* one. Think of the ladder-through-the-hallway problem, or the truck-going-under-bridges problem.

1. Motivation and basic terminology

In the previous lecture, we discussed how to compute points and values of absolute maximum and absolute minimum. Our focus now shifts to using these tools and techniques for real-world (or pseudo-real-world) optimization problems. Because the techniques we have developed are so limited, we will be very selective about the nature of the real-world problems that we pick. Nonetheless, we'll see that even with the modest machinery we have built, we have ways of effectively understanding and tackling many real-world problems.

1.1. Notion of constraints and objective function.

In a typical real-world situation, we usually have multiple things interacting. Many of these items can be measured quantitatively, i.e., they can be measured using real numbers. The values of these real numbers may be subjected to further constraints. Those making decisions may have control over some of the variables. Those making decisions are also tasked with trying to maximize some kind of utility that is dependent on these variables, or minimize some kind of cost function that is dependent on these variables. The task is to *choose the variables subject to constraints in the manner that best maximizes that particular utility function or minimizes that particular cost function.*

For example, think of money management, something that you might be familiar with. You have a certain limited amount of money, and there are various things you want to buy with that money. Each thing that you buy gives you a certain amount of satisfaction; however, for most things, the amount of satisfaction varies with how much you buy it. The question is: how do you allocate money between the many competing things in the market so as to get the best deal for yourself? The number of variables in this case is just the number of different items that you can buy in variable quantities. At a broad level, you may choose to spend A on food, B on clothing, and C on extra books to study calculus. If the total money quantity with you is M, then, assuming that you're not one of those who likes to live on credit, you'll have the constraints $A + B + C \leq M$.

Now, there are going to be three functions f, g, and h, where $f(A)$ is the happiness that spending A on food gives you, $g(B)$ is the happiness that spending B on clothing gives you, and $h(C)$ is the happiness (?)

that spending on extra calculus books gives you. Assuming that happiness is additive, what you want to maximize is $f(A) + g(B) + h(C)$. Given specific functional forms for f, g, and h, we hope to use the tools of calculus to tackle this problem.

1.2. **The extremes and the middle path.** There are two schools of philosophical thought that shall contend for our attention here: the school that says that extremes are good, and the school that urges you to follow a middle path – a bit of this and a bit of that. Which of them is right? Depends.

The extremists would say that among the three things: food, clothing, and calculus books, one of them is the best value for money (for instance, in our case, it may be calculus books). This means that every additional unit of money that you spend should go on calculus books. Thus suggests that the way you'll be "happiest" is if you spend all your money on calculus books.

In reality, however, we know that that isn't how things work. The problem? We need a bit of food, a bit of clothing, and a bit of calculus, but beyond a point, more food, clothing and calculus is less helpful. This is obvious in the case of food – too much food at your disposal means that you either eat more than your body can handle or throw food away. It is also obvious in the case of clothing. It may not be that obvious in the case of calculus, but you'll have to take my word for it that there does come a point after which more calculus may not be worth it.

So, basically, this is a three-way trade-off game, and we need to figure out where to make the trade-offs between food, clothing, and calculus. This comes somewhere in between – a local maximum, where shifting resources from any one sector to any other sector reduces utility.

On the other hand, there are situations where extremes are better. Those are situations where it's just no contest between the two options – more of one thing is better than more of the other no matter how much of either you have. So the extremes could be the best option.

Thus, the maximum could occur at the endpoints, but it could also occur in between, as a local maximum, where diverting resources a bit in either direction makes things worse.

1.3. **Humbler matters: one-variable ambitions.** After suggesting that I could help you with managing money, I have to retreat to humbler ground. All the tools we have developed so far are tools that specifically deal with one variable – we've talked of *functions of one variable*, and developed concepts of limits, continuity, and differentiation all in this context. Thus, the kind of budgeting and allocation problems that we encounter in the real world, that involve a plethora of variables, are simply too hard for us to handle with this machinery. This is also a reason why you shouldn't just stop with the 150s, and should go on to study multivariable calculus, but let's now talk of what *can* be done using the one-variable approach.

A priori, you might expect that the one-variable approach can only work when there is only one variable involved. It's actually a little more general.

The one-variable approach can be used for situations where we can use some of the constraints to express all variables in terms of a single variable, wherein the optimization problem simply becomes a problem in terms of that variable. So, even though the problem has more than one variable, we are able to tackle it as a one-variable problem. Here is one example.

For instance, consider the following problem: For all the rectangles with diagonal length c, find the dimensions of the one with the largest area.

To solve this problem, we try to figure out what variables we have control over, and what constraints these variables satisfy. A rectangle is specified by specifying its length and breadth, i.e., the two side lengths. If we call these l and b, the goal is to maximize lb. Also, l and b are subject to the constraint $l^2 + b^2 = c^2$, and $l > 0, b > 0$.

The problem is that we have two variables, and we only know how to tackle situations with one variable. In order to solve the problem, we need to write one of the variables in terms of the other one. Note that the relation $l^2 + b^2 = c^2$, along with the fact that $l > 0$ and $b > 0$, allows us to write $b = \sqrt{c^2 - l^2}$. Thus, the area is given by a function of l, namely:

$$A(l) := l\sqrt{c^2 - l^2}$$

Note that $\sqrt{c^2 - l^2} > 0$ implies that $l < c$. Thus, the goal is to maximize this function on the interval $(0, c)$.

We compute the derivative:

$$A'(l) = \sqrt{c^2 - l^2} - \frac{l^2}{\sqrt{c^2 - l^2}}$$

Setting $A'(l) = 0$, we obtain that:

$$\sqrt{c^2 - l^2} = \frac{l}{\sqrt{c^2 - l^2}}$$

Simplifying, we obtain:

$$l = b = c/\sqrt{2}$$

and thus:

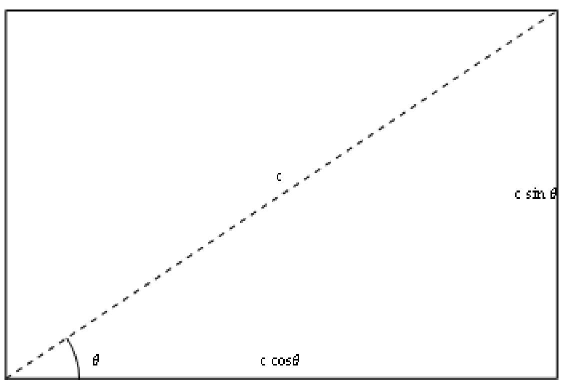
$$A(l) = c^2/2$$

Thus, the only critical point for the function is at $l = c/\sqrt{2}$. Note that for $l < c/\sqrt{2}$, we have $l < \sqrt{c^2 - l^2}$, so the expression for $A'(l)$ is positive, and for $l > c/\sqrt{2}$, the expression is negative. Thus, the point $l = c/\sqrt{2}$ is a point of local maximum.

To determine whether it is a point of absolute maximum, we need to verify that the value of the function at this point is greater than the limits at the two endpoints. It is easy to see that the limits at both endpoints are zero, so indeed, we have a local maximum at $l = c/\sqrt{2}$.

Here is an alternative approach, that involves a different way of choosing variables. Let θ be the angle made by the diagonal with the base of the rectangle. Then, $0 < \theta < \pi/2$, and the two sides of the rectangle have length $c \cos\theta$ and $c \sin\theta$. Thus, the area is given by the function:

$$f(\theta) = c^2 \sin\theta\cos\theta = (c^2/2)\sin(2\theta)$$



This attains an absolute maximum at $\theta = \pi/4$, where $2\theta = \pi/2$, so that $\sin(2\theta) = 1$. Note that we can solve the problem in this case even without using calculus, but if you don't notice that $\sin\theta\cos\theta = (1/2)\sin(2\theta)$, you can solve the problem the calculus way and obtain that the absolute maximum is at $\theta = \pi/4$. Thus, the area is $c^2/2$ and the two side lengths are $c/\sqrt{2}$.

Thus, the maximum occurs for a square.

1.4. **Geometrical and visual optimization.** In most of the situations that we'll be dealing with, it is helpful to draw a figure, label all the lengths and/or angles involved in the figure, and then write down the various constraints as well as the objective function that needs to be maximized. Then, try to get everything in terms of one variable, using the constraints, and finally, do the maximization for that variable. The book gives the following five-point procedure on Page 183:

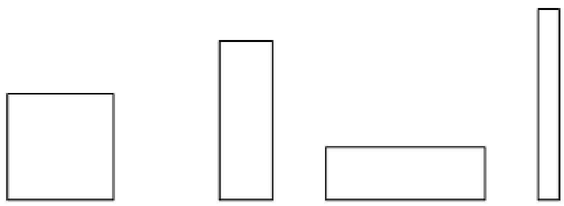
(1) Draw a representative figure and assign labels to the relevant quantities.
(2) Identify the quantity to be maximized or minimized and find a formula for it.
(3) Express the quantity to be maximized or minimized in terms of a single variable; use the conditions given in the problem to eliminate the other variable(s).
(4) Determine the domain of the function generated by Step 3.

(5) Apply the techniques of the preceding sections to find the extreme value(s).

One of the important things in this is to notice that we usually need to maximize the function with the input variable restricted to a certain domain. Thus, there are often situations where the absolute maximum or minimum occurs at an endpoint of the domain, i.e., it is an endpoint maximum/minimum.

Here are some examples that you should keep in mind, both in terms of the final results and the methods we use to get them:

(1) Of all the rectangles with a given perimeter, the square has the largest area. This boils down to maximizing $l((p/2) - l)$. There is no rectangle with minimum area – the minimum area occurs in the *degenerate rectangle* where one of the sides has zero length and the other side has length half the perimeter. The degenerate rectangle isn't ordinarily considered a rectangle. Here are pictures of a collection of rectangles with the same perimeter. It is visually clear that the square has the largest area.



(2) Conversely, of all the rectangles with a given area, the square has the smallest perimeter. This boils down to minimizing $2(l + (A/l))$. There is no rectangle with the largest perimeter – we can keep getting longer and thinner rectangles.
(3) Of all the rectangles with a given diagonal length, the square is the one with the largest area. This boils down to maximizing $l\sqrt{c^2 - l^2}$. Trigonometrically, it involves maximizing $\cos\theta \sin\theta$. The minimum again occurs for the degenerate rectangle, hence does not occur for any actual rectangle.
(4) Of all the rectangles with a given diagonal length, the square is the one with the largest perimeter. This boils down to maximizing $l + \sqrt{c^2 - l^2}$. Trigonometrically, it involves maximizing $\cos\theta + \sin\theta$. The minimum again occurs for the degenerate rectangle, hence does not occur for any actual rectangle.

1.5. **Applications to real-world physical situations.** We see a common concern that is apparent with all these maximization/minimization problems. Maximizing the area for a given perimeter, or minimizing the perimeter for a given area, is a concern that arises when trying to create containers with as little material used for the boundary as possible. Maximizing the area for a given diagonal length or constraints on lengths occurs in situations where concerns of space availability and fitting stuff are paramount.

Here are some of the quantities and formulas that are useful:

(1) For a right circular cylinder with base radius r and height h, the total volume (or capacity) is $\pi r^2 h$. The curved surface area is $2\pi rh$. Each of the disks at the ends has area πr^2. Thus, a right circular cylinder closed at one end has surface area $\pi r(2h + r)$ and a right circular cylinder closed at both ends has surface area $2\pi r(r + h)$. Based on the situation at hand, we need to figure out which of these three surface areas is being refered to.
(2) For a right circular cone with base radius r, vertical height h and slant height l, the volume is $(1/3)\pi r^2 h$. The curved surface area is πrl and the surface area of the base is πr^2, so the total surface area is $\pi r(r + l)$. Again, we need to figure out, based on the situation, which of the surface areas is being refered to. Also note that r, h, and l are related by the Pythagorean theorem: $l^2 = r^2 + h^2$.
(3) For a semicircle of radius r, the area is $(1/2)\pi r^2$. The length of the curved part is πr and the length of the straight part (the diameter) is $2r$, so the total perimeter is $r(\pi + 2)$. More generally, for a sector of the circle bounded by two radii and an arc, where the radii make an angle of θ, the perimeter is $r(2 + \theta)$ and the area is $(1/2)\theta r^2$.
(4) For a sphere, the surface area is $4\pi r^2$ and the volume is $(4/3)\pi r^3$. For a hemisphere, the surface area is $3\pi r^2$ ($2\pi r^2$ for the curved part and πr^2 for the bounding disk) and the volume is $(2/3)\pi r^3$.

Here are some important results on optimization in these various examples:

(1) For a right circular cylinder with volume V, there is no minimum and no maximum for the curved surface area. This is because for given radius r, the expression for the curved surface area is $2V/r$, which approaches ∞ as $r \to 0$ (smaller and smaller radius, larger and larger height) and approaches 0 as $r \to \infty$ (larger and larger radius, smaller and smaller height). If, however, we have additional boundary constraints on the radius or height, the maximum/minimum will occur at these boundaries.
(2) For a right circular cylinder with volume V, there is an absolute minimum for the surface area of the base plus curved part (i.e., only one bounding disk is included). The expression is $2V/r + \pi r^2$. As $r \to 0$ or $r \to \infty$, this expression tends to ∞. The absolute minimum occurs at the point $r = (2V/\pi)^{1/3}$. (see also Example 1 from the book).
(3) For a right circular cylinder with volume V, there is an absolute minimum for the total surface area (including both disks). The expression is $2V/r + 2\pi r^2$. As $r \to 0$ or $r \to \infty$, this tends to infinity. The absolute minimum occurs at the point $r = (V/\pi)^{1/3}$.

2. Important tricks in real-world problems

2.1. **The maximum is determined by the tightest constraint.** Let me first state this mathematically (where it's obvious) and then non-mathematically (where again it's obvious).

Suppose x is a real number subject to the constraints $x \le a_1$, $x \le a_2$, and $x \le a_3$. What is the *maximum* value that x can take? Clearly, it is the *minimum* among a_1, a_2, and a_3, because that is the tightest, most limiting constraint on x.

Here are some non-mathematical formulations:

(1) A truck has to go on a highway. As part of its journey, the truck needs to negotiate three underpasses, with clearances of 10 feet, 9 feet, and 11 feet respectively. What is the *maximum* possible height of the truck? (Hint: You want to make sure you don't get into a problem anywhere).
(2) Hydrogen and oxygen combine in a ratio of 1 : 8 by mass to produce water. Assuming that we have 50 grams of hydrogen and 220 grams of oxygen, what is the *maximum* possible amount of water that can be produced from these? (Hint: Limiting reagent).

To repeat: *the maximum value that something can take is determined by the tightest of the upper bounds on it.* The importance of this idea cannot be over-emphasized. On the one hand, it is a staple of a whole branch of graph theory/network theory results called max-min theorems. All of them have the flavor that the upper end of what's possibility coincides with the lower end of the constraints. On the other hand, it is also the whole basis for the theory of least upper bounds and greatest lower bounds that we will see in 153 and that forms the basis for a rigorous study of the reals (which you might see if you proceed to study real analysis).

2.2. **Some random tricks.** A real-world optimization problem is not usually given in a ready-to-solve form. Rather, some decisions and judgments need to be made about the procedure and the general form of the solution in order to obtain a mathematical setup.

The initial judgment may use general rules: for instance, the rule that straight line paths, where possible, are shorter than non-straight line paths. Thus, when asked to find a shortest path subject to certain constraints, we may be able to narrow it down to a straight line path and then do the optimization within that.

As a somwhat trickier example, consider the following problem, which appears on your homework:

> Two hallways, one 8 feet wide and the other 6 feet wide, meet at right angles. Determine the length of the longest ladder that can be carried horizontally from one hallway to the other.

Here, the significance of *horizontally* is simply that the ladder cannot be tilted vertically, a strategy that would enable one to carry a longer ladder. This problem is a hard one because the nature of the constraint is not clear. How does the width of the hallways constrain the length of the ladder that can be passed through?

We need to role-play the *process of carrying the ladder.* When a ladder is being carried along a corridor, it makes the most sense to align the ladder parallel to the walls of the corridor. When the direction of the corridor changes, the ladder needs to be rotated to align it with the new corridor. We must be able to rotate the ladder through every angle. This leads to the constraint: for every angle, the ladder must fit in. We then try to find, for every angle θ, the maximum length of ladder that can fit in at the junction between the two corridors. Each of these imposes a constraint on the length of ladder. The most relevant binding constraint is the *minimum* of these lengths.

2.3. **The intuition of tangency.** Let's now look at another problem that also appears on your homework:

> A tapestry 7 feet high hangs on a wall. The lower edge is 9 feet above an observer's eye. How far from the wall should the observer stand in order to obtain the most favorable view? Namely, what distance from the wall maximizes the visual angle of the observer?

Here's the intuition behind this problem. If you stand right under the tapestry, it seems *foreshortened*. If, however, you go very far, then it simply seems small. The quantity that measures how large the tapestry appears is the visual angle, or the angle between the lines joining your eyes to the top and bottom of the tapestry. This angle is zero if you are right under the tapestry, and it approaches zero as you go out far from the tapestry. Where is it maximum? Somewhere in between. But where exactly?

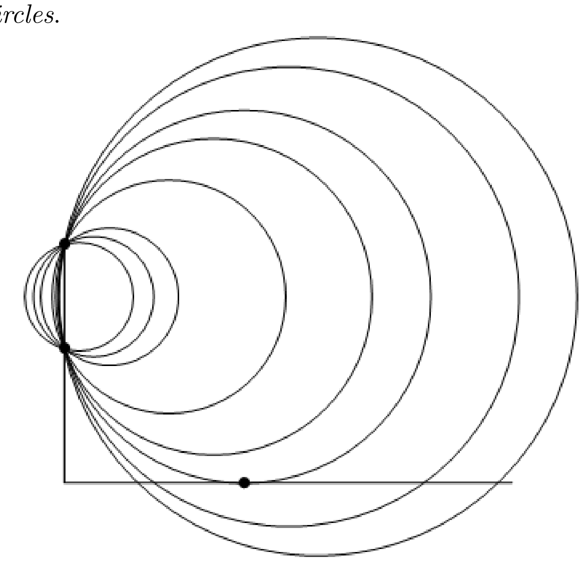
You can find this using calculus – which is what you are expected to do in this homework. But there is an alternative, related approach that is more geometric.

The main geometric fact used is that the angle subtended by a chord of a circle at any two points on the circle on the same side of the chord is the same.

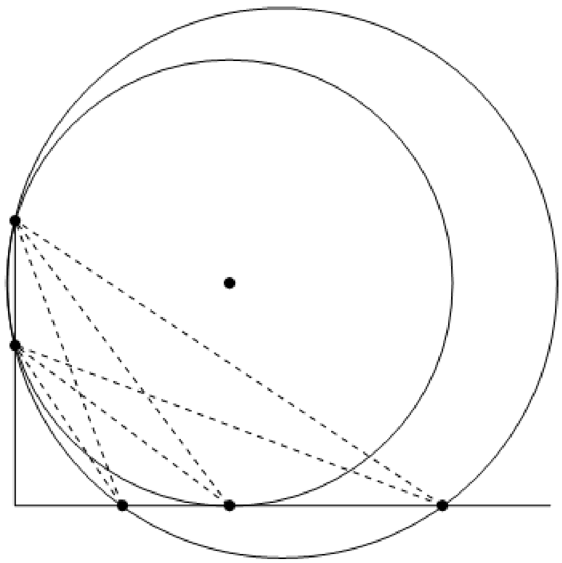
Now, consider a circle passing through the two ends of the tapestry. If this circle intersects the horizontal line of possible locations of your eye at two points P and Q, then by the result I mentioned, the visual angle at P equals the visual angle at Q. Note that for a very large circle, P is very close to the base of the tapestry and Q is very far away. This helps explain why small visual angles are achieved both very close and very far away from the tapestry.

We also see that the smaller the circle, the larger the visual angle. Thus, the goal is to find the smallest circle passing through the two ends of the tapestry that intersects the horizontal line of possible locations of the eye. A little thought reveals that this occurs when the circle is tangent to the horizontal line. If you imagine starting with a very large circle and shrinking it further and further, the circle that is tangent to the horizontal line is the one at which the circle just leaves the horizontal line. (Having deduced this, it is possible to determine the precise point using geometry and algebra, without any calculus. You can verify the answer you obtain using calculus via this method).

Here's the picture with lots of such circles drawn. Such a system of circles is called a *coaxial system of circles*.



Here's the same picture with just the circle of tangency and another circle drawn:



Note: We can use another result of geometry to calculate the distance of the point of tangency from the foot of the tapestry. Namely, the result says that if P is a point outside a circle, PT is a tangent to the circle with point of tangency T, and a secant line through P intersects the circle at A and B, then $PA \cdot PB = PT^2$. We can use this to calculate PT as the square root of the product of the distances from the base of the bottom and top of the tapestry.

2.4. **The heuristic of multiple uses.** Suppose a resource (such as fencing wire, which plays the role of perimeter) is to be divided among two alternative uses: say a square fence, and a circular fence. It is a fact that, of all possible shapes with a given perimeter, the circle encloses the largest area. (This is called the *isoperimetric problem*, and although we will not show it, it is useful to remember). In particular, devoting all the fencing to the circle yields a larger area than devoting all the fencing to the square. (This can be checked easily by algebra, and the fact that $\pi < 4$).

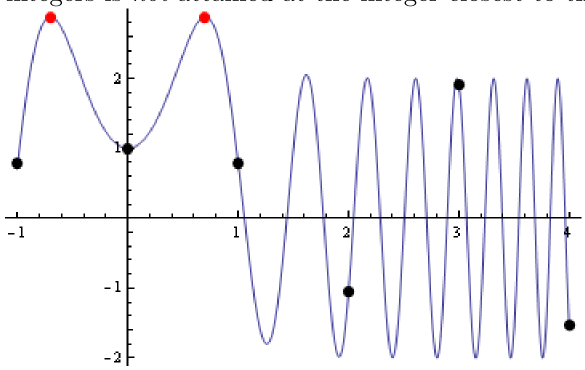
Given this, what is the way of allocating fencing so as to get the maximum and minimum possible total area? It turns out that for the *maximum possible*, we allocate all resources to the hands-down better use, which is in this case the circle. However, for the *minimum possible*, the strategy is *not* that of allocating everything to the square. Why not? It turns out that we can do even worse by providing some fencing to the square and some fencing to the circle? Why? Because there is some wastage that arises simply from having two fences. Even though a square is less efficient than a circle, devoting everything to the square is a little more efficient than devoting mostly to the square and a bit for the circle. A problem of this kind appears in Homework 5.

In the good old days when everybody farmed, each time a farmer with multiple sons died, his land was divided among the sons. As a result, fencing costs and wastage kept increasing. One solution to this was *primogeniture laws*, which stated that the eldest son was entitled to the land. While not fair, these laws helped combat the problem of fragmentation of land holdings.

2.5. **Integer optimization.** A few brief notes on integer optimization may be worthwhile. In many real-world situations, the possible values that a variable can take are constrained to be integers. For instance: *how many passengers can travel on this vehicle*? The optimization here thus requires one to optimize *subject to the integer value constraint on the variables.*

It may seem reasonable at first to believe that the best integer solution is the integer closest to the best real solution. This is not always the case. In fact, computer scientists have shown that even solving systems of linear equations and inequalities in integer variables has no general-purpose algorithm that runs quickly (subject to a long-standing conjecture called $P \neq NP$). This is despite the fact that the analogous problem is very easy to solve for real variables.

The problem here is that the value of a function can change very rapidly between a real number and the integers closest to it. See, for instance, this picture for a function where the maximum value among values at integers is *not* attained at the integer closest to the absolute maximum:



However, it is useful to look at the behavior of a function over all real numbers in order to determine the integers where it attains maxima and minima. For instance, if we can determine where the function is increasing and decreasing, we can use this information along with testing some values in order to find out the maxima and minima. Specifically, what we first do is *extend the function to all real numbers* (by considering the definition of the function applied to all real numbers) and find the intervals where the function is increasing and decreasing as a function with real inputs. Then:

(1) If f is increasing on an interval, then the minimum of f on that interval occurs at the smallest integer in the interval and the maximum occurs at the largest integer in the interval.
(2) If f is decreasing on an interval, then the minimum of f on that interval occurs at the largest integer in the interval and the minimum occurs at the smallest integer in the interval.
(3) If we need to find the absolute maximum of f over all integers, we can first break up into intervals where f is increasing/decreasing, find the maximum over each of those intervals, and finally compare the values of all these maxima to find which is the largest one.

Here are some simple examples:

(1) Consider a function that is decreasing on $(-\infty, 1.3]$ and increasing on $[1.3, \infty)$. Then, if viewed as a function on reals, this function has a unique absolute minimum at 1.3. As a function on integers, we know that the function is increasing from 2 onwards, so the value for any integer greater than 2 is greater than the value at 2. Similarly, we know that the value at any integer less than 1 is greater than the value at 1. So, there are two candidates for the absolute minimum among integers: the values at 1 and 2. We now calculate the values at 1 and 2 and find which one is smaller.
(2) Consider a function that decreases on $(-\infty, -1.1]$, increases on $[-1.1, 0.1]$, decreases on $[0.1, 0.9]$, and then increases on $[0.9, \infty)$. On the interval $(-\infty, -1.1]$, the minimum among integers is at -2. On the interval $[-1.1, 0.1]$, the minimum among integers is at -1. On the interval $[0.1, 0.9]$, there are no integers. On the interval $[0.9, \infty)$, the minimum among integers is at 1. Thus, the three candidates for the point of absolute minimum are -2, -1, and 1.

Note that it is *not* necessarily true that the integer where the absolute minimum among integers is attained is the closest integer to the real number where the absolute minimum among real numbers is attained. This is because the function can change very rapidly between a real number and the closest integer. For certain special kinds of functions (such as quadratic functions), it *is* true that the integer for absolute minimum is the closest integer to the real number for absolute minimum. But this is due to the symmetric nature of quadratic functions – the graph of a quadratic function with positive leading coefficient is symmetric about the vertical line through its point of absolute minimum.[1]

[1]For negative leading coefficient, the corresponding statement is true if we replace absolute minimum by absolute maximum.

3. Some notes from social and natural sciences

3.1. **An important maximization: Cobb-Douglas, fair share, and kinetics.** Let's now go to a question considered by some economists in the early twentieth century. We'll then talk about how a similar question comes up in chemical reactions.

Suppose a factory is producing some goods using two kinds of inputs: labor and capital. For a given production process, if the factory spends L on labor and K on capital, the output of the factory is given by $L^a K^b$, where a and b are positive numbers. The goal of the factory is to maximize output for a given expenditure $(L + K)$. In other words, if the factory is spending a total E on labor and capital put together, how should it allocate E between labor and capital to obtain the maximum output?

Since E is fixed, we can choose L as the variable and write $K = E - L$. We thus get that the output is $L^a(E - L)^b$. If we further let $x = L/E$ (the fraction on labor), then the output is given by $E^{a+b}x^a(1 - x)^b$. Thus, in order to maximize output, we need to maximize $x^a(1 - x)^b$, where $x \in [0, 1]$.

A maximization of this sort appears on your homework, and we find there that the absolute maximum on the interval $[0, 1]$ occurs at the point $a/(a + b)$. Thus, the maximum occurs when $L = Ea/(a + b)$ and $K = Eb/(a+b)$. Thus, the labor-to-capital expenditure ratio L/K is a/b – the same as the ratio of exponents.

What this result shows is that the ratio of exponents on labor and capital represents the relative contributions of labor and capital to production. Optimization occurs when the allocation of resources is done according to these relative contributions: a fraction of $a/(a+b)$ to labor and a fraction of $b/(a+b)$ to capital. In hindsight, this makes intuitive sense: the larger the value of a, the more sense it makes to invest in labor, because the return on investment in labor is higher. However, after some point, it also makes sense to invest a bit in capital, otherwise that becomes a bottleneck. The proportion should have something to do with the ratio of a and b. Mathematically, we have shown that these two proportions in fact coincide.

This raises the question of what determines a and b in the first place. This has something to do with the nature of the production process. A *labor-intensive process* would be one where a dominates and a capital-intensive process would be one where b dominates.

All production functions do not look like the function above. However, it was the argument of Cobb and Douglas that assuming functions to be of the above form is a useful simplification and many phenomena of relative allocation of resources to factors of production can be understood this way. In many parts of economics and the social sciences, people wanting to do a simple analysis often begin by assuming that a given production function is Cobb-Douglas, in order to get a clear handle on the relative contribution of different factors.

Another place where a similar formulation pops up is chemical kinetics. Suppose we have a chemical reaction between two substances A and B, with equation of the form $mA + nB \to$ products. The theory of chemical kinetics suggests that, assuming this reaction is elementary, the rate of forward reaction is given by $k_f[A]^m[B]^n$ where k_f is a constant (with suitable dimensions), $[A]$ is the concentration of A and $[B]$ is the concentration of B.

Now, the question may be: for a given total concentration, how do you decide the proportions in which to mix A and B to get the fastest reaction? This is the same problem in a new guise, and it turns out that the maximum occurs when $[A] : [B] = m : n$. This is poetic justice, because this is precisely the *right* ratio from the *stoichiometric* viewpoint.

3.2. **Frontier curves and optimal allocation.** An important concept, which you may first see in economics courses, but which also occurs elsewhere, is that of a *production possibility frontier* or *production possibility curve*. Let's understand these curves in the language of optimization.

Suppose you are running a farm that can produce only two things: wheat and rice. Now, let's say that you decide to produce $50,000$ bushels of wheat. Given this constraint, your goal is to produce as much rice as possible. This is now an optimization problem and you somehow solve it and find out that you can produce at most $25,000$ bushels of rice if you want to produce $50,000$ bushels of wheat.

Now, if you instead wanted to produce only $40,000$ bushels of wheat, it is possible that you can produce more – say $40,000$ bushels of rice. Thus, *for each quantity of wheat that you choose to produce*, there is a maximum quantity of rice you can produce with the given resources. We can thus define a *function* that takes as input the quantity of wheat and outputs the maximum quantity of rice that can be produced alongside. This is a *decreasing* function (the more wheat you produce, the less resources you can devote to producing rice) and its domain is from 0 to the maximum amount of wheat that you can produce. The largest value in the domain is the maximum amount of wheat you can produce if you devote all resources to wheat production, and the value of the function at 0 if the maximum amount of rice you can produce if you devote all your resources to rice production.

The graph of this function is called the *production possibility curve* or *production possibility frontier*. The important thing to note about this graph is that *every point on the graph is an optimal point in some sense* – there is no way to unambiguously improve from any of these points. Any point below a point on the curve, or on the inside of the curve, is achievable but non-optimal, in the sense that it is possible to increase the production of one or both the outputs without decreasing the other one. A point above or outside the production possibility frontier is a point that cannot be achieved, reflecting the *reality of scarcity* or the *limitations of current technology*, depending on your perspective.

3.3. **Can spontaneous processes solve optimization problems?** First, a little clarification on what the question means. In all the situations we have seen so far, there is a conscious agent that is using calculus to find an optimal allocation or optimal value by explicitly considering constraints. But optimization has been a goal for living creatures and for nature long before the advent of calculus. How did they do it?

For instance, bubbles tend to be spherical in order to minimize their surface area. More generally, the shapes that soap films can take are minimal surfaces – they minimize surface area. But are bubbles and soap films solving a complicated optimization problem by choosing a spherical shape? Do we need to posit a theory of consciousness and calculus ability every time we see such optimization in the physical or biological world?

No. Physical entities (and most primitive biological entities) are not trying to reach an optimal state – they simply *keep moving around until* they hit upon a *stable equilibrium*, which is *locally optimal*. In fact, the same is true for humans interacting in a large market. This point is extremely important.

For instance, here are some crude heuristics:

(1) In the world of physics, the reason why mechanical or physical systems tend to certain "optimal" configurations is that in these configurations, there are no forces rending them apart or causing them further change.
(2) In the world of chemistry, materials keep reacting until they reach a configuration where the push to the reaction in one direction equals the push to reaction in the other direction.
(3) In the world of biology, living creatures explore the space around them till they hit on something that's better than the stuff around it.
(4) In the word of economics, each individual keeps making changes in the variables under his/her economic control until reaching a situation where a change in either direction is not to his/her advantage.

The upshot is that local optima tend to be places of stability simply because there isn't a tendency to deviate either way, not because anybody did calculus. You can think of it like an ant moving along the graph of a curve and stopping when it gets to a peak and would need to go down both ways.

But this also has a flip side:

- Local optima need not be absolute optima. That was the whole point of our earlier lecture on the subject! But given their stability, natural systems may stay stuck at these local optima. To get to an even bigger global optima, a *push* may be needed. (For instance, activation energy in the context of a chemical reaction, or the entry of a new competitor in a stagnating and non-innovating industry).
- In some cases, there may be so many different local optima, or the situation may be so shaky, that there is never any place to settle down at. In some cases, inertia may prevent settling down. This causes such phenomena as *oscillatory* and *chaotic* behavior.

CONCAVITY, INFLECTIONS, CUSPS, TANGENTS, AND ASYMPTOTES

MATH 152, SECTION 55 (VIPUL NAIK)

Corresponding material in the book: Section 4.6, 4.7.

Difficulty level: Moderate to hard. If you have seen these topics in AP Calculus, then moderate difficulty; if you haven't, then hard.

What students should definitely get: The definitions of *concave up*, *concave down*, and *point of inflection*. The strategies to determine *limits at infinity*, *limits valued at infinity*, *vertical tangents*, *cusps*, *vertical asymptotes*, and *horizontal asymptotes*.

What students should hopefully get: The intuitive meanings of these concepts, important examples and boundary cases, the significance of concavity in determining local extrema, the use of higher derivative tests. Important tricks for calculating limits at infinity.

Executive summary

0.1. **Concavity and points of inflection.** Words ...

(1) A function is called *concave up* on an interval if it is continuous and its first derivative is continuous and increasing on the interval. If the function is twice differentiable, this is equivalent to requiring that the second derivative be positive except possibly at isolated points, where it can be zero. (Think x^4, whose first derivative, $4x^3$, is increasing, and the second derivative is positive everywhere except at 0, where it is zero).
(2) A function is called *concave down* on an interval if it is continuous and its first derivative is continuous and decreasing on the interval. If the function is twice differentiable, this is equivalent to requiring that the second derivative be negative except possibly at isolated points, where it can be zero.
(3) A *point of inflection* is a point where the sense of concavity of the function changes. A point of inflection for a function is a point of local extremum for the first derivative.
(4) Geometrically, at a point of inflection, the tangent line to the graph of the function *cuts through* the graph.

Actions ...

(1) To determine points of inflection, we first find critical points for the first derivative (which are points where this derivative is zero or undefined) and then use the first or second derivative test at these points. Note that these derivative tests are applied to the first derivative, so the first derivative here is the second derivative and the second derivative here is the third derivative.
(2) In particular, if the second derivative is zero and the third derivative exists and is nonzero, we have a point of inflection.
(3) A point where the first two derivatives are zero could be a point of local extremum or a point of inflection. To find out which one it is, we either use sign changes of the derivatives, or we use higher derivatives.
(4) Most importantly, the second derivative being zero does *not* automatically imply that we have a point of inflection.

0.2. **Tangents, cusps, and asymptotes.** Words...

(1) We say that f has a horizontal asymptote with value L if $\lim_{x\to\infty} f(x) = L$ or $\lim_{x\to-\infty} f(x) = L$. Sometimes, both might occur. (In fact, in almost all the examples you have seen, the limits at $\pm\infty$, if finite, are both equal).
(2) We say that f has a vertical asymptote at c if $\lim_{x\to c^-} f(x) = \pm\infty$ and/or $\lim_{x\to c^+} f(x) = \pm\infty$. Note that in this case, it usually also happens that $f'(x) \to \pm\infty$ on the relevant side, with the sign

the same as that of $f(x)$'s approach if the approach is from the left and opposite to that of $f(x)$'s approach if the approach is from the right. However, this is not a foregone conclusion.

(3) We say that f has a vertical tangent at the point c if f is continuous (and finite) at c and $\lim_{x\to c} f'(x) = \pm\infty$, with the *same sign of infinity* from both sides. If f is increasing, this sign is $+\infty$, and if f is decreasing, this sign is $-\infty$. Geometrically, points of vertical tangent behave a lot like points of inflection (in the sense that the tangent line cuts through the graph). Think $x^{1/3}$.

(4) We say that f has a vertical cusp at the point c if f is continuous (and finite) at c and $\lim_{x\to c^-} f'(x)$ and $\lim_{x\to c^+} f'(x)$ are infinities of opposite sign. In other words, f takes a sharp about-turn at the x-value of c. Think $x^{2/3}$.

(5) We say that f is asymptotic to g if $\lim_{x\to\infty} f(x) - g(x) = \lim_{x\to-\infty} f(x) - g(x) = 0$. In other words, the graphs of f and g come progressively closer as $|x|$ becomes larger. (We can also talk of one-sided asymptoticity, i.e., asymptotic only in the positive direction or only in the negative direction). When g is a *nonconstant linear function*, we say that f has an *oblique asymptote*. Horizontal asymptotes are a special case, where one of the functions is a constant function.

Actions...

(1) Finding the horizontal asymptotes involves computing limits as the domain value goes to infinity. Finding the vertical asymptotes involves locating points in the domain, or the boundary of the domain, where the function limits off to infinity. For both of these, it is useful to remember the various rules for limits related to infinities.

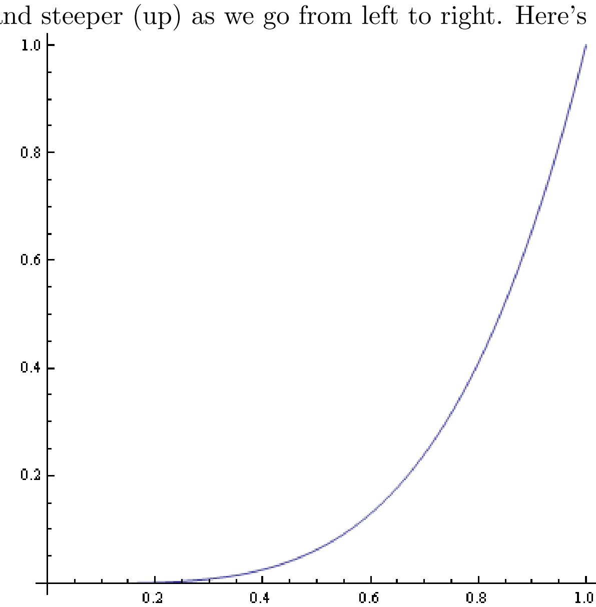
(2) Remember that for a vertical tangent or vertical cusp at a point, it is necessary that the function be continuous (and take a finite value). So, we not only need to find the points where the derivative goes off to infinity, we also need to make sure those are points where the function is continuous. Thus, for the function $f(x) = 1/x$, $f'(x) \to -\infty$ on both sides as $x \to 0$, but we do *not* obtain a vertical tangent – rather, we obtain a vertical asymptote.

1. Concavity and points of inflection

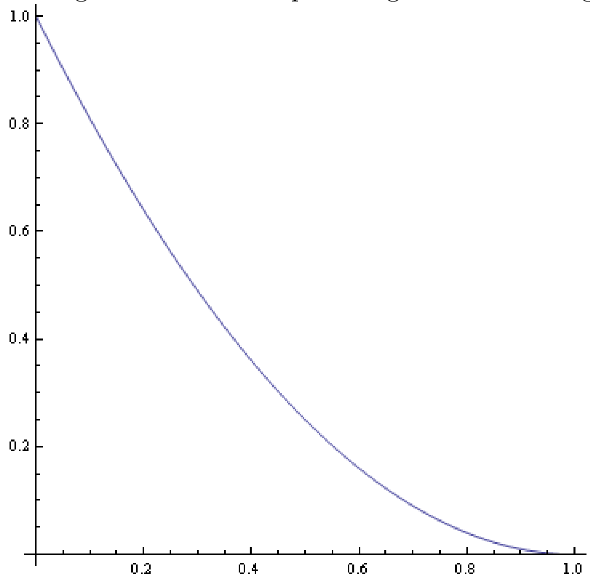
1.1. **Concavity.** *Concave up* means that the derivative of the function (which measures its rate of change) is itself increasing. Formally, a function f differentiable on an open interval I is termed *concave up* on I if f' is increasing on I. I hope you remember the definition of an *increasing function*: it means that for two points $x_1, x_2 \in I$, with $x_1 < x_2$, we have $f'(x_1) < f'(x_2)$.

Here's three points:

(1) If f itself is increasing (so that f' is positive), then being concave up means that f is increasing *at an increasing rate*. In other words, the slope of the tangent line to the graph of f becomes steeper and steeper (up) as we go from left to right. Here's a typical picture:



(2) If f itself is decreasing (so that f' is negative), then being concave up means that f is decreasing *at a decreasing rate*. In other words, the slope of the tangent line to the graph is negative, but it is becoming less and less steep as we go from left to right. Here's a typical picture:

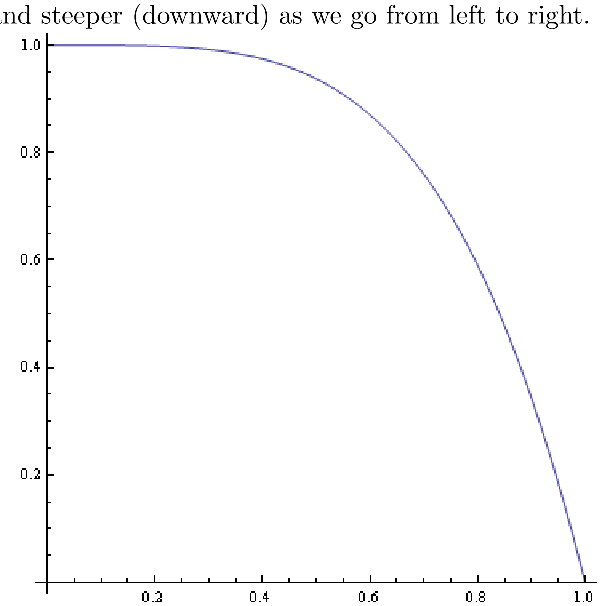


(3) If f is *twice differentiable*, i.e., f' is differentiable, then we can deduce whether f' is increasing by looking at f''. Specifically, if f' is continuous on I, and $f'' > 0$ everywhere on I except at a few isolated points, then f is concave up throughout.

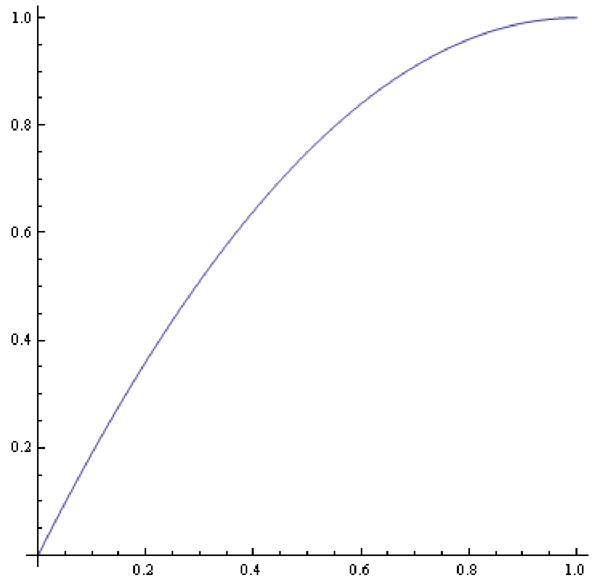
Similarly, if f is differentiable on an open interval I, we say that f is *concave down* on I if f' is *decreasing* on the interval I. I hope you remember the definition of a *decreasing function*: it means that for two points $x_1, x_2 \in I$, with $x_1 < x_2$, we have $f'(x_1) > f'(x_2)$.

Here's three points:

(1) If f itself is decreasing (so that f' is negative), then being concave down means that f is decreasing *at an increasing rate*. In other words, the slope of the tangent line to the graph of f becomes steeper and steeper (downward) as we go from left to right.



(2) If f itself is increasing (so that f' is positive), then being concave down means that f is increasing *at a decreasing rate*. In other words, the slope of the tangent line to the graph is positive, but it is becoming less and less steep as we go from left to right.



(3) If f is *twice differentiable*, i.e., f' is differentiable, then we can deduce whether f' is increasing by looking at f''. Specifically, if f' is continuous on I, and $f'' < 0$ everywhere on I except at a few isolated points, then f is concave down throughout.

1.2. **Points of inflection.** A point of inflection is a point c in the interior of the domain of a differentiable function (i.e., the function is defined and differentiable on an open interval containing that point) such that the function is concave in one sense to the immediate left of c and concave in the other sense to the immediate right of c.

Another way of thinking of this is that points of inflection of a function are points where the derivative is increasing to the immediate left and decreasing to the immediate right, or decreasing to the immediate left and increasing to the immediate right. In other words, it is a point of local maximum or a point of local minimum for the derivative of the function.

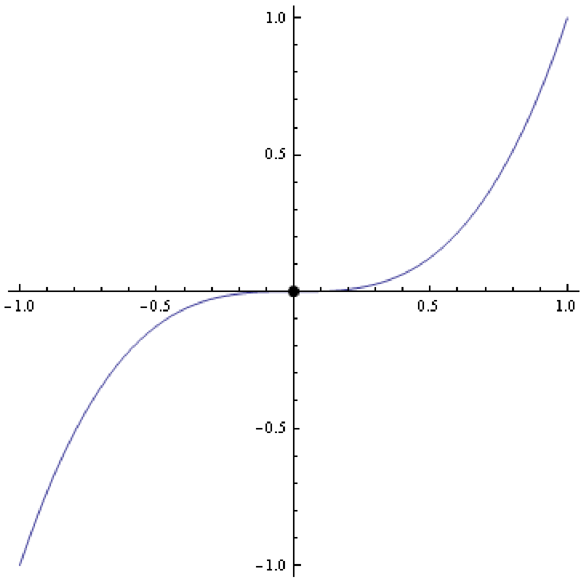
Recall that earlier, we noted that for a point of local maximum or a point of local minimum, either the derivative is zero or the derivative does not exist. Since everything we're talking about now is related to f', we have that for a point of inflection, either $f'' = 0$ or f'' does not exist.

So the upshot: concave up means the derivative is increasing, concave down means the derivative is decreasing, point of inflection means the sense in which the derivative is changing changes at the point.

1.3. **A point of inflection where the first two derivatives are zero.** We now consider one kind of point of inflection: where the first derivative and the second derivative are both zero. Let's begin with the example.

Consider the function $f(x) := x^3$. Recall first that since f is a cubic function, it has odd degree, so as $x \to -\infty$, $f(x) \to -\infty$, and as $x \to \infty$, we also have $f(x) \to \infty$. Further, if we compute $f'(x)$, we get $3x^2$. Note that the function $3x^2$ is positive for $x \neq 0$, and is 0 at $x = 0$. So, from our prior discussion of increasing and decreasing functions, we see that f is increasing on $(-\infty, 0]$ and then again on $[0, \infty)$. And since the point 0 is common to the two intervals, f is in fact increasing everywhere on $(-\infty, \infty)$.

If you remember, this was an important and somewhat weird example because, although $f'(0) = 0$, f does *not* attain a local extreme value at 0. This is because the derivative of f is positive *on both sides* of 0.



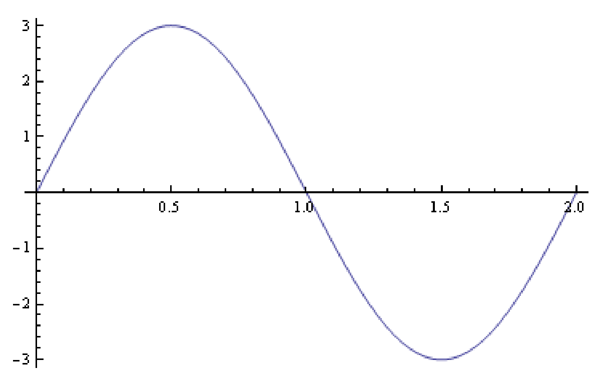
This was the picture we had from our earlier analysis. But now, with the concepts of concave up, concave down, and points of inflection, we can get a better understanding of what's going on. Specifically, we see that the second derivative f'' is $6x$, which is negative for $x < 0$, zero for $x = 0$, and positive for $x > 0$. Thus, f is concave down for $x < 0$, $x = 0$ is a point of inflection, and f is concave up for $x > 0$.

So, here's the picture: for $x < 0$, f is negative, f' is positive, and f'' is again negative. Thus, the graph of f is below the x-axis (approaching 0), it is going upward, and it is going up at a decreasing rate. So, as $x \to 0$, the graph becomes flatter and flatter.

For $x > 0$, f is positive, f' is positive, and f'' is again positive. Thus, the graph of f is above the x-axis (starting from the origin), it is going upward, and it is going up at an increasing rate. The graph starts out from flat and becomes steeper and steeper.

So, $x = 0$ is a *no sign change* point for f', which is why it is not a point of local maximum or local minimum. This is because on both sides of 0, f' is positive. What happens is that it is going down from positive to zero and then up again from zero to positive. But on a related note, because f' itself dips down to zero and then goes back up, the point 0 is a point of local minimum for f', so it is a point of inflection for f.

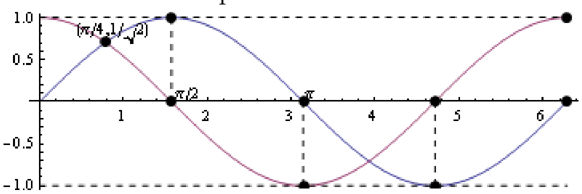
The main thing you should remember is that when we have a critical point for a function, where the derivative is zero, but it is neither a point of local maximum nor a point of local minimum, then it is *likely to be* a point of inflection. In other words, this idea of something that is increasing (or decreasing) and momentarily stops in its tracks, is the picture of neither a local maximum nor a local minimum but a point of inflection.



1.4. **Points of inflection where the derivative is not zero.**

Let's review the graph of the sine function. The sine function starts with $\sin(0) = 0$, goes up from 0 to $\pi/2$, where it reaches the value 1, then drops down to 0, drops down further to -1 at $3\pi/2$, and then turns back up to reach 0 at 2π. And this pattern repeats periodically.

So far, you've taken me on faith about the way the graph curves. But we can now start looking at things in terms of concave up and concave down.

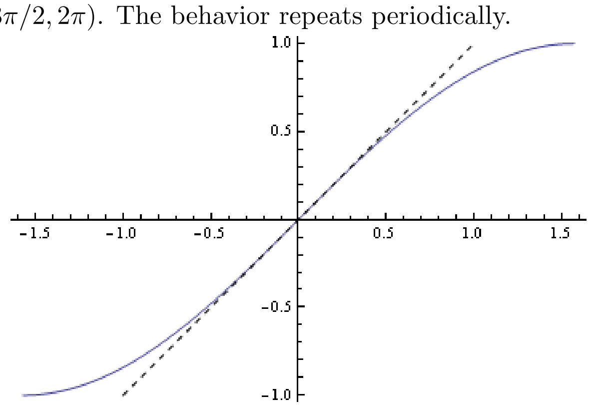


The derivative of the sin function is the cos function. Let's graph the cos function. This starts with the value 1 at $x = 0$, goes down to zero at $x = \pi/2$, dips down to -1 at $x = \pi$, goes back up to 0 at $x = 3\pi/2$, and then goes up to 1 at $x = 2\pi$.

We see that cos is positive from 0 to $\pi/2$, and sin is increasing on that interval. cos is negative from $\pi/2$ to $3\pi/2$, and sin is decreasing on that interval. cos is again positive from $3\pi/2$ to 2π, and sin is again increasing on that interval.

Next, we want to know where sin is concave up and where it is concave down. And for this, we look at the second derivative of sin, which is the function $-\sin$. As you know, the graph of this function is the same as the graph of sin, but flipped about the x-axis. This means that where sin is positive, its second derivative is negative, and where sin is negative, its second derivative is positive.

So, from the interval between 0 and π, sin is concave down and on the interval between π and 2π, sin is concave up. Breaking the interval down further, sin is increasing and concave down on $(0, \pi/2)$, decreasing and concave down on $(\pi/2, \pi)$, decreasing and concave up on $(\pi, 3\pi/2)$, and increasing and concave up on $(3\pi/2, 2\pi)$. The behavior repeats periodically.



Now, let's concentrate on the points of inflection. Note that the sense of concavity changes at multiples of π – at the point 0, the function changes from concave up to concave down. At the point π, the function changes from concave down to concave up. Another way of thinking about this is that just before π, the function is decreasing at an increasing rate – it is becoming progressively steeper. But from π onwards, it starts decreasing at a decreasing rate, in the sense that it starts becoming less steep. So π is the point where the way the tangent line is turning starts changing.

1.5. **A graphical characterization of inflection points.** Inflection points are graphically special because they are points where the way the tangent line is turning changes sense. There's a related characterization. If you draw the tangent line through an inflection point, the tangent line *cuts through* the curve. Equivalently, the curve *crosses* the tangent line. This is opposed to any other point, where the curve *locally* lies to one side of the tangent line.

For instance, for the cube function $f(x) := x^3$, the tangent line is the x-axis, and the curve crosses the x-axis at $x = 0$. We see something similar for the tangent lines at the points of inflection 0 and π for the sin function.

1.6. **Third and higher derivatives: exploration.** (I may not get time to cover this in class).

A while ago, we had developed criteria to determine whether a critical point is a point where a local extreme value is attained. We discussed two tests that could be used: the *first derivative test* and the *second derivative test*. The first derivative test said that if f' changes sign across the critical point, it is a point where a local extreme value occurs: a local maximum if the sign change is from positive on the left to negative on the right, and a local minimum if the sign change is from negative on the left to positive on the right.

The second derivative test was a test specially suited for functions that are twice differentiable at the critical point. This test states that if the second derivative at a critical point is negative, the function attains a local maximum, and if the second derivative is positive, the function attains a local minimum. This leaves one case open: what happens if the second derivative at the critical point is zero?

In this case, things are inconclusive. We might have a point of local maximum, a point of local minimum, an inflection point, or none of the above. How do we figure this out? I will give two general principles of alternation, and then we will look at some examples:

(1) If c is a point of inflection for f' and and $f'(c) = 0$, then c is a point of local extremum for f. If the point of inflection is a change from concave up to concave down, we get a local maximum and if the change is from concave down to concave up, we get a local minimum.
(2) If c is a point of local maximum or minimum for f', then c is a point of inflection for f. Local maximum implies a change from concave up to concave down and local minimum implies a change from concave down to concave up.

Let's illustrate this with the function $f(x) := x^5$ and the point $c = 0$. Let's also assume you knew nothing except differentiation and applying the derivative tests. We have $f'(x) = 5x^4$, $f''(x) = 20x^3$, $f'''(x) = 60x^2$, $f^{(4)}(x) = 120x$, and $f^{(5)}(x) = 120$. At $c = 0$, $f^{(5)}$ is the first nonzero derivative.

Now, 0 is a point at which $f^{(4)} = 0$ and $f^{(5)} > 0$. Thus, by the second derivative test, 0 is a point of local minimum for $f^{(3)}$. So, 0 is a point of inflection for $f^{(2)}$, by point (2) above. Thus, 0 is a point of local minimum for f', by point (1) above. Thus, 0 is a point of inflection for f, by point (2) above.

So, the upshot of this is the alternating behavior between derivatives.

1.7. **Higher derivative tests.** The discussion above gives a practical criterion to simply use evaluation of derivatives to determine whether a critical point, where a function is infinitely differentiable, is a point of local maximum, point of local minimum, point of inflection, or none of these.

Suppose f is an infinitely differentiable function around a critical point c for f.. Let k be the smallest integer for which $f^{(k)}(c) \neq 0$ and let L be the nonzero value of the k^{th} derivative. Then:

(1) If k is odd, then c is a point of inflection for f and hence neither a point of local maximum nor a point of local minimum.
(2) If k is even and $L > 0$, then c is a point of local minimum for f.
(3) If k is even and $L < 0$, then c is a point of local maximum for f.

For instance, for the power function $f(x) := x^n, n \geq 2$. 0 is a critical point and f is infinitely differentiable. In this case, $c = 0$, $k = n$, and $L = n! > 0$. Thus, if n is even, then f does attain a local minimum at 0. if n is odd, 0 is a point of inflection. In this simple situation, we could have deduced this directly from the first derivative test – for n even, the first derivative changes sign fron negative to positive at 0, because $n - 1$ is odd. For n odd, $n - 1$ is even, so the first derivative has positive sign on both sides of 0. However, the good news is that this general method is applicable for other situations where the first derivative test is harder to apply.

1.8. **Notion of concave up and concave down for one-sided differentiable.** So far, we have defined the notion of concave up and concave down on an interval assmuing the function is differentiable everywhere on the interval. In higher mathematics, a somewhat more general definition is used, and this makes sense for functions that have one-sided derivatives everywhere.

Note: Please, please, please, please make sure you understand this clearly: we can calculate the left-hand derivative and right-hand derivative using the formal expressions *only* after we have checked that the function is continuous from that side! If there is a piecewise description of the function and it is not continuous from one side where the definition is changing, then the corresponding one-sided derivative is *not*, repeat *not* defined.

Suppose f is a function on an interval $I = (a, b)$ such that both the left-hand derivative and the right-hand derivative of f are defined everywhere on I. Note that *both* one-sided derivatives being defined at every point in particular means that the function is continuous at each point, and hence on I. However, f need not be differentiable at every point, because it is possible that the left-hand derivative and right-hand derivative differ at different points.

We then say that:

(1) f is concave up if, at every point, the right-hand derivative is greater than or equal to the left-hand derivative, and both one-sided derivative are increasing functions on $\mathbb{R}$.
(2) f is concave down if, at every point, the right-hand derivative is less than or equal to the left-hand derivative, and both one-sided derivatives are decreasing functions on $\mathbb{R}$.

For instance, first consider the function f on $(0,\infty)$:

$$f(x) := \{ \begin{matrix} x^2, & 0 < x \le 1 \\ x^3, & 1 < x \end{matrix}$$

Before we proceed further, we check/note that the function is continuous at 1. Indeed it is. Hence, to calculate the left-hand derivative and right-hand derivative at 1, we can formally differentiate the expressions at 1. We obtain that the left-hand derivative at 1 is $2 \cdot 1 = 2$ and the right-hand derivative at 1 is $3 \cdot 1^2 = 3$. We thus obtain the following piecewise definitions for the left-hand derivative and right-hand derivative:

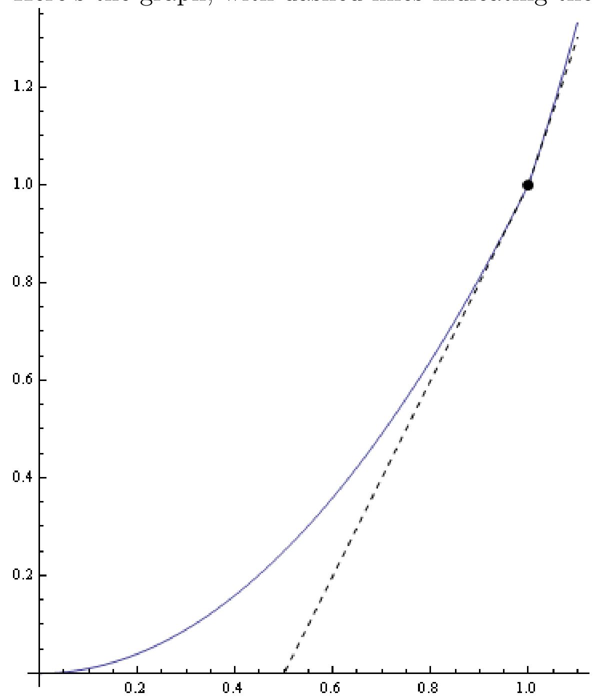
$$\text{LHD of } f \text{ at } x = \{ \begin{matrix} 2x, 0 < x \le 1 \\ 3x^2, \quad 1 < x \end{matrix}$$

and:

$$\text{RHD of } f \text{ at } x = \{ \begin{matrix} 2x, 0 < x < 1 \\ 3x^2, \quad 1 \le x \end{matrix}$$

The derivative is undefined at 1. Note that both one-sided derivatives are increasing everywhere, and at the point 1, where the function is not differentiable, the right-hand derivative is bigger. Thus, the function is concave up on $(0,\infty)$.

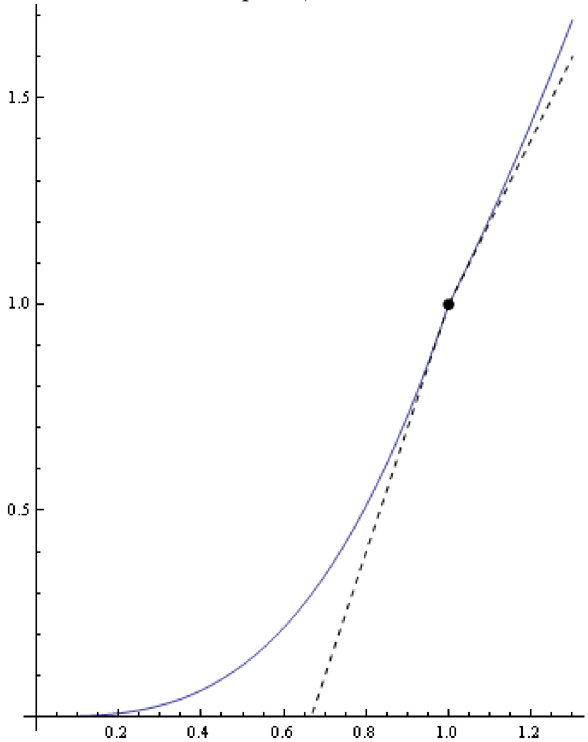
Here's the graph, with dashed lines indicating the one-sided derivatives:



On the other hand, consider the function g on $(0,\infty)$:

$$g(x) := \{ \begin{matrix} x^3, & 0 < x \le 1 \\ x^2, & 1 < x \end{matrix}$$

The function g is continuous and has one-sided derivatives everywhere. Also note that on the intervals $(0,1)$ and $(1,\infty)$, g is concave up. However, at the critical point 1 where g' is undefined, the right-hand derivative is smaller than the left-hand derivative. Thus, the function is not concave up overall on $(0,\infty)$, because at the critical point, its rate of increase takes a plunge for the worse. Here's the picture of g:



1.9. **Graphical properties of concave functions.** Here are some properties of the graphs of functions that are concave up, which are particularly important in the context of optimization. You should be able to do suitable role changes and obtain corresponding properties for concave down functions. In all the points below, f is a continuous function on an interval $[a,b]$ and is concave up on the interior (a,b).

(1) The only possibilities for the increase-decrease behavior of f are: increasing throughout, decreasing throughout, or decreasing first and then increasing.
(2) In particular, f either has exactly one local minimum or exactly one endpoint minimum, and this local or endpoint minimum is also the absolute minimum.
(3) Also, f cannot have a local maximum in its interior. It has exactly one endpoint maximum, and this is also the absolute maximum.
(4) For any two points x_1, x_2 in the domain of f, the part of the graph of f between $(x_1, f(x_1))$ and $(x_2, f(x_2))$ lies below the chord joining the points $(x_1, f(x_1))$ and $(x_2, f(x_2))$.
(5) If we assume that f is differentiable on (a,b), the tangent line through any point $(x, f(x))$ for $a < x < b$ does *not* intersect the curve at any other point. In the more general notion where f has one-sided derivatives, both the left and right tangent line satisfy this property.

For concave down functions, the role of minimum and maximum gets interchanged, and in point (3) above, the graph is now above the chord rather than below.

1.10. **Addendum: concave and convex.** The book uses the terminology *concave up* and *concave down*, but it's worth knowing that in much of mathematics as well as applications of mathematics, the term *convex* is used for concave up and the term *concave* is used for concave down. However, there is some confusion about this since some people use *concave* for concave up and *convex* for concave down.

2. Infinity and asymptotes

2.1. **Limits to infinity and vertical asymptotes.** We have already discussed what it means to say $\lim_{x \to c} f(x) = +\infty$, but here's a friendly reminder. It means that as x comes closer and closer to c (from either side), $f(x)$ goes above every finite value and does not then come back down. You can similarly understand what it means to say that $\lim_{x \to c} f(x) = -\infty$. You should also be able to understand the *one-sided versions* of these concepts: $\lim_{x \to c^-} f(x) = \infty$, $\lim_{x \to c^+} f(x) = \infty$, $\lim_{x \to c^-} f(x) = -\infty$, and $\lim_{x \to c^+} f(x) = -\infty$.

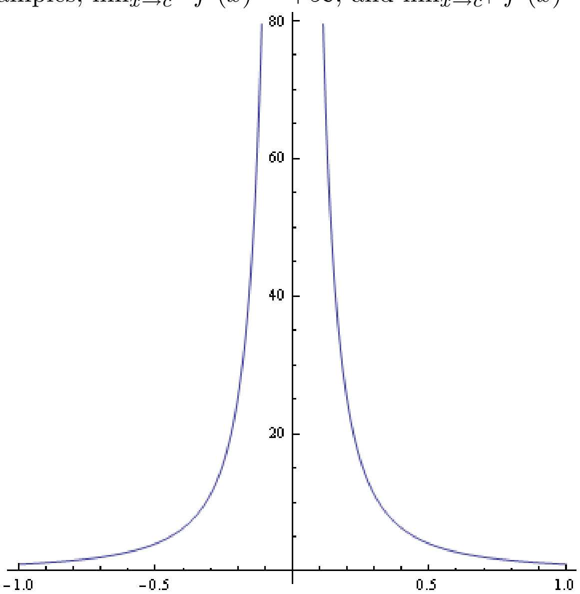
Now, here's a little guesswork question: if $\lim_{x \to c^-} f(x) = \infty$, what can you say about $\lim_{x \to c^-} f'(x)$? As a general rule, nothing, but in most of the situations that we see, it turns out that $f'(x)$ also approaches $+\infty$[1]. The reason is easy to see graphically: for $f(x)$ to head to $+\infty$ as x approaches a finite value, f needs to climb faster and faster and faster.

Similarly, it is usually the case that if $\lim_{x \to c^-} f(x) = -\infty$, then $\lim_{x \to c^-} f'(x) = -\infty$. Also, if $\lim_{x \to c^+} f(x) = \infty$, then it is likely that $\lim_{x \to c^+} f'(x) = -\infty$ (because it has to drop very quickly down from infinity) and if $\lim_{x \to c^+} f(x) = -\infty$, then $\lim_{x \to c^+} f'(x) = \infty$ (because it has to rise very quickly back up from $-\infty$). Again, these things are not always true, but for most of the typical examples you'll see, they will be.

Now here's the meaning of *vertical asymptote*. If $\lim_{x \to c^-} f(x) = \pm\infty$ and/or $\lim_{x \to c^+} f(x) = \pm\infty$, then the line $x = c$ is termed a *vertical asymptote* for f. This is because the graph of f is approaching the vertical line $x = c$. In some sense, if we think of $f(c) = +\infty$ or $-\infty$ as the case may be, the vertical line becomes the tangent line to the curve at that infinite point.

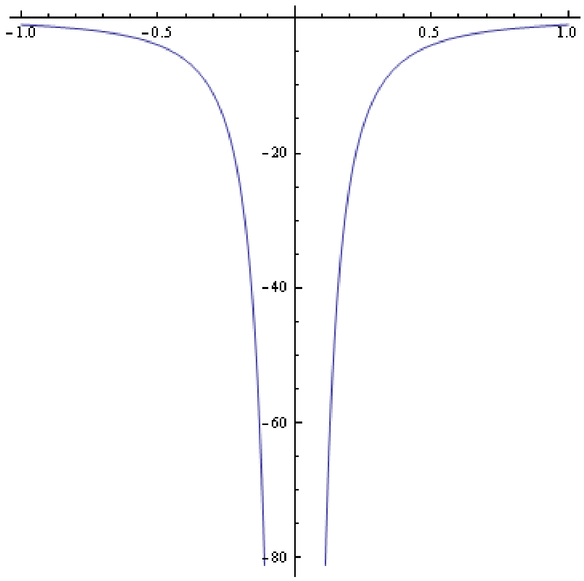
Some of the typical situations worth noting are:

(1) $\lim_{x \to c} f(x) = +\infty$ from both sides. An example of this is the function $f(x) = 1/x^2$ with $c = 0$. The vertical asymptote is the y-axis, i.e., the line $x = 0$. In this case, and in most other representative examples, $\lim_{x \to c^-} f'(x) = +\infty$, and $\lim_{x \to c^+} f'(x) = -\infty$.

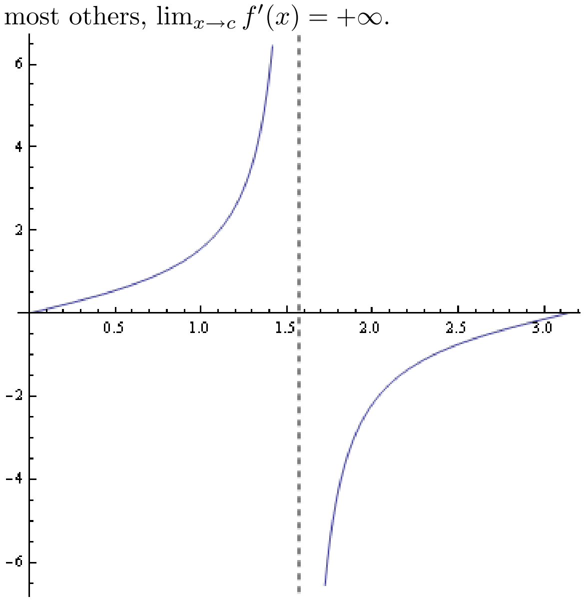


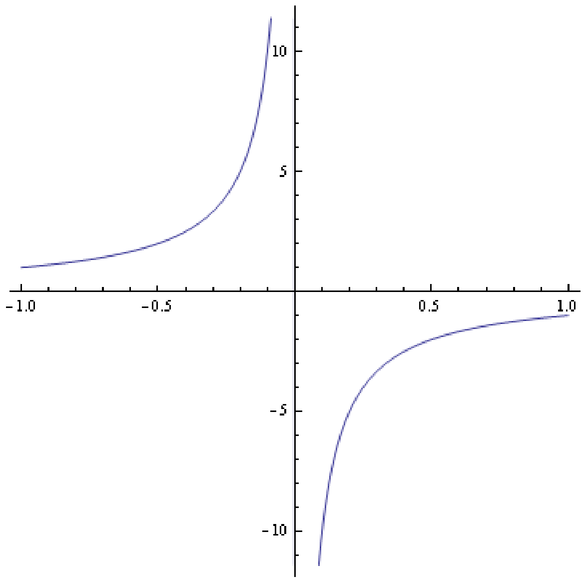
(2) $\lim_{x \to c} f(x) = -\infty$ from both sides. An example of this is the function $f(x) = -1/x^2$ with $c = 0$. The vertical asymptote is the y-axis, i.e., the line $x = 0$. In this case, and in most other representative examples, $\lim_{x \to c^-} f(x) = -\infty$, and $\lim_{x \to c^+} f(x) = +\infty$.

[1] More precisely, it turns out that if f' is continuous and *does* approach something, that something must be $+\infty$. However, there are weird examples where it oscillates

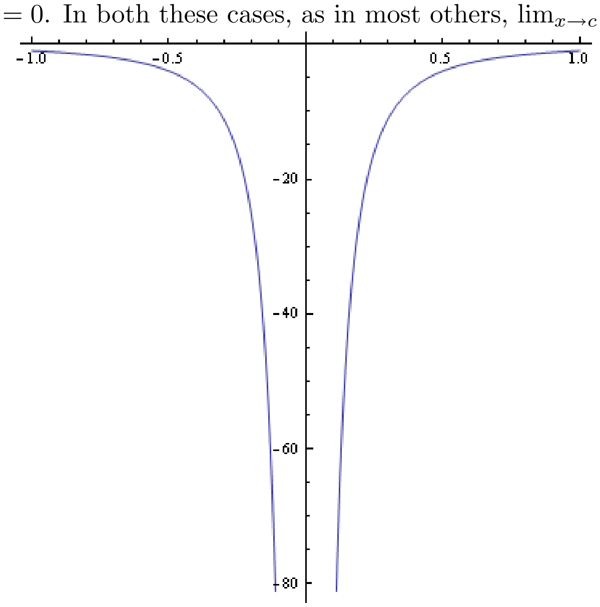


(3) $\lim_{x\to c^-} f(x) = \infty$ and $\lim_{x\to c^+} f(x) = -\infty$. Examples include $f = \tan$ at $c = \pi/2$ (vertical asymptote $x = \pi/2$) and $f(x) = -1/x$ at $c = 0$ (vertical asymptote $x = 0$). In both these cases, as in most others, $\lim_{x\to c} f'(x) = +\infty$.





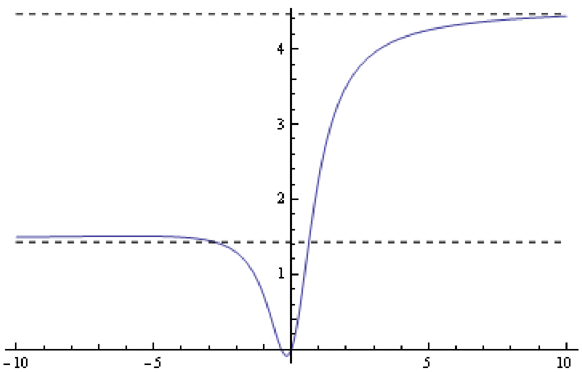
(4) $\lim_{x\to c^-} f(x) = -\infty$ and $\lim_{x\to c^+} f(x) = \infty$. Examples include $f = \cot$ at $c = 0$ and $f(x) := 1/x$ at $c = 0$. In both these cases, as in most others, $\lim_{x\to c} f'(x) = -\infty$.

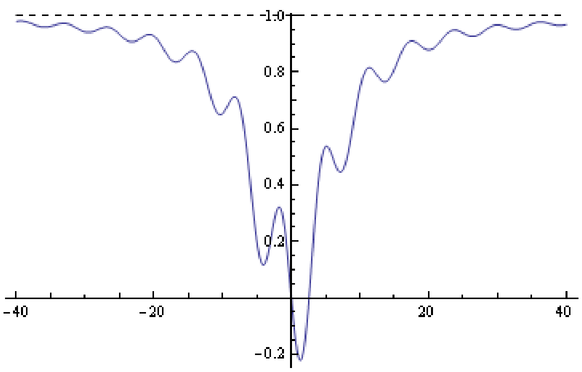


2.2. **Horizontal asymptotes.** Horizontal asymptotes are horizontal lines that the graph comes closer and closer to, just as vertical asymptotes are vertical lines that the graph comes closer and closer to.

We saw that vertical asymptotes arose when the *range value* was approaching $\pm\infty$ for a finite limiting value of the domain. Horizontal asymptotes arise where the *domain value* approaches $\pm\infty$ for a finite limiting value of the range.

Explicitly, if $\lim_{x\to\infty} f(x) = L$ (with L a finite number), then the line $y = L$ is a horizontal asymptote for the graph of f, because as $x \to \infty$, the graph comes closer and closer to this horizontal line. Similarly, if $\lim_{x\to-\infty} f(x) = M$, then the line $y = M$ is a horizontal asymptote for the graph of f. Thus, a function whose domain extends to infinity in both directions could have zero, one, or two horizontal asymptotes.

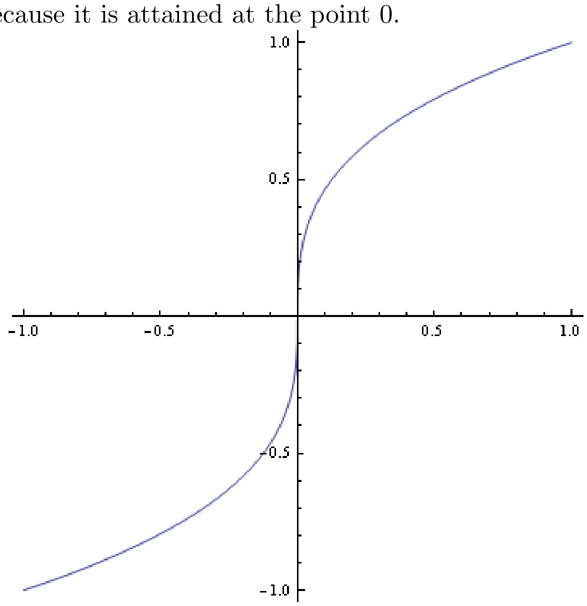




We will discuss some of the computational aspects of vertical and horizontal asymptotes in the problem sessions. Later in the lecture, and in the addendum, we look at some computational tips and guidelines over and above what is there in the book.

2.3. **Vertical tangents.** A vertical tangent to the graph of a function f occurs at a point $(c, f(c))$ if f is continuous but not differentiable at c, and $\lim_{x \to c} f'(x) = +\infty$ or $\lim_{x \to c} f'(x) = -\infty$. *It is important that the sign of infinity in the limit is the same from both the left and the right side.*

An example is the function $f(x) := x^{1/3}$ at the point $c = 0$. The function is continuous at 0. The derivative functions is $(1/3)x^{-2/3}$, and the limit of this as $x \to 0$ (from either side) is $+\infty$. Graphically, what this means is that the tangent is vertical. In this case, the vertical tangent coincides with the y-axis, because it is attained at the point 0.



Points of vertical tangent are points of inflection, as we can see from the $x^{1/3}$ example. Recall that the horizontal tangent case of the point of inflection was typified by x^3, and the general slogan was that the function slows down for an instant to speed zero. For vertical tangents, we can think of it as the function speeding up instantaneously to speed infinity before returning to the realm of finite speed.

It is important to note that the situation of a vertical tangent requires that the function itself be defined and continuous, and hence finite-valued, at the point. Thus, for instance, the function $f(x) := 1/x$ satisfies $\lim_{x \to 0} f'(x) = -\infty$ but does *not* have a vertical tangent at zero because the function is undefined at zero.

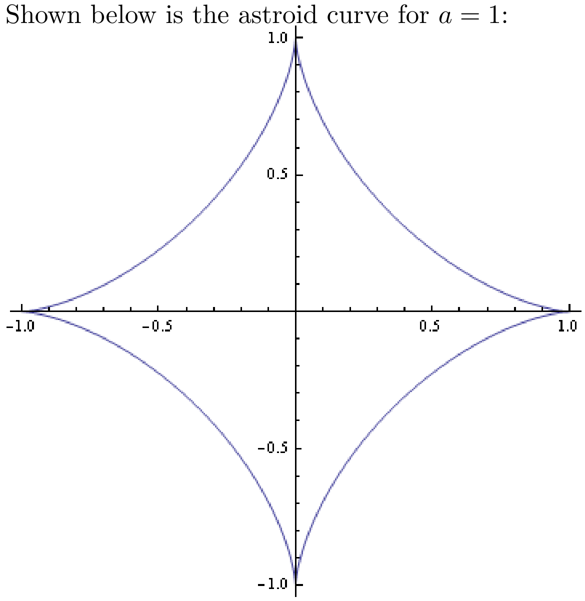
2.4. **Vertical cusps.** A vertical cusp in the graph of f occurs at a point c if f is continuous at c, and both one-sided limits of f' at c are infinities of *opposite* sign. There are two possibilities:

(1) The left-hand limit of the derivative is $+\infty$ and the right-hand limit of the derivative is $-\infty$. Then, $(c, f(c))$ is a point of local maximum. An example is $f(x) := -x^{2/3}$ and $c = 0$ What happens in this situation is the the graph has a sharp peak (picking up to speed infinity) at the point c, after which it rapidly starts dropping.

(2) The left-hand limit of the derivative is $-\infty$ and the right-hand limit of the derivative is $+\infty$. In this case we get a local minimum. An example is $f(x) := x^{2/3}$ and $c = 0$.

There is a special curve called the *astroid curve* (I had planned to put this on the homework, but it went on the chopping block when I needed to trim the homeworks to size), given by the equation $x^{2/3} + y^{2/3} = a^{2/3}$. This curve is not the graph of a function, since every value of x in $(-a, a)$ has two corresponding values of y. Nonetheless, the curve is a good illustration of the concept of cusps: there are two vertical cusps at the points $(0, a)$ and $(0, -a)$ respectively, and two horizontal cusps at the points $(a, 0)$ and $(-a, 0)$ respectively.

Shown below is the astroid curve for $a = 1$:



Note that for the graph of a function, the only kind of cusp that can occur is a vertical cusp, because a horizontal or oblique cusp would result in the curve intersecting a vertical line at multiple points, which would contradict the meaning of a function.

It is important to note that the situation of a vertical cusp requires that the function itself be defined and continuous, and hence finite-valued, at the point. Thus, for instance, the function $f(x) := 1/x^2$ satisfies $\lim_{x \to 0^-} f'(x) = +\infty$ and $\lim_{x \to 0^+} f'(x) = -\infty$ but does *not* have a vertical tangent at zero because the function is undefined at zero.

3. Computational aspects

3.1. **Computing limits at infinity: a review.** We review the main results that you have probably seen and add some more:

(1) $(\to \infty)(\to \infty) = \to \infty$. In other words, if $\lim_{x \to c} f(x) = \infty$ and $\lim_{x \to c} g(x) = \infty$, then $\lim_{x \to c} f(x)g(x) = \infty$. c could be finite or $\pm\infty$ here, and we could take one-sided limits instead.

(2) $(\to \infty)(\to -\infty) =\to -\infty$, and $(\to -\infty)(\to -\infty) =\to \infty$.
(3) $(\to a)(\to \infty) =\to \infty$ if $a > 0$ and $(\to a)(\to \infty) =\to -\infty$ if $a < 0$. Similarly, $(\to a)(\to -\infty) =\to -\infty$ if $a > 0$ and $(\to a)(\to -\infty) =\to \infty$ if $a < 0$.
(4) The previous point can be generalized somewhat: $(\to \infty)$, times a function that eventually has a positive lower bound (even if it keeps oscillating), is also $\to \infty$. Analogous results hold for negative upper bounds.
(5) $(\to 0)(\to \infty)$ is an indeterminate form: it is not clear what it tends to without doing more work.
(6) $(\to \infty) + (\to \infty) =\to \infty$.
(7) $(\to \infty) + (\to a) =\to \infty$ where a is finite. More generally, $\to \infty$ plus anything that is bounded from below is also $\to \infty$.
(8) $(\to \infty) - (\to \infty)$ and $(\to \infty) + (\to -\infty)$ are indeterminate forms.

Apart from this, the main facts you need to remember are that if $a > 0$, then $\lim_{x\to\infty} x^a = \infty$ and $\lim_{x\to\infty} x^{-a} = \lim_{x\to-\infty} x^{-a} = 0$. Note that this holds regardless of whether a is an integer.

When a is an odd integer or a rational number with odd numerator and odd denominator, $\lim_{x\to-\infty} x^a = -\infty$. When a is an even integer or a rational number with even numerator and odd denominator, $\lim_{x\to-\infty} x^a = \infty$.

Also worth noting: $\lim_{x\to 0^+} x^{-a} = \infty$ for $a > 0$ and $\lim_{x\to 0^-} x^{-a} = \lim_{x\to-\infty} x^a$, which is computed by the rule above.

We can use these facts to explain most of the limits involving polynomial and rational functions. Earlier, we had noted that when calculating the limit of a polynomial, it is enough to calculate the limit of its leading monomial. Let's now see why.

Consider the function $f(x) := x^7 - 5x^5 + 3x + 2$. Then, we can write $f(x) = x^7\left[1 - 5x^{-2} + 3x^{-6} + 2x^{-7}\right]$. The expression on the inside is 1 plus various negative powers of x. Each of those negative powers of x goes to 0 as $x \to \infty$. So, we obtain:

$$\lim_{x\to\infty}[1 - 5x^{-2} + 3x^{-6} + 2x^{-7}] = 1$$

We also have $\lim_{x\to\infty} x^7 = \infty$. Thus, the limit of the product is ∞.

Let's now consider an example of a rational function:

$$\frac{9x^3 - 3x + 2}{103x^2 - 17x - 99}$$

Earlier, we had discussed that when computing such limits at $\pm\infty$, we can simply calculate thel imits of the leading terms and ignore the rest. We now have a better understanding of the rationale behind this. Formally:

$$\begin{aligned}
&\lim_{x\to\infty} \frac{9x^3 - 3x + 2}{103x^2 - 17x - 99} \\
=&\lim_{x\to\infty} \frac{x^3(9 - 3x^{-2} + 2x^{-3})}{x^2(103 - 17x^{-1} - 99x^{-2})} \\
=&\lim x \to \infty x \lim_{x\to\infty} \frac{9 - 3x^{-2} + 2x^{-3}}{103 - 17x^{-1} - 99x^{-2}} \\
=&\lim_{x\to\infty} x \cdot \frac{9}{103} \\
=&+\infty
\end{aligned}$$

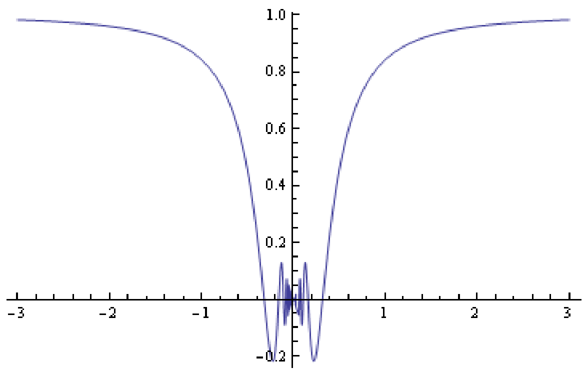
More generally, we see that if the degree of the numerator is greater than the degree of the denominator, the fraction approaches $\pm\infty$ as $x \to \pm\infty$, with the sign depending on the signs of the leading coefficients and the parity (even versus odd) of the exponents.

If the numerator and denominator have equal degree, the limit is a finite number. For $x \to \pm\infty$, it is the ratio of the leading coefficients (notice that it is the same on both sides). This is the case where we get horizontal asymptotes. In this case, the horizontal asymptotes on both ends coincide.

For instance:

$$\begin{aligned}
&\lim_{x\to-\infty} \frac{2x^2-3x+5}{23x^2-x-1} \\
=&\lim_{x\to-\infty} \frac{x^2(2-3x^{-1}+5x^{-2})}{x^2(23-x^{-1}-x^{-2})} \\
=&\lim_{x\to-\infty} \frac{x^2}{x^2} \lim_{x\to-\infty} \frac{2-3x^{-1}+5x^{-2}}{23-x^{-1}-x^{-2}} \\
=&1 \cdot \frac{2}{23} \\
=&\frac{2}{23}
\end{aligned}$$

Finally, when the degree of the numerator is less than the degree of the denominator, then the fraction tends to 0 as $x \to \infty$ and also tends to 0 as $x \to -\infty$. Thus, in this case, we get the x-axis as the horizontal asymptote on both sides.



3.2. **The $1/x$ substitution trick.**

Consider the limit:

$$\lim_{x\to\infty} x \sin(1/x)$$

This limit cannot be computed by plugging in values, because $x \to \infty$ and $1/x \to 0$, so $\sin(1/x) \to 0$, and we get the indeterminate form $(\to \infty)(\to 0)$. The approach we use here is to set $t = 1/x$. As $x \to \infty$, $t \to 0^+$. Since $t = 1/x$, we get $x = 1/t$. Plugging in, we get:

$$\lim_{t\to 0^+} \frac{\sin t}{t}$$

This limit is 1, as we know well.

Note that with this general substitution, limits to infinity correspond to right-hand limits at 0 for the reciprocal and limits at $-\infty$ correspond to left-hand limits at 0 for the reciprocal. If there is a two-sided limit at 0 for the reciprocal, the limits at $\pm\infty$ are the same. In fact, in the $x \sin(1/x)$ example, the limits at ∞ and $-\infty$ are both 1 since $\lim_{t\to 0} \sin t/t = 1$.

3.3. **Difference of square roots.** Consider the limit:

$$\lim_{x\to\infty} (\sqrt{x+1} - \sqrt{x})$$

There are many ways to compute this limit, but the easiest is to use the general $1/x$ substitution trick. Let $t = 1/x$. Then the above limit becomes:

$$\lim_{t\to 0^+} \frac{\sqrt{t+1} - 1}{\sqrt{t}}$$

This is an indeterminate form (specifically, a $0/0$ form). However, we can do the rationalization trick and rewrite this as:

$$\lim_{t \to 0^+} \frac{t}{\sqrt{t}(\sqrt{t+1}+1)}$$

The t and $\sqrt{t}$ cancel to give a $\sqrt{t}$ in the numerator, and we can evaluate and find the limit to be 0. A similar approach can be used to handle, for instance, a difference of cube roots.

3.4. **Combinations of polynomial and trigonometric functions.** We illustrate using some examples:

(1) Consider the function $f(x) := x + 25\sin x$. As $x \to \infty$, this is the sum of a function that tends to infinity and a function that oscillates. The oscillating component, however, has a finite lower bound, and hence, $\lim_{x\to\infty} f(x) = \infty$. Similarly, $\lim_{x\to-\infty} f(x) = -\infty$.

(2) Consider the function $f(x) := x \sin x$. As $x \to \infty$, this is the product of a function that goes to ∞ and a function that oscillates between -1 and 1. The oscillating part causes the sign of the whole expression to shift, and so as $x \to \infty$, $f(x)$ is oscillating with an ever-increasing magnitude of oscillation. A similar observation holds for $x \to -\infty$.

(3) Consider the function $f(x) := x(3 + \sin x)$ As $x \to \infty$, this is the product of a function that tends to ∞ and a function that oscillates between 2 and 4. The important point here is that the latter oscillation has a *positive lower bound*, so the product still tends to ∞.

(4) Consider the function $f(x) := x\sin(1/x)$. As $x \to \infty$, this is the product of a function that tends to ∞ and a function that tends to 0, so it is an indeterminate form. We already discussed above how this particular indeterminate form can be handled.

GRAPHING

MATH 152, SECTION 55 (VIPUL NAIK)

Corresponding material in the book: Section 4.8

Difficulty level: Hard.

What students should definitely get: The main concerns in graphing a function, how to figure out what needs figuring out. It is important for students to go through all the graphing examples in the book and do more hands-on practice. Transformations of graphs. Quickly graphing constant, linear, quadratic graphs.

What students should hopefully get: How all the issues of symmetry, concavity, inflections, periodicity, and derivative signs fit together in the grand scheme of graphing. The qualitative characteristics of polynomial function and rational function graphs, as well as graphs involving a mix of trigonometric and polynomial functions.

Weird feature: Ironically, there are very few pictures in this document. The naive explanation is that I didn't have time to add many pictures. The more sophisticated explanation is that since the purpose here is to review how to graph functions, having actual pictures drawn perfectly is counterproductive. Please keep a paper and pencil handy and sketch pictures as you feel the need.

Executive summary

0.1. **Symmetry yet again.** Words...

(1) All mathematics is the study of symmetry (well, not all).
(2) One interesting kind of symmetry that we often see in the graph of a function is *mirror symmetry* about a vertical line. This means that the graph of the function equals its reflection about the vertical line. If the vertical line is $x = c$ and the function is f, this is equivalent to asserting that $f(x) = f(2c - x)$ for all x in the domain, or equivalently, $f(c + h) = f(c - h)$ whenever $c + h$ is in the domain. In particular, the domain itself must be symmetric about c.
(3) A special case of mirror symmetry is the case of an *even function*. An even function is a function with mirror symmetry about the y-axis. In other words, $f(x) = f(-x)$ for all x in the domain. (Even also implies that the domain should be symmetric about 0).
(4) Another interesting kind of symmetry that we often see in the graph of a function is *half-turn symmetry* about a point on the graph. This means that the graph equals the figure obtained by rotating it by an angle of π about that point. A point (c, d) is a point of half-turn symmetry if $f(x) + f(2c - x) = 2d$ for all x in the domain. In particular, the domain itself must be symmetric about c. If f is defined at c, then $d = f(c)$.
(5) A special case of half-turn symmetry is an odd function, which is a function having half-turn symmetry about the origin.
(6) Another symmetry is *translation symmetry*. A function is *periodic* if there exists $h > 0$ such that $f(x + h) = f(x)$ for all x in the domain of the function (in particular, the domain itself should be invariant under translation by h). If a smallest such h exists, then such an h is termed the period of f.
(7) A related notion is that of a function with *periodic derivative*. If f is differentiable for all real numbers, and f' is periodic with period h, then $f(x + h) - f(x)$ is constant. If this constant value is k, then the graph of f has a two-dimensional translational symmetry by (h, k) and its multiples.

Cute facts...

(1) Constant functions enjoy mirror symmetry about every vertical line and half-turn symmetry about every point on the graph (can't get better).

(2) Nonconstant linear functions enjoy half-turn symmetry about every point on their graph. They do not enjoy any mirror symmetry because they are everywhere increasing or everywhere decreasing.
(3) Quadratic (nonlinear) functions enjoy mirror symmetry about the line passing through the vertex (which is the unique absolute maximum/minimum, depending on the sign of the leading coefficient). They do not enjoy any half-turn symmetry.
(4) Cubic functions enjoy half-turn symmetry about the point of inflection, and no mirror symmetry. Either the first derivative does not change sign anywhere, or it becomes zero at exactly one point, or there is exactly one local maximum and one local minimum, symmetric about the point of inflection.
(5) Functions of higher degree do not necessarily have either half-turn symmetry or mirror symmetry.
(6) More generally, we can say the following for sure: a nonconstant polynomial of even degree greater than zero can have at most one line of mirror symmetry and no point of half-turn symmetry. A nonconstant polynomial of odd degree greater than one can have at most one point of half-turn symmetry and no line of mirror symmetry.
(7) If a function is continuously differentiable and the first derivative has only finitely many zeros in any bounded interval, then the intersection of its graph with any vertical line of mirror symmetry is a point of local maximum or local minimum. The converse does not hold, i.e., points where local extreme values are attained do *not* usually give axes of mirror symmetry.
(8) If a function is twice differentiable and the second derivative has only finitely many zeros in any bounded interval, then any point of half-turn symmetry is a point of inflection. The converse does not hold, i.e., points of inflection do *not* usually give rise to half-turn symmetries.
(9) The sine function is an example of a function where the points of inflection and the points of half-turn symmetry are the same: the multiples of π. Similarly, the points with vertical axis of symmetry are the same as the points of local extrema: odd multiples of $\pi/2$.
(10) For a periodic function, any translate by a multiple of the period of a point of half-turn symmetry is again a point of half-turn symmetry. (In fact, any translate by a multiple of half the period is also a point of half-turn symmetry).
(11) For a periodic function, any translate by a multiple of the period of an axis of mirror symmetry is also an axis of mirror symmetry. (In fact, translation by multiples of half the period also preserve mirror symmetry).
(12) A polynomial is an even function iff all its terms have even degree. Such a polynomial is termed an *even polynomial*. A polynomial is an odd function iff all its terms have odd degree. Such a polynomial is termed an *odd polynomial*.
(13) Also, the derivative of an even function (if it exists) is odd; the derivative of an odd function (if it exists) is even.

Actions ...

(1) Worried about periodicity? Don't be worried if you only see polynomials and rational functions. Trigonometric functions should make you alert. Try to fit in the nicest choices of period. Check if smaller periods can work (e.g., for $\sin^2$, the period is π). Even if the function in and of itself is not periodic, it might have a periodic derivative or a periodic second derivative. The sum of a linear function and a periodic function has periodic derivative, and the sum of a quadratic function and a periodic function has a periodic second derivative.
(2) Want to milk periodicity? Use the fact that for a periodic function, the behavior everywhere is just the behavior over one period translates over and over again. If the first derivative is periodic, the increase/decrease behavior is periodic. If the second derivative is periodic, the concave up/down behavior is periodic.
(3) Worried about even and odd, and half-turn symmetry and mirror symmetry? If you are dealing with a quadratic polynomial, or a function constructed largely from a quadratic polynomial, you are probably seeing some kind of mirror symmetry. For cubic polynomials and related constructions, think half-turn symmetry.
(4) Use also the cues about even and odd polynomials.

0.2. **Graphing a function.** Actions ...

(1) To graph a function, a useful first step is finding the domain of the function.

(2) It is useful to find the intercepts and plot a few additional points.
(3) Try to look for symmetry: even, odd, periodic, mirror symmetry, half-turn symmetry, and periodic derivative.
(4) Compute the derivative. Use that to find the critical points, the local extreme values, and the intervals where the function increases and decreases.
(5) Compute the second derivative. Use that to find the points of inflection and the intervals where the function is concave up and concave down.
(6) Look for vertical tangents and vertical cusps. Look for vertical asymptotes and horizontal asymptotes. For this, you may need to compute some limits.
(7) Connect the dots formed by the points of interest. Use the information on increase/decrease and concave up/down to join these points. To make your graph a little better, compute the first derivative (possibly one-sided) at each of these points and start off your graph appropriately at that point.

Subtler points... (see the "More on graphing" notes for an elaboration of these points; not all of them were covered in class):

(1) When graphing a function, there may be many steps where you need to do some calculations and solve equations and you are unable to carry them out effectively. You can skip some of the steps and come back to them later.
(2) If you cannot solve an equation exactly, try to approximate the locations of roots using the intermediate value theorem or other results such as Rolle's theorem.
(3) In some cases, it is helpful to graph multiple functions together, on the same graph. For instance, we may be interested in graphing a function and its second and higher derivatives. There are other examples, such as graphing a function and its translates, or a function and its multiplicative shifts.
(4) A graph can be used to suggest things about a function that are not obvious otherwise. However, the graph should not be used as conclusive evidence. Rather, the steps used in drawing the graph should be retraced and used to give an algebraic proof.
(5) We are sometimes interested in sketching curves that are not graphs of functions. This can be done by locally expressing the curve piecewise as the graph of a function. Or, we could use many techniques similar to those for graphing functions.
(6) For a function with a piecewise description, we plot each piece within its domain. At the points where the definition changes, determine the one-sided limits of the function and its first and second derivatives. Use this to make the appropriate open circles, asymptotes, etc.

1. Graphing in general

The goal of this lecture is to make you more familiar with the tools and techniques that can be used to graph a function. The book has a list of points that you should keep in mind. The list in the book isn't complete – there are a number of additional points that tend to come up for functions of particular kinds, but it is a good starting point. But in this lecture, we'll focus on something more than just the techniques – we'll focus on the broad picture of why we want to draw graphs and what information about the function we want the graph to convey. Working from that, we will be able to reconstruct much of the book's strategy.

1.1. Graphs – utility, sketching and plotting.

The graph of a function f on a subset of the real numbers is the set of points in $\mathbb{R}^2$ (the plane) of the form $(x, f(x))$, where x is in the domain of f. The graph of f gives a geometric description of f, and it completely determines f. For a given $x = x_0$, $f(x_0)$ is the y-coordinate of the unique point of the graph that is also on the line $x = x_0$.

Graphs are useful because they allow us to see many things about the function at the same time, and enable us to use our visual instincts to answer questions about the function. It is usually easy to look at the graph and spot, without precise measurement, phenomena such as periodicity, symmetry, increase, decrease, discontinuity, change in direction, etc. Thus, the graph of a function, *if correctly drawn*, is not only equivalent in information content to the function itself, it makes that information content much more easy to read.

The problem is with the caveat *if correctly drawn*. The domains of most of the functions we consider are unions of intervals, so they contain infinitely many points. *Plotting the graph* in a complete sense would involve evaluating the function at these infinitely many points. In practice, *graph plotting* works by dividing the domain into very small intervals (say, of length 10^{-3}), calculating the values of the function (up to some level of accuracy, say 10^{-4}) at the endpoints of the intervals, and then drawing a curve that passes through all the graph points thus obtained. This last joining step is typically done using straight line segments.[1]

Unfortunately, although softwares such as Mathematica are good for plotting graphs, we humans would take too long to do the millions of evaluations necessary to plot graphs. However, we have another asset, which is our brains. We need to use our brains to find some substitute for plotting the graph that still gives a reasonable approximation of the graph and *captures the qualitative characteristics that make the graph such an informative representation of the function.* The process that we perform is called *graph sketching.*

A sketch of a graph is good if any information that is visually compelling from the sketch (without requiring precise measurement) is actually *correct* for the function. In other words, a good sketch may mislead people into thinking that $f(2) = 2.4$ while it is actually 2.5, but it should not make people think that f is increasing on the interval $(2, 3)$ if it is actually decreasing on the interval.[2]

1.2. **The domain of a function.** The domain of a function is easy to determine from its graph. Namely, the domain is the subset of the x-axis obtained by orthogonally projecting the graph onto the x-axis. In other words, it is the set of possible x-coordinates of points on the graph.

So, the first step in drawing the graph is finding the domain. We consider two main issues here:

(1) Sometimes, the domain may contain an open interval without containing one or both of the endpoints of that interval. In other words, there may be points in the boundary of the domain but not in the domain. In such cases, try to determine the limits (left and/or right, as applicable) of the function at these boundary points. If finite, we have open circles. If equal to $+\infty$ or $-\infty$, we have vertical asymptotes.
(2) In cases where the domain of the function stretches to $+\infty$ and/or $-\infty$, determine the limit(s). Any finite limit thus obtained corresponds to a horizontal asymptote.

Intercepts and a bit of plotting. So that the graph is not completely wrong, it is helpful to make it realistic using a bit of plotting. The book suggests computing the x-intercepts and the y-intercept.

The x-intercepts are the points where the graph intersects the x-axis, i.e., the points of the form $(x, 0)$ where $f(x) = 0$. There may be zero, one, or more than one x-intercepts. The y-intercept is the unique point where the graph intersects the y-axis, i.e., the point $(0, f(0))$. Note that if 0 is not in the domain of the function, then the y-intercept does not make sense.

In addition to finding the intercepts, it may also be useful to do a bit of plotting, e.g., finding $f(x)$ for some values of x, or finding solutions to $f(x) = y$ for a few values of y. The intercepts are the bare minimum of plotting. They're important to compute mainly because the values of the intercepts are visually obvious and it would be misleading to people viewing the graph if these values were obtained wrong.

1.3. **Symmetry/periodicity.** Another thing that is visually obvious from the graph is *patterns of repetition.* There are two kinds of patterns of repetition that we are interested in:

(1) *Periodicity*: The existence of $h > 0$ such that $f(x + h) = f(x)$ for all x in $\mathbb{R}$. Periodicity is graphically visible – the shape of the graph repeats after an interval of length h. Note that we can talk of periodicity even for functions that are not defined for all real numbers, as long as it is true that the domain itself is invariant under the addition of h. For instance, tan has a period of π.
(2) *Symmetry: even and odd*: An even function ($f(x) = f(-x)$ for all $x \in \mathbb{R}$) exhibits a particular kind of symmetry: symmetry about the y-axis. An odd function ($f(x) = -f(-x)$ for all $x \in \mathbb{R}$) exhibits *half-turn symmetry* about the origin. Both these properties are geometrically visible. Note that we can talk of even and odd for functions not defined for all real numbers, as long as the domain is symmetric about 0. For instance, $f(x) := 1/x$ is odd and $f(x) := 1/x^2$ is even.

There are somewhat more sophisticated versions of this:

[1] If you have seen computer graphics in the old days where computer memory and processing speed was limited, you would have seen that computer renderings of geometric figures such as circles was done using small line segments. As we improve the resolution, the line segments become smaller and smaller until our eyes cannot make out the difference.

[2] This does raise an interesting point, which is that the reason why sketches seem adequate even when inaccurate is because of our limited observational power – the correctly plotted graph would not look too different in terms of compelling visual information, and hence, we find the sketch good enough for our purposes.

(1) Periodicity with shift: This happens when there exists $h > 0$ and $k \in \mathbb{R}$ such that $f(x+h) = f(x)+k$ for all $x \in \mathbb{R}$. Thus, the graph of f repeats after an interval of length h, but it is shifted vertically by k. Note that the case of shift 0 is precisely the case where f itself is periodic. If f is also differentiable, this is equivalent to the derivative being periodic.

   A function is periodic with shift if and only if it is the sum of a periodic function and a linear function. The breakup as a sum is unique up to constants. The periodic function part can be thought of as representing the seasonal trend and the linear function part can be thought of as representing the secular trend.
(2) Half turn symmetry about axes other than the y-axis.
(3) Mirror symmetry about points other than the origin.

With the exception of *periodicity with shift*, all the other notions are discussed in detail in the second set of lecture notes on functions (Functions: A Rapid Review (Part 2)) so we will not repeat that discussion. Since the mirror symmetry and half turn symmetry material was not covered in class at the time, we'll take a short detour in class to cover that material.

1.4. **First derivative.** The next step in getting a better picture of the function is to use the derivative. The derivative helps us find the intervals on which the function is increasing and decreasing, the critical points, and other related phenomena. We shall return in some time to the application of this information to graph-sketching.

1.5. **Second derivative.** If the function is twice differentiable (at most points) the second derivative is another useful tool. We can use the second derivative to find intervals where the function is concave up, intervals where the function is concave down, and inflection points of the function. Combining this with information about the first derivative, we can determine intervals where the function is increasing and concave up (i.e., increasing at an increasing rate), increasing and concave down (i.e., increasing at a decreasing rate), decreasing and concave up (i.e., decreasing at a decreasing rate), or decreasing and concave down (i.e., decreasing at an increasing rate).

1.6. **Classifying and understanding points of interest.** Some of the cases of interest are:

(1) Point of discontinuity: Separately compute the left-hand limit, right-hand limit and value. If either one-sided limit is $\pm\infty$, we have a vertical asymptote. If a one-sided limit equals the value, the graph has a closed circle. If a one-sided limit exists but does not equal the value, the graph has an open circle.
(2) Critical point where the function is continuous and not differentiable: Determine whether the left-hand derivative and right-hand derivative individually exist. If so, determine the values of these derivatives. If the left-hand and right-hand derivatives do not exist as finite values, try determining the left-hand limit and right-hand limit of the derivative. If the limit of the derivative is $+\infty$ from both sides or $-\infty$ from both sides, we have a vertical tangent at the point. if the limit of the derivative is $+\infty$ from one side and $-\infty$ from the other side, we have a vertical cusp at the point. In all cases, determine the value of the function at the point.
(3) Critical points where the derivative of the function is zero: Determine whether this is a point of local maximum, a point of local minimum, a point of inflection, or none of these. In any case, determine the value of the function at the point.
(4) Point of inflection: Determine the value of the function as well as the value of the first derivative at the point. Also, determine whether the graph switches from concave up to concave down or concave down to concave up at the point.

Critical points, and phenomena related to the first derivative, are usually geometrically compelling, so it is important to focus on getting them right so as not to paint a misleading picture. The precise location of points of inflection is less geometrically compelling, except when such a point is also a critical point. Generally, it is geometrically clear that there exists an inflection point in the interval between two points, because the graph is concave up at one point and concave down at the other. However, the precise location of the critical point may be hard to determine. Thus, getting the precise details of inflection points correct is desirable but not as basic as getting the critical points correct.

1.7. **Sketching the graph.** We first plot the points of interest and values (including $\pm\infty$, corresponding to vertical asymptotes), as well as the horizontal asymptotes for points at infinity. Here, *points of interest* includes the critical points and inflection points, intercept points, and a few other points added in to get a preliminary plot. In addition to plotting the graph points (which is the pair $(x, f(x))$ where x is the point of interest in the domain), it is also useful to compute the one-sided derivatives at each of the points of interest, and draw a short segment of the tangent line (or half-line, if only one-sided derivatives exist) corresponding to that.

Next, we use the increase/decrase and concave up/down information, as well as the tangent half-lines, to make the portions of the graph between these points of interest. This is the step that involves some guesswork. The idea is that because we are sure that the main qualitative characteristics (increase versus decrease, concave up versus concave down) are correct, errors in further shape details are not a big problem.

Since these actual shapes are the result of guesswork, it is particularly important that the issues of symmetry and periodicity be taken into account while sketching. For a periodic function, it is better to have a *somewhat less accurate shape repeated faithfully in each period* than a number of different-looking shapes in different periods. Similar remarks apply for symmetry and even/odd functions.

The book has a number of worked out examples, and you should go through them. To keep your homework set of manageable size, I haven't included graph sketching problems in the portion of the homework to be submitted. But I have recommended a few graph sketching problems from the book's exercises and you should try these (and others if you want) and can check your answer against a graphing calculator or software.

2. Graphing particular functions

Here we discuss various simple classes of functions and how they can be graphed. For functions in these well behaved classes, we do not need to go through the entire rigmarole for graphing.

2.1. **Constant and linear functions.** We begin by looking at the constant function $f(x) := k$. This function is soporific, because you know the graph of the function is a straight horizontal line, the derivative of the function is zero everywhere, it is constant everywhere. Every point is a local minimum and a local maximum in the trivial sense. The limits $\lim_{x\to\infty} f(x)$ and $\lim_{x\to-\infty} f(x)$ are also both equal to k.

Next, we look at the linear function $f(x) := ax + b$ where $a \neq 0$. This function has graph a straight line. The tangent line at any point on the graph is the same straight line. The slope of the straight line is a. If $a > 0$, the function is increasing everywhere, and if $a < 0$, the function is decreasing everywhere. The derivative is the constant a and the second derivative is 0.

If $a > 0$, then $\lim_{x\to\infty} f(x) = +\infty$ and $\lim_{x\to-\infty} f(x) = -\infty$. If $a < 0$, then $\lim_{x\to\infty} f(x) = -\infty$ and $\lim_{x\to-\infty} f(x) = +\infty$.

2.2. **Quadratic functions.** Consider the function $f(x) := ax^2 + bx + c$, where $a \neq 0$. This is a quadratic function. The derivative function $f'(x)$ is equal to $2ax + b$, the second derivative $f''(x)$ is the constant function $2a$, and the third derivative is 0 everywhere. In other words, the slope of the tangent line to the graph of this function is not constant, but it is changing at a constant rate.

The graph of this function is called a *parabola*. We describe the graph separately for the cases $a > 0$ and $a < 0$.

In the case $a > 0$, we have $\lim_{x\to\infty} f(x) = \lim_{x\to-\infty} f(x) = \infty$. The function attains a local as well as an absolute minimum at the point $x = -b/2a$, and the value of the minimum is $(4ac - b^2)/4a$. The point $(-b/2a, (4ac - b^2)/4a)$ is termed the *vertex* of the parabola. f is decreasing on the interval $(-\infty, -b/2a]$ and increasing on the interval $[-b/2a, \infty)$. Also, the graph of f is symmetric (i.e., a *mirror symmetry*) about the vertical line $x = -b/2a$.

In the case $a < 0$, we have $\lim_{x\to\infty} f(x) = \lim_{x\to-\infty} f(x) = -\infty$. The function attains a local as well as an absolute maximum at the point $x = -b/2a$, and the value of the maximum is $(4ac - b^2)/4a$. The point $(-b/2a, (4ac - b^2)/4a)$ is termed the *vertex* of the parabola. The function is increasing on the interval $(-\infty, -b/2a]$ and decreasing on the interval $[-b/2a, \infty)$.

Finally, note the following about the existence of zeros, based on cases about the sign of the discriminant $b^2 - 4ac$:

(1) Case $b^2 - 4ac > 0$ or $b^2 > 4ac$: In this case, there are two zeros, and they are located symmetrically about $-b/2a$. If $a > 0$, the function f is positive to the left of the smaller root, negative between the roots, and positive to the right of the larger root. If $a < 0$, the function f is negative to the left of the smaller root, positive between the roots, and negative to the right of the larger root.
(2) Case $b^2 - 4ac = 0$ or $b^2 = 4ac$: In this case, $-b/2a$ is a zero of multiplicity two. The vertex is thus a point on the x-axis with the x-axis a tangent line to it. Note that if $a > 0$, the parabola lies in the upper half-plane and if $a < 0$, the parabola lies in the lower half-plane.
(3) Case $b^2 - 4ac < 0$ or $b^2 < 4ac$: In this case there are no zeros. If $a > 0$, the parabola lies completely in the upper half-plane, and if $a < 0$, the parabola lies completely in the lower half-plane.

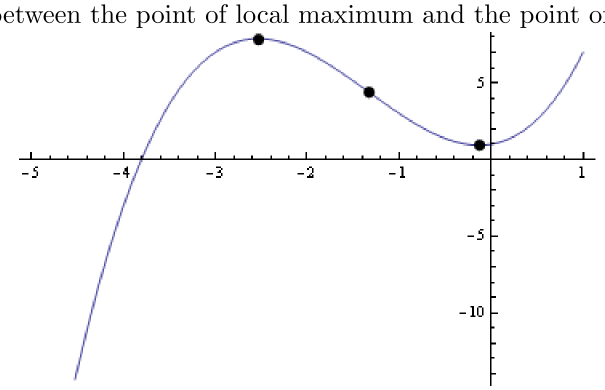
2.3. **Cubic functions.** We next look at the case of a cubic polynomial, $f(x) := ax^3 + bx^2 + cx + d$, where $a \neq 0$. We carry out the discussion assuming $a > 0$. In case $a < 0$, maxima and minima get interchanged and the sign of infinities on limits get flipped.[3]

So let's discuss the case $a > 0$. We have $\lim_{x\to-\infty} f(x) = -\infty$ and $\lim_{x\to\infty} f(x) = \infty$. Notice that, by the intermediate-value theorem, the cubic polynomial takes all real values. The derivative of the function is $f'(x) = 3ax^2 + 2bx + c$, the second derivative is $f''(x) = 6ax + 2b$, the third derivative is $f'''(x) = 6a$ and the fourth derivative is zero. This means that not only is the slope changing, but it is changing at a changing rate, but the rate at which that rate is changing isn't changing (yes, you read that right).

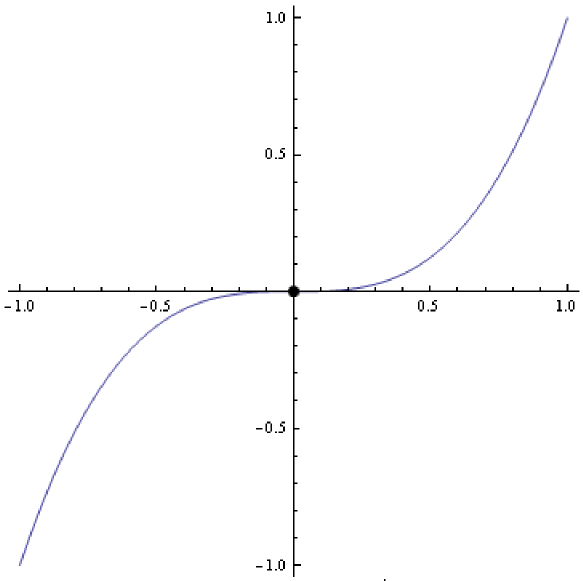
We now try to determine where the function has local maxima and minima, and where it is increasing or decreasing. For this, we need to first find the critical points. The critical points are solutions to $f'(x) = 0$. The discriminant of the quadratic polynomial f' is $4b^2 - 12ac$. We make three cases based on the sign of the discriminant.

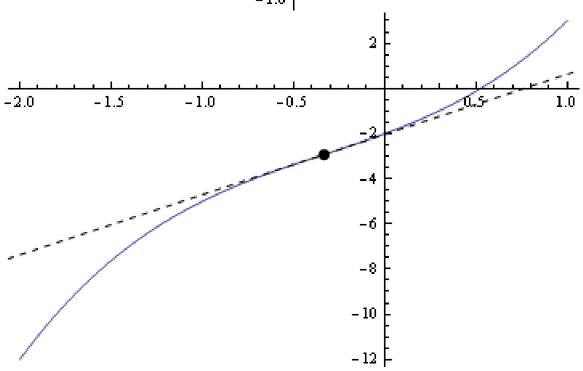
(1) $4b^2 - 12ac > 0$, or $b^2 > 3ac$: In this case, there are two critical points, given by the two solutions to the quadratic equation. We also see that, since $a > 0$, f' is positive to the left of the smaller root, negative between the two roots, and positive to the right of the larger root. Thus, f is increasing from $-\infty$ to the smaller root, decreasing between the two roots, and increasing from the larger root to ∞. The smaller root is thus a point of local maximum and the larger root is a point of local minimum.
(2) $4b^2 - 12ac = 0$, or $b^2 = 3ac$: In this case, there is one critical point, namely $-b/3a$. The function is increasing all the way through, so although this is a critical point, it is neither a local maximum nor a local minimum. In fact, it is a point of inflection, where both the first and the second derivative become zero.
(3) $4b^2 - 12ac < 0$, or $b^2 < 3ac$: In this case, the function has no critical points and is increasing all the way through.

Any cubic polynomial enjoys a half-turn symmetry about the point $(-b/3a, f(-b/3a))$, i.e., the graph is invariant under a rotation by π about this point. This center of half-turn symmetry is also the unique point of inflection for the graph. In the case that $b^2 > 3ac$, the point of half-turn symmetry is the exact midpoint between the point of local maximum and the point of local minimum.



[3]Another way of thinking of it is that we can first plot the graph by taking out a minus sign on the whole expression, then flip it about the x-axis.





2.4. **Polynomials of higher degree.** Here are some general guidelines to understanding polynomials of higher degree:

(1) The limits at $\pm\infty$ are determined by whether the polynomial has even or odd degree, and the sign of the leading coefficient. Positive leading coefficient and even degree mean a limit of $+\infty$ on both sides. Negative leading coefficient and even degree mean a limit of $-\infty$ on both sides. Positive leading coefficient and odd degree mean a limit of $+\infty$ as $x \to \infty$ and $-\infty$ as $x \to -\infty$. Negative leading coefficient and odd degree mean a limit of $+\infty$ as $x \to -\infty$ and $-\infty$ as $x \to \infty$.
(2) The points where the function could potentially change direction are the zeros of the first derivative. For a polynomial of degree n, there are at most $n-1$ of these points. For such a point, we can use the first-derivative test and/or second-derivative test to determine whether the point is a point of local maximum, local minimum, or a point of inflection. *Note that for polynomial functions, any critical point must be a point of local maximum, local minimum, or a point of inflection.* There are no other possibilities for polynomial functions, because the number of times the first and/or second derivative switch sign is finite, hence we cannot construct all those weird counterexamples involving oscillations when dealing with polynomial functions.
(3) Between any two zeros of the polynomial there exists at least one zero of the derivative (this follows from Rolle's theorem). This can help us bound the number of zeros of a polynomial using information we have about the number of zeros of the derivative of that polynomial.
(4) A polynomial of odd degree takes all real values, and in particular, intersects every horizontal line at least once.
(5) A polynomial of even degree and positive leading coefficient has an absolute minimum value, and takes all values greater than or equal to that absolute minimum value at least once. A polynomial

of even degree and negative leading coefficient has an absolute maximum value, and takes all values less than or equal to that absolute maximum value at least once.

2.5. **Rational functions: the many concerns.** A lot of things are going on with rational functions, so we need to think about them more carefully than we thought about polynomials.

Graphing the function requires putting these pieces together, each of which we have dealt with separately:

(1) Determine where the rational function is positive, negative, zero, and not defined.
(2) At the points where the rational function is not defined, determine the left-hand and right-hand limits. In most cases, these limits are $\pm\infty$. The exceptions are for cases such as the $FORGET$ function, defined as $FORGET(x) = x/x$, which is not defined at $x = 0$, but has a finite limit at that point. *These exceptions only occur in situations where the rational function as originally expressed is not in reduced form.*
(3) Determine the limits of the rational function at $\pm\infty$. Note that this depends on how the degrees of the numerator and denominator compare and the signs of the leading coefficients.
(4) Consider the derivative f', and do a similar analysis on the derivative. The regions where the derivative is positive are the regions where f is increasing. The regions where the derivative is negative are the regions where f is decreasing.
(5) Consider the second derivative f'', and do a similar analysis on this. Use this to find the regions where f is concave up and the regions where f is concave down.

We combine all of these to draw the graph of f. We can also use all this information to determine where the function attains its local maxima and local minima.

2.6. **Piecewise functions.** Let's now deal with functions that are piecewise polynomial or rational functions. We'll also use this occasion to discuss general strategies for handling functions with piecewise definitions.

First, we need a clear piecewise definition, i.e., a definition that gives a polynomial or rational function expression on each part of the domain. The original definition may not be in that form. Here are some things we need to do:

(1) Whenever the whole expression, or some component of it, is in the absolute value, we make cases based on whether the expression whose absolute value is being evaluated is positive or negative. The transitions usually occur either at points where the expression is not defined, or at points where the absolute value is zero.
(2) Whenever the expression involves something like $\max\{f(x), g(x)\}$, then we make cases based on whether $f(x) > g(x)$ or $f(x) < g(x)$. The transition occurs at points where $f(x) = g(x)$ or at points where one or both of f and g is undefined.

Once we have the definition in piecewise form, we can differentiate, with the rule being to use the formula for differentiating in each piece where we have the expression. If the function is continuous at the points where the definition changes, we can use these formal expressions to calculate the left-hand derivative and right-hand derivative. We can then combine all this information to get a comprehensive picture of the function.

2.7. **A max-of-two-functions example.** *Note*: This or a very similar example appeared in a past homework. You might want to revisit that homework problem.

Consider $f(x) := \max\{x - 1, \frac{x}{x+1}\}$. We first need to get a piecewise description of f. For this, we need to determine where $x - 1 > x/(x + 1)$ and where $x - 1 < x/(x + 1)$. This reduces to determining where $(x^2 - x - 1)/(x + 1)$ is positive, zero, and negative.

The expression is positive on $((1 + \sqrt{5})/2, \infty) \cup (-1, (1 - \sqrt{5})/2)$, negative on $((1 - \sqrt{5})/2, (1 + \sqrt{5})/2) \cup (-\infty, -1)$, zero at $(1 \pm \sqrt{5})/2$, and undefined at -1. Thus, we get that:

$$f(x) = \{ \begin{array}{ll} x - 1, & x \in ((1 + \sqrt{5})/2, \infty) \cup (-1, (1 - \sqrt{5})/2) \\ \frac{x}{x+1}, & x \in (-\infty, -1) \cup [(1 - \sqrt{5})/2, 1 + \sqrt{5}/2] \end{array}$$

Next, we want to determine the limits of f as $x \to \pm\infty$. Since the definition for $x > (1 + \sqrt{5})/2$ is $x - 1$, $\lim_{x\to\infty} f(x) = \lim_{x\to\infty} x - 1 = \infty$. On the other hand, the definition for $x < -1$ is $x/(x + 1)$, so

$\lim_{x \to -\infty} f(x) = \lim_{x \to -\infty} x/(x+1)$. This is a rational function where the numerator and denominator have equal degrees, and the leading coefficients are both 1, so the limit as $x \to -\infty$ is 1.

Next, we want to find out the left-hand limit and right-hand limit at the point $x = -1$. The definition from the left side is $x/(x+1)$. The denominator approaches 0 from the left side and the numerator approaches a negative number, so the quotient approaches $+\infty$. The right-hand limit is $\lim_{x \to -1} x - 1 = -2$.

Next, let us try to determine where the function is increasing and decreasing. For this, we need to differentiate the function on each interval.

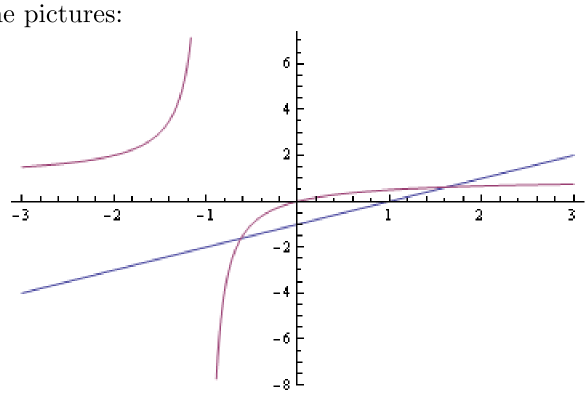
On the intervals $(-\infty, -1)$ and $[(1-\sqrt{5})/2, (1+\sqrt{5})/2]$, f is equal to $x/(x+1)$. The derivative is thus $1/(x+1)^2$, which is positive everywhere, and hence, in particular, on this region. Thus, f is increasing on $(-\infty, -1)$ as well as on $[(1-\sqrt{5})/2, (1+\sqrt{5})/2]$. On the intervals $(-1, (1-\sqrt{5})/2)$ and $(1+\sqrt{5}/2, \infty)$, f is defined as $x - 1$. The derivative is 1, so f is increasing on these intervals as well. In fact, since f is continuous at $1 + \sqrt{5}/2$, f is increasing on $[(1+\sqrt{5})/2, \infty)$. Combining all this information, we obtain that f is increasing on $(-\infty, -1)$ and on $(-1, \infty)$.

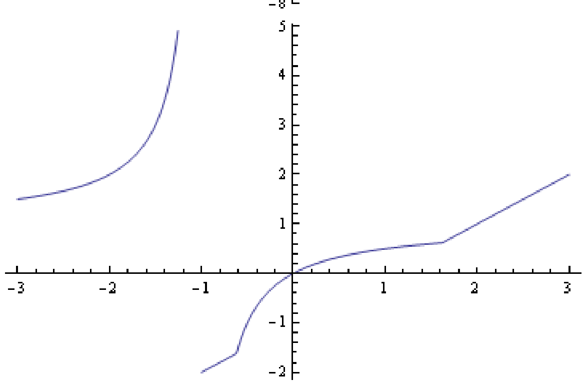
We can now understand and graph f better. As $x \to -\infty$, $f(x) \to 1$, and as $x \to -1^-$, $f(x) \to +\infty$. Thus, on the interval $(-\infty, -1)$, f increases from 1 to ∞. Since the right-hand limit at -1 is -2 and the limit at ∞ is ∞, we see that on the interval $(-1, \infty)$, f increases from -2 to ∞. There are two intermediate points where the definition changes: $(1 \pm \sqrt{5})/2$. From -1 to $(1-\sqrt{5})/2$, f increases from -2 to $(-1-\sqrt{5})/2$ in a straight line. Between $(1-\sqrt{5})/2$ and $(1+\sqrt{5})/2$, f increases from $(-1-\sqrt{5})/2$ to $(-1+\sqrt{5})/2$, but not in a straight line. From $(1+\sqrt{5})/2$ onward, f increases in a straight line again.

The critical points are $(1 \pm \sqrt{5})/2$. Neither of these is a local minimum or a local maximum. There is no absolute maximum, because the left-hand limit at -1 is ∞, so the function takes arbitrarily large positive values.

The function does not take arbitrarily small values. In fact, a lower bound on the function is -2. Despite this, the function has no absolute minimum, because -2 arises only as the right-hand limit at -1 and not as the value of the function at any specific point.

Note that more careful graphing of the function would also take into account concavity issues. Here are the pictures:





2.8. **Trigonometric functions.** Trigonometric functions are somewhat more difficult to study because, unlike the case of polynomials and rational functions, there could be infinitely many zeros.

One technique that is sometimes helpful when dealing with periodic functions is to concentrate on the behavior in an interval the length of one period, draw conclusions from there, and then use that to determine what happens everywhere. A very useful fact here is that f is a periodic function with period p, then f' (wherever it exists) also has period p. Similarly, the points and values of local maxima, local minima, absolute maxima and absolute minima all repeat after period p. In particular, in order to find the absolute maximum or absolute minimum, it suffices to find the absolute maximum or absolute minimum over a closed interval whose length is one period.

Consider, for instance, the function $f(x) := \sin x \cos x$. Since both sin and cos have a period of 2π, f repeats after 2π (so the period divides 2π). So, it suffices to find maxima and minima over the interval $[0, 2\pi]$. At the endpoints the value is 0. The derivative of the function is $\cos^2 x - \sin^2 x = \cos(2x)$. For this to be zero, we need $2x$ to be an odd multiple of $\pi/2$, so $x = \pi/4, 3\pi/4, 5\pi/4, 7\pi/4$. We can use the second-derivative test to see that the points $\pi/4, 5\pi/4$ are points of local maximum and the points $3\pi/4, 7\pi/4$ are points of local minimum. The value of the local maximum is $1/2$ and the value of the local minimum is $-1/2$.

(It turns out that the function $\sin x \cos x$ has a period of π, and can also be thought of as $(1/2)\sin(2x)$.)

2.9. **Mix of polynomial and trigonometric functions.** When the function is a mix involving polynomial and trigonometric functions, it is not usually periodic, nor is it a polynomial, so we need to do some ad hoc work.

For instance, consider the function $f(x) := x - 2\sin x$. The derivative is $f'(x) = 1 - 2\cos x$. Note that although f is not periodic, f' is periodic, so in order to find out where $f' > 0$, $f' = 0$, and $f' < 0$, we can restrict attention to the interval $[-\pi, \pi]$.

We have $f'(x) < 0$ for $x \in (-\pi/3, \pi/3)$, $f'(x) = 0$ for $x \in \{-\pi/3, \pi/3\}$, and $f'(x) > 0$ for $x \in (\pi/3, \pi) \cup (-\pi, -\pi/3)$.

Translating this by multiples of 2π, we obtain that $f'(x) < 0$ for $x \in (2n\pi - \pi/3, 2n\pi + \pi/3)$ for n an integer, $f'(x) = 0$ for $x \in \{2n\pi - \pi/3, 2n\pi + \pi/3\}$, and $f'(x) > 0$ at other points. Thus, f keeps shifting between increasing and decreasing.

On the other hand, for the function $f(x) := 2x - \sin x$, the derivative is $f'(x) = 2 - \cos x$. This is always positive, so f is increasing.

2.10. **Functions involving square roots and fractional powers.** For functions involving squareroots or other fractional powers, we first need to figure out the domain. Then, we use the usual techniques to handle things.

Consider, for instance, the function:

$$f(x) := \sqrt{x} + \sqrt{1-x}$$

The domain of this function is the set of values of x for which both $\sqrt{x}$ and $\sqrt{1-x}$ is defined. This turns out to be the set $[0, 1]$, since we need both $x \geq 0$ and $1 - x \geq 0$. We can differentiate f to get:

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

Note that although f is defined on the closed interval $[0, 1]$, f' is defined on the *open* interval $(0, 1)$ – it is not defined at the endpoints. In fact, the right-hand limit at 0 is $+\infty$ and the left-hand limit at 1 is $-\infty$.

Next, we want to determine where $f'(x) = 0$. Solving this, we get $x = 1/2$. Thus, $x = 1/2$ is a critical point. We also see that for $x < 1/2$, $\sqrt{x} < \sqrt{1-x}$, so the reciprocal $1/2\sqrt{x}$ is greater than the reciprocal $1/2\sqrt{1-x}$. Thus, the expression for $f'(x)$ is greater than 0. On the other hand, to the right of $1/2$, $f'(x) < 0$. Thus, f' is positive to the left of $1/2$ and negative to the right of $1/2$, yielding that f is increasing on $[0, 1/2]$ and decreasing on $[1/2, 1]$. Thus, f attains a unique absolute maximum at $1/2$, with value $\sqrt{2}$.

A more complicated version of the coffee shop problem. Remember the coffee shop problem, where there are two coffee shops located at points $a < b$ on a two-way street, and our task was to construct the function that describes distance to the nearest coffee shop. Let's now look at a somewhat different version, where the coffee shops are both located off the main street.

Suppose coffee shop A is located at the point $(0,1)$ and coffee shop B is located at the point $(2,2)$, and our two-way street is the x-axis. The goal is similar to before: write as a piecewise function the distance from the nearest coffee shop.

Define $p(x)$ as the distance from A and $q(x)$ as the distance from B. Then, we have $p(x) = \sqrt{x^2+1}$ and $q(x) = \sqrt{(x-2)^2+4} = \sqrt{x^2-4x+8}$. Our goal is to write down explicitly the function $f(x) := \min\{p(x), q(x)\}$.

In order to do this, we need to consider the function $p(x) - q(x)$ and determine where it is positive, zero and negative. Define $g(x) := p(x) - q(x)$. Then, for $g(x) = 0$, we need:

$$\sqrt{x^2+1} = \sqrt{x^2-4x+8}$$

Squaring both sides and simplifying, we obtain that $x = 7/4$. Since g is continuous, we can see that it has constant sign to the left of $7/4$ (which turns out to be negative, as we see by evaluating at 0) and constant sign to the right of $7/4$ (which turns out to be positive, as we see by evaluating at 2). Thus, our expression for f is given by:

$$f(x) = \{ \begin{array}{rl} \sqrt{x^2+1}, & x \in (-\infty, 7/4] \\ \sqrt{x^2-4x+8}, & x \in (7/4, \infty) \end{array}$$

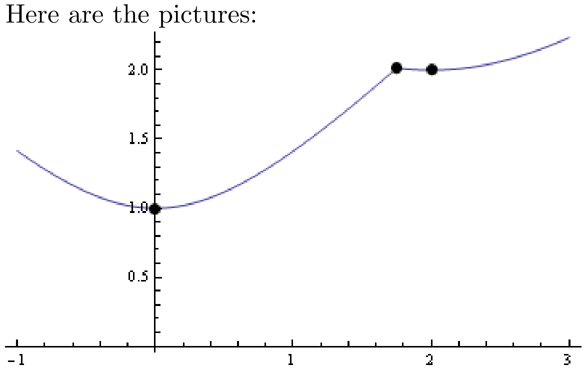
We can now use this to calculate f'. $f'(x) = x/\sqrt{x^2+1}$ to the left of $7/4$ and $(x-2)/\sqrt{x^2-4x+8}$ to the right of $7/4$. At the point $7/4$, the left-hand derivative is $7/\sqrt{65}$ and the right-hand derivative is $-1/\sqrt{65}$. The function is not differentiable at $7/4$.

Next, we want to determine where $f' > 0$, $f' = 0$ and $f' < 0$. For $x < 7/4$, we see that $f'(x) < 0$ for $x \in (-\infty, 0)$, $f'(0) = 0$, and $f'(x) > 0$ for $x \in (0, 7/4)$. For $x > 7/4$, we see that $f'(x) < 0$ for $x \in (7/4, 2)$, $f'(2) = 0$, and $f'(x) > 0$ for $x \in (2, \infty)$ (this should again be clear by looking at the picture geometrically. The distance to coffee shop A decreases till we get to the same x-coordinate as A, then it increases. At some point, B starts becoming closer, whence the distance to B starts decreasing, till we reach the point with the same x-coordinate as B, and then it starts increasing).

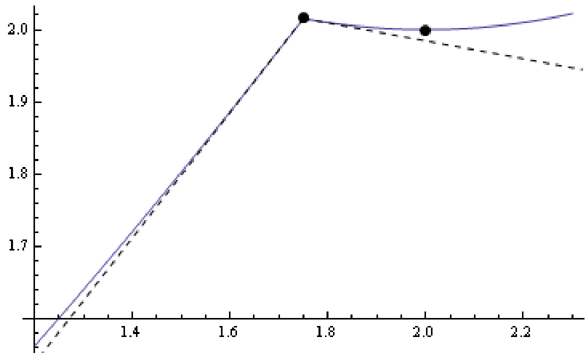
Thus, f is decreasing on $(-\infty, 0]$, increasing on $[0, 7/4]$, decreasing on $[7/4, 2]$, and increasing on $[2, \infty)$. The critical points are 0, $7/4$, and 2. There are local minima at 0 (with value 1) and 2 (with value 2) and a local maximum at $7/4$ (with value $\sqrt{65}/4$). The limits at $\pm\infty$ are both ∞. Thus, there is no absolute maximum, but the absolute maximum occurs at 0, and it has value 1.

Notice that although the picture here is qualitatively somewhat similar to the case where both coffee shops are on the x-axis, there are also some small differences – the graph never touches the x-axis, and the function is differentiable with derivative zero at two of the three critical points.

Here are the pictures:



Hre is the picture zoomed in (note: axes not centered at origin) near the value $x = 7/4$.

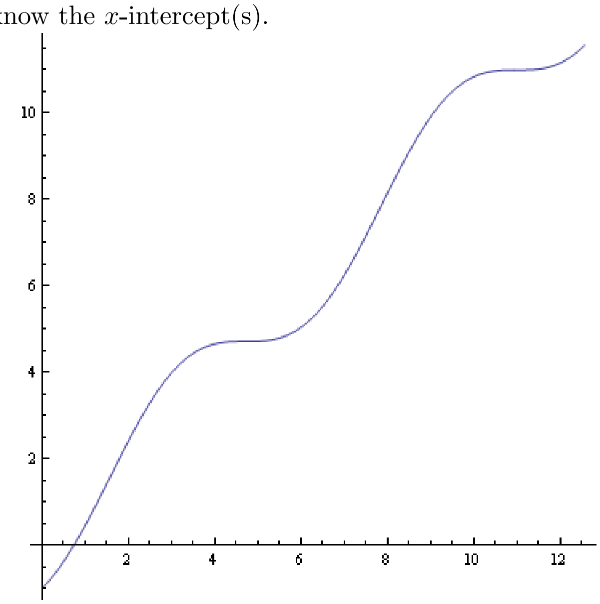


3. Subtle issues

3.1. **Equation-solving troubles.** In some cases, it is not computationally easy to do each of the suggested steps. For instance, we may not have any known method for solving $f(x) = 0$ for the given function f. Similarly, we may not have any known method for solving $f'(x) = 0$ or $f''(x) = 0$.

In cases where we do not have exact solutions, what we should do is try to find the number of solutions and the intervals in which these solutions lie, to as close an approximation as possible. Two useful tools in this are the *intermediate-value theorem* and *Rolle's theorem*.

For instance, consider the function $f(x) := x - \cos x$. f is an infinitely differentiable function, and its derivative, $1 + \sin x$, is periodic with period 2π. Thus, the graph of f repeats after 2π, with a vertical upward shift of 2π. We can further find that f is increasing everywhere, because $1 + \sin x \geq 0$ for all x, with equality occurring only at isolated points. $f''(x) = \cos x$, so f is concave up on $(-\pi/2, \pi/2)$ and its 2π-translates, and f is concave down on $(\pi/2, 3\pi/2)$ and its 2π-translates. The inflection points of f are precisely the odd multiples of $\pi/2$. The x-intercept is -1.We thus have a fairly complete picture of f, except that we do not know the x-intercept(s).



Although we do not know the x-intercept(s) precisely, we have some qualitative information. First, there can be at most one x-intercept, because f is increasing on $\mathbb{R}$. The intermediate-value theorem now reveals that the x-value must be somewhere between 0 ($f(0) = -1$) and $\pi/2$ ($f(\pi/2) = \pi/2$). In other words, the zero occurs in the segment between the y-intercept and the first inflection point after that. This is fine for a rough visual guide, but for a more accurate graph, we might like to narrow the location of the zero further. We can narrow it down further to $(\pi/6, /\pi/4)$ using elementary trigonometric computations. Further narrowing is best done with the aid of a computer.

Note that even if we did not bother about knowing the x-intercept before sketching the graph, our graph sketch would have been quite okay and would in fact have *suggested* the location of the x-intercept. This is an example of a general principle: *Often, even if we are computationally unable to handle all the suggested steps for graph-sketching, a preliminary sketch based on the steps we could successfully execute gives enough valuable hints.* The moral of the story is to not be discouraged about not executing a few steps and instead to do as much as possible with the steps already executed, and then seek alternative ways of tackling the recalcitrant steps.

See also Example 5 in the book.

3.2. **Graphing multiple functions together.** In many situations, it is necessary to be able to graph multiple functions together. This is sometimes necessary to compare and contrast these functions. Some examples include:

(1) Graphing a function and its first, second and higher derivatives together: This is often visually useful in discerning patterns about the function, and helps with rapid switching between the global and local behavior of a function.
(2) Graphing a function and another function obtained by scaling or shifting it: For instance, it may be helpful to graph $f(x)$ and $g(x) := f(x+h)$ on the same graph. This allows for easy visual insight into how the value of f changes after an interval of length h.
(3) Graphing two functions to determine their intersection points, angles of intersection, etc.

When graphing multiple functions together, the procedure is similar to that when graphing a single function, but the following additional point needs to be kept in mind: It is important to make sure that any visually obvious inferences made about the comparison of values of the functions are correct. For instance, it is important to get right which function is bigger where. The ideal way to do this is to find precisely the points of intersection – however, that may not be possible because the equation involved cannot be solved. Nonetheless, try to bound the locations of intersection points in small intervals using the intermediate value theorem. (Note that for functions obtained as derivatives, we can use Rolle's theorem and the mean value theorem.)

3.3. **Transformations of functions/graphs.** Also, if the two functions are related in terms of a transform, it is important that the geometric picture suggested by the transform is the correct one. Here are some examples:

(1) Suppose we have two functions f and g where $g(x) := f(x+h)$. Then, the graph of g should be the graph of f shifted left by h. If h is negative, it is the graph of f shifted right by $-h = |h|$.
(2) Suppose we have $g(x) := f(x) + C$. Then, the graph of g equals the graph of f shifted upward by C. If C is negative, it is the graph of f shifted downward by $-C = |C|$.
(3) Suppose $g(x) := f(\alpha x)$. Then, the graph of g should be the graph of f shrunk along the x-dimension by a factor of α. If α is negative, then this shrinking is a composite of a shrinking by $|\alpha|$ and a flip about the y-axis.
(4) Suppose $g(x) := \alpha f(x)$. Then the graph of g should be the graph of f expanded along the y-dimension by a factor of α. If α is negative, this involves an expansion by $|\alpha|$ and a flip about the x-axis.

3.4. **Can a graph be used to prove things about a function?** Yes and no. Remember that the way we drew the graph was using algebraic information about the function. So anything we deduce from the graph, we could directly deduce from that algebraic information, without drawing the graph.

The importance of graphs is that *they suggest good guesses that may not be obvious simply by looking at the algebra.* In other words, they allow visual and spatial intuition to complement the formal, symbolic intuition of mathematics. However, once the guess is made, it should be possible to justify without resort to the graph. Such justifications may use theorems such as the intermediate value theorem, Rolle's theorem, the extreme value theorem, and the mean value theorem. *In cases where things suggested by the graph cannot be verified algebraically, it is possible that some unstated and unjustified assumption was made while drawing the graph.*

3.5. **Sketching curves that are not graphs of functions.** Some curves are not in the form of functions, and cannot be expressed in that form because there are multiple y-values for a given x-value. To sketch such curves, we follow similar guidelines, but there are some changes:

(1) There is no clear concept of domain. However, it is still useful to determine the possible x-values for the curve and the possible y-values for the curve. This allows us to bound the curve in a rectangle or strip. For instance, consider the curve $x^4 + y^4 = 16$. Then, the x-value is in the interval $[-2, 2]$ and the y-value is in the interval $[-2, 2]$.
(2) We can use the techniques of sketching graphs of functions by breaking the curve down into graphs of functions. For instance, the curve $x^4 + y^4 = 16$ can be broken down as a union of graphs of two functions: $y = (16 - x^4)^{1/4}$ and $y = -(16 - x^4)^{1/4}$. We can sketch both graphs using the techniques of graph-sketching (in fact, it suffices to sketch the first graph and then construct the second graph as the reflection of the first graph about the x-axis).
(3) In cases where this separation is not easy to do, we can still try to draw the graph using the general techniques: use implicit differentiation to find the first derivative and second derivative, determine the critical points, local extreme values, points of inflection, regions of increase and decrease, regions of concave up and concave down, and so on.

3.6. **Piecewise descriptions, absolute values and max/min of two functions.** To graph a function explicitly given in piecewise form, we need to keep in mind the following things:

(1) Within the domain of each definition, plot the graph of the function the usual way.
(2) At the points where the definition changes, determine the one-sided limits, one-sided limits of first derivatives, and one-sided limits of second derivatives. These points are likely candidates for discontinuity of the function, likely candidates for discontinuity of the derivative, and likely candidates for discontinuity of the second derivative of the function.
(3) Piece this information together to draw the overall graph. Use open circles, closed circles etc. to mark clearly the limits at the points of definition changes.

In some cases, it is helpful to draw the graphs of each of the pieces over *all real values* and then pick out the requisite pieces from the relevant domains of definition.

If a function is defined as the maximum of two functions or the minimum of two functions, or in terms of absolute values, then we can first express it as a piecewise function and then graph it. Alternatively, we can graph both the functions (taking care of the points of intersection) and then use a combination of visual insight and algebra to graph the maximum and/or minimum of the two functions.

4. Addenda

4.1. **Addendum: Plotting graphs using Mathematica.** It is possible to plot the graph of a function using Mathematica. Doing a few such plots can help reinforce your intuition about the shape of graphs.

The Mathematica syntax is:

```
Plot[f[x],{x,a,b}]
```

This plots the graph of $f(x)$ for $x \in [a, b]$.

For instance:

```
Plot[x^2,{x,0,1}]
```

plots the graph of x^2 for $x \in [0, 1]$.

The command:

```
Plot[x - Sin[x],{x,-3*Pi,3*Pi}]
```

plots the graph of the function $x - \sin x$ on the interval $[-3\pi, 3\pi]$. Note that it is not possible to graph a function from $-\infty$ to ∞, so we have to stay content with finite plots.

It is also possible to plot the graphs of multiple functions together. For instance:

```
Plot[{Sin[x],(Sin[x])^2},{x, -Pi,Pi}]
```

This plots the graphs of the fnuctions sin and $\sin^2$ on the interval $[-\pi, \pi]$. To learn more, see the Mathematica documentation on the Plot function.

We can also use Mathematica to find where a function is positive, zero, and negative. You can use the Solve, Reduce, and FindRoot functions in Mathematica:

(1) The Solve function only solves equalities, and may not find all solutions. It also uses formal methods, so may not find the solutions numerically. However, it will give a formal solution saying π instead of $3.14\ldots$, for instance).
(2) The Reduce function is more powerful. It solves both equalities and inequalities, and finds all solutions. Like Solve, it only works for certain kinds of functions where these analytical and formal methods can be applied.
(3) The FindRoot function can be used to find points where a function is zero numerically. It is applicable to functions that involve a mixture of algebra and trigonometry. However, since it uses numerical methods, it may not give the exactly correct answer (for instance, it may compute 0.998 instead of 1).

For instance, we can do:

```
Reduce[x^3 - x - 6 > 0,x]
```

and find that the solution set to this is $x > 2$.

For something non-algebraic, we can find the roots:

```
FindRoot[x - Cos[x],{x,1}]
```

find a solution to $\cos x = x$. Note that Solve and Reduce do not work here because of the mixture of algebra and trigonometry. See the documentation on Solve, Reduce, and FindRoot.

We can also find the derivative of a function. First, define the function, e.g.:

```
f[x_] := x - Sin[x]
```

We can then refer to the derivative of f as f' and the second derivative as f''. Thus, we can do:

```
Reduce[{f''[x] > 0,-Pi < x, x < Pi},x]
```

This finds all solutions to $f''(x) > 0$ for x in the open interval $(-\pi, \pi)$.

These commands allow us to execute most of the computational aspects needed for graph-sketching using Mathematica.

Addendum: using a graphing software or graphing calculator. When using a graphing software or graphing calculator to plot the graph of a function, please make sure you zoom in and out enough to make sure that you are not fooled because of the scale chosen by the calculator. For instance, plotting the graph of $x^2 \sin(1/x)$ using a graphing software makes it seem like it crosses the x-axis at only finitely many points. However, zooming in closer to zero shows a lot of oscillation close to zero, and the more you zoom in, the more oscillation you see. Thus, it is important to use graphing software as a complement rather than a substitute for basic mathematical common sense.

INTEGRATION AND DEFINITE INTEGRAL: INTRODUCTION

MATH 152, SECTION 55 (VIPUL NAIK)

Corresponding material in the book: Section 5.1, 5.2.

Difficulty level: Hard if you have not seen this before. Medium if you have.

What students should definitely get: The definitions of partition, upper sum, and lower sum. A rough idea of what it means to take finer partitions and how this limiting process can be used to define integrals.

What students should hopefully get: The intuition behind an integral as an infinite summation; how it measures cumulative quantities. The intuitive relation with the area of a curve.

Executive summary

Words ...

(1) The definite integral of a continuous (though somewhat weaker conditions also work) function f on an interval $[a, b]$ is a measure of the signed area between the graph of f and the x-axis. It measures the *total value* of the function.
(2) For a partition P of $[a, b]$, the lower sum $L_f(P)$ adds up, for each subinterval of the partition, the length of that interval times the minimum value of f over that interval. The upper sum adds up, for each subinterval of the partition, the length of that interval times the maximum value of f on that subinterval.
(3) Every lower sum of f is less than or equal to every upper sum of f.
(4) The *norm* or *size* of a partition P, denoted $\|P\|$, is defined as the maximum of the lengths of its subintervals.
(5) If P_1 is a finer partition than P_2, i.e., every interval of P_1 is contained in an interval of P_2, then the following three things are true: (a) $L_f(P_2) \le L_f(P_1)$, (b) $U_f(P_1) \le U_f(P_2)$, and (c) $\|P_1\| \le \|P_2\|$.
(6) If $\lim_{\|P\|\to 0} L_f(P) = \lim_{\|P\|\to 0} U_f(P)$, then this common limit is termed the *integral* of f on the interval $[a, b]$.
(7) We can define $\int_a^b f(x)\,dx$ as above if $a < b$. If $a = b$ the integral is defined to be 0. If $a > b$, the integral is defined as $-\int_b^a f(x)\,dx$.
(8) A continuous function on $[a, b]$ has an integral on $[a, b]$. A piecewise continuous function where one-sided limits exist and are finite at every point is also integrable.

Actions ...

(1) For constant functions, the integral is just the product of the value of the function and the length of the interval.
(2) Points don't matter. So, if we change the value of a function at one point while leaving the other values unaffected, the integral does not change.
(3) A first-cut lower and upper bound on the integral can be obtained using the *trivial* partition, where we do not subdivide the interval at all. The upper bound is thus the maximum value times the length of the interval, and the lower bound is the minimum value times the length of the interval.
(4) The finer the partition, the closer the lower and upper bounds, and the better the approximation we obtain for the integral.
(5) A very useful kind of partition is a *regular partition*, which is a partition where all the parts have the same length. If the integral exists, we can calculate the actual integral as $\lim_{n\to\infty}$ of the upper sums or the lower sums for a regular partition into n parts.
(6) When a function is increasing on some parts of the interval and decreasing on other parts, it is useful to choose the partition in such a way that on each piece of the partition, the function is either

increasing throughout or decreasing throughout. This way, the maximum and minimum occur at the endpoints in each piece. In particular, try to choose all points of local extrema as points of partition.

1. Motivation and basics

In this lecture, we will introduce some of the ideas behind integration. Integration is a continuous analogue of summation (or adding things up) with a few additional complications because of the infinitely divisible nature of the real line.

1.1. **Summation: numerically.** Suppose we have a real-valued function f defined on all integers. Given any two integers $a < b$, we can legitimately ask for the sum of the values of $f(n)$ for all n in the interval $[a, b)$ (including a, excluding b). This can be interpreted as the total value of f on this interval. This summation poses no problems because we are adding finitely many real numbers. We could alternatively be interested in the sum of the values of $f(n)$ for all n in the interval $(a, b]$ (excluding a, including b).

To make things simpler, we introduce a notation for summation. This notation is something we will pick up again much later, so for now this is just as a temporary device, and not something you need to learn. The notation is:

$$\sum_{n=a}^{b-1} f(n)$$

This notation means that we add up the values of $f(n)$ for all n starting from $n = a$ and ending at $n = b - 1$. In this case, a is the lower limit of the summation, $b - 1$ is the upper limit of the summation, and $f(n)$ is the summand.

This is also sometimes written as:

$$\sum_{a \leq n < b} f(n)$$

This means that the sum is over all the integers n satisfying $a \leq n < b$. The expression $a \leq n < b$ can be replaced by any condition that restricts the n to certain integers.

1.2. **Summations: graphically.** We can think of these summations graphically as *areas*. For the first area (the summation on $[a, b)$), consider the following: for each integer n, draw a rectangle with base on the x-axis from n to $n + 1$ and height $f(n)$. The total area above $[a, b)$ is the summation of the values of $f(n)$ on $[a, b)$. Note that there's a little caveat: rectangles with negative height are given a negative area.

For the other case ($(a, b]$), we make the rectangle from $n - 1$ to n with height $f(n)$, i.e., the rectangle height is given by the value of the function at the *right end* of the rectangle.

This suggests some relationship between summations and areas. Here's one way to think about it. The area is the sum of the lengths of all the vertical slices of the figure, with each vertical slice length weighed by how much horizontal length it continues for. Thus, if the vertical length is 3 for a horizontal length of 2 and then 4 for a horizontal length of 1, the total area is $(3 * 2) + (4 * 1) = 10$.

1.3. **Piecewise constant functions: integration.** We now try to define a notion of integration for piecewise constant functions. What this notion of integration should do is measure the total value of the function, based on the ideas that we discussed above. Geometrically, it measures the *signed area* between the graph of the function and the x-axis, with a negative sign when the graph of the function is below the x-axis.

(1) For each interval $[a, b]$ where the function takes a constant value L, the integral on that interval is $L(b - a)$.
(2) The overall integral is the sum of the integrals on each of the pieces where it is constant.

This makes sense geometrically – we are breaking the area to be measured into rectangles and then finding the area of each rectangle as the product of its height L and base length $b - a$.

For instance, consider the signum function, which is -1 for x negative, 0 at 0, and $+1$ for x positive. The integral of this function on the interval $[-3, 7]$ is $(-1) * (0 - (-3)) + (1) * (7 - 0) = -3 + 7 = 4$.

1.4. **Extending the idea to other functions.** We want to define a notion of integration for a function over an interval when the function is not piecewise constant, such that:

(1) This notion measures what we intuitively think of as the area between the curve and the x-axis, with suitable signs: a positive contribution for the regions where the curve is above the x-axis and a negative contribution for the regions where the curve is below the x-axis.
(2) This notion measures some kind of *total value* of the function.
(3) If we subdivide the interval into smaller intervals, the integral over the whole interval is the sum of the integrals over the smaller intervals.

Our goal is to find something that roughly satisfies all these properties. We do, however, need to qualify the kinds of functions that we are willing to consider, because it is not possible to define a notion of integral for every function in a consistent and intuitive manner. One thing that seems to be desirable when trying to integrate is *continuity* – for a well-defined region to take the area of, the graph of the function should not randomly jump about. A slightly weaker formulation, *piecewise continuity*, will also do. Piecewise continuous means that there are only finitely many points of discontinuity. A piecewise continuous function can be integrated if it has the property that one-sided limits exist and are finite at all points of discontinuity. (Some other piecewise continuous functions can be integrated

The way we integrate them is to break the interval into subintervals where the function is continuous, integrate the function on those subintervals, and then add up the values.

1.5. **Points and zero length idea.** In the case of finite sums, changing the value at any single point changes the final sum. However, when dealing with integration, the picture is a little different. The value of the function at a particular point a makes a very small contribution – in fact a zero contribution, to the integral. This is because the rectangle corresponding to the interval $[a, a]$ has base length zero. Thus, changing the value of the function at just one point, without changing it elsewhere, has no effect on the integral. Another way of saying this is that our sample size, or base of aggregation, is so large, that measurement errors in one data point have no effect on the final answer.

1.6. **Brief note on terminology and notation.** If $a < b$ and f is a function defined on $[a, b]$, we use the notation:

$$\int_a^b f(x)\, dx$$

Here, f is termed the *integrand* or the *function being integrated*, a and b are termed the *limits of integration*, with a the *lower limit* and b the *upper limit*, x is the variable of integration, and $[a, b]$ is the *interval of integration* (also called the *domain of integration* or *region of integration*). The answer that we get is termed the *integral* of f over $[a, b]$ or the integral of f from a to b. This integral is also sometimes called a *definite integral*, to distinguish it from indefinite integrals, that we will encounter later.

As already noted, the value of the function at any one point is irrelevant, so we often do not care much if the function is not defined at finitely many of the points on $[a, b]$. Similarly, we do not care whether the function is defined at the endpoints a and b. As far as integration is concerned, we shall not make very fine distinctions between the open, closed, left-open right-closed, and right-open left-closed intervals.

We now proceed to make sense of $\int_a^b f(x)\, dx$. In a later lecture, we will extend the meaning so that we can interpret $\int_a^b f(x) dx$ for $a = b$ and $a > b$ as well.

2. Partitions: technical details begin

2.1. **Partitions, upper sums, and lower sums.** Consider a closed interval $[a, b]$. By a *partition* of $[a, b]$ we mean a sequence of points $x_0 < x_1 < \cdots < x_n$ with $a = x_0$ and $b = x_n$. The nontrivial cases of partitions are when $n \geq 2$. We use the term *partition* because given the x_i, we can divide $[a, b]$ into the *parts* $[x_0, x_1]$, $[x_1, x_2]$, and so on, right till $[x_{n-1}, x_n]$. The union of these parts is $[a, b]$. Moreover, two adjacent parts intersect at a single point, and two non-adjacent parts do not intersect. *For our purposes, single points are too small to matter, as discussed above.* So, for our purposes, this is a partition into (almost) disjoint pieces.

The idea behind using partitions is to break up the behavior of the function into smaller intervals, wherein the variation in the value of the function within each interval is smaller than the overall variation in the value. Thus, if we choose a partition with small enough parts, and find reasonable approximations for the integral on each part, adding those approximations up should give a reasonable approximation of the overall area.

2.2. **Upper bounds and lower bounds.** For the notion of integral to be reasonable, it should be true that if $f(x) \leq g(x)$ for all $x \in [a,b]$, then the integral of f is less than or equal to the integral of g. Verbally, if the function gets bigger everywhere on the interval, its total value should also get bigger. Thus, we can try determining upper and lower bounds on the integral of f by finding functions slightly smaller and slightly larger than f that we know how to integrate. The integral of f is bounded between those two integrals.

Now, the only kinds of functions that we have already decided how to integrate are the piecewise constant functions, so we need to find good piecewise constant functions. We do this using the partition.

Suppose $P = \{x_0, x_1, \dots, x_n\}$ is a partition of the interval $[a,b]$. Define piecewise constant functions f_l and f_u as follows: on each interval (x_{i-1}, x_i), f_l is constant at the minimum (more precisely infimum) of f over the interval $[x_{i-1}, x_i]$ and f_u is constant at the maximum (more precisely supremum) of f over the interval $[x_{i-1}, x_i]$. So, both f_l and f_u are piecewise constant functions (define them whatever way you want at the points x_i – as mentioned earlier, the values at individual points do not matter). Note that for continuous functions, the extreme-value theorem guarantees that the function attains its maximum and minimum over any closed interval, so we do not need to make fine distinctions between infimum and minimum or between supremum and maximum.

The integral of f_l is given by the summation, for $1 \leq i \leq n$, of the product of $(x_i - x_{i-1})$ and the minimum value of f over $[x_{i-1}, x_i]$. This value is known as the *lower sum* of f for the partition P, and it is denoted $L_f(P)$. In symbols:

$$L_f(P) = \sum_{i=1}^{n} (x_i - x_{i-1}) * (\text{minimum value for } f \text{ over } [x_{i-1}, x_i])$$

The integral of f_u is given by the summation, for $1 \leq i \leq n$, of the product of $(x_i - x_{i-1})$ and the maximum value of f over $[x_{i-1}, x_i]$. This value is known as the *upper sum* of f ofor the partition P, and it is denoted $U_f(P)$. For obvious reasons, $L_f(P) \leq U_f(P)$.

$$U_f(P) = \sum_{i=1}^{n} (x_i - x_{i-1}) * (\text{maximum value for } f \text{ over } [x_{i-1}, x_i])$$

2.3. **Finer partitions and integral as limiting value.** Given two partitions P_1 and P_2, we say that P_2 is *finer* than P_1 if the points of P_1 form a subset of the points of P_2. In other words, P_2 has all the points of P_1 and perhaps more. This means that each interval for the partition P_2 is contained in an interval for the partition P_1. The finer the partition, the *better* in some sense, since the smaller the interval, the more legitimate the process of approximating by a constant function on that interval.

If P_2 is finer than P_1, then it turns out that $U_f(P_2) \leq U_f(P_1)$ and $L_f(P_2) \geq L_f(P_1)$. In other words, the upper sums get smaller (though not necessarily strictly smaller) and the lower sums get bigger (though not necessarily strictly bigger) as the partition becomes finer. This can be seen formally as well. The idea is that when one part is subdivided further, the maximum over the entire part is greater than or equal to the maximum over each subpart. Thus, after subdivision, we are multiplying potentially smaller numbers with the same interval lengths, and the overall upper sum thus either remains the same or becomes smaller.

What we hope is that, as the partition gets finer and finer, the lower sums converge upward and the upper sums converge downward to a particular value, and we can then declare that value to be the integral of the function. Formally, for a function f on $[a,b]$ and partitions P of $[a,b]$:

If $\lim_{\|P\| \to 0} U_f(P) = \lim_{\|P\| \to 0} L_f(P)$, then this common value is termed the *integral* of f over the interval $[a,b]$, and is denoted $\int_a^b f(x)\,dx$.

What precisely does $\lim_{\|P\| \to 0}$ mean? For $P = \{x_0, x_1, x_2, \dots, x_n\}$, we define $\|P\| = \max_{1 \leq i \leq n}(x_i - x_{i-1})$. In other words, it is the maximum of the lengths of the intervals in the partition P. Sending this limit to zero means that we are considering partitions that get smaller and smaller in the sense that their largest part's size approaches zero.

This is a kind of limiting process that you have not seen in the past. So far, you have only seen limits as one real-valued variable approaches one constant value. But a partition P is not a real number; it is a more complex collection of information. In order to make sense of limiting to zero, we invent a way of measuring the size of the partition (by looking at the maximum of the sizes of the parts) and then apply the constraint that this size needs to go to zero. The limit is being taken over the space of all partitions, which is not a line.

To make matter simpler, we can restrict attention to what are called *regular partitions*. A regular partition is a partition where all the parts have equal size. For an interval $[a, b]$, there is a unique regular partition with n parts, and in that, each part has size $(b-a)/n$. Restricted to regular partitions, the above just means that we are sending n to ∞.

2.4. **Integrating the identity function.** We illustrate the technique of using partitions to integrate the function $f(x) = x$ over the interval $[0, 1]$.

We begin by looking at the trivial partition $P_1 = \{0, 1\}$. This basically means that we do not subdivide the interval into smaller pieces. For the function $f(x) = x$, the maximum value over the interval $[0, 1]$ is 1 and the minimum value is 0. Thus, $U_f(P_1) = 1(1-0) = 1$ and $L_f(P_1) = 0(1-0) = 0$. Thus, even without breaking the interval up further, we already know that the integral is somewhere between 0 and 1.

Next, consider $P_2 = \{0, 1/2, 1\}$. In this case, we have the two intervals $[0, 1/2]$ and $[1/2, 1]$. On the first interval, the minimum value is 0 and the maximum value is $1/2$. And on the second interval, the minimum value is $1/2$ and the maximum value is 1.

We thus get $L_f(P_2) = (0)(1/2-0)+(1/2)(1-1/2) = 1/4$ and $U_f(P_2) = (1/2)(1/2-0)+(1)(1-1/2) = 3/4$. Thus, the integral is somewhere between $1/4$ and $3/4$. We have thus narrowed the value of the integral to within a smaller interval.

Let us now consider a regular partition into n pieces, i.e., the partition $P_n = \left\{0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}, 1\right\}$. In each interval $[(i-1)/n, i/n]$, the maximum is i/n and the minimum is $(i-1)/n$. Thus, we get:

$$L_f(P_n) = \sum_{i=1}^{n} \frac{i-1}{n} \left(\frac{i}{n} - \frac{i-1}{n} \right)$$

That summation is given by:

$$L_f(P_n) = \frac{1}{n^2} \sum_{i=1}^{n} (i-1)$$

The summation inside is the sum of the numbers $0, 1, \dots, n-1$. The summation (which we proved by induction in the first quarter) is $n(n-1)/2$, and we thus get:

$$L_f(P_n) = \frac{n-1}{2n} = \frac{1}{2} - \frac{1}{2n}$$

Similarly, we can calculate that:

$$U_f(P_n) = \frac{n+1}{2n} = \frac{1}{2} + \frac{1}{2n}$$

As $n \to \infty$, the fraction $1/2n$ tends to zero, and we obtain that both $L_f(P_n)$ and $U_f(P_n)$ tend to $1/2$ (with $L_f(P_n)$ approaching from the left and $U_f(P_n)$ approaching from the right). Thus, the integral of the identity function on $[0, 1]$ equals $1/2$.

More generally, it turns out that the integral $\int_a^b f(x)dx = (b^2 - a^2)/2$. In a later lecture, we will look at general ways of finding the integral.

2.5. **Brief note: integral of piecewise constant functions.** As mentioned earlier, the integral of a piecewise constant function is given by the sum of the signed areas of the rectangles corresponding to each interval where it is constant. For instance, consider the function f on $[0, 3]$ such that $f(x) = 5$ on $[0, 1)$ and and $f(x) = -7$ on $[1, 3]$. Then, the integral of f is given by:

$$5 * (1 - 0) + (-7) * (3 - 1) = -9$$

For a piecewise constant function, it turns out that we can (almost) choose a partition such that both the upper and lower sum for the partition equal the value of the integral.

Here's the rough idea: we can choose a partition such that the function is constant on each part. Thus, on each of those parts, the maximum and minimum of the function are equal to the constant value, hence the contributions to both the upper sum and the lower sum are equal.

The problem is that because the partitions use closed intervals, we run into issues at places where the function changes value. If we used open intervals instead of closed intervals, this problem would not arise and we would be fine.

3. Additional notes on partitions, norms, and regular partitions

3.1. **Norm of a partition and its significance.** Recall that we defined the norm $\|P\|$ of a partition $P = \{x_0, x_1, x_2, \dots, x_n\}$ as the largest of the part sizes, i.e., $\max_{1 \le i \le n} |x_i - x_{i-1}|$. What is the significance of this norm?

The norm is not important per se, but its main significance is as follows: we want a norm P with the property that $\|P\| \to 0$ forces *all* the parts to become small. Thus, if instead of the largest of the part sizes, we took the *average* part size or the *smallest* part size, then that norm could be made arbitrarily small while keeping some of the pieces in the partition very large.

This shifts the question: why do we want a partition where *all* the parts become small? The intuition is that the smaller the part, the less the variation (hopefully) in the value of the function within each part. If we do not shrink everywhere, it may so happen that the portion of the interval where the size is large is precisely the portion where there is huge variation in the function value, so that the upper and lower sum estimates are grossly off.

3.2. **For wild functions: what can happen with lower and upper sums?** When a function is continuous on a closed interval, the integral always exists and is finite. The same holds for piecewise continuous functions. What happens for a function that is not continuous or even piecewise continuous? What if the function is discontinuous on a dense subset of the reals? In these cases, the integral does not exist.

If the function is bounded, the $\lim_{\|P\| \to 0} L_f(P)$ and $\lim_{\|P\| \to 0} U_f(P)$ both exist but they are not equal, i.e., the first limit is strictly smaller than the second limit. An example is for f the Dirichlet function that takes the value 1 at rationals and the value 0 at irrationals. Here, on any interval, the maximum value is 1 and the minimum value is 0, hence over any interval $[a, b]$ and any partition P of the interval, $U_f(P) = b - a$ and $L_f(P) = 0$.

3.3. **Finer partitions, norm, and incomparability.** We know that if P_2 is a finer partition than P_1, then (i) $\|P_2\| \le \|P_1\|$, (ii) $L_f(P_2) \ge L_f(P_1)$, and (iii) $U_f(P_2) \le U_f(P_1)$. In other words, the norm becomes smaller, the upper sum becomes smaller, and the lower sum becomes larger. However, none of (i), (ii), or (iii) individually imply that P_2 is finer than P_1. It is very much possible that two partitions are incomparable (i.e., neither is finer than the other). The mathematical jargon for this is that the relation of being finer is a partial order and not a total order on the collection of all partitions – many pairs of partitions are incomparable. In contrast, any numerical value associated with a partition has a real value and these numerical values can be totally ordered.

Despite this, any two partitions always have a common refinement that is finer than both of them.

3.4. **Regular partitions.** As mentioned, a regular partition of $[a, b]$ into n parts is a partition with n intervals each of size $(b - a)/n$. The norm of a regular partition is $(b - a)/n$. Taking the limit as $n \to \infty$, we get that the norm goes to 0. The sequence of regular partitions with n parts, with n varying over the natural numbers, thus is a natural sequence to use when computing integrals using upper and lower sums.

The larger the n, the smaller the norm of a regular partition. However, it is *not* true that the regular partition for any larger n is *finer*. For instance, the partitions of $[0, 1]$ for $n = 2$ and $n = 3$ are incomparable. A partition into n parts is finer than a partition into m parts if m divides n, i.e., n is a multiple of m.

DEFINITE INTEGRALS, FUNDAMENTAL THEOREM OF CALCULUS, ANTIDERIVATIVES

MATH 152, SECTION 55 (VIPUL NAIK)

Corresponding material in the book: Section 5.3, 5.4

Difficulty level: Hard.

What students should definitely get: Some results leading to and including the fundamental theorem of integral calculus, the definition of antiderivative and how to calculate antiderivatives for polynomials and the sine and cosine functions.

What students should hopefully get: The intuition behind the way differentiation and integration relate; the concept of indeterminacy up to constants when we integrate. The reason for making assumptions such as continuity.

Executive summary

0.1. **Definite integral, antiderivative, and indefinite integral.** Words ..

(1) We have $\int_a^b f(x)\,dx + \int_b^c f(x)\,dx = \int_a^c f(x)\,dx$.
(2) We say that F is an antiderivative for f if $F' = f$.
(3) For a continuous function f defined on a closed interval $[a, b]$, and for a point $c \in [a, b]$, the function F given by $F(x) = \int_c^x f(t)\,dt$ is an antiderivative for f.
(4) If f is continuous on $[a, b]$ and F is a function continuous on $[a, b]$ such that $F' = f$ on (a, b), then $\int_a^b f(x)\,dx = F(b) - F(a)$.
(5) The two results above essentially state that differentiation and integration are opposite operations.
(6) For a function f on an interval $[a, b]$, if F and G are antiderivatives, then $F - G$ is constant on $[a, b]$. Conversely, if F is an antiderivative of f, so is F plus any constant.
(7) The *indefinite integral* of a function f is the collection of all antiderivatives for the function. This is typically written by writing one antiderivative plus C, where C is an arbitrary constant. We write $\int f(x)\,dx$ for the indefinite integral. Note that there are no upper and lower limits.
(8) Both the definite and the indefinite integral are additive. In other words, $\int f(x)\,dx + \int g(x)\,dx = \int f(x) + g(x)\,dx$. The analogue holds for definite integrals, with limits.
(9) We can also pull constants multiplicatively out of integrals.

Actions ...

(1) To do a definite integral, find any one antiderivative and evaluate it between limits.
(2) An important caveat: when using antiderivatives to do a definite integral, it is important to make sure that the antiderivative is defined and continuous everywhere on the interval of integration. (Think of the $1/x^3$ example).
(3) To do an indefinite integral, find any antiderivative and put a $+C$.
(4) To find an antiderivative, use the additive splitting and pulling constants out, and the fact that $\int x^r\,dx = x^{r+1}/(r + 1)$.

0.2. **Higher derivatives, multiple integrals, and initial/boundary conditions.** Actions ...

(1) The simplest kind of *initial value problem* (a notion we will encounter again when we study differential equations) is as follows. The k^{th} derivative of a function is given on the entire domain. Next, the *values* of the function and the first $k - 1$ derivatives are given at a single point of the domain. We can use this data to find the function. Step by step, we find derivatives of lower orders. First, we integrate the k^{th} derivative to get that the $(k - 1)^{th}$ derivative is of the form $F(x) + C$, where C is unknown. We now use the value of the $(k - 1)^{th}$ derivative at the given point to find C. Now, we have the $(k - 1)^{th}$ derivative. We proceed now to find the $(k - 2)^{th}$ derivative, and so on.

(2) Sometimes, we may be interested in finding *all* functions with a given second derivative f. For this, we have to perform an indefinite integration twice. The net result will be a general expression of the form $F(x) + C_1 x + C_2$, where F is a function with $F'' = f$, and C_1 and C_2 are arbitrary constants. In other words, we now have *up to constants or linear functions* instead of *up to constants* as our degree of ambiguity.
(3) More generally, if the k^{th} derivative of a function is given, the function is uniquely determined up to additive differences of polynomials of degree strictly less than k. The number of free constants that can take arbitrary real values is k (namely, the coefficients of the polynomial).
(4) This general expression is useful if, instead of an initial value problem, we have a boundary value problem. Suppose we are given G'' as a function, and we are given the value of G at two points. We can then first find the general expression for G as $F + C_1 x + C_2$. Next, we plug in the values to get a system of two linear equations, that we solve in order to determine C_1 and C_2, and hence G.

1. Statements of main results

1.1. **The definite integral: recall and more details.** Recall from last time that for a *continuous* (or piecewise continuous where all discontinuities are jump discontinuities) function f on an open interval $[a, b]$, the *integral* of f over the interval $[a, b]$, denoted:

$$\int_a^b f(x)\,dx$$

is a kind of summation for f for all the real numbers from a to b. This integral is also called a *definite integral*. The function is often termed the *integrand*. The number b is termed the *upper limit* of the integration, and the number a is termed the *lower limit* of the integration. The variable x is termed the *variable of integration*.

So far, we have made sense of the expression as described above with $a < b$. We now add in a few details on how to make sense of two other possibilities:

- If $a = b$, then, by *definition*, the integral is defined to be zero.
- If $a > b$, then, by *definition*, the integral is defined as the *negative* of the integral $\int_b^a f(x)\,dx$.

Now, it makes sense to consider the symbol $\int_a^b f(x)\,dx$ without any ordering conditions on a and b. With these definitions, we have, for any $a, b, c \in \mathbb{R}$:

$$\int_a^b f(x)\,dx + \int_b^c f(x)\,dx = \int_a^c f(x)\,dx$$

1.2. **Definite integrals do exist for piecewise continuous functions.** It is useful to know the following:

(1) For a continuous function f on a closed and bounded interval $[a, b]$, the integral exists and is finite. In fact, the integral over the interval $[a, b]$ is bounded from above by $(b - a)$ times the maximum value of the function and from below by $(b - a)$ times the minimum value of the function. (Both of these exist by the extreme value theorem).
(2) For a piecewise continuous function f on a closed and bounded interval $[a, b]$ such that all the one-sided limits exist and are finite at points of discontinuity, the integral exists and is finite. This follows from the previous part, via the intermediate step of breaking $[a, b]$ into parts such that the restriction of the function to each part is continuous and extends continuously to the boundary of that part.

1.3. **The definite integral and differentiation.** There is also a clear relationship between the definite integral and differentiation. In some sense, the integral and derivative are inverses (opposites) of each other. Let $[a, b]$ be an interval. Suppose f is a continuous function on $[a, b]$ and $c \in [a, b]$ is any number. Define the following function F on $[a, b]$:

$$F(x) := \int_c^x f(t)\,dt$$

Note the way the function is defined. t is the variable of integration, and F depends on x in the sense that the *upper limit* of the interval of integration is x, whereas the lower limit is fixed at c.

Continuous functions are integrable, as discussed above, so F turns out to be well-defined.

Further, F is continuous on $[a, b]$, differentiable on (a, b), and has derivative

$$F'(x) = f(x) \qquad \text{for all } x \in (a, b)$$

This is Theorem 5.3.5.

1.4. **Concept of antiderivative.** Suppose f is continuous on $[a, b]$. An *antiderivative* for f, or *primitive* for f, or *indefinite integral* for f, is a function G on $[a, b]$ such that:

- G is continuous on $[a, b]$.
- $G'(x) = f(x)$ for all $x \in (a, b)$.

A little while back, we had seen the following result: if F, G are functions on an interval I such that $F' = G'$ for all points in the interior of I, then $F - G$ is a constant function on I. In other words, F and G differ by a constant.

Conversely, if F and G are functions on an interval I with F differentiable on the interior of I, and $F - G$ is constant, then G is also differentiable on I and $F' = G'$ on the interior of I.

Thus, the antiderivative of a function is not unique – we can always add a constant function to one antiderivative to obtain another antiderivative. However, the antiderivative is unique up to differing by constants. In other words, any two antiderivatives differ by a constant.

We are now in a position to state the fundamental theorem of calculus.

Suppose f is a continuous function on the interval $[a, b]$. If G is an antiderivative for f on $[a, b]$, then we have:

$$\int_a^b f(t)\, dt = G(b) - G(a)$$

Also, as we already noted, any two antiderivatives differ by a constant, so if we replace G by another antiderivative, the right side remains the same because both $G(a)$ and $G(b)$ get shifted by the same amount.

For notational convenience, this is sometimes written as:

$$\int_a^b f(t)\, dt = [G(t)]_a^b$$

Here, the right side is interpreted as the difference between the values of $G(t)$ for $t = b$ and $t = a$, which simplifies to $G(b) - G(a)$.

2. Computing antiderivatives and integrals: easy facts

2.1. **Computing some antiderivatives.** We now compute some common expressions for antiderivatives of functions.

(1) If $f(x) = x^r$, and $r \neq -1$, then we can set $G(x) = x^{r+1}/(r+1)$. The factor of $1/(r+1)$ is intended to cancel the factor of $r + 1$ that appears as a coefficient when we differentiate x^{r+1}. In particular, if $f(x) = x$, $G(x) = x^2/2$, and if $f(x) = x^2$, $G(x) = x^3/3$. Most importantly, if $f(x) = 1$, then $G(x) = x$.

(2) An antiderivative for sin is $-\cos$ and an antiderivative for cos is sin. Note the sign differences between these formulas and those for the derivative. The derivative of sin is cos but the antiderivative of sin is $-\cos$. The derivative of cos is $-\sin$ and the antiderivative of cos is sin. (We will see more trigonometric antiderivatives later).

2.2. **Linearity of the integral.** The integral is *linear*, in the sense of being additive and allowing for the factoring out of scalars. Specifically:

$$\int_a^b [f(x) + g(x)]\, dx = \int_a^b f(x)\, dx + \int_a^b g(x)\, dx$$

and

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

Thus, we can pull out scalars and split sums additively when computing integrals, just as we did for derivatives.

2.3. **Linearity of the antiderivative.** The linearity of the integral turns out to be closely related to the linearity of the antiderivative. Of course, it is not precise to say "the" antiderivative, since the antiderivative is defined only up to differences of constants. What we mean is the following:

(1) If F is an antiderivative for f and G is an antiderivative for g, then $F + G$ is an antiderivative for $f + g$.
(2) If F is an antiderivative for f and α is a real number, then αF is an antiderivative for αf.

(These statements are immediate corollaries of the corresponding statements for derivatives).

2.4. **General expression for indefinite integral.** Once we have computed one antiderivative for the integral, the general expression for the indefinite integral is obtained by taking that antiderivative and writing a "+ C" at the end, where C is a freely varying real parameter. What this means is that *every specific choice* of numerical value for C gives yet another antiderivative for the original function.

Note that the letter C is used conventionally, but there is nothing special about this latter. If the situation at hand already uses the letter C in some other context, please use another letter.

For instance:

$$\int (x - \sin x)\,dx = (x^2/2) + \cos x + C$$

2.5. **Getting our hands dirty.** We are now in a position to do some straightforward computations of integrals for polynomials and some basic trigonometric functions. For instance:

$$\int_0^1 (x^2 - x + 1)\,dx$$

We can find an antiderivative for this function, by finding antiderivatives for the individual functions x^2, $-x$, and 1, and then adding up. An antiderivative that works is $x^3/3 - x^2/2 + x$. Now, to calculate the definite integral, we need to calculate the difference between the values of the antiderivative at the upper and lower limit. We write this as:

$$\left[\frac{x^3}{3} - \frac{x^2}{2} + x\right]_0^1$$

Next, we do the calculation:

$$\left(\frac{1}{3} - \frac{1}{2} + 1\right) - (0 - 0 + 0) = \frac{5}{6}$$

Thus, the value of the definite integral is $5/6$.

In other words, the signed area between the graph of the function $x^2 - x + 1$ and the x-axis, between the x-values 0 and 1, is $5/6$.

Some people prefer to split the definite integral as a sum first and then compute antiderivatives for each piece. The work would then appear as follows:

$$\int_0^1 (x^2 - x + 1)\,dx = \int_0^1 x^2\,dx - \int_0^1 x\,dx + \int_0^1 1\,dx = [x^3/3]_0^1 - [x^2/2]_0^1 + [x]_0^1 = 1/3 - 1/2 + 1 = 5/6$$

There is no substantive difference in the computations.

3. Higher derivatives and repeated integration

3.1. **Finding all functions with given k^{th} derivative.** Suppose the second derivative of a function is given. What are all the possibilities for the original function? In order to answer this question, we need to integrate twice. For instance, suppose $f''(x) = \cos x$. Then, we know that:

$$f'(x) = \int \cos x \, dx = (\sin x) + C_1$$

where C_1 is an arbitrary real number.

Integrating again, we get:

$$f(x) = \int f'(x) \, dx = \int [(\sin x) + C_1] \, dx = (-\cos x) + C_1 x + C_2$$

Here, both C_1 and C_2 are arbitrary real numbers. Thus, the family of all possible fs that work is described by two parameters, freely varying over the real numbers.

More generally, if the k^{th} derivative of a function is known, then the original function is known up to additive difference of a polynomial fo degree at most $k - 1$. Each coefficient of that polynomial is a freely varying real parameter, and there are k such coefficients: the constant term, the coefficient of x, and so on till the coefficient of x^{k-1}.

3.2. **Degree of freedom and initial/boundary values.** One way of thinking of the preceding material is that each time we integrate, we introduce one more degree of freedom. Thus, integrating thrice introduces a total of three degrees of freedom.

In practice, when we are asked to find a function f in the real world, we know the k^{th} derivative of f, but we also have information about the values of f at some points. The two typical ways this information is packaged are:

- *Initial value problem* packaging: Here, the value of f and all its derivatives, up to the $(k-1)^{th}$ derivative, at a single point c are provided, along with the general expression for the k^{th} derivative. For this kind of problem, we can, at each stage of antidifferentiation, determine the value of the constant we get, and thus we get a *single* function at the end.
- *Boundary value problem* packaging: Here, the value of f at k distinct points is specified. To solve this kind of problem, we first find the general expression for f with k unknown constants, then use the values at k distinct points to get a system of k linear equations in k variables, which we then proceed to solve.

4. Subtle issues/additional notes

4.1. **Variable of integration – don't reuse!** When writing something like $\int_a^b f(t) \, dt$, please remember that the letter t, which is used locally as a variable of integration, *cannot* be used outside the expression.

4.2. **Definite integral as a size or norm of function.** To completely describe a function f on a closed interval $[a, b]$ requires a lot of work, since it requires specifying the function value at infinitely many points. On the other hand, the value of the integral of f on $[a, b]$, given by $\int_a^b f(x) \, dx$, is a single real number. Since numbers are easier to grasp than functions, we often use the integral of a function on an interval to get an approximate estimate of its size.

More generally, we are often interested in expressions of the form $\int_a^b f(x)g(x) \, dx$ where $g(x)$ plays the role of a weighting function. Usually, we have a bunch of two or three functions g and we are interested in the above integral on $[a, b]$ for each of those gs. We use the collection of two or three numbers we get that way to say profound things about the function f, even without knowing f directly.

4.3. **Linear algebra interpretation of antiderivative.** (This material is not necessary for this course, but is useful for subsequent mathematics – we'll see it again in 153 and you'll see more of these ideas if you take Math 196/199 or advanced courses in the social and/or physical sciences).

Denote by C^1 the set of functions on $\mathbb{R}$ that are continuously differentiable everywhere. Denote by C^0 the set of continuous functions on $\mathbb{R}$.

First, note that C^0 and C^1 are both *vector spaces* over $\mathbb{R}$. Here's what this means for C^0: the sum of two continuous functions is continuous, and any scalar multiple of a continuous function is continuous. Here's what this means for C^1: the sum of two continuously differentiable functions is continuously differentiable, and any scalar multiple of a continuously differentiable function is continuously differentiable.

Differentiation is a *linear* operator from C^1 to C^0 in the following sense: first, for any $f \in C^1$, f' is an element of C^0. Second, we have the rules $(f+g)' = f' + g'$ and $(\alpha f)' = \alpha f'$. In other words, differentiation respects the vector space structure.

The *kernel* of a linear operator is the set of a functions which go to zero. For any linear operator between vector spaces, the kernel is a subspace.

Our basic result is that the kernel of the differentiation operator is the space of constant functions. Two elements in a vector space have the same image under a linear operator iff their difference is in the kernel of that operator. In our context, this translates to the statement that two functions have the same derivative iff their difference is a constant function.

Our second basic result is that any function in C^0 arises as the derivative of something in C^1. This something can be computed using a definite integral.

Thus, the kernel of the differentiation operator is a copy of the real line inside C^1, given by the scalar functions. For any element of C^0, the set of elements of C^1 which map to it is a line inside C^1 parallel to the line of constant functions.

Instead of looking at functions on the entire real line, we can also restrict attention to functions on an open interval inside the real line – qualitatively, all our results hold.

CHAIN RULE, u-SUBSTITUTION, SYMMETRY, MEAN VALUE THEOREM

MATH 152, SECTION 55 (VIPUL NAIK)

Corresponding material in the book: Section 5.6, 5.7, 5.8, 5.9.

Difficulty level: Hard.

What students should definitely get: The idea of using differentiation rules to determine antiderivative, the application of the chain rule to indefinite integration, and the idea of the u-substitution. The application of the u-substitution to definite integrals, the idea that definite integrals can be computed in light of certain kinds of symmetry even without computing an antiderivative. Bounding definite integrals via other definite integrals. The mean value theorem for integrals.

What students should eventually get: A grasp and clear memory of all the rules for computing definite integrals for functions with a certain kind of symmetry. The subtleties of u-substitutions.

Executive summary

0.1. **Reversing the chain rule.** Actions ...

(1) The chain rule states that $(f \circ g)' = (f' \circ g) \cdot g'$.
(2) Some integrations require us to reverse the chain rule. For this, we need to realize the integrand that we have in the form of the right-hand side of the chain rule.
(3) The first step usually is to find the correct function g, which is the *inner function* of the composition, then to adjust constants suitably so that the remaining term is g', and then figure out what f' is. Finally, we find an antiderivative for f', which we can call f, and then compute $f \circ g$.
(4) A slight variant of this method (which is essentially the same) is the substitution method, where we identify g just as before, try to spot g' in the integrand as before, and then put $u = g(x)$ and rewrite the integral in terms of u.

0.2. **u-substitutions and transformations.** Words ... (try to recall the numerical formulations)

(1) When doing the u-substitution for definite integrals, we transform the upper and lower limits of integration by the u-function.
(2) Note that the u-substitution is valid only when the u-function is well-defined on the entire interval of integration.
(3) The integral of a translate of a function is the integral of a function with the interval of integration suitably translated.
(4) The integral of a multiplicative transform of a function is the integral of the function with the interval of integration transformed by the same multiplicative factor, scaled by that multiplicative factor.

0.3. **Symmetry and integration.** Words ...

(1) If a function is continuous and even, its integral on $[-a, 0]$ equals its integral on $[0, a]$. More generally, its integrals on any two segments that are reflections of each other about the origin are equal. As a corollary, the integral on $[-a, a]$ is twice the integral on $[0, a]$.
(2) If a function is continuous and odd, its integral on $[-a, 0]$ is the negative of its integral on $[0, a]$. More generally, its integrals on any two segments that are reflections of each other about the origin are negatives of each other. As a corollary, the integral on $[-a, a]$ is zero.
(3) If a function is continuous and has mirror symmetry about the line $x = c$, its integral on $[c - h, c]$ equals its integral on $[c, c + h]$.
(4) If a function is continuous and has half-turn symmetry about $(c, f(c))$, its integral on any interval of the form $[c - h, c + h]$ is $2hf(c)$. Basically, all the variation about $f(c)$ *cancels out* and the *average value* is $f(c)$.

(5) Suppose f is continuous and periodic with period h and F is an antiderivative of f. The integral of f over any interval of length h is constant. Thus, $F(x+h) - F(x)$ is the same constant for all x. (We saw this fact long ago, without proof).
(6) The constant mentioned above is zero iff F is periodic, i.e., f has a periodic antiderivative.
(7) There is thus a well-defined *average value* of a continuous periodic function on a period. This is also the average value of the same periodic function on any interval whose length is a nonzero integer multiple of the period. This is also the limit of the average value over very large intervals.

Actions...

(1) All this even-odd-periodic stuff is useful for trivializing some integral calculations without computing antiderivatives. This is more than an idle observation, since in a lot of real-world situations, we get functions that have some obvious symmetry, even though we know very little about the concrete form of the functions. We use this obvious symmetry to compute the integral.
(2) Even if the whole integrand does not succumb to the lure of symmetry, it may be that the integrand can be written as (something nice and symmetric) + (something computable). The (nice and symmetric) part is then tackled using the ideas of symmetry, and the computable part is computed.

0.4. **Mean-value theorem.** Words ...

(1) The *average value*, or *mean value*, of a continuous function on an interval is the quotient of the integral of the function on the interval by the length of the interval.
(2) The mean value theorem for integrals says that a continuous function must attain its mean value somewhere on the interior of the interval.
(3) For periodic functions, the mean value over any interval whose length is a multiple of the period is the same. Also, the mean value over a very large interval approaches this value.

1. Philosophical remarks on hardness

1.1. **A fundamental asymmetry between differentiation and integration.** A little while back in the course, we saw how to differentiate functions. In order to carry out differentation, we learned how to differentiate all the basic building block functions (the polynomial functions and the sine and cosine functions) and then we learned a bunch of rules that allowed us to differentiate any function built from these elementary functions using either function composition or pointwise combination.

This means that for any function built from the elementary functions, if we know how to write it, we know how to compute its derivative. The strategy is to keep breaking down the task using the rules for combination and composition until we get to differentiating the elementary functions, fo which we have formulas.

The analogue is *not* true for finding antiderivatives. In other words, there is no foolproof procedure to break down operations such as combination and composition and ultimately reduce the problem to computing antiderivatives of the basic building blocks. Thus, *even though there are formulas* for the antiderivatives of all the basic building blocks, there exist functions constructed from these that do not have antiderivatives that can be written as elementary functions.

One important point should be made here. Just because the antiderivative of f cannot be expressed as an elementary function (or, it can but we're not able to determine that elementary function) does *not* mean that the antiderivative does not exist. Rather, it means that the existing pool of functions that we have is not large enough to contain that function, and we may need to introduce new classes of functions.

1.2. **Dealing with failure, getting used to it.** Examples of functions that do not have antiderivatives in the classes of functions we have seen so far include $1/(x^2+1)$, $1/\sqrt{x^2+1}$, $1/x$, and $1/\sqrt{x^2-1}$. Later in the course, we shall introduce new classes of functions, and it turns out that these functions are integrable within those new classes of functions. (The new classes include logarithmic functions and inverse trigonometric functions). Functions such as $1/\sqrt{x^3+2x+7}$ cannot be integrated even in this larger collection of functions – to integrate these functions, we would need to introduce *elliptic functions* and *inverse elliptic functions* which are an analogue of trigonometric and inverse trigonometric functions. We won't formally introduce those functions in the 150s, and you probably will not see them ever in a formal way.

Similarly, the functions $\sin(x^2)$ and $(\sin x)/x$ do not have indefinite integrals expressible in our current vocabulary (the integral of the latter is particularly important and is called the *sine integral*, even though it has no easy expression). When we later introduce the logarithmic function, we will see that $1/\log x$ has no antiderivative in the classes of functions we are dealing with, though one of its antiderivatives, called the *logarithmic integral*, is extremely important in number theory and in the distribution of prime numbers.

When we later introduce the exponential function, we shall see that e^{-x^2} has no antiderivative expressible in terms of elementary functions. However, the antiderivative of e^{-x^2} is *extremely important* in statistics, since e^{-x^2} corresponds to the shape of the Gaussian or normal distribution (a shape often called a *Bell curve*) and its integral measures the area under the curve for such a distribution. The integral is so important in statistics that there are tables of the values of the definite integral from 0 to a for different numerical values of a. These tables can be used to calculate the definite integral between any two points. As an interesting aside, it is true that the integral of e^{-x^2} over the entire real line (a concept we will see later) is $\sqrt{\pi}$.

In general, the use of antiderivatives and indefinite integration is a powerful tool in performing definite integration. Recall that if F is an antiderivative for f, then $\int_a^b f(x)\,dx = F(b) - F(a)$. So, to integrate f between limits, all we need to do is find an antiderivative, evaluate it at the limits, and subtract. However, there are three problems that we encounter as soon as we start trying this approach for nontrivial functions:

(1) An antiderivative may not even exist within the class of functions that we are familiar with. In other words, we may need to define and introduce new classes of functions to fit in the antiderivative. This is not very helpful for computational purposes.
(2) Even if the antiderivative exists, it may require considerable ingenuity to find it. This is because there is no clear and short step-by-step reductive algorithm to find an antiderivative. This is in sharp contrast with the situation for derivatives, where we can reduce step by step.
(3) Even if we successfully calculate the antiderivative, it may not be much use to us computationally if we cannot evaluate the antiderivative at the two endpoints. This is more of a problem when dealing with functions that involve trigonometric functions (and inverse trigonometric, exponential, and logarithmic functions – new classes of functions you have been sheltered from so far).

In all these cases, one tool still remains at our disposal – the *back-to-basics* definition of the definite integral using upper sums and lower sums. This definition can usually allow us to quickly obtain crude upper bounds and crude lower bounds. Such bounds are not as good as an exact answer but they may be good enough.

2. Breaking the differentiation code: reverse engineering

2.1. **Recalling the rules for differentiation.** Our next stop is the rules for differentiation. Broadly, our strategy for computing antiderivatives is *working backward*: starting from rules that we know for differentiation and trying to guess what the antiderivative must have been so that differentiating it gives the function we have at hand.

We have already seen the rules for sums and scalar multiples for differentiation. In technical terminology, we say that the antiderivative is *linear* – the antiderivative of the sum is the sum of the antiderivatives, and the antiderivative of a scalar multiple is the same scalar multiple of the antiderivative (I'm being imprecise by using *the*, but you should get the idea).

There are two other rules for differentiation that are somewhat more complicated: the *product rule* and the *chain rule*. In this lecture, we concentrate on the chain rule. The product rule manifests itself in a technique, called *integration by parts*, that we will see next quarter.

2.2. **The chain rule.** Let's look at the chain rule.

$$(f \circ g)' = (f' \circ g) \cdot g'$$

Equivalently:

$$\frac{d}{dx}[f(g(x))] = f'(g(x))g'(x)$$

In order to use the chain rule to integrate a function p, we need to do what's called *pattern matching* – we need to find functions f and g such that we can write $p = (f' \circ g) \cdot g'$. In some cases, the way the function is written is reasonably suggestive of what f and g are. In other cases, we need to do a little work. We look at some examples.

Consider:

$$\int \sin(\cos x)\sin x\,dx$$

Comparing this with the general expression, we see that we should have $g(x) = \cos x$, since cos is the *inner function* of the composition in this expression. If $g(x) = \cos x$, then we obtain $g'(x) = -\sin x$. We notice that the expression we have is $\sin x$, not $-\sin x$, so we put a negative sign on the outside, and obtain:

$$-\int \sin(\cos x)(-\sin x)\,dx$$

If $g(x) = \cos x$, then $g'(x) = -\sin x$. Looking again at the pattern we are trying to match, we see that we must have $f'(t) = \sin t$. We thus see that a possible candidate for $f(t)$ is $-\cos t$, since that is an antiderivative for sin. Using the chain rule in reverse, we thus obtain that $-(-\cos(\cos x)) = \cos(\cos x)$ is an antiderivative. The indefinite integral is thus:

$$\cos(\cos x) + C$$

where C is an arbitrary real constant.

It might be worthwhile to differentiate this and check that the derivative we get is the original integrand.

Here is another example. Consider:

$$\int (x^2+1)^{45}x\,dx$$

We can perform this integration by first computing $(x^2+1)^{45}$ as a polynomial, then multiplying each term by x, then integrating termwise. However, this is impractical. Instead, we try to use the chain rule.

The composite function of interest is $(x^2+1)^{45}$. This is a composite of the function $g(x) = x^2 + 1$ and the function $h(t) = t^{45}$. The derivative of g is $g'(x) = 2x$, which is twice of the expression we have (simply x). Thus, we need to multiply and divide by 2:

$$\frac{1}{2}\int (x^2+1)^{45}2x\,dx$$

Now, we see that $f'(t) = h(t) = t^{45}$, so f is an antiderivative for that. We could take $f(t) = t^{46}/46$. The overall antiderivative then simplifies to:

$$\frac{1}{2}\frac{(x^2+1)^{46}}{46} + C = \frac{(x^2+1)^{46}}{92} + C$$

Let's look at another example:

$$\int x^3\,dx$$

We already know that an antiderivative for this is $x^4/4$ and the general expression for the indefinite integral is $(x^4/4)+C$. We now see how this result can be obtained using the chain rule. We write $x^3 = x^2 \cdot x$. We then set $g(x) = x^2$, and $f'(t) = t$ (so $f(t) = t^2/2$), so that $x^2 = f'(g(x))$. We also have $g'(x) = 2x$, which is twice of x, so we get:

$$\frac{1}{2}\int (x^2)\cdot(2x)\,dx$$

The integral is thus:

$$\frac{1}{2}\frac{(x^2)^2}{2} + C = \frac{x^4}{4} + C$$

Let us look at one more example:

$$\int \frac{2x}{(x^2+1)^2}\,dx$$

Here, we notice that the derivative of $x^2 + 1$ is $2x$. Thus, we set $g(x) = x^2 + 1$. We obtain $g'(x) = 2x$. Also, we have $f'(g(x)) = 1/(x^2 + 1)^2$, so $f'(t) = 1/t^2$. Thus, $f(t)$ is an antiderivative for $1/t^2$, so we can set $f(t) = -1/t$. Plugging these in, we obtain that $f(g(x)) = -1/(x^2 + 1)$, so we obtain that the indefinite integral is:

$$\frac{-1}{x^2+1} + C$$

A glimpse into the u-substitution. One drawback of the approach outlined above for reverse-engineering the chain rule is that we have to do a lot of rough work and this becomes tedious for harder problems. An alternative way of presenting this, that makes things easier to handle in harder situations, is by using a substitution. Here, we identify $g(x)$ in the same way as we did earlier, and we then try to write our integral as:

$$\int h(g(x))g'(x)\,dx$$

The main difference is that instead of trying to find *a priori* a function f such that $f' = h$, we instead postpone that for later. We perform a substitution $u = g(x)$, whereby we replace $g'(x)dx$ with du, and obtain:

$$\int h(u)\,du$$

which we now proceed to integrate (which is essentially the same as finding an antiderivative for h, which is the function we called f). Finally, we substitute in $g(x)$ for u in the expression we obtain.

This seems like more steps. However, the main advantage is that one of the steps that we had to do as scratch work, namely finding f using the expression we have for f', is now done in the open. This is particularly useful if the function $h = f'$ is complicated and integrating it requires many steps.

There is also a slight variant of substitution for definite integrals. We now turn to that.

3. Magic of definite integrals with chain rule

3.1. **The u-substitution revisited.** Recall the u-substitution, which is a variant of the procedure to reverse the chain rule, but has the advantage that it breaks our work more clearly into two steps: first find g, then reduce the problem to a new problem that involves finding the antiderivative for h.

How do we use this to compute a definite integral? We can use the procedure outlined above to compute an antiderivative in terms of x, and then evaluate it between limits. For instance, consider:

$$\int_0^{\pi} \cos(\sin x)\cos x\,dx$$

We first try to compute the antiderivative:

$$\int \cos(\sin x)\cos x\,dx$$

Set $u = \sin x$. Then, the above integral becomes:

$$\int \cos u\,du$$

which is $\sin u$. Since $u = \sin x$, we obtain that $\sin(\sin x)$ is an antiderivative. The definite integral is thus $\sin(\sin\pi) - \sin(\sin 0) = 0$.

There is an alternative way of doing things, which involves *changing the limits of integration with each u-substitution*. The idea here is that every time we make a substitution of the form $u = g(x)$, we replace the lower and upper limits by their images under g. In other words, if the function is being integrated from a to b, the new function is being integrated from $g(a)$ to $g(b)$. In symbols:

$$\int_a^b h(g(x))g'(x)\,dx = \int_{g(a)}^{g(b)} h(u)\,du$$

The advantage of this is that after we find an antiderivative for h, say f, we do not need to compute the function $f \circ g$, i.e., we do not need to find an antiderivative for the original integrand. We simply evaluate the new antiderivative between the new limits $g(a)$ and $g(b)$.

The approach has other advantages, namely, in situations where it is difficult or impossible to get explicit expressions for antiderivatives, but a definite integral can be computed due to symmetry considerations or for other degenerate reasons. For instance, consider:

$$\int_0^\pi \cos(\sin x)\cos x \, dx$$

Set $u = \sin x$. The limits now become $\sin 0$ and $\sin \pi$, so the integral becomes:

$$\int_0^0 \cos u \, du$$

Note that with this u-substitution method, we do not even need to find an antiderivative for the integrand: we can straightaway compute that the integral is zero, because the upper and lower limits for integration coincide.

3.2. **Inequalities involving the definite integral.** We'll now review some of the properties of the definite integral that are discussed in Section 5.8 in the book. We begin with properties 5.8.1 – 5.8.4. These are fairly straightforward, and are expected from the notion of integral as a total value, or from the formal definition involving lower and upper sums. Note that by default, all integrals are over intervals of positive length, taken from left to right, i.e., the lower limit of the integral is strictly smaller than the upper limit of the integral.

(1) The integral of a nonnegative continuous function is nonnegative. (5.8.1)
(2) The integral of an everywhere positive function is positive. (5.8.2)
(3) If $f(x) \le g(x)$ for all $x \in [a, b]$, then $\int_a^b f(x)\,dx \le \int_a^b g(x)\,dx$. (5.8.3)
(4) If $f(x) < g(x)$ for all $x \in [a, b]$, then $\int_a^b f(x)\,dx < \int_a^b g(x)\,dx$. (5.8.4)

These inequalities give us a new tool for bounding an integral from above and below. We now turn to that tool.

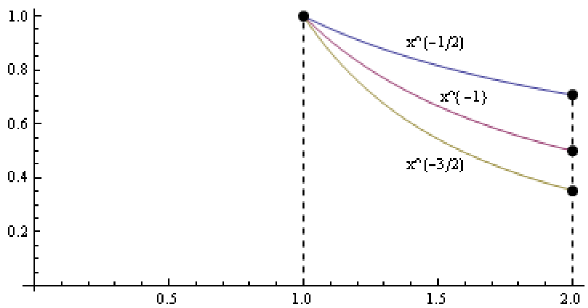
3.3. **Bounding an integral.** It is not always possible to find an explicit expression for an antiderivative. Hence, we cannot always compute the definite integral of a function via the antiderivative route. One strategy we had for overcoming this was the use of *upper and lower sums*. These sums, however, can get tedious to compute. An alternative strategy is to bound the function between two other functions, and hence bound its integral between the integrals of those two functions.

For instance, consider the integral:

$$\int_1^2 \frac{dx}{x}$$

We consider the function $f(x) := x^{-1}$ on the interval $[1, 2]$. Since $x \ge 1$, this function is bounded from above by $g(x) := x^{-1/2}$, and from below by $h(x) := x^{-3/2}$. Thus, the integral of f is bounded between the integrals of g and of h.

An antiderivative for g is $2\sqrt{x}$, and evaluating it between limits gives $2(\sqrt{2} - 1)$, which is slightly less than 0.83. An antiderivative for h is $-2/\sqrt{x}$, and evaluating it between limits gives $2 - \sqrt{2}$, which is slightly greater than 0.58. Thus, the integral of f is somewhere between 0.58 and 0.83. (the actual value is about 0.693, as you will see later). Note how we were able to get a very reasonable estimate without computing an antiderivative or using upper and lower sums.



In fact, upper and lower sums are a special case of this bounding procedure where the two bounding functions that we choose are *piecewise constant functions*.

3.4. **Other inequalities.** Recall the triangle inequality, which states that for any two real numbers x and y, we have:

$$|x + y| \leq |x| + |y|$$

This can be generalized to more than two variables. The general form reads as:

$$|x_1 + x_2 + \cdots + x_n| \leq |x_1| + |x_2| + \cdots + |x_n|$$

Since an integration is an infinite analogue of a sum, the triangle inequality must have an analogue for integration. This reads as follows:

$$\left| \int_a^b f(x)\,dx \right| \leq \int_a^b |f(x)|\,dx$$

Note that we already have a geometric intepretation of both sides. The right side is the total unsigned area between the graph of f and the x-axis from point a to point b. The left side is the magnitude of the signed area from point a to point b. On the left side, we are adding the areas with signs (leading to possible cancellations) and then taking the absolute value in the end. On the right side, we are adding the absolute values to begin with. Thus, there is no scope for cancellation.

4. The beauty of symmetry

4.1. **The role of symmetry.** Before proceeding to the role of symmetry, we first explore how various transformations of the real line affect the value of the integral.

First, how does shifting by h affect integration?

$$\int_a^b f(x+h)\,dx = \int_{a+h}^{b+h} f(x)\,dx$$

This is an example of the chain rule in action, or the u-substitution. What we did is the following: start from the left side, and express $u = x + h$. Then $du/dx = 1$, and $f(x+h)$ becomes $f(u)$. The limits become $a + h$ and $b + h$, so we get:

$$\int_a^b f(x+h)\,dx = \int_{a+h}^{b+h} f(u)\,du$$

Now, however, u is a *dummy variable*, so we can replace this dummy variable by the dummy variable x. The term *dummy variable* is used for a variable that appears as the variable of integration or summation which hence cannot appear anywhere else. The dummy variable is by nature *local* to the integration or summation operation and hence its representing letter can be changed.

Graphically, the area staked out by f between $a + h$ and $b + h$ is the same as the area staked out by $f(x+h)$ between a and b. This is intuitively clear, because the graph of $f(x+h)$ is obtained from the graph of f via shifting left by h.

The other kind of operation that is of interest here is the flip-over, namely, sending x to $-x$. The relevant identity here is:

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

This again follows from a u-substitution.

These two basic ideas give us the interesting results we have on even, odd, and periodic functions:

(1) Suppose f is an odd continuous function on the interval $[-a, a]$. Then, its integral on $[-a, a]$ is 0. Roughly, this is because the integral on the interval $[-a, 0]$ cancels out the integral on the interval $[0, a]$, with each $f(x)$ being canceled by the corresponding $f(-x)$.

(2) More generally, if f is a continuous function on $[p, q]$ with half-turn symmetry about $((p+q)/2, f((p+q)/2))$, then the integral of f on $[p, q]$ is $(q-p)$ times the value $f((p+q)/2)$. Intuitively, this is the average value, and for every deviation above the value, there is a corresponding deviation below the value on the other side.

(3) Suppose f is an even continuous function on the interval $[-a, a]$. Then, its integral on $[-a, a]$ is twice its integral on $[0, a]$. This is because the picture of the function on $[-a, 0]$ is the same as the picture on $[0, a]$ (in the reverse order from left to right, but this does not affect area).

(4) More generally, if f enjoys mirror symmetry about $x = c$, the integral on $[c, c+h]$ equals the integral on $[c-h, c]$.

(5) If f is a periodic function that is continuous and defined for all real numbers, the integral of f over any interval of length equal to the period is the same. If f has a periodic antiderivative, then this integral is zero. If the period is h, and the integral over one period is k, then we can think of k/h as the long-run average value of f. More on this in the next section and in homework problems.

5. Mean-value theorem for integrals

5.1. **Statement of the theorem.** This result states that if f is a continuous function on a closed interval $[a, b]$, then there exists $c \in (a, b)$ such that

$$f(c) = \frac{\int_a^b f(x)\,dx}{b-a}$$

The right side of this expression is the *mean value*, or *average value*, of f on the interval $[a, b]$. Thus, this result simply states that a function attains its mean value somewhere on the interval.

Recall the earlier mean-value theorem:

If F is a function that is continuous on $[a, b]$ and differentiable on (a, b), then there exists $c \in (a, b)$ such that:

$$F'(c) = \frac{F(b) - F(a)}{b-a}$$

We now explain how the mean-value theorem for integrals follows from the (original) mean-value theorem. The idea is to pick F as an antiderivative for f. Then, $F' = f$, and F satisfies the hypotheses needed to apply the mean-value theorem for derivatives.

The left side of the (original) mean-value theorem is $F'(c)$, which equals $f(c)$. The numerator on the right side is $F(b) - F(a)$, which, by the fundamental theorem of integral calculus, is the same as $\int_a^b f(x)\,dx$. Thus, we get the necessary expression for the mean-value theorem for integrals.

5.2. **Mean value of periodic functions.** For a continuous function defined on all of $\mathbb{R}$, we can define the mean value of the function *over an interval*, but it does not make sense to define an *overall* mean value. For functions that go off to infinity in either direction, the mean value also goes off to infinity as we shift the intervals farther and farther off. On the other hand, for functions that are bounded, there is some hope in talking of a mean value.

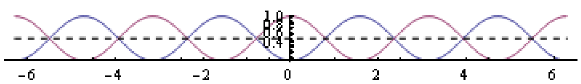
One class of functions for which a mean value makes eminent sense are periodic functions. As mentioned earlier, if f is periodic with period h, the integral of f over any interval of length h is a constant. Call this constant k. If F is an antiderivative of f, then F can be expressed as the sum of a periodic and a linear function. The linear part of F has slope k/h. Graphically, F is periodic with shift: the graph of F repeats after a length of h, but is vertically shifted by k.

Thus, there is a strong case to declare that the average value of f is k/h. Note that when f has a periodic antiderivative, then its average value is 0. For instance, sin and cos have average value 0, as we can see from the fact that they are symmetrically distributed above and below the x-axis.

On the other hand, the $\sin^2$ function has positive average value, and its antiderivative has a nontrivial linear component. We'll get back to this function in a short while.

For a periodic function f, it is *not* true that its mean value over *every* interval is k/h. However, any deviation from k/h is due to periodic, or seasonal fluctuation. As far as secular trends go, the mean value is k/h. In particular, if $k \neq 0$ (so that there is a nontrivial linear component) then, in the limit, as interval length becomes large, the mean value approaches k/h, even if the interval length is not a multiple of h. Intuitively, imagine that the period is 1, the average value on an interval of length 1 is k, and we take an interval of length 29417.3. Of this length, if we just took a sub-interval of length 29417, we would get average value k. The remaining interval length of 0.3 can upset things. But the integral on this remaining part will be divided by an interval length of 29417.3, so the deviation it causes will be small. The limit of the average value over an interval, as the interval length goes to ∞, is k/h.

5.3. The $\sin^2$ and $\cos^2$ functions.



A brief note on graphing and integrating the $\sin^2$ and $\cos^2$ functions. Although these functions can be integrated by a method called integration by parts, we will for now use another approach: the double angle formula. This states that:

$$\begin{aligned} \sin^2 x &= (1 - \cos(2x))/2 \\ \cos^2 x &= (1 + \cos(2x))/2 \end{aligned}$$

As a sanity check, note that if we add the right sides, we get 1, as we should.

We can now do graph transformations to plot $\sin^2 x$ and $\cos^2 x$. Note that it is now pictorially clear, even before we bother with actual integration, that both these functions have an average value of 1/2. This stands to reason: the sum of $\sin^2 x$ and $\cos^2 x$ is 1, and they're both the same graph shifted over, so on average, 1/2 should belong to $\sin^2 x$ and the other 1/2 should belong to $\cos^2 x$.

We can also formally integrate these functions:

$$\begin{aligned} \int \sin^2 x \, dx &= (x/2) - (\sin(2x))/4 + C \\ \int \cos^2 x \, dx &= (x/2) + (\sin(2x))/4 + C \end{aligned}$$

We see that the linear part of the antiderivative has slope 1/2, as expected, and the periodic part has periodicity π, again as expected, since $\sin^2$ and $\cos^2$ both have a periodicity of π.

Now, what the discussion about the mean value of periodic functions states is that, over a very long interval, the average value of the $\sin^2$ function is almost 1/2, even if the length of the interval is not a multiples of π.

These formulas for the average value of $\sin^2$ and $\cos^2$ appear in the context of waves. To calculate the energy of a wave involves integrating the square of the wave function over an interval. Since the wave function is of the form $A\sin(mx + \varphi)$, a slight generalization of the above calculations shows that the average energy per unit length of the wave is $A^2/2$. Similarly, if it is a time wave (so $A\sin(kt + \varphi)$) then the average energy per unit time is $A^2/2$. Note that the value of m doesn't affect this energy computation at all, because it is the value of A that affects the average value. (Note: There are different concepts of wave energy, and they usually do depend on the frequency, but the point here is that if the energy is simply defined as the integral of the square of the wave function, then the average value does not depend on the frequency).

6. Fun appendix: statistics application

We will not cover this in class, due to time considerations, but it is suggested you read through this while attempting the advanced homework problems related to this material.

6.1. The Gini coefficient setup.

Recall the setup that we had for the Gini coefficient. We arranged our huge population in increasing order of income. Then, for $x \in [0, 1]$, we defined $f(x)$ as the fraction of the income earned by the bottom x fraction of the population. With reasonable assumptions and using continuous approximations, we obtained that f is continuous and increasing, $f(0) = 0$, $f(1) = 1$, and $f(x) \leq x$ for all $x \in [0, 1]$. These were the observations that were necessary for doing the homework problems.

Another observation that was not necessary for doing the homework problems, but is nonetheless true, is that about the significance of f'. $f'(x)$ measures the fractional contribution of a person at level x (i.e., with x fraction of the population earning less). More precisely, we have:

$$f'(x) = \frac{\text{Income of person at level } x}{\text{Mean income}}$$

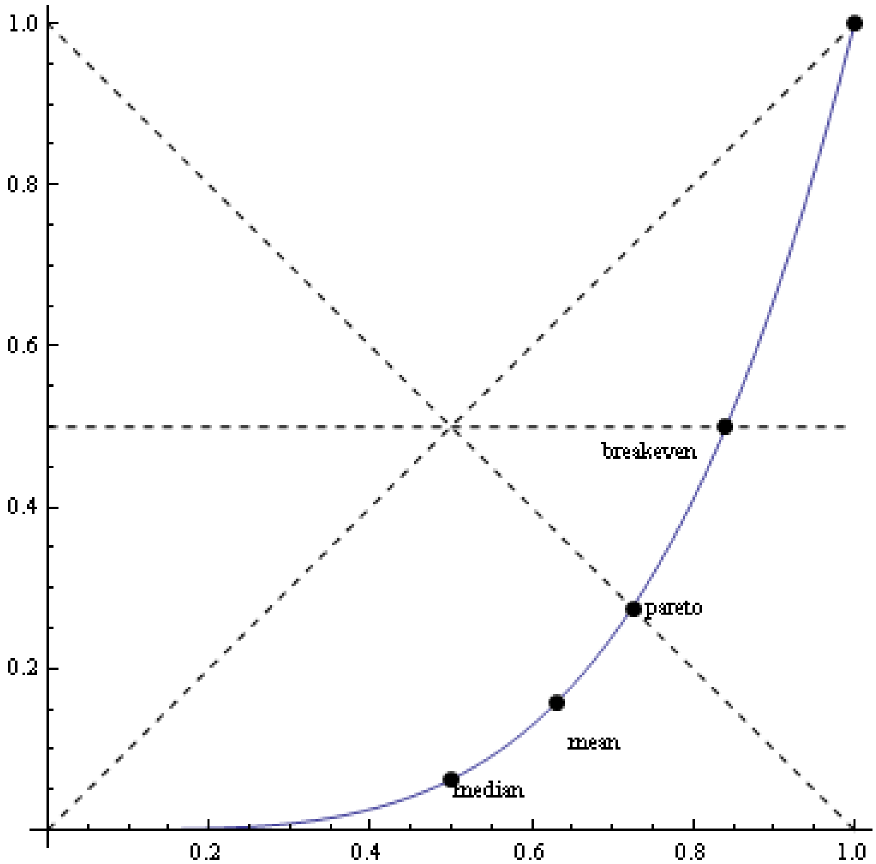
The reason why we need to normalize by mean income is that we have normalized things to $[0, 1]$. Here are some corollaries:

(1) f' is itself increasing, so f is concave up. In other words, people at a higher income level earn more. (In a degenerate case, f' may be constant in an interval, and f linear on that interval. This is when multiple people earn the same income. Unless otherwise stated, we'll assume no degeneracy).
(2) There is a c such that $f'(c) = 1$. This follows from the mean-value theorem. In other words, there is a person who earns the mean income.

6.2. Positions of interest.

What are all the positions x and values $f(x)$ of interest? Here are some of them:

(1) The value x such that the person at level x earns the *mean income*. Mathematically, this means that $f'(x) = 1$. Note that the existence of this value is guaranteed by the mean-value theorem while its uniqueness is guaranteed by the fact that the function is concave up.
(2) The value $1/2$. $f'(1/2)$ is the *median income*. $f(1/2)$ is the fraction of total income earned by the *bottom half* of the population.
(3) The *break-even point*, i.e., the value x such that $f(x) = 1/2$. This is the level x at which the bottom fraction earns the same total income as the remaining top fraction. The break-even point is always bigger than $1/2$ because of the concave up nature of the function. The income earned at this point (given by $f'(x)$) may also be of interest. The existence of a break-even point is guaranteed by the intermediate-value theorem and its uniqueness is guaranteed by the fact that f is increasing.
(4) The *Pareto point*, i.e., the value x such that $f(x) = 1 - x$. This is greater than $1/2$, but less than the break-even point. The income earned at this point may also be of interest. (You proved the existence and uniqueness of this point in your homework).



6.3. **Mean versus median? Mode? Looking at the third derivative.** Is the mean income greater than the median income? Equivalently, is the value x for which $f'(x) = 1$ greater than 1/2? There is no clear-cut answer. It turns out that the answer depends on whether the distribution of incomes is skewed more toward lower incomes or toward higher incomes.

A third statistical concept that comes up is that of the *mode*. Roughly speaking, the mode is the region where there is maximum clustering of incomes.

We thus want mathematical tools that will help answer the questions: (a) how can we compare mean and median? (b) how can we define mode in this situation?

The answer, interestingly, has something to do with the third derivative.

6.4. **The first, second and third derivatives.** You might remember that, when discussing how to graph functions to understand them better, one useful technique we discussed was to graph the function as well as its first and second derivative (and perhaps higher derivatives as well). Let us put this technique to use here.

Note that the graph of f measures the *cumulative income* earned by certain fractions of the population. This is good for some purposes, but for other purposes, we are interested in individual incomes. Though the graph of f contains this information, it is hidden in that graph. To see the information on individual incomes better, we consider the graph of f'.

As discussed above, the first derivative of f, denoted f', is the ratio of the income of the person at level x to the mean income. We know that f' is a continuous and increasing function on $[0, 1]$. We also know that $f'(0) \geq 0$ and that there is some $c \in (0, 1)$ such that $f'(c) = 1$. Thus, $f'(0) \leq 1$ and $f'(1) \geq 1$. We cannot say anything more conclusive.

Thus, f' is a continuous increasing function on $[0, 1]$ with $0 \leq f'(0) \leq 1$ and $f'(1) \geq 1$. The fact that f' is increasing corresponds to the fact that f is concave up. The value $f'(1/2)$ is the median income, and the point c where $f'(c) = 1$ is the point where the mean income is attained. We can see that the graph of f', subject to the given constraints, could be of many kinds. In particular, the median may occur before the mean or it may occur after the mean.

One advantage of drawing the graph of f' is that, compared to the graph of f, we can focus more in-depth on the way f' increases. We see that f'' measures the rate at which income increases (relative to mean income) as we move from the poorest to the richest. However, we also see that there are many unanswered questions. Where is f' concave up and concave down? Where does it rise most quickly and where does it rise most slowly? We see that the answers to these questions depend on f'''. In the regions where f''' is positive, f' is concave up, which means that the gain in income by moving to the right increases as we move to the right. In the regions where f''' is negative, f' is concave down, which means that the gain in income by moving to the right decreases as we move to the right.

We see that if f''' is positive throughout, that means that the relative gain in income for every slight increase in position goes up as we go from poorer to richer people. This means that the growth of f' is initially sluggish and picks up pace later. Such situations typically correspond to larger values of the mean.

On the other hand, if f''' is negative, that means that the relative gain in income for every slight increase in position goes down as we go from poorer to richer people. In other words, a small step up in the relative ranking means more in income gain terms for poor people than the same small step means for rich people. In this case, the growth of f' is sluggish for rich people and large for poor people. These situations correspond to the mean occurring relatively early.

A final question of interest is about the modal income. What is the income range that most people have? This corresponds to:

- The parts where the graph of f is closest to linear, i.e.,
- The parts where the graph of f' is closest to horizontal, i.e.,
- The parts where the graph of f'' attains its minimum values.
- (Probably) the parts where $f''' = 0$ and $f^{(4))} > 0$.

In other words, the modal segment is the segment where people's income is changing as little as possible with x.

6.5. **The peril of numbers.** Before you entered the world of functions and calculus, the only type of mathematical object you dealt with was a number. But once you entered the world of functions and calculus, you saw yourself dealing regularly with mathematical objects that were more complicated than mere numbers: for instance, sets of numbers, functions, collections of functions, points in the plane (which are ordered pairs of numbers) and so on. Some of these objects are so complicated that it is not possible to describe them using one or two or three numbers.

For instance, we saw that a partition of the interval $[a, b]$ is given by an increasing finite sequence of numbers starting at a and ending at b. Unfortunately, the finite sequence may be arbitrarily large. How do we compare different partitions? We saw two ideas for comparing partitions: (a) The notion of *finer* partition, whereby one refines the other. Unfortunately, given two partitions, it isn't necessary that either one be finer than the other. (b) The notion of the *norm* of a partition, which measures the size of the largest part. We can use the norm to compare two partitions. Unfortunately, a partition with smaller norm may not always behave like a *smaller* partition as far as the upper and lower sums of a particular function are concerned, as you discovered in the midterm.

So, one powerful idea is to use single numbers that measure *size* for complicated objects and reflect some underlying reality of those objects that is empirically useful. The drawback with that idea is that when we look only at that single number, we lose a lot of information about the original object. We may not be able to answer every question that comes up.

The distribution of incomes is another such complicated construct. It is described, as we saw, by this function $f : [0, 1] \to [0, 1]$. But a function cannot be described by a single number. So, instead we ask for single numbers that we can obtain from the function that measure some empirically useful reality about the function. One such number, which tries to measure the *extent of inequality*, is the Gini coefficient. But one problem with the Gini coefficient is that it only measures total inequality, and is not sensitive to inequalities within subpopulations. For instance, if everybody earns roughly the same income and a few people at the top earn a much much larger income, the Gini coefficient is close to 1, even though in some sense there is not much inequality among most people. In other words, the Gini coefficient is sensitive to *huge outliers* in the high-income direction.

That is why it is useful to have a number of different size measures that we can use, and to look at all of them. For instance, the break-even point and Pareto point are useful single numbers that give some intuition about the skew in the distribution of incomes. The median income or the level at which the mean is attained are also useful numbers. When you learn statistics, you will learn many other single numbers that capture useful information about aggregates and distributions. Keep in mind that for any single measure that you choose, there will always be examples of distributions where that measure does not seem to capture what you would like it to intuitively capture.

6.6. **Averages and compositional effects.** As some fun unwinding, here is a trick question. Suppose you have two countries A and B. Is it possible that the mean income in both A and B goes down, but the mean of no *individual* in either country goes down, and in fact, there are individuals whose mean income goes *up*?

Yes, it is possible. Suppose the mean income in country A is 100 money units and the mean income in country B is 400 money units. Imagine that there is a person in country A earning an income of 200 money units who chooses to migrate to country B and gets her income boosted to 300 money units. Assume that nobody else migrates, and nobody else's income changes.

The mean income of country A has gone down, because a person earning above the mean left the country. The mean income of country B has *also* goes down, because it just took in a person earning less than the mean income. The net effect is that both countries see a decline in their mean, but no individual is worse off – and at least one individual is better off! This is just one reason why *group averages and aggregates are not always reflective of individuals*. What we have described here is an example of a *compositional effect* – changes in group compositions affecting averages that reflect the opposite of what is happening at the individual level.

Of course, the group averages might still be useful in their own right, but the statistical error would be to *deduce things about individuals using group averages without taking into account compositional effects and the fluidity of group boundaries*.

Here are some other examples of the same phenomenon:

(1) Inter-sectoral migration: In rapidly industrializing nations such as China, agricultural productivity and industrial productivity are both rising about 5% per year. Yet, overall productivity is rising by something like 8% How is this happening? This is because the industrial sector is much more productive than the agricultural sector. As agricultural productivity increases, less people are needed in agriculture, and so people move from the (comparatively less productive) agricultural sector to the (comparatively more productive) industrial sector. This shifting of people from a less productive to a more productive sector itself causes an increase in productivity independent of the increase in productivity within each sector. Here, agriculture plays the role of the poorer nation A and industry plays the role of the richer nation B.

(2) Inter-level migration in calculus: Imagine that one of you, who is doing badly in the 150s, drops down to the 130s, which are a cakewalk for you. Then, the average mathematical skill of the 150s students increases, the average mathematical skill of the 130s student increases, yet there may probably be a net *decrease* in the overall average mathematical skill of the population, if your mathematical skills decline after you're no longer subjected to the rigors of the 150s.

AREA COMPUTATIONS USING INTEGRATION

MATH 152, SECTION 55 (VIPUL NAIK)

Corresponding material in the book: Section 5.5, 6.1.

Difficulty level: Moderate. The basic computational ideas are of easy to moderate difficulty, but some of the slicing ideas at the end are somewhat hard.

What students should definitely get: The application of integral to computing areas. The idea of slicing and integration of slice lengths.

Executive summary

Words ...

(1) We can use integration to determine the area of the region between the graph of a function f and the x-axis from $x = a$ to $x = b$: this integral is $\int_a^b f(x)\,dx$. The integral measures the signed area: parts where $f \geq 0$ make positive contributions and parts where $f \leq 0$ make negative contributions. The magnitude-only area is given as $\int_a^b |f(x)|\,dx$. The best way of calculating this is to split $[a, b]$ into sub-intervals such that f has constant sign on each sub-interval, and add up the areas on each sub-interval.

(2) Given two functions f and g, we can measure the area between f and g between $x = a$ and $x = b$ as $\int_a^b |f(x) - g(x)|\,dx$. For practical purposes, we divide into sub-intervals so that on each sub-interval one function is bigger than the other. We then use integration to find the magnitude of the area on each sub-interval and add up. If f and g are both continuous, the points where the functions *cross* each other are points where $f = g$.

(3) Sometimes, we may want to compute areas against the y-axis. The typical strategy for doing this is to interchange the roles of x and y in the above discussion. In particular, we try to express x as a function of y.

(4) An alternative strategy for computing areas against the y-axis is to use formulas for computing areas against the x-axis, and then compute differences of regions.

(5) A general approach for thinking of integration is in terms of slicing and integration. Here, integration along the x-axis is based on the following idea: divide the region into vertical slices, and then integrate the lengths of these slices along the horizontal dimension. Regions for which this works best are the regions called *Type I regions*. These are the regions for which the intersection with any vertical line is either empty or a point or a line segment, hence it has a well-defined length.

(6) Correspondingly, integration along the y-axis is based on dividing the region into horizontal slices, and integrating the lengths of these slices along the vertical dimension. Regions for which this works best are the regions called *Type II regions*. These are the regions for which the intersection with any horizontal line is either empty or a point or a line segment, hence it has a well-defined length.

(7) Generalizing from both of these, we see that our general strategy is to choose two perpendicular directions in the plane, one being the direction of our slices and the other being the direction of integration.

Actions ...

(1) In some situations we are directly given functions and/or curves and are asked to find areas. In others, we are given real-world situations where we need to find areas of regions. Here, we have to find functions and set up the integration problem as an intermediate step.

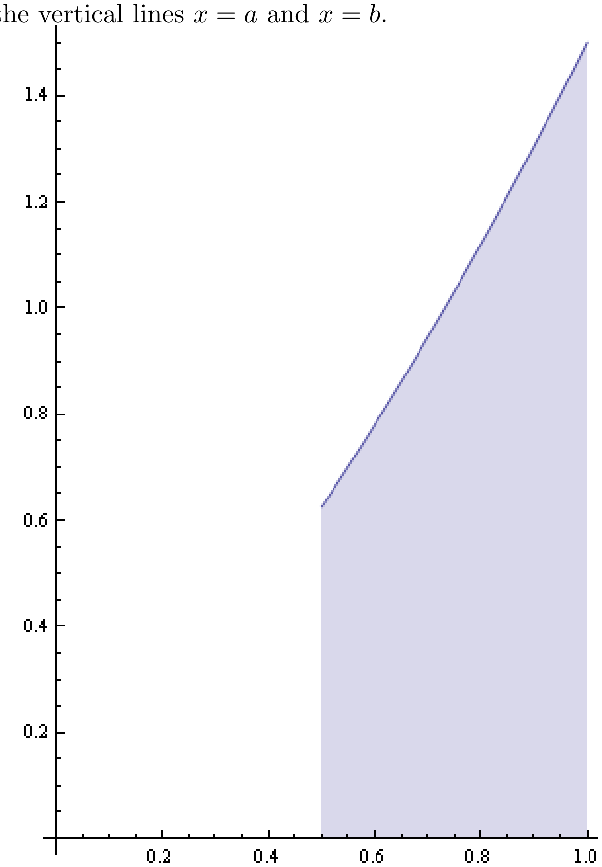
(2) In all these situations, it is important to draw the graphs in a reasonably correct way. This brings us to all the ideas that are contained in graph drawing. Remember, here we may be interested in simultaneously graphing more than one function. Thus, in addition to being careful about each function, we should also correctly estimate where one function is bigger than the other, and find

(approximately or exactly) the intersection points. (Go over the notes on graph-drawing, and some additional notes on graphing that weren't completely covered in class).

(3) In some situations, we are asked to find the area(s) of region(s) bounded by the graphs of one, two, three, or more functions. Here, we first need to sketch the figure. Then, we need to find the interval of integration, and if necessary, split this interval into sub-intervals, such that on each sub-interval, we know exactly what integral we need to do. For instance, consider the region between the graphs of sin, cos, and the x-axis. Basically, the idea is to find, for all the vertical slices, the upper and lower limits of the slice.

1. Integral and area: against the x-axis

1.1. **Definite integral as the signed area between the graph and the x-axis.** Suppose f is a continuous function on a closed interval $[a, b]$. The graph of f forms a curve in the plane $\mathbb{R}^2$. Consider the signed area between this curve and the x-axis. This is the area of the region bounded by the graph, the x-axis, and the vertical lines $x = a$ and $x = b$.



The basic result of integration is that this area equals the definite integral

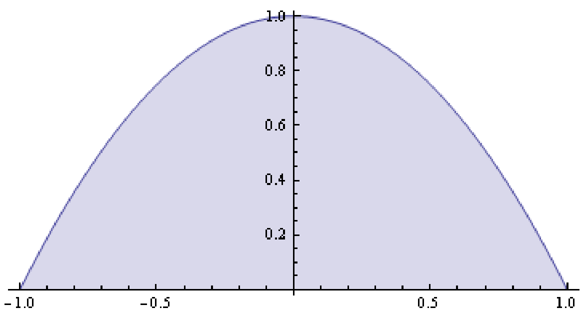
$$\int_a^b f(x)\,dx$$

If $f(x) \geq 0$ for all $x \in [a, b]$, i.e., if the graph is entirely in the upper half-plane (possibly hitting the boundary x-axis), then this integral is nonnegative, and its value is the magnitude of the area. If $f(x) \leq 0$ for all $x \in [a, b]$, i.e., if the graph is entirely in the lower half-plane (possibly hitting the boundary x-axis), then this integral is zero or negative, and its value is the *negative* of the magnitude of the area. If the function has parts where it is positive and parts where it is negative, then the parts where it is positive make positive contributions and the parts where it is negative make negative contributions.

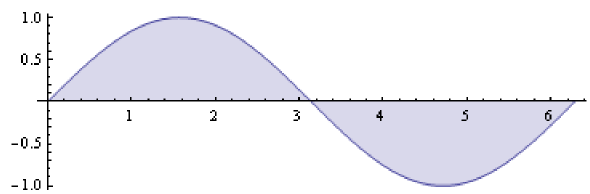
For instance, consider the function $f(x) := 1 - x^2$. We want to find the area between the x-axis and the graph of the part of this function that is above the x-axis.

First, note that the graph is above the x-axis on $(-1, 1)$. Thus, in order to find the area, we need to perform the integration:

$$\int_{-1}^{1} (1 - x^2)\, dx$$



We can do this integration by finding an antiderivative and evaluating it between limits. We take $x - x^3/3$ as the antiderivative. Evaluating it between limits gives the value $4/3$. Thus, the area of the region we are interested in is $4/3$.



1.2. **Measuring unsigned area.**

Suppose we want to measure the total area between the graph of the sine curve and the x-axis over one period, say $[0, 2\pi]$. In other words, we want to compute the integral

$$\int_{0}^{2\pi} \sin x\, dx$$

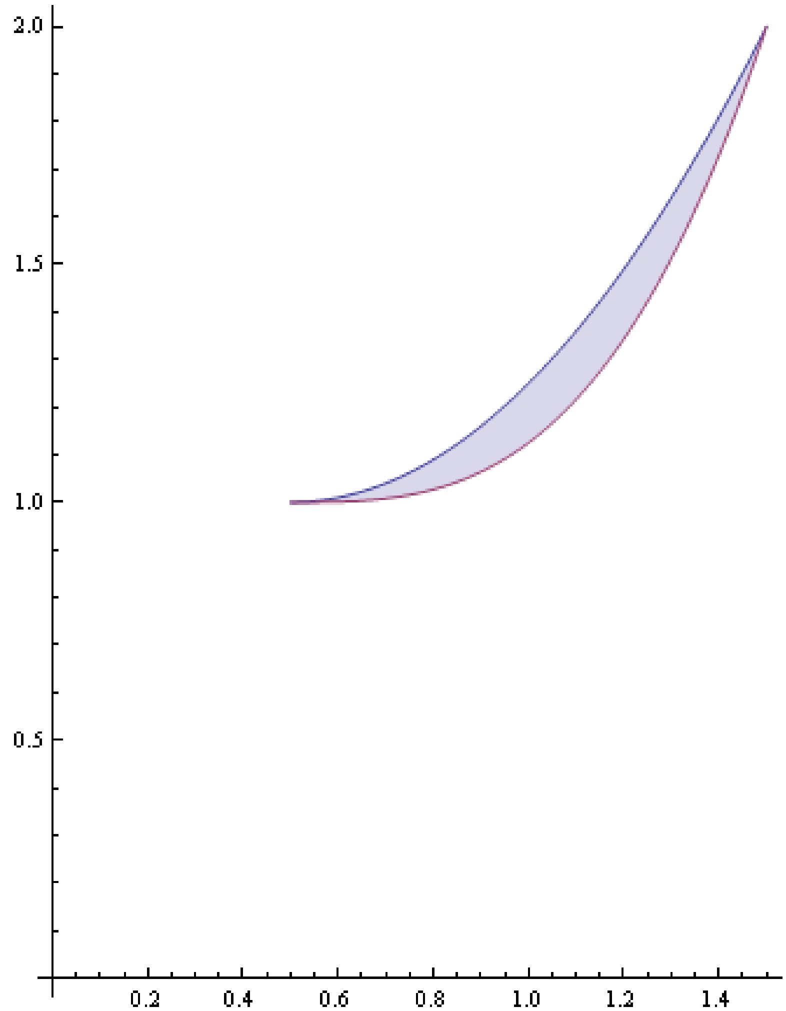
We know that $-\cos$, which is an antiderivative for sin, also has a period of 2π. Hence, its value between limits is zero, so the above integral is zero. Thus, the total *signed area* between the graph of the sine curve and the x-axis is zero. This makes sense graphically. The positive area between the sine curve and the x-axis on the interval $[0, \pi]$ is canceled by a negative area of equal magnitude between the x-axis and the sine curve on the interval $[\pi, 2\pi]$. Why are the two areas the same? There are plenty of ways of seeing this geometrically. For instance, we have $\sin(\pi + \theta) = -\sin\theta$ for all angles θ.

Suppose now that, instead of measuring the signed area, we are interested in measuring the unsigned area. The *unsigned area* between the graph of a function f and the x-axis on an interval $[a, b]$ is given by

$$\int_{a}^{b} |f(x)|\, dx$$

Equivalently, we break the interval $[a, b]$ into subintervals such that $f \geq 0$ or $f \leq 0$ on each subinterval. Then we calculate the magnitude of the integral on each subinterval and add these magnitudes.

In the case of the sine function, we can partition $[0, 2\pi]$ at π, to get the subintervals $[0, \pi]$ and $[\pi, 2\pi]$. On $[0, \pi]$, the integral is $[-\cos x]_0^{\pi}$, which simplifies to 2. On $[\pi, 2\pi]$, the integral is -2, and its magnitude is 2. The total magnitude of the integral is thus $2 + 2$, and we know that $2 + 2 = 4$. Hence, the unsigned area between the graph of sin and the x-axis on $[0, 2\pi]$ is 4.



1.3. **Area between two graphs.**

Suppose f and g are two continuous functions. To measure the signed area between the graphs of f and g between the points a and b, we compute the integral

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

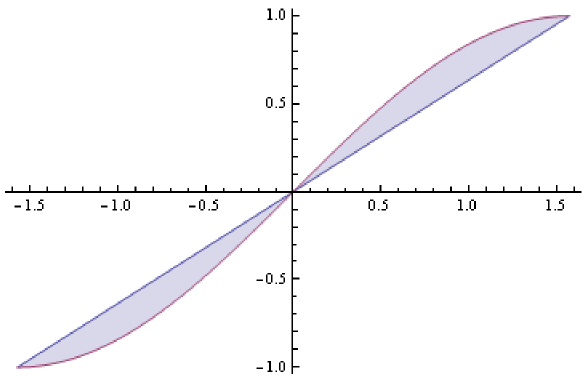
Here, the subintervals where f is bigger than g make positive contributions and the subintervals where g is bigger than f make negative contributions. If we are interested in the unsigned area, whereby we want positive contributions regardless of which function is bigger, we consider the integral

$$\int_a^b |f(x) - g(x)|\, dx$$

To compute this, we break up the interval $[a, b]$ into subintervals based on whether f or g is smaller (the overtaking can happen at points where $f(x) = g(x)$). We then compute the integral of $f - g$ (or $g - f$, depending on which is bigger) on each subinterval and add up the magnitudes.

For instance, consider the unsigned area between the graphs of $f(x) = 2x/\pi$ and $g(x) = \sin x$ on the interval $[-\pi/2, \pi/2]$. We see that $f(x) = g(x)$ at $-\pi/2, 0, \pi/2$. On $(-\pi/2, 0)$, $f(x) > g(x)$, and on $(0, \pi/2)$, $g(x) > f(x)$. Thus, the integral is:

$$\int_{-\pi/2}^{0} (2x/\pi - \sin x)\, dx + \int_0^{\pi/2} (\sin x - 2x/\pi)\, dx$$

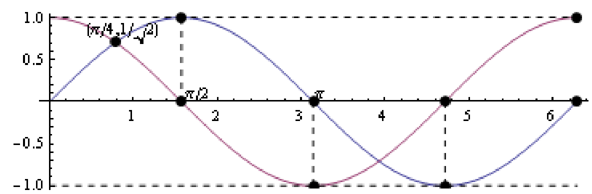


We can calculate and simplify both these integrals. Note that instead of computing indefinite integrals for both separately, we can note that the two functions are negatives of each other, so if we compute an antiderivative for the first, the antiderivative for the second is its negative. We get:

$$[x^2/\pi + \cos x]_{-\pi/2}^{0} + [-\cos x - x^2/\pi]_0^{\pi/2}$$

Both parts are $1 - \pi/4$, and we thus get $2 - \pi/2$. Since π is approximately 3.14, this is approximately 0.43.

Why are the two integrals the same? This can be seen geometrically from the fact that both f and g are odd, so the picture from $-\pi/2$ to 0 is the same as the picture from 0 to $\pi/2$, subjected to a half-turn about the origin. Thus, the magnitude of the two areas is the same.



1.4. **Areas bounded by graphs of different functions.**

Sometimes, the bounding curves for an area come from different functions. In this case, it makes sense to split up the interval of integration into subintervals so that we are dealing with only one function in each subinterval. For instance, consider computing the area of the region between the x-axis and the graphs of sin and cos on the interval $[0, \pi/2]$. On the interval $[0, \pi/4]$, this is the definite integral of the sin function, and on the interval $[\pi/4, \pi/2]$, this is the definite integral of the cos function. The total area is the sum of the values of these two definite integrals.

As we can see, both integrals are $1 - 1/\sqrt{2}$, and the total integral is $2 - \sqrt{2}$, which is approximately 0.59.

Why are the two integrals the same? We can see graphically that the two areas being measured are mirror images of each other about the line $x = \pi/4$. This is because for any angle θ, $\cos(\pi/2 - \theta) = \sin\theta$.

For the second midterm, you are responsible only for the material till this point.

2. Area computations as slicing, and other methods

This way of thinking about area computations will turn out to be useful for the subsequent topic, which is volume computations. It also makes it possible to compute areas of shapes oriented somewhat differently from before.

2.1. **Vertical slicing.**

So far, the situations where we've been computing areas are: area between the graph of a function and the x-axis, area bounded between graph of a function, the x-axis, and two vertical lines, area between the graphs of two functions, area bounded by the graphs of two functions and two vertical lines.

In all these situations, the region Ω whose area we need to compute has the property that the intersection of Ω with any vertical line is either empty or a line segment. Regions of this kind are sometimes called Type I regions. For Type I regions, the general formula for the unsigned area is:

$$\int (\text{Length of the line segment as a function of } x)\, dx$$

This process can be thought of as *vertical slicing*. We are dividing the area that we want to measure into vertical slices, and then integrating the length along the perpendicular axis (which is horizontal).

2.2. **Horizontal slicing.** Horizontal slicing is a lot like vertical slicing, but works for regions where the role of vertical and horizontal is replaced.

Horizontal slicing works for regions Ω which have the property: the intersection of Ω with any horizontal line is either empty or a line segment. Regions of this type are sometimes called Type II regions. The formula for the area of a Type II region is

$$\int (\text{Length of the line segment as a function of } y)\, dy$$

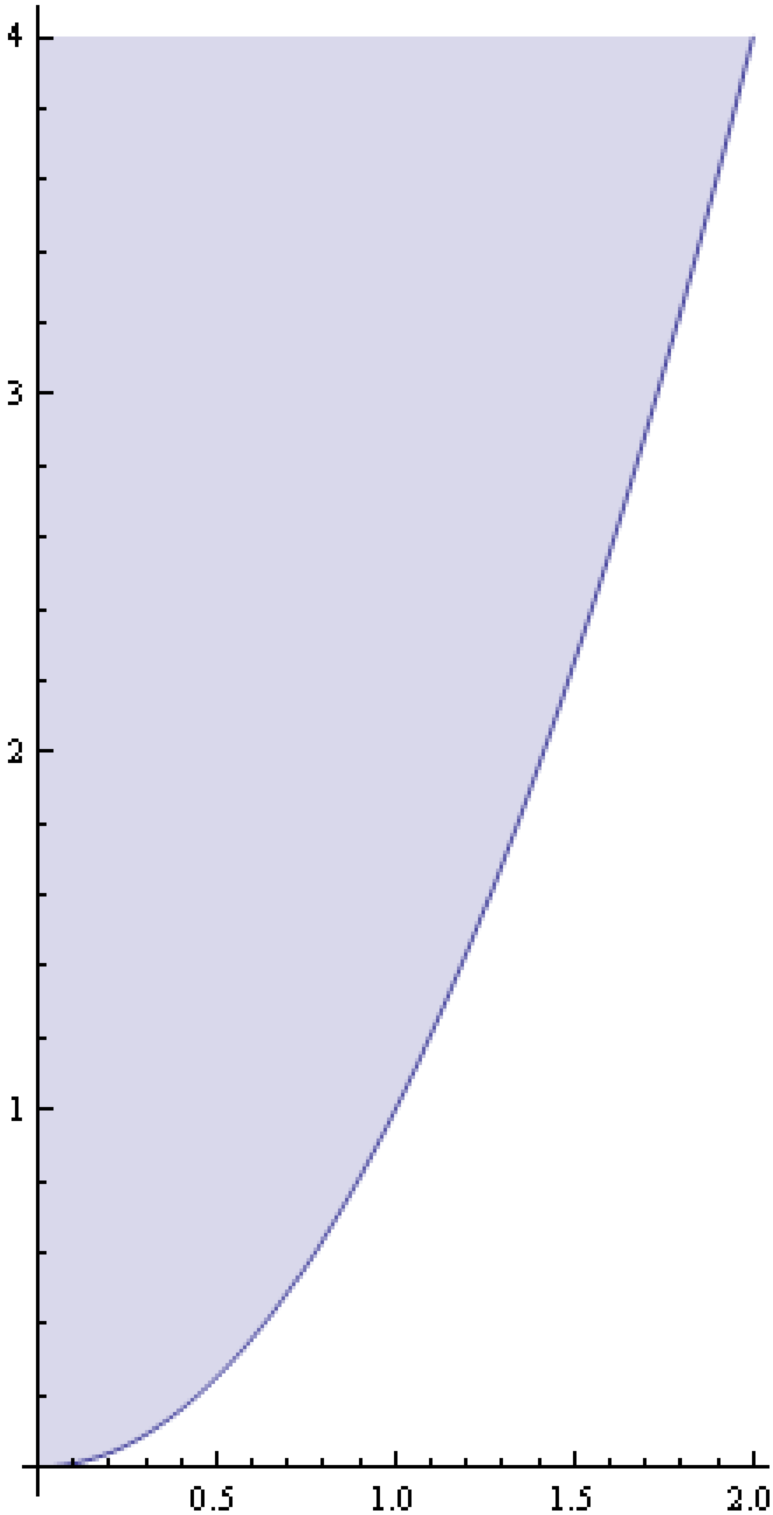
This process can be thought of as *horizontal slicing*. We are dividing the area that we want to measure into horizontal slices, and the integrating the length along the perpendicular axis (which is vertical).

Thus, we have seen two processes of breaking up an area into slices: vertical slicing (where we integrate the lengths along a horizontal axis) and horizontal slicing (where we integrate the lengths along a vertical axis).

Notice that both these procedures are variants of the same basic procedure: choose two mutually perpendicular directions, such that all lines in one direction have intersection with the region that is either empty or a line segment. Then, integrate the length of the line segment along the perpendicular direction.

Note also that the extreme case of both these occurs in rectangles. Here, whether we use horizontal or vertical slicing, we are integrating a constant function.

2.3. **Regions whose area can be computed by integration in multiple ways.** Consider the region bounded by the line $y = 4$, $y = x^2$, and the y-axis. This is both a Type I and a Type II region, so we can determine its area by vertical slicing as well as by horizontal slicing. Let's first compute the area by vertical slicing.



By vertical slicing, the interval is $[0, 2]$, and the lower and upper functions are x^2 and 4 respectively. Thus, the length of the line segment in each vertical slice is $4 - x^2$. The area is thus:

$$\int_0^2 (4 - x^2)\, dx = [4x - (x^3/3)]_0^2 = 8 - 8/3 = 16/3$$

We could also integrate using horizontal slicing. For this, we express x in terms of y. We get $x = \sqrt{y}$, with $y \in [0, 4]$. Measuring the area between this and the y-axis, we get:

$$\int_0^4 \sqrt{y}\, dy = [y^{3/2}/(3/2)]_0^4 = 8/(3/2) = 16/3$$

When we later introduce the concept of *inverse function*, we will notice that what we've just done is moved from integrating one function to integrating its inverse function. We'll also see a relationship between this and integration by parts next quarter.

3. Areas of regions given by inequalities

Suppose a region of the plane is defined by a set of inequalities. In other words, the region is defined as the set of all points in the plane that satisfy a given system of inequalities. How do we find its area?

The first step is to identify the *bounding* lines/curves for this region. The bounding lines are typically the lines given by the case where equality holds instead of inequality. Once we have found these boundary curves, we can then try to use horizontal or vertical slicing to determine the area. In some cases, it makes sense to divide the region into sub-regions so that it is easy to tackle each sub-region separately by slicing.

Another complication is that the boundary curves may not be graphs of functions. Often, they may be graphs of relations, i.e., the set of points (x, y) satisfying $F(x, y) = 0$ for some two-variable function F. In these cases, we try to break it up into functions. We consider some examples.

3.1. **The example of the circular disk.** Consider the region $1 \le x^2 + y^2 \le 2$. In other words, we are looking at the set of points (x, y) such that $x^2 + y^2 \in [1, 2]$. We easily see graphically that this region is bounded on the inside by the circle $x^2 + y^2 = 1$ and on the outside by $x^2 + y^2 = 2$. The region is called a *circular annulus*. To find the area of the annulus, we thus need to subtract the area of the disk $x^2 + y^2 \le 1$ from the area of the disk $x^2 + y^2 \le 2$.

Now, it so happens that we know formulas for the areas of these disks: they are π and 2π respectively, so the difference of areas is $2\pi - \pi = \pi$. If we did not know these formulas, we would need to break up the circle into graphs of functions $\pm\sqrt{r^2 - x^2}$. Unfortunately, integrating these functions requires a trigonometric substitutions, so illustrating this idea would take us too far afield.

VOLUME COMPUTATIONS USING INTEGRALS

MATH 152, SECTION 55 (VIPUL NAIK)

Corresponding material in the book: Section 6.2, 6.3.

Difficulty level: Hard (degree of hardness depends on your visuo-spatial skills and prior exposure to these ideas).

What students should definitely get: The basic constructive ideas for volume: cylinders with constant and varying cross section, surfaces of revolution (disks and washers). The formula and mechanics for the shell method.

Note: I haven't included pictures, since these are hard to draw. I suggest that you look at pictures in the book, which are pretty well done, and of course, pay attention in class.

Executive summary

Words ...

(1) The cross section method for computing volume is an analogue of the two-dimensional area computation method: our slices are replaced by cross sections by planes parallel to a fixed plane, and the line of integration is a line perpendicular to the planes. One-dimensional slices are replaced by two-dimensional cross sections.

(2) Suppose Ω is a region in the plane. We can construct a right cylinder with base Ω and height h. This is obtained by translating Ω in a direction perpendicular to its plane by a length of h. The cross section of this right cylinder along any plane parallel to the original plane looks like Ω if that plane is within range. The volume is the product of the area of Ω and the height h. This is also called the right cylinder with constant cross section Ω.

(3) We can also construct an oblique cylinder. Here, the direction of translation is not perpendicular to the original plane. The total volume is the product of the area of Ω and the height perpendicular to Ω. Oblique cylinders are to right cylinders what parallelograms are to rectangles.

(4) More generally, the volume of a solid can be computed using the cross section method. Here, we choose a direction. We measure areas of cross sections along planes perpendicular to that direction, and integrate these areas along that direction.

(5) This general approach has another special case that is perhaps as important as right cylinders. These are the *cones* (there are right cones and oblique cones). A cone is obtained by taking a region in a plane and connecting all points in it to a point outside the plane. It is a right cone if that point is directly above the center of the region. The volume of a cone is 1/3 times the product of the base area and the height, i.e., the perpendicular distance from the outside point to the plane. In particular, a cone has one-third the volume of a cylinder of the same base and height.

(6) A solid of revolution is a solid obtained by revolving a region in a plane about a line (called the axis of revolution). The volume of a solid of revolution can be computed by choosing the axis as the axis of integration and using the planes of cross section as planes perpendicular to it. These cross sections are either circular disks or annuli.

(7) The *disk method* is a special case of the above, where the region between revolved is supported on the axis of revolution. For instance, consider the region between the x-axis, the graph of a function f, and the lines $x = a$ and $x = b$. The volume of the corresponding solid of revolution is $\pi \int_a^b [f(x)]^2 \, dx$. This is because the radius of the cross section disk at $x = x_0$ is $|f(x_0)|$.

(8) The *washer method* is the more general case where the region need not adhere to the axis of revolution. For instance, consider two nonnegative functions f, g and suppose $0 \le g \le f$. Consider the region bounded by the graphs of these two functions and the lines $x = a$ and $x = b$. The volume of the corresponding solid of revolution is $\pi \int_a^b ([f(x)]^2 - [g(x)]^2) \, dx$. Note that in the more general case

where the functions cross each other, we may need to split into sub-intervals so that we can apply the washer method on each sub-interval.

(9) The shell method works for situations where we revolve about the y-axis the region made between the graph of a function and the x-axis. The formula here is $2\pi \int_a^b xf(x)\,dx$ for f nonnegative and $0 < a < b$. If f could be positive or negative, we use $2\pi \int_a^b x|f(x)|\,dx$.. More generally, if we are looking at the region between the graphs of f and g (vertically) with $g \le f$, we get $2\pi \int_a^b x[f(x) - g(x)]\,dx$. If we don't know which one is bigger where, we use $2\pi \int_a^b x|f(x) - g(x)|\,dx$.

Actions ...

(1) To compute the volume using cross sections, we first need to set things up so that we know the cross section areas as a function of the position of the plane. For this, it is usually necessary to use either coordinate geometry or basic trigonometry, or a combination.
(2) A solid occurs as a solid of revolution if it has complete rotational symmetry about some axis. In that case, that axis is the axis of revolution and the original region that we need is obtained by taking a cross section in any plane containing the axis of revolution and looking at the part of that cross section that is on one side of the axis of revolution.
(3) For solids of revolution, be particularly wary if the original figure being revolved has parts on both sides of the axis of revolution. If it is symmetric about the axis of revolution, delete one side.
(4) Be careful about the situations where you have to be sign-sensitive and the situations where you do not. In the disk method sensitivity to signs is not important. In the washer method and shell method, it is.
(5) The farther the shape being revolved is from the axis, the greater the volume of the solid of revolution.
(6) The average value point of view is sometimes useful for understanding such situations.

1. Motivation: from area to volume

1.1. **What are we trying to do?** Our purpose right now is to find formulas for the volumes of various three-dimensional figures. This is a little like our attempts at finding areas of regions, which we successfully did, at least for some regions. Wait, what?

The process that we are going through is something whose broad outlines should be familiar to you. Think back, for instance, to how we dealt with differentiation. We first computed formulas for the derivatives of a few functions. Then, we considered all the ways that new functions can be created from old functions. Finally, we found formulas that tackled each way of creating a new function from an old function. Combined with the knowledge of how to differentiate the basic functions, this allowed us to differentiate any function given to us using a simple set of rules.

Similarly, when trying to figure out general strategies for finding limits, we started out by computing a few basic limits, and looking at rules for computing limits of functions created from simpler functions.

We did a similar thing for integration: we learned rules for finding antiderivatives for some basic functions, and then we learned various processes of combination (something we're not quite done with yet). The overall strategy is:

(1) Find out how to deal with basic situations.
(2) Identify the typical ways that basic situations are combined to create more complicated situations.
(3) For each such process of combining basic situations into more complicated situations, identify a way of reducing the problem for the complicated situation in terms of the basic situations.

We shall consider how to deal with volumes. Our main difficulty in calculating volumes is with step (2) – we don't have an understanding of the systemic processes whereby new three-dimensional figures can be created. Once we do, we can try to find a volume formula for each such process, and use these formulas to calculate the areas of a number of figures.

1.2. **A recapitulation of how we handled area computations.** We have so far dealt with two kinds of area computations. The first is computing areas *against the x-axis*. Here, we are measuring the area bounded between a curve and the x-axis, or the area between two curves and two vertical lines.

Let us reflect more carefully on how we can characterize these situations geometrically. In all these situations, the region Ω that we have has the property that the intersection of Ω with any vertical line is either empty or a line segment. Regions of this kind are sometimes called Type I regions. For Type I regions, the general formula for the unsigned area is:

$$\int (\text{Length of the line segment as a function of } x)\, dx$$

This process can be thought of as *vertical slicing*. We are dividing the area that we want to measure into vertical slices, and then integrating the length along the perpendicular axis (which is horizontal).

The other procedure that we saw for integration is integration against the y-axis. This kind of integration works for regions Ω which have the property: the intersection of Ω with any horizontal line is either empty or a line segment. Regions of this type are sometimes called Type II regions. The formula for the area of a Type II region is

$$\int (\text{Length of the line segment as a function of } y)\, dy$$

This process can be thought of as *horizontal slicing*. We are dividing the area that we want to measure into horizontal slices, and the integrating the length along the perpendicular axis (which is vertical).

Thus, we have seen two processes of breaking up an area into slices: vertical slicing (where we integrate the lengths along a horizontal axis) and horizontal slicing (where we integrate the lengths along a vertical axis).

Notice that both these procedures are variants of the same basic procedure: choose two mutually perpendicular directions, such that all lines in one direction have intersection with the region that is either empty or a line segment. Then, integrate the length of the line segment along the perpendicular direction.

Note also that the extreme case of both these occurs in rectangles. Here, whether we use horizontal or vertical slicing, we are integrating a constant function.

1.3. **How does this general idea carry over to three dimensions?** $1 + 1 = 2$, but $1 + 1 \neq 3$. So, the idea of choosing two mutually perpendicular directions, one for the slices and the other as the direction of integration, does not work directly for computing volumes. However, it *is* true that $2 + 1 = 3$. This suggests a slightly different strategy to measure the volume of a three-dimensional region Ω: choose a plane π and a line ℓ perpendicular to π. Now, measure the areas of the intersection of Ω with regions perpendicular to π, and integrate this area along ℓ.

In other words, the *slices* are two-dimensional and parallel to each other, and the direction of the line along which we integrate is perpendicular to those planes.[1]

In the forthcoming section, we look at some systemic processes for creating three-dimensional structures and for slicing them suitably.

Sidenote: Distinction between a disk and a circle. Henceforth, when I refer to a *circle*, I refer to the *boundary*, i.e., the set of points whose distance from the center equals the radius. When I want to talk of the circle along with the interior region, I will use the term *circular disk*, or, more briefly, *disk*. When I want to simply look at the interior and exclude the boundary, I will use the term *interior of the disk* or *open disk*.

I will, however, switch between a circle and its disk easily, hence when I talk about the center, radius, or diameter of a disk, I am referring to those notions for its boundary circle.

2. Creating three-dimensional structures

2.1. **The general concept of a right cylinder.** What you may have been told is a *cylinder* is more appropriately termed a *right circular cylinder*. The adjectival qualifier *circular* indicates that the base is a circle (more precisely, the boundary is a circle and the base is a circular disk). The term *right cylinder*, in general, means something like a right circular cylinder except that the base need not be circular.

Basically, we take a region Ω in the plane with boundary Λ and then translate Ω along a direction perpendicular to the plane for a fixed length. That fixed length is called the *height* of the right cylinder. This gives the (solid) right cylinder with cross section Ω. The curved surface of the cylinder is the boundary

[1]The reason why we are forced to use $2 + 1 = 3$ rather than $1 + 2 = 3$ is because the only kind of integration that we have explicitly dealt with is integration in one variable, i.e., along a line.

of this, which is obtained by translating Λ in a direction perpendicular to the plane of Ω. The two *caps* are the two copies of Ω located at the two ends.

The term *cross section* here refers to the fact that if we take any plane parallel to the plane of Ω, its intersection with the right cylinder is a copy of Ω if the plane is located in the relevant region; otherwise it is empty.

You may have heard the term *cross section* arising in different contexts. It basically means the intersection with a given plane. For instance, in biology, when studying things ranging from tree trunks to micro-organisms and cells, we take cross-sections in various directions.

The right cylinder has a constant cross section. In this sense, it is similar to a rectangle in two dimensions, which has constant cross sections.

The volume of a right cylinder is given by:

$$\text{Volume of right cylinder} = \text{Area of cross section} \times \text{Height}$$

Some particular cases of interest:

- When the base cross section is a circular disk, we get a *right circular cylinder*.
- When the base cross section is a polygon, we get what is often called a *prism*. In particular, when the base is a rectangle, we get a rectangular prism.

2.2. **Oblique cylinders.** A slight variant on right cylinder is oblique cylinder. Oblique cylinders are to parallelograms what right cylinders are to rectangles. Here is the construction of an oblique cylinder.

Start with a region Ω in a plane π. Now, choose a direction in space that is not parallel to the plane π. Translate Ω by a length l along this direction. The region traced this way is termed an oblique cylinder.

The volume of an oblique cylinder is given by:

$$\text{Volume of oblique cylinder} = \text{Area of cross section} \times \text{Height perpendicular to cross section}$$

Equivalently:

$$\text{Volume of oblique cylinder} = \text{Area of cross section} \times \text{Length } l \times \sin\theta$$

where θ is the angle between the plane π and the direction of translation. In particular, when $\theta = \pi/2$, we get a right cylinder.

If we consider cross sections of oblique cylinder parallel to π, each of these cross sections looks like Ω. However, unlike the right cylinder case, the *location* of the Ω in the cross section plane keeps changing.

2.3. **More oblique than oblique.** In fact, it is possible to get eve nmore oblique than oblique – we translate a shape in a plane along a direction other than the plane, but we keep changing the direction. Thus, each cross section still has the same shape, but its location changes rather unpredictable. We'll see some such situations in a homework/quiz/test.

2.4. **Variable cross sections.** We next consider a situation where the cross sections are variable. This is no longer a right cylinder, but we can use the idea mentioned a little while ago – integrating the area function. Earlier, we multiplied a constant with the height over which that constant was valid. Now, we integrate a variable function over an interval. Remember, integration is like multiplication where the thing you're trying to multiply keeps changing. The important thing is that the area of each cross section should be something we know how to measure. The general formula is:

$$\text{Volume} = \int (\text{Area of cross section perpendicular to } x)\, dx$$

2.5. **Cones.** One case of particular importance, where it is useful to remember a general approach as well as the specific answer, is that of the *cone*. A cone is defined as follows. Suppose Ω is a region in a plane π and P is a pont not in π. The cone corresponding to Ω and P is the union of all the line segments joining P to points in Ω.

Some examples of cones are:

(1) A *tetrahedron* is a cone where the base is a triangular region.

(2) A *right circular cone* is a cone where the base is a circular disk.
(3) A *pyramid* is a cone where the base is some polygon.

When we set up the cross section integration for the cone, we see that the shape of any cross section parallel to π is the same as that of Ω, but the size is different. We can use similar triangles to determine the size. If we define:

$$\alpha = \frac{\text{Distance from } P \text{ to cross section}}{\text{Distance from } P \text{ to } \pi}$$

Then the linear measurements for the cross section are α times the corresponding linear measurements for Ω. Since area is two-dimensional, the area of the cross section is α^2 times the area of Ω. We now get that the overall volume is:

$$\int_0^h (x/h)^2 \operatorname{Ar}(\Omega)\, dx$$

Plugging in $x = \alpha h$, we get:

$$\int_0^1 \alpha^2 \operatorname{Ar}(\Omega) h \, d\alpha$$

We pull out the constants, and get:

$$\operatorname{Ar}(\Omega) h \int_0^1 \alpha^2 \, d\alpha$$

The integral now gives $1/3$, and we thus get:

$$\text{Volume of cone} = \frac{1}{3} \text{Area of base region} \times \text{Height}$$

Now you understand why you have that $1/3$ in the formula for the volume of a cone: $(1/3)(\pi r^2)(h)$.

But not completely. Why $1/3$? Well, let's think back to the two-dimensional analogue of this. What's a two-dimensional analogue of a cone? It's just a triangular region. The analogue of the two-dimensional base is a one-dimensional line segment. And we remember that:

$$\text{Area of triangle} = \frac{1}{2} \text{Length of base line segment} \times \text{Height}$$

So why do we get $1/2$ in the two-dimensional case and $1/3$ in the three-dimensional case? Well, you might guess that we basically get $1/n$ in the n-dimensional case. And then you go back and look at the proof, and see that it essentially works this way:

$$\int_0^1 \alpha^{n-1} \, d\alpha = [\alpha^n/n]_0^1 = 1/n$$

3. Solids of revolution: the disk and washer method

3.1. **Definition of solid of revolution.** There is another procedure for constructing three-dimensional figures. Three-dimensional figures constructed this way are called *solids of revolution*. This is obtained as follows: we start with a region Ω and a line ℓ. Next, we rotate Ω about the line ℓ in three dimensions. The region obtained in this way is termed the *solid of revolution* of Ω.

For simplicity, we will assume that Ω lies completely to one side of ℓ. We study such surfaces in two steps. First, we study the special case where one boundary of Ω is along ℓ. After that, we study the case where all of Ω could lie on one side of ℓ. The method for the first case is termed the *disk method* and the method for the second case is termed the *washer method*.

Aside: The surface of a solid of revolution includes two capping disks. The remaining part of this surface is the curved surface, and this is often called a *surface of revolution*. Surfaces of revolution turn out to be very important in a variety of natural processes.

3.2. **Disk method.** Consider the area bounded by the graph of the function $y = f(x)$ and the x-axis between $x = a$ and $x = b$ (with $a < b$). Assume, for now, that the graph of f lies completely on the positive side of the x-axis. So, the picture looks something like Figure 6.2.8 (left) of the book. Revolving this about the x-axis gives a solid of revolution as shown in fiure 6.2.8 (right) of the book.

We now consider how to apply the method of parallel cross sections to this volume computation. We consider the axis as the x-axis and the cross sections are thus in the yz-plane. In particular, we see by our construction that all the cross-sections are disks and the disk for a cross section at $x = x_0$ has radius $f(x_0)$ and area $\pi(f(x_0))^2$. The area is thus:

$$\int_a^b \pi(f(x))^2 \, dx$$

We can pull the π out of the integral if we want. This is the general formula for calculating the area.

It turns out that the formula is also valid for a function that crosses the x-axis. In this case, the parts abover the x-axis and the parts below the x-axis are out of phase by π as we revolve them. However, the overall analysis remains the same, with the radius being $|f(x_0)|$ instead of $f(x_0)$. Since we are squaring it anyway, the final answer remains the same.

Here are some particular cases of solids of revolution whose volume can be computed using the disk method:

(1) The right circular cylinder with radius r and height h can be realized as the solid of revolution for the region between the x-axis and the graph of a constant function with value r (bounded by vertical lines) over an interval of length h. The region being rotated is thus a rectangle with dimensions r, h, and h is the fixed side.

(2) The right circular cone with radius r and height h can be realized as the solid of revolution for the region between the x-axis and the graph of the function $y = rx/h$ on the interval $[0, h]$ (bounded by a vertical line at $x = h$). The region being rotated is thus a right triangle with legs r and h and h is the fixed side.

We can verify that we get the same answer as usual when we apply the disk method.

3.3. **Solids of revolution: the washer method.** What if the region being rotated is completely on one side of the axis of rotation? For instance, imagine a disk far away from the x-axis being revolved about the x-axis. The corresponding solid is sometimes called a *filled torus* or *solid torus* (the boundary of this, which is a surface obtained by revolving the boundary circle, is usually simply called a *torus*).

The washer method is a method that allows us to compute the areas of such solids. Again, the idea is to use parallel cross sections. In this case, the cross sections are not disks, but regions called *annuli*. Given a point P and two concentric circles centered at P (in the same plane) with radii $r < R$, the annulus for these two radii is the set of points in the bigger disk that are not there in the interior of the smaller disk. Thus, it is the region between the circles of radii r and R, along with the two boundary circles.

The area of such an annulus is given by $\pi(R^2 - r^2)$.

The upshot of this is that the volume of the solid of revolution obtained by revolving the region between $y = g(x)$ and $y = f(x)$, with $0 \le g(x) \le f(x)$, on $[a, b]$, is:

$$\int_a^b \pi[(f(x))^2 - (g(x))^2] \, dx$$

If the two functions cross each other, then if we are interested in the unsigned volume, we need to split into intervals based on which one is bigger where, calculate the volumes of the solids of revolution corresponding to each interval, and add up. In other words, we need to compute:

$$\int_a^b \pi|(f(x))^2 - (g(x))^2| \, dx$$

3.4. **Solids of revolution: the tale of the receding axis.** The first thing worth noticing about the volumes of solids of revolution is that the volume is *not determined* by the area of th region being rotated. It *also depends* on the choice of axis. As a general rule, the farther the axis from the region being rotated, the bigger the volume.

To understand this, consider the question: given a fixed number $h > 0$, what can we say about the area of the annulus of thickness h, i.e., where the outer radius is h more than the inner radius? For fixed h, this number increases as we increase the two radii. This is because the area is $\pi[(r+h)^2 - r^2] = \pi(2r+h)h$. The $2r + h$ term increases as r increases.

To give you some intuition about this, here is something that might strike you as visually counterintuitive: the area of the annulus with inner radius 4 and outer radius 5 equals the area of the disk of radius 3 (since $5^2 - 4^2 = 3^2$) even though the former has a much smaller thickness. The smaller thickness is compensated for (roughly) by the larger circumference.

The calculations that we did for the annulus show that as we move our axis farther and farther from Ω, the solid of revolution becomes larger and larger in volume. Remember: to calculate the volume of the solid of revolution, we create slices perpendicular to the axis of revolution, but we are not integrating the length of these slices; we are integrating the differences of *squares* of the endpoints of the slices. And this difference of squares increases as both numbers get bigger, even when the actual difference between them is constant.

3.5. **Solids of revolution: when the axis straddles the region.** So far, we have considered a situation where the region being revolved is completely on one side of the axis of revolution.

In the case that the region being revolved is partly on one side and partly on the other side of the solid of revolution, we must keep the following things in mind:

(1) If the region has *mirror symmetry* about the axis of revolution, then we can simply delete the half of the region on one side and consider the solid of revolution for the other half.
(2) Otherwise, in general, we must *fold* the region being rotated along the axis, i.e., reflect all the stuff on one side to the other, while keeping the stuff on the other side unchanged. Note that in the case of mirror symmetry, the reflected material overlaps with the original material. In some cases, such as the graph of a function about the x-axis, the reflected portion has no area of overlap with the stuff already there. In yet other cases, part of the reflected region overlaps, and the rest doesn't.

4. The shell method

4.1. **Formula.** The shell method applies to situations where we revolve about the y-axis the region made between a graph $y = f(x)$ and the x-axis. As before, we work with a nonnegative continuous function f on a closed interval $[a, b]$ with $0 < a < b$. Consider the region bounded by the graph of f, the x-axis, and the vertical lines $x = a$ and $x = b$. Now, consider the solid of revolution obtained by revolving this region about the y-axis. The volume of this solid of revolution is given by the formula

$$\int_a^b 2\pi x f(x)\,dx = 2\pi \int_a^b x f(x)\,dx$$

In the case that the function is not nonnegative throughout, we can use the more general formula:

$$\int_a^b 2\pi x |f(x)|\,dx$$

The best way of doing this is to partition the interval according to the sign of f.

4.2. **Slight generalization of this formula.** Consider now a slightly more general situation: we are looking at the region between the graphs of the functions f and g between $x = a$ and $x = b$. We consider the solid of revolution obtained by revolving this region about the y-axis. If $g \le f$ on $[a, b]$, then the volume is given by:

$$\int_a^b 2\pi x[f(x) - g(x)]\,dx = 2\pi \int_a^b x[f(x) - g(x)]\,dx$$

(Note: We don't need any conditions on the nonnegativity of f and g here).

If f and g cross each other, we can use the general formula:

$$\int_a^b 2\pi x|f(x) - g(x)|\,dx$$

This is best handled by partitioning the interval according to where f is greater and where g is greater.

5. Average value point of view

5.1. **Overview.** For the various approaches we have seen so far for volume computation, there is an *average value point of view.* This can be thought of as a process whereby we compare our actual imperfect solid to a more perfect solid which is more uniform, and where the volume is given as a simple product. Let's illustrate this by beginning with our interpretation of volume as the integral of a variable cross section area.

Here, our *ideal* figure is a right cylinder, where the cross section area does not change for the cross sections (more generally, this is also true for oblique cylinders). In these ideal figures, the volume is the product of the constant cross section area and the height.

The volume in general can be thought of as the product of the *average* cross section area and the height. Here, the *average* cross section area is *defined* the way we calculate the average value for a function: we integrate it over the entire interval, and then divide by the length of the interval. In other words, the average cross section area is defined so that a right cylinder with that cross section and the height of our current figure has the same volume.

How does the average value point of view help? Computationally, it doesn't, but it gives us some intuition as to what kind of answers to expect. This is because, looking at the figure, we have some ideas about the average value: it must be somewhere between the minimum and the maximum value, for instance. This provides a reality check on the computations that we do.

5.2. **Average value for shell method.** Here, the ideal function is a constant function f on $[a, b]$ with constant value C. Revolving it about the y-axis yields a cylindrical shell with inner radius a, outer radius b, and height C. The volume is $\pi C(b^2 - a^2) = \pi C(b + a)(b - a)$. The value $\pi C(b + a) = 2\pi C(b + a)/2$ is the curved surface area of the cylinder whose radius is $(b + a)/2$, which is the cylinder whose radius is halfway between the inner and outer radius. We thus see that:

Volume of cylindrical shell = Curved surface area of mid-value cylinder × Difference of outer and inner radii

This is the ideal situation. In the real situation, we define the *average curved surface area* as:

$$\text{Average curved surface area} = \frac{\text{Volume of solid of revolution}}{\text{Difference of upper and lower limits}}$$

In our notation, the denominator is $b - a$. Thus, we obtain that the volume of the solid of revolution is $b - a$ times the *average curved surface area.* As before, this is not really computationally useful, but it might give us some intuition.

5.3. **Average value for disk method: different notions of average!** The average value point can also be used to understand the disk method.

Recall that the volume of a solid of revolution obtained by revolving about the x-axis the region between the x-axis and the graph of f from $x = a$ to $x = b$ is given by $\pi \int_a^b (f(x))^2 \, dx$. Recall that we proved this formula by taking cross sections perpendicular to the x-axis. The area of a cross section at the value x is $\pi(f(x))^2$, because the cross section is a disk of radius $|f(x)|$.

Under the average value point of view, we are interested in the average value of this cross section area. There's a little subtlety in this.

To find the *area* between the graph of f and the x-axis from $x = a$ to $x = b$, we perform a simple integration $\int_a^b f(x) \, dx$ (or $\int_a^b |f(x)| \, dx$). On the other hand, to find the volume of the solid of revolution, we perform the integration $\int_a^b (f(x))^2 \, dx$.

In other words, when finding the volume of the solid of revolution, we give a lot more weight to larger radii – because the radius is being squared. Remember the discussion from last time where we saw that an annulus with inner and outer radii 4 and 5 has the same area as the disk of radius 3. This is because the square of a number grows much faster than the number itself.

Our averaging process is also correspondingly biased. When we are calculating the average value in the ordinary sense, we do $\int_a^b f(x) \, dx/(b - a)$. However, when calculating the average of the areas of the disks, we are doing $\pi \int_a^b (f(x))^2\, dx/(b-a)$. The latter average value is *usually not the same* as the area of the disk whose radius is the average radius. Rather, it is usually larger, because taking the squares assigns greater weight to the bigger radii.[2]

6. Stock-taking

We have now seen some formulas and general approaches that use the ideas of integration to compute areas and volumes. Later in the course and/or in later life, you will encounter formulas to do a lot of the other things you've always wanted to do, such as formulas for arc lengths and surface areas. We are not getting into those formulas right now for two reasons: (i) they require more conceptual apparatus to understand, (ii) the kind of expressions that you typically get to integrate are expressions that you do not know how to deal with.

This brings us to one of the things that differentiates (pun!) differentiation from integration. Differentiation was based on a set of rules that we could apply blindly, because for every way of combining and composing existing functions, we had a corresponding way of breaking down the differentiation problem. With integration, however, we are in more wild territory, since there are no easy hard-and-fast rules and a lot depends on creativity and spotting persuasive patterns. This makes integration more fascinating, but it also means that ever so often, we come across a situation from the real world that boils down to computing an integral, and we don't really have an idea how to go about it.

Nonetheless, reducing a geometric problem of volume computation into a purely arithmetic/algebraic problem of evaluating a definite integral should be seen as a major step forward. Even if we have no clue about what an antiderivative might be, we can still use the upper sum/lower sum method to approximate this integral.

7. More computational intuition

7.1. **Stretching, shrinking, and scaling.** We can use the "solid of revolution" idea to compute the volume of a sphere. A solid sphere of radius r is obtained by revolving a semicircular region of radius r about its diameter. The volume formula is thus:

$$\pi \int_{-r}^{r} (r^2 - x^2)\, dx$$

This gives the familiar formula $(4\pi/3)r^3$.

Note that a sphere is also the solid of revolution of a circular disk about its diameter. As noted earlier, since a circular disk has mirror symmetry about its diameter, so we can delete one of the semicircular pieces and still get the same solid of revolution.

Now, let's think about what happens if, instead of revolving a circular (or semicircular) disk, we revolve the region enclosed by an ellipse about its major or minor axis. An ellipse oriented along the axes and centered at the origin is a curve given by the equation:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

with a, b positive.

If $a > b$ then the x-axis is the major axis and the y-axis is the minor axis. Note that both the x-axis and y-axis are axes of mirror symmetry; however, unlike the case of the circle, it is no longer true that every line through the origin is an axis of mirror symmetry.

Now, we could do the calculations pretty easily to compute the volume of the solid of revolution about either axis, but let's give an intuitive explanation that allows us to get at the answer. If we start with a circle centered at the origin and of radius b and stretch it by a factor of a/b in the x-direction, we get an ellipse. Clearly, the *area* of the ellipse is therefore a/b times the area of the circle, hence it is πab. What about the volume? We note that the axis along which we integrate gets stretched by a factor of a/b. A little thought now tells us that the answer will be $(4\pi/3)ab^2$.

[2]It turns out that the two averages are equal only for a constant function. The inequality being alluded to here indirectly is known as the arithmeti mean-quadratic mean (AM-QM) inequality or the arithmetic mean-root mean square (AM-RMS) inequality.

More generally, we see that:

(1) If the region being revolved is stretched by a factor of λ *along* the axis of revolution, the volume is multiplied by a factor of λ.
(2) If the region being revolved is stretch by a factor of μ *along* the axis of revolution, the volume is multiplied by a factor of μ^2. The square happens because when we revolve, the area contribution in each slice is proportional to the square of the radius or difference of squares of radius.

Thus, if the same ellipse were revolved about its minor axis, we'd get a volume of $(4\pi/3)a^2b$.

7.2. **Brief mention: Pappus' theorem.** Pappus' theorem is in a later section of the chapter that we're not including in this course, but it's a theorem worth taking a look at and understanding at least temporarily. For Exercise 6.3.44 (featuring in Homework 8), Pappus' theorem gives an alternative solution approach that is much shorter than the disk and shell methods that we will use to solve the problem. It also tells us what the answer will be – in this case $2\pi^2a^3$. The reason it is easier is because for the case of the circle, we know exactly where the center (centroid) is.

ONE-ONE FUNCTIONS AND INVERSES

MATH 152, SECTION 55 (VIPUL NAIK)

Corresponding material in the book: Section 7.1.

What students should definitely get: The definition of one-to-one function, the computational and checking procedures for checking that a function is one-to-one, computing the inverse of such a function, and relating the derivative of a function to that of its inverse.

What students should hopefully get: The subtleties of domain and range issues, the distinction between the algebraic and the calculus approaches.

Executive summary

0.1. Vague generalities. Words...

(1) Old hat: Given two sets A and B, a function $f : A \to B$ is something that takes inputs in A and gives outputs in B. The *domain* of a function is the set of possible inputs, while the *range* of a function is the set of possible outputs. The notation $f : A \to B$ typically means that the domain of the function is A. However, the whole of B need not be the range; rather, all we know is that the range is a *subset* of B. One way of thinking of functions is that *equal inputs give equal outputs*.

(2) A function f is one-to-one if $f(x_1) = f(x_2) \implies x_1 = x_2$. In other words, *unequal inputs give unequal outputs*. Another way of thinking of this is that *equal outputs could only arise from equal inputs*. Or, *knowledge of the output allows us to determine the input uniquely*. One-to-one functions are also called one-one functions or injective functions.

(3) Suppose f is a function with domain A and range B. If f is one-to-one, there is a *unique* function g with domain B and range A such that $f(g(x)) = x$ for all $x \in B$. This function is denoted f^{-1}. We further have that g is also one-to-one, and that $f = g^{-1}$. Note that f^{-1} differs from the reciprocal function of f.

(4) Suppose $f : A \to B$ and $g : B \to C$ are one-to-one functions. Then $g \circ f$ is also one-to-one, and its inverse is the function $f^{-1} \circ g^{-1}$.

Actions ...

(1) To determine whether a function is one-to-one, solve $f(x) = f(a)$ for x in terms of a. If, for every a in the domain, the only solution is $x = a$, the function is one-to-one. If, on the other hand, there are some values of a for which there is a solution $x \neq a$, the function is not one-to-one.

(2) To compute the inverse of a one-to-one function, solve $f(x) = y$ and the expression for x in terms of y is the inverse function.

0.2. In graph terms. Thousand words ...

(1) A picture in a coordinatized plane is the graph of a function if every vertical line intersects the picture at most once. The vertical lines that intersect it exactly once correspond to the x-values in the domain. This is known as the *vertical line test*.

(2) A function is one-to-one if and only if its graph satisfies the *horizontal line test*: every horizontal line intersects the graph at most once. The horizontal lines that intersect the graph exactly once correspond to y-values in the range.

(3) For a one-to-one function, the graph of the inverse function is obtained by reflecting the graph of the function about the $y = x$ line. In particular, a function equals its own inverse iff its graph is symmetric about the $y = x$ line.

(4) Many of the results on inverse functions and their properties have graphical interpretations. For instance, the fact that the derivative of the inverse function is the reciprocal of the derivative corresponds to the geometrical fact that reflection about the $y = x$ line inverts slopes of tangent lines.

Similarly, the results relating increase/decrease and concave up/down for a function and its inverse function can all be deduced graphically.

0.3. **In the real world.** Words... (from now on, we restrict ourselves to functions whose domain and range are both subsets of the real numbers)

(1) An increasing function is one-to-one. A decreasing function is one-to-one.
(2) A *continuous* function on an *interval* is one-to-one if and only if it is either increasing throughout the interval or decreasing throughout the interval.
(3) If the derivative of a continuous function on an interval is of constant sign everywhere, except possibly at a few isolated points where it is either zero or undefined, then the function is one-to-one on the interval. Note that we need the function to be continuous *everywhere* on the interval, even though it is tolerable for the derivative to be undefined at a few isolated points.
(4) In particular, a one-to-one function cannot have local extreme values.
(5) A continuous one-to-one function is increasing if and only if its inverse function is increasing, and is decreasing if and only if its inverse function is decreasing.
(6) If f is one-to-one and differentiable at a point a with $f'(a) \neq 0$, with $f(a) = b$, then $(f^{-1})'(b) = 1/f'(a)$. This agrees with the previous point and also shows that the rates of relative increase are inversely proportional.
(7) Two extreme cases of interest are: $f'(a) = 0$, $f(a) = b$. In this case, f has a horizontal tangent at a and f^{-1} has a vertical tangent at b. The horizontal tangent is typically also a point of inflection. It is definitely *not* a point of local extremum. Similarly, if $(f^{-1})'(b) = 0$, then f^{-1} has a horizontal tangent at b and f has a vertical tangent at a.
(8) A slight complication occurs when f has one-sided derivatives but is not differentiable. If both one-sided derivatives of f exist and are nonzero, then both one-sided derivatives of f^{-1} (at the image point) exist and are nonzero. When f is increasing, the left hand derivative of f^{-1} is the reciprocal of the left hand derivative of f, and the right hand derivative of f^{-1} is the reciprocal of the right hand derivative of f. When f is decreasing, the right hand derivative of f^{-1} is the reciprocal of the left hand derivative of f, and the left hand derivative of f^{-1} is the reciprocal of the right hand derivative of f.
(9) The second derivative of f^{-1} at $f(a)$ is $-f''(a)/(f'(a))^3$. In particular, the second derivative of the inverse function at the image point depends on the values of both the first and the second derivatives of the function at the point.
(10) If f is increasing, the sense of concavity of f^{-1} is opposite to that of f. If f is decreasing, the sense of concavity of f^{-1} is the same as that of f.

Actions ...

(1) For functions on intervals, *to check if the function is one-to-one*, we can compute the derivative and check if it has constant sign everywhere except possibly at isolated points.
(2) In order to find $(f^{-1})'$ at a particular point, given an explicit description of f, it is *not* necessary to find an explicit description of f^{-1}. Rather, it is enough to find f^{-1} at that particular point and then calculate the derivative using the above formula. The same is true for $(f^{-1})''$, except that now we need to compute the values of both f' and f''.
(3) The idea can be extended somewhat to finding $(f^{-1})'$ when f satisfies a differential equation that expresses $f'(x)$ in terms of $f(x)$ (with no direct appearance of x).

0.4. **In graph terms.** Thousand words ...

(1) A picture in a coordinatized plane is the graph of a function if every vertical line intersects the picture at most once. The vertical lines that intersect it exactly once correspond to the x-values in the domain. This is known as the *vertical line test*.
(2) A function is one-to-one if and only if its graph satisfies the *horizontal line test*: every horizontal line intersects the graph at most once. The horizontal lines that intersect the graph exactly once correspond to y-values in the range.

(3) For a one-to-one function, the graph of the inverse function is obtained by reflecting the graph of the function about the $y = x$ line. In particular, a function equals its own inverse iff its graph is symmetric about the $y = x$ line.
(4) Many of the results on inverse functions and their properties have graphical interpretations. For instance, the fact that the derivative of the inverse function is the reciprocal of the derivative corresponds to the geometrical fact that reflection about the $y = x$ line inverts slopes of tangent lines. Similarly, the results relating increase/decrease and concave up/down for a function and its inverse function can all be deduced graphically.

1. Warm-up

1.1. **What is/was a function?** Let's recall some of the terminology associated with the concept of functions. A *function* is some thing that allowed you to take certain kind of inputs and spit out certain kinds of outputs, with the main constraint being that *equal inputs give equal outputs*.

The set of permissible inputs for a function is called the *domain* of the function. If the input fed into the function is in the domain, the fnuction processes it and give an output. If the input is not in the domain, the function cannot process it. The set of possible values that the function could spit out was called the *range* of the function. We think of a function as a *black box* that takes an input at one end and emits the output at the other end.

When we say that $f : A \to B$ is a function, what we mean is that the domain of f is A (i.e., f takes as inputs precisely the elements of A) and the range is a *subset* of B. In other words, the notation $f : A \to B$ does *not* imply that everything in B is in the range. This is a useful notational convenience because we would often like to define functions that takes values in some large set (such as the real numbers) without trying to locate the *precise* range.

Another thing that we saw long ago is that a function is *not* the same thing as an expression for the function. There are two aspects to this:

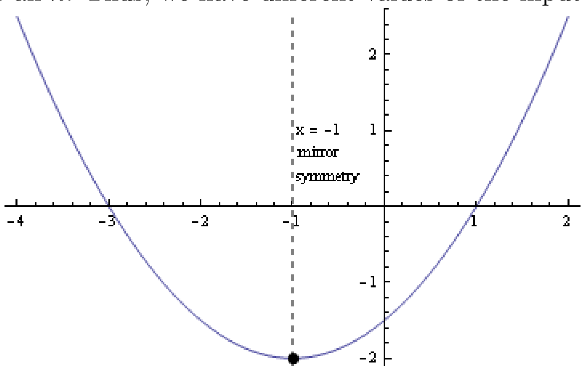
(1) The same *expression* could define different functions, depending on the domain where we are considering the function. For instance, the expression x^2 could be considered on the positive reals, on the negative reals, on the positive integers, on the negative integers, or on the open interval $(0, 1)$. These are all different functions in the technical sense. To avoid this confusion, we posited that if the domain is not explicitly specified or otherwise clear from the context, it is taken to be the largest subset of the real numbers where the expression makes sense (this is the *maximal possible domain*).
(2) Different expressions could specify the same function. For instance, x^2 is the same, as a function, as $2x(x/2)$, even though the literal expressions are different. Similarly, $\sin(\pi x)$ and 0 are the same as functions when restricted to the set of integers.

1.2. **The role of expressions.** There is a fundamental difference between thinking of *functions* and thinking of *expressions*. When we are thinking of a function, we are thinking of a very specific input-output relationship, which may be expressed using an algebraic expression, a table of values, or a graph. The algebraic expression has the advantage of being compact, succinct, and unambiguous, as well as easy to manipulate for many purposes. The graphical expression allows us to use our visual instincts. The table method is something we have been giving short shrift for good reason: most of our functions have infinite domains, and tables just don't work. If you were taking discrete mathematics rather than calculus, we might have been using tables instead of graphs because we were dealing with finite domains.

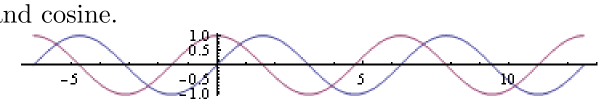
Expressions are useful for a multitude of reasons. With the algebraic expressions, we are able to formally differentiate the function once, twice, and more times. We can calculate its value, the value of its derivatives, find the domain, find the critical points, find the points of inflection, etc., all just starting from a compact formal expression. Another point worth noting is that some of the formal manipulations of expressions can be done even without having any idea of how the graph of the function looks like.

1.3. **Thinking of inverting the function.** Equal inputs for a function give equal outputs, but unequal inputs may give the same output. The extreme example of this is the constant function, whose output is completely indifferent to the input. But there are other examples. For instance:

(1) For *even functions* f, such as the absolute value function, the square function, and the cosine function, we have the relation $f(x) = f(-x)$. Thus, two different inputs give rise to the same output.
(2) More generally, consider a function f with *mirror symmetry* about the line $x = c$. This is a function whose graph is symmetric about the vertical line $x = c$. In particular, we have $f(c + h) = f(c - h)$ for all h. Thus, we have different values of the input giving the same value of the output.



(3) For a *periodic function* with period h, we have the relation $f(x + h) = f(x)$. Thus, two different inputs give rise to the same output. The typical examples are trigonometric functions, such as sine and cosine.



(4) If we have a continuous function with a *local maximum* or a *local minimum*, then there are multiple inputs close to the point of attainment of the local extreme value where the function values are equal.

2. Getting into one-to-one functions

2.1. One-to-one functions. A function $f : A \to B$ is termed *one-to-one*, *one-one*, or *injective* if $f(x) = f(y) \implies x = y$. In other words, it is a function having the property that unequal inputs give unequal outputs. Equivalently, we can *reverse* the function in the sense that knowing the output allows us to deduce the input.

Note that this general definition is set-theoretic, and makes sense for functions between arbitrary sets; however, in this course, all functions that we consider are between subsets of reals. So, we are looking at functions $f : A \to \mathbb{R}$ where $A \subseteq \mathbb{R}$.

Next note: Whether a function is one-to-one depends on what domain we are considering for that function. For instance, the squaring function is one-to-one on $[0, \infty)$ but not on the whole real line. Similarly, the greatest integer function is one-to-one when restricted to integers but not on the whole real line. The sine function is one-to-one on the interval $(-\pi/2, \pi/2)$ but not on the whole real line. Of course, if we are just given an *expression* and asked whether the function corresponding to that expression is one-to-one, we consider the domain to be the maximum possible subset of the real line.

2.2. The horizontal line test. Consider a function f and now consider the graph of $y = f(x)$. The *horiozontal line test* says that f is one-to-one if and only if every horizontal line intersects the graph of f at most once. Further, the horizontal lines for which the intersection occurs once are precisely those corresponding to the range. This makes sense, because a horizontal line corresponds to a particular value of y, and the intersections with the graph correspond to the values x such that $f(x) = y$.

Remember the *vertical line test*? This states that a given picture arises as the graph of a function if and only if its intersection with every vertical line has at most one point. Further, the vertical lines that intersect it at one point are the vertical lines corresponding to the domain. The rationales behind the vertical line test and horizontal line test are similar.

2.3. How do we find out if a function is one-to-one? The purely algebraic way. To determine whether a function is one-to-one, we can use a purely algebraic way – except that it usually doesn't work. We pick two letters, x and a, then write $f(x) = f(a)$ and try to solve algebraically to see if we get a solution with $x \neq a$. For instance, consider the function $f(x) := x^2$. The general equation would be:

$$x^2 = a^2$$

This simplifies to:

$$(x - a)(x + a) = 0$$

This has two solutions: $x = a$ and $x = -a$. The two solutions coincide when $a = 0$ and are distinct otherwise. Thus, the function is *not* one-to-one.

Now consider the function $f(x) := x^3$.

The general expression would be:

$$x^3 = a^3$$

This simplifies to:

$$(x - a)(x^2 + ax + a^2) = 0$$

The second quadratic factor has negative discriminant, so has no real solution, and the only solution is $x = a$. Thus, the function is one-to-one.

What about something more complicated, such as $f(x) := x^3 + x$? We again set $f(x) = f(a)$ and simplify:

$$x^3 + x = a^3 + a$$

Moving everything to one side:

$$(x^3 - a^3) + (x - a) = 0$$

This simplifies to:

$$(x - a)(x^2 + ax + a^2) + (x - a)(1) = 0$$

We combine terms:

$$(x - a)(x^2 + ax + a^2 + 1) = 0$$

The quadratic factor has negative discriminant, so the only solution is $x = a$.

This purely algebraic approach works for quadratic and cubic functions, but it starts getting tedious for more complicated functions. For instance, how do we handle functions such as $f(x) := x - \sin x$? It is hard to solve:

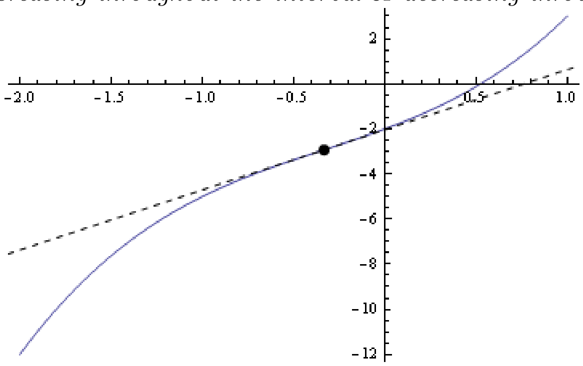
$$x - \sin x = a - \sin a$$

Even for algebraic functions, the approach could be less tractable when the functions are more complicated. This suggests that we need to supplement the *algebraic approach* (with its attendant focus on looking at points of the domain separately) with the *calculus approach* (with its attendant focus/stress on moving along the real line and thinking in terms of limits and continuity). What can the calculus approach tell us that mere algebra cannot?

2.4. **Remember the range computations.** It might be useful to draw a parallel and remember a related generic problem we tackled a while ago. That was the problem of *finding the range*. The *algebraic method* of finding the range of a function f is to set $a = f(x)$ and *solve for* x. We are not interested in actually *finding* a solution – we are interested in determining the conditions on a such that *at least* one solution exists. For instance, for linear and quadratic polynomials and rational functions of small degree, this often reduces to a condition on the discriminant of a quadratic polynomial.

That was the algebraic approach, and it was limited to a small number of functions that were algebraically tractable. But then we saw the calculus approach, which essentially allowed us to graph any reasonably nice function. Once we have the entire graph, we can find the range. For a continuous function, this is simply the interval between the minimum value of the function and the maximum value of the function. This allowed us to determine the range of a much larger class of functions, particularly those that are continuous and once or twice differentiable.

2.5. **The calculus interpretation of one-to-one.** Consider a *continuous* function f on a (possibly open, closed, half-open, half-closed, or infinite) interval I. Continuous means that the function cannot jump about suddenly. Under what conditions is the function one-to-one? Clearly, if it changes direction somewhere, i.e., has a local extreme value, then there are horizontal lines close by that intersect the graph at two points. Thus, there are no local extreme values. Or, another way of putting this is that the function must either be *increasing throughout the interval* or *decreasing throughout the interval.*



Thus, for a continuous function on an interval, being one-to-one is equivalent to being increasing throughout or decreasing throughout. *If you change direction, you repeat points.* Remember that both parts of the statement: *continuous* and *interval*, are needed in order to conclude that a one-to-one function *must* be increasing throughout or decreasing throughout. However, the *other direction of implication* is always true: a function that is increasing throughout on its domain is one-to-one, and so is a function that is decreasing throughout on its domain.

Let's see some counterexamples:

(1) Can you think of a discontinuous function on an interval that is one-to-one but not increasing or decreasing?
(2) Can you think of a function on a set that is a union of two or more intervals that is continuous on each piece and is one-to-one but is not increasing or decreasing when viewed on the whole domain? The problem is that although we can deduce that the function is increasing throughout or decreasing throughout separately on each of the intervals, we cannot compare across intervals.

2.6. **The proof that a continuous function on an interval is one-to-one iff its increasing or decreasing.** Let us now try to prove the statement that a continuous function on an interval is one-to-one if and only if it is either increasing throughout on the interval or decreasing throughout on the interval. Note that increasing throughout or decreasing throughout obviously implies one-to-one, so we concentrate on proving the other direction of implication.

Suppose f is a continuous function on an interval I and there are points x_1, x_2, x_3 with $x_1 < x_2 < x_3$ and $f(x_1) < f(x_2) > f(x_3)$. Then, suppose M is a number greater than both $f(x_1)$ and $f(x_3)$ but less than $f(x_2)$. By the intermediate value theorem, there exists $x_4 \in (x_1, x_2)$ and $x_5 \in (x_2, x_3)$ such that $f(x_4) = f(x_5) = M$. This forces f to *not be one-to-one*, a contradiction. Thus, we cannot have a situation where $x_1 < x_2 < x_3$ but $f(x_1) < f(x_2) > f(x_3)$. A similar argument shows that we cannot have $x_1 < x_2 < x_3$ but $f(x_1) > f(x_2) < f(x_3)$. This forces f to be increasing throughout or decreasing throughout.

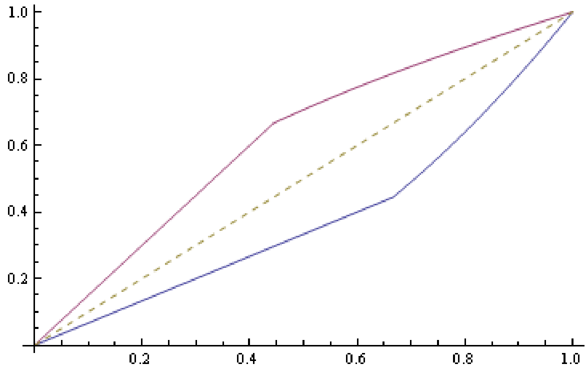
Note that to apply the intermediate value theorem, we applied both the continuity of f and the fact that the domain is an interval, hence contains $[x_1, x_2]$ and $[x_2, x_3]$.

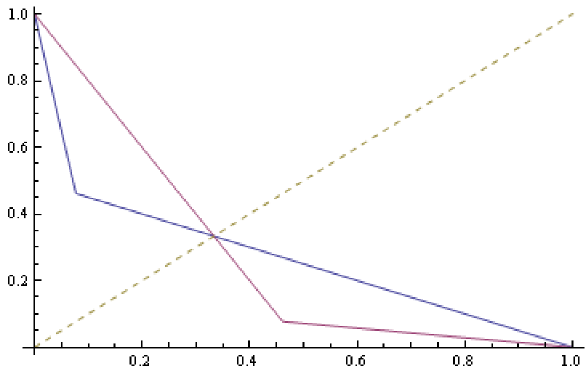
3. A two-way street: inverse functions

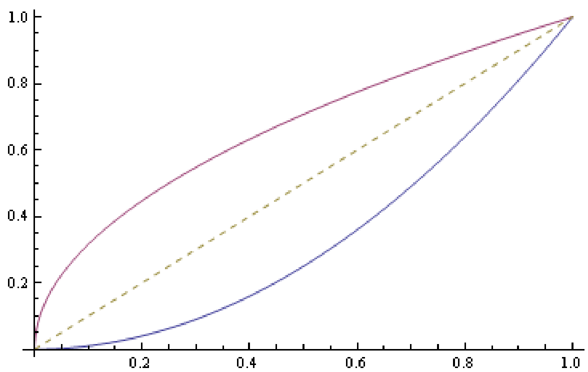
3.1. **The inverse function.** Suppose we have a function f with domain A and range B. If f is one-to-one, we can define an *inverse function* g such $f \circ g$ is the identity map on B. Moreover, this function g is unique. This function is called the *inverse function* to f, since it *reverses* or *inverts* the action of f. Note that here, we really do need B to be precisely the range, because we cannot define the function g on points outside the range.

This function g is denoted as f^{-1} and is termed the *inverse function* to f. Note that this is not the same as the *pointwise multiplicative inverse* of f, which is the function $1/f(x)$. The latter may be denoted as

$[f(x)]^{-1}$ or as $1/f$, as opposed to $f^{-1}(x)$. I might also use the word *reciprocal function* for the pointwise multiplicative inverse.







What happens if f is not one-to-one? In this case, there are many different candidates for g that work, and it is not clear which one to pick. To avoid confusion, we do not talk of *the inverse function* any more. For instance, when f is the squaring function on the reals, we could take g to be the positive square root or the negative squareroot, or to sometimes be the positive squareroot and sometimes the negative squareroot. For instance, one candidate would be a function that is the positive square root for nonnegative rationals and the negative square root for positive irrationals.

This is a very rich and deep question that we shall return to later, when we study things such as inverse trigonometric functions.

3.2. **The inverse operation is involutive.** The operation of taking the inverse is an *involutive* operation, in that it has the following two properties:

(1) $(f^{-1})^{-1} = f$. In other words, if $g = f^{-1}$, the $f = g^{-1}$.
(2) $(f_1 \circ f_2)^{-1} = f_2^{-1} \circ f_1^{-1}$. In other words, the inverse of the composite is the composite of the inverses, but the sequence of composition flips over.

You'll be asked to show this in a forthcoming homework.

3.3. **Finding the inverse function: like finding the range.** As we just recalled, to find the range of a function f, we consider the equation $y = f(x)$ and solve for x in terms of y. When we were trying to compute the range, our sole purpose was to find the set of y for which there exists at least one value of x that solves the equation. When our goal is to determine the inverse function, we are interested in the actual *expression* for x in terms of y since that is the inverse function.

Note that in this process, we can discard the values of y for which *no solution exists*. But if we find that for some y, there are multiple values of x, then we've gone down a bad path: the function wasn't one-to-one, so we shouldn't have been trying to find an inverse at all.

3.4. **Situations where the algebraic procedure works.** It works for nonconstant linear functions. Given a function $y = mx + c, m \neq 0$, we can rewrite $x = (y - c)/m$. It also works for functions of the form $y = x^{p/q}$ where both p and q are odd integers. The inverse function to $y = x^{p/q}$ is $y = x^{q/p}$. And it works for functions that are obtained by composing such power functions and linear functions. For instance, see Example 3 in the book.

3.5. **Graphical interpretation of inverse function.** Suppose g is the inverse function of a one-to-one function f. For every point (x, y) in the graph of f, we have $y = f(x)$, hence $x = g(y)$. Thus, the point (y, x) is in the graph of g. In other words, the graph of g is obtained by taking the graph of f and sending each point to the point obtained by interchanging its coordinates. The *coordinate interchange operation* is equivalent to the geometric operation of reflection about the $y = x$ line. Thus, the graph of g is obtained by reflecting the graph of f about the $y = x$ line. This geometrical interpretation is useful for understanding the relationship between derivatives.

It also helps us identify a new kind of symmetry that some functions possess. The graph of a function f is symmetric about the $y = x$ line iff $f = f^{-1}$. Examples of such functions are $y = x$, $x + y = C$ for some constant C, and *implicit* functions given by $p(x) + p(y) = C$ where p is a one-to-one function from $\mathbb{R}$ to $\mathbb{R}$. For instance, $x^3 + y^3 = 1$ is an implicit description of y as a function of x – the explicit description is $y = (1 - x^3)^{1/3}$, and the inverse is exactly what we'd expect.

3.6. **The inverse of a continuous function is continuous.** From the previous result about reflection, it should be reasonably intuitive that the inverse of a continuous function is continuous. Proving this using the $\epsilon - \delta$ definition is a nice exercise, but not one that we shall undertake in class. This is Theorem 7.1.7 of the book, and you are encouraged to read the proof for your understanding and also to refresh your memories of ϵs and δs.

There is something qualitative that we can say, though, that will shed some light. Remember that the general $\epsilon - \delta$ definition is *not* symmetric in the roles of x and $f(x)$. The skeptic starts with choosing an $\epsilon > 0$ which determines an open interval around the claimed limit for the $f(x)$-value. The prover then has to come up with an interval around the domain value (of radius δ) such that the function value is within the ϵ-interval for every input value in the δ-interval. This definition is asymmetric, because we insist that if the x-value is really close, the $f(x)$-value is also really close, but we do not insist that if the $f(x)$-value is really close, the x-value is also really close.

For one-to-one functions, however, this inherent definitional asymmetry is automatically overcome, because a given $f(x)$-value can be realized only by one x-value. This is the reason that, even though the definition is not inherently symmetrical, the one-to-one nature allows us to show that it is in effect symmetric in the roles of domain and range.

4. Bijective functions and infinite sets

This is a concept you will see in somewhat more detail if you take higher mathematics courses, so I'll briefly mention it here.

A function $f : A \to B$ is termed a *bijection* or *bijective function* from A to B if its range is precisely B and it is one-one. Bijective functions are the same thing as one-one functions considered as functions *to their range* and ignoring the rest of the stuff. The notion of inverse function that we introduced is best viewed in the context of a bijective function, because the inverse is defined only on the range of the function.

Note that the theorem proved earlier basically states that for a continuous bijective function, the inverse is also continuous.

An interesting question now might be: can we define continuous bijective functions between various typical infinite subsets of $\mathbb{R}$? We note some obvious positive results in this direction:

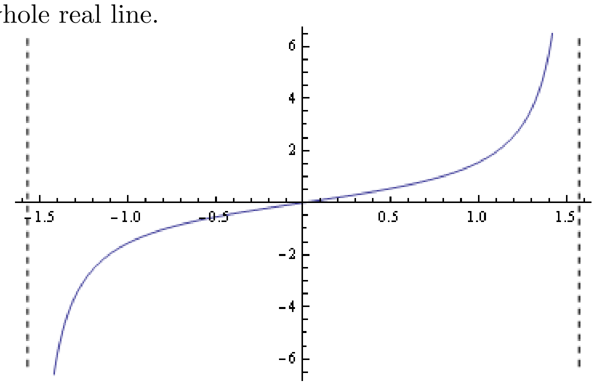
(1) For the open intervals (a, b) and (c, d), where $a < b$ and $c < d$ are real numbers, there is a *linear* bijection between the open intervals. Namely, there is a unique linear function that sends a to c and b to d, and this function gives a bijection of the intervals in between. We could also pick the linear map that would send a to d and b to c.

    In particular, this bijection is continuous and infinitely differentiable, since it involves only translation and scaling.

(2) Interestingly, there is a bijection between any finite open interval and an open interval going to infinity in one or both directions. Since (1) above shows that any two finite open intervals look the same, we just give one bijection of each kind.

    The map $x \mapsto \tan x$ gives a bijection between the open interval $(0, \pi/2)$ and the one-sided infinite open interval $(0, \infty)$.

    The map $x \mapsto \tan x$ also gives a bijection between the finite open interval $(-\pi/2, \pi/2)$ and the whole real line.



Thus, for infinite sets, small finite open intervals can be in bijection with larger finite open intervals and even with infinite open intervals. The fact that a subset of a set can be in bijection with the whole set perturbed the mathematician Cantor when he first discovered and pondered about it. He eventually went crazy, but before doing so, made a key observation: being in bijection with a proper subset is a *defining characteristic* of infinite sets.

5. A new process and new responsibilities

So far, we have seen the following processes for creating new functions from old:

(1) Pointwise addition, subtraction, multiplication, and division.
(2) Function composition.
(3) Piecing together different definitions (using piecewise definitions).

We have now added a new process: inverting a one-to-one function. Hence, it is our job to now describe how to do all the things we used to do in the past for new functions created from old functions using this new process.

5.1. **The derivative of a one-to-one function and its inverse.** uppose we have a one-to-one function f on an interval I with inverse function g. The intermediate value theorem tells us that the range of f (and hence the domain of g) is also an interval. If the domain of f is a closed bounded interval, the extreme value theorem tell us that the range of f (and hence the domain of g) is also closed and bounded.

Since f is increasing throughout or decreasing throughout, what can we say about its derivative (assuming it exists everywhere)? If f is differentiable, then:

(1) f is increasing throughout iff f' is positive everywhere except possibly at isolated points, where it can be zero.
(2) f is decreasing throughout iff f' is negative everywhere except possibly at isolated points, where it can be zero.

Thus, a differentiable f is one-to-one iff f' is of constant sign throughout except possibly at isolated points, where it can be zero. Examples are x^3 and $x - \sin x$.

More generally, if f' has constant sign everywhere except at isolated points where it is either zero or undefined, f is one-to-one.

Let us now bring g into the picture. It turns out that the following is true: if f is one-to-one, and g is its inverse, then if $f(a) = b$, and $f'(a)$ exists and is nonzero, then $g'(b) = 1/f'(a)$. Thus, we have the general formula:

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

Equivalently:

$$g'(f(x)) = \frac{1}{f'(x)}$$

This can be seen in three ways:

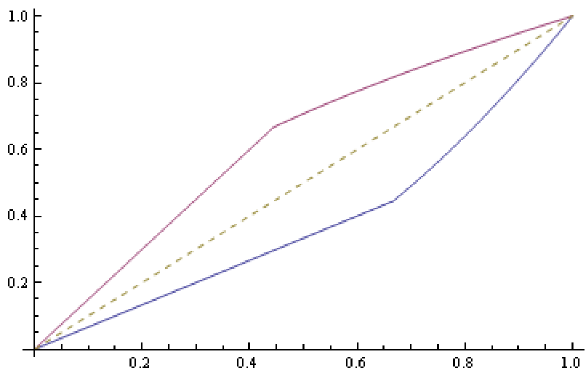
(1) *From first principles*: The difference quotient whose limit gives the value of f' is the reciprocal of the difference quotient whose limit gives the value of g'.
(2) *Using the chain rule*: Use that $f \circ g$ is the identity map and hence obtain that $(f' \circ g) \cdot g' = 1$.
(3) *Graphically*: We know that the graphs of f and g are reflections of each other about the line $y = x$, and the points we are interested in are images of each other under this reflection. The tangent lines through these points are thus also reflections of each other about $y = x$. The slopes of these tangent lines are thus reciprocals of each other.

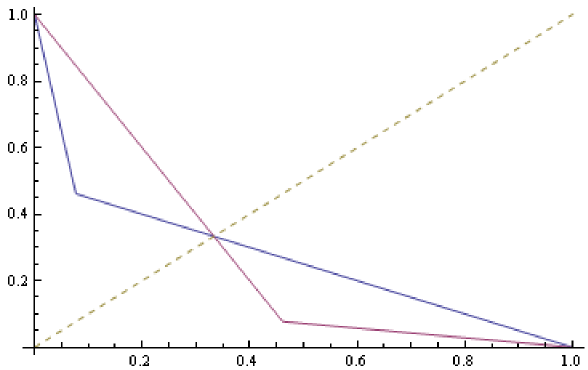
Let $f(a) = b$. There are the following cases of interest:

(1) $f'(a)$ exists and is nonzero. This happens if and only if $g'(b)$ exists and is nonzero, and $f'(a)$ and $g'(b)$ are multiplicative inverses of each other. Pictorially, this means that the tangent lines for the graphs of f and g are neither vertical nor horizontal. $f'(a)$ is positive iff $g'(b)$ is positive, in which case both f and g are locally increasing. $f'(a)$ is negative iff $g'(b)$ is negative.
(2) $f'(a)$ exists and is equal to zero: In this case, $g'(b)$ is undefined. The graph of f has a horizontal tangent at the point a, and the graph of g has a vertical tangent. Moreover, we can deduce from the one-to-one nature of f that the horizontal tangent for f cannot be a local extreme value type – hence (with suitable further differentiability assumptions) it must be the *point of inflection* type.[1]
(3) $g'(b)$ exists and is equal to zero: In this case, $f'(a)$ is undefined. The remarks of the previous point apply with the roles of f and g interchanged, as well as the roles of a and b.
(4) Both the left-hand and the right-hand derivative for f exist at a, but they are not equal: In this case, the left-hand derivative and the right-hand derivative exist for g. Further, if f is increasing, so is g, in which case the left-hand derivative of g is the multiplicative inverse for the left-hand derivative of f, and the right-hand derivative of g is the multiplicative inverse of the right-hand derivative of f. If f is decreasing, so is g, in which case the left-hand derivative of g is the multiplicative inverse of the right-hand derivative of f, and the right-hand derivative of g is the multiplicative inverse of the left-hand derivative of f.

Re-read that last point a few times till you understand it. The interplay between one-sidedness and increase/decrease behavior is extremely important and potentially confusing. Here are some pictures:

[1]The tangent line cutting through the graph is the typical geometric description of a point of inflection; however, it is not strictly correct since there do exist weird situations where we have a point that is not a point of inflection but the tangent line still cuts through the graph. Nonetheless, this is the typical case to keep in mind.





5.2. **Full details of the difference quotient derivation.** To understand this proof, it is helpful to recall that there are two different ways of thinking about functions and about differentiation. The first, which is the typical way, is to think about a function as a machine that takes in an input and gives out an output. There is another, slightly different, way of thinking about functions. Here, the focus is not on the function but on the input and the output. We think of the function as the process relating the input quantity and the output quantity. For instance, we may think of the position x of a particle as a function of time t. Here, the *output quantity* is viewed as a function of the input quantity.

When we switch back and forth between these two ideas of functions, there is a slight abuse of notation. For instance, when we are trying to write the position of a particle as a function of time, we often use the same letter x for the position *function* $x(t)$ and for the actual position variable. This is a bit like saying that instead of writing a function $y = f(x)$, we write $y = y(x)$. This really is an abuse of notation, but it is an abuse that comes with some advantages. For instance, in the book's description of the u-substitution, the book used the letter u both for the function and the variable name.

Recall now that for a function f, the *difference quotient* between the input values x_1 and x_2 is the value:

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

If we write $y_1 = f(x_1)$ and $y_2 = f(x_2)$, we can rewrite this as:

$$\frac{y_2 - y_1}{x_2 - x_1}$$

which can be written in shorthand as:

$$\frac{\Delta y}{\Delta x}$$

With the interpretation as a relationship between quantities, we are interested in the question of how much a specific change in x-values leads to a change in the y-values. The limit of this is at $x = x_0$ is defined as the derivative $f'(x_0)$. With this notation, we also see that:

$$\frac{\Delta y}{\Delta x} = \frac{1}{\frac{\Delta x}{\Delta y}}$$

Thus, we see the intuitive reason why, when we pass to the limits, we should get that dy/dx and dx/dy are multiplicative inverses at any particular pair (x, y).

Let us make this formal. Suppose f is one-to-one. Then, for any a with $f'(a)$ finite and with $f(a) = b$, we have:

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} = \lim_{y \to f(a)} \frac{y - f(a)}{g(y) - a} = \lim_{y \to b} \frac{y - b}{g(y) - g(b)} = \frac{1}{g'(b)}$$

This explains why $f'(a)$ and $g'(b)$ are multiplicative inverses of each other.

The multiplicative inverse relationship can also be verified using the chain rule. Here, we use the fact that $f \circ g$ is the identity map, and apply the chain rule to get:

$$(f \circ g)'(b) = f'(g(b))g'(b) \implies 1 = f'(a)g'(b)$$

5.3. **Higher derivatives.** Recall that we can compute higher derivatives as well for various ways of creating new functions from old. For sums, differences, and scalar multiples, the rule is simple: since differentiation is a linear operator, the k^{th} derivative of the sum/difference/scalar multiple is the sum/difference/scalar multiple of the k^{th} derivatives. For products, the k^{th} derivative, as we saw in some quizzes, has a binomial formula, which we can discover by iteration. In particular, for instance:

$$\begin{aligned}
(f \cdot g)' &= (f' \cdot g) + (f \cdot g') \\
(f \cdot g)'' &= (f'' \cdot g) + 2(f' \cdot g') + (f \cdot g'') \\
(f \cdot g)''' &= (f''' \cdot g) + 3(f'' \cdot g') + 3(f' \cdot g'') + (f \cdot g''')
\end{aligned}$$

The story is trickiest for composites, where, in order to compute the second derivative of a composite, we need to use the chain rule *and* the product rule. The formula we get, which you saw in a past quiz, was:

$$(f \circ g)'' = (f'' \circ g) \cdot (g')^2 + (f' \circ g) \cdot (g'')$$

We have to do something similar to calculate the second derivative of the inverse function. However, this time we need to use the *quotient rule*. Note that:

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

To find the second derivative, we must differentiate both sides and use the quotient rule or equivalently, the rule for differentiating a reciprocal function. The upshot is that we get:

$$(f^{-1})''(x) = \frac{-f''(f^{-1}(x))}{(f'(f^{-1}(x)))^3}$$

You'll be working out the full details of this in a homework problem.

LOGARITHM, EXPONENTIAL, DERIVATIVE, AND INTEGRAL

MATH 152, SECTION 55 (VIPUL NAIK)

Corresponding material in the book: Section 7.2, 7.3, 7.4.

What students should definitely get: The definition of logarithm as an integral, its key properties. The differentiation and integration formulas for logarithm and exponential, the key ideas behind combining these with the chain rule and u-substitution to carry out other integrals.

Executive summary

0.1. **Logarithm and exponential: basics.**

(1) The *natural logarithm* is a one-to-one function with domain $(0, \infty)$ and range $\mathbb{R}$, and is defined as $\ln(x) := \int_1^x (dt/t)$.
(2) The natural logarithm is an increasing function that is concave down. It satisfies the identities $\ln(1) = 0$, $\ln(ab) = \ln(a) + \ln(b)$, $\ln(a^r) = r \ln a$, and $\ln(1/a) = -\ln a$.
(3) The limit $\lim_{x\to 0} \ln(x)$ is $-\infty$ and the limit $\lim_{x\to\infty} \ln(x)$ is $+\infty$. Note that ln goes off to $+\infty$ at ∞ even though its derivative goes to zero as $x \to +\infty$.
(4) The derivative of $\ln(x)$ is $1/x$ and the derivative of $\ln(kx)$ is also $1/x$. The derivative of $\ln(x^r)$ is r/x.
(5) The antiderivative of $1/x$ is $\ln|x| + C$. What this really means is that the antiderivative is $\ln(-x)+C$ when x is negative and $\ln(x)+C$ when x is positive. If we consider $1/x$ on both positive and negative reals, the constant on the negative side is unrelated to the constant on the positive side.
(6) e is defined as the unique number x such that $\ln(x) = 1$. e is approximately 2.718. In particular, it is between 2 and 3.
(7) The inverse of the natural logarithm function is denoted exp, and $\exp(x)$ is also written as e^x. When x is a rational number, $e^x = e^x$ (i.e., the two definitions of exponentiation coincide). In particular, $e^1 = e$, $e^0 = 1$, etc.
(8) The function exp equals its own derivative and hence also its own antiderivative. Further, the derivative of $x \mapsto e^{mx}$ is me^{mx}. Similarly, the integral of e^{mx} is $(1/m)e^{mx} + C$.
(9) We have $\exp(x + y) = \exp(x)\exp(y)$, $\exp(rx) = (\exp(x))^r$, $\exp(0) = 1$, and $\exp(-x) = 1/\exp(x)$. All of these follow from the corresponding identities for ln.

Actions...

(1) We can calculate $\ln(x)$ for given x by using the usual methods of estimating the values of integrals, applied to the function $1/x$. We can also use the known properties of logarithms, as well as approximate ln values for some specific x values, to estimate $\ln x$ to a reasonable approximation. For this, it helps to remember $\ln 2$, $\ln 3$, and $\ln 5$ or $\ln 10$.
(2) Since both ln and exp are one-to-one, we can *cancel* ln from both sides of an equation and similarly *cancel* exp. Technically, we cancel ln by applying exp to both sides, and we cancel exp by applying ln to both sides.

0.2. **Integrations involving logarithms and exponents.** Words/actions ...

(1) If the numerator is the derivative of the denominator, the integral is the logarithm of the (absolute value of) the denominator. In symbols, $\int g'(x)/g(x)\,dx = \ln|g(x)| + C$.
(2) More generally, whenever we see an expression of the form $g'(x)/g(x)$ inside the integrand, we should consider the substitution $u = \ln|g(x)|$. Thus, $\int f(\ln|g(x)|)g'(x)/g(x)\,dx = \int f(u)\,du$ where $u = \ln|g(x)|$.
(3) $\int f(e^x)e^x\,dx = \int f(u)\,du$ where $u = e^x$.
(4) $\int e^x[f(x) + f'(x)]\,dx = e^x f(x) + C$.

(5) $\int e^{f(x)} f'(x)\, dx = e^{f(x)} + C$.
(6) Trigonometric integrals: $\int \tan x\, dx = -\ln|\cos x| + C$, and similar integration formulas for cot, sec and csc: $\int \cot x\, dx = \ln|\sin x| + C$, $\int \sec x = \ln|\sec x + \tan x| + C$, and $\int \csc x\, dx = \ln|\csc x - \cot x| + C$.

1. Logarithms: the adventure begins

1.1. **Finding an antiderivative of the reciprocal function.** Recall that the process of differentiation never gave us fundamentally new functions, because the derivatives of all the basic functions that we knew were expressible in terms of other basic functions, and using the operations of pointwise combination and composition did not allow us to break ground into new functions. The situation differs somewhat for integration. We have seen that we often come across functions for which we have no clue as to how to find an antiderivative. We now discuss how to handle one such function.

This function is the function $1/x$, which, for now, we will assume to be a function on $(0, \infty)$. We want to find an antiderivative for this function.

The basic results of integration tell us that one way of defining an antiderivative is by using a definite integral from a fixed value to x, as long as that fixed value is in the domain. For reasons that are not immediately obvious, we choose the fixed value (the reference point) as 1. We thus define the following function:

$$L(x) := \int_1^x \frac{dt}{t}$$

Note that this is the *unique* antiderivative which has the property that its value at 1 is 0. By definition, $L'(x) = 1/x$ for all x. What further information can we derive about L?

1.2. **Using the multiplicative transform.** By the u-substitution method, we can readily verify that, for $a, b > 0$:

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

The key thing that is special about $1/x$ is that the multiplicative factor on the dt part cancels the multiplicative factor on the t part.

This gives us that:

$$L(a) - L(1) = L(ab) - L(b)$$

Since $L(1) = 0$, we obtain that L is a function satisfying the property:

$$L(ab) = L(a) + L(b)\ \forall\ a, b > 0$$

Thus, even though we do not have an explicit description of L, we know that L converts products to sums. In particular, we also see, for instance, that:

$$L(a^n) = nL(a)\ \forall a > 0, n \in \mathbb{Z}$$

In particular, $L(1/a) = -L(a)$.

We can further see that for any rational number r, we have:

$$L(a^r) = rL(a)\ \forall a > 0, r \in \mathbb{Q}$$

In other words, the function L converts products to sums and pulls the exponent into a multiple. We also know that since $L'(x) > 0$ for all $x > 0$, L is continuous and increasing. In particular, we see that L is a one-to-one map on $(0, \infty)$.

What is the range of L? Consider $a = 2$. Then, $L(a) = L(2) > 0$. As $n \to \infty$, $L(a^n) = nL(a) \to \infty$, and as $n \to -\infty$, $L(a^n) = nL(a) \to -\infty$. Since L is increasing, we can use this to see that $\lim_{x\to\infty} L(x) = \infty$ and $\lim_{x\to 0} L(x) = -\infty$. Further, by the intermediate value theorem, we see that the range of L is $\mathbb{R}$.

The upshot: L is a continuous increasing one-to-one function from $(0, \infty)$ to $\mathbb{R}$ that sends 1 to 0 and converts products to sums.

1.3. **L for (natural) logarithm.** The function L that is described above is termed the *natural logarithm* function. It is ubiquitous in mathematics, and is denoted ln. Thus, we have the definition:

$$\ln x := \int_1^x \frac{dt}{t} \ \forall \ x > 0$$

It turns out that this natural logarithm behaves in ways very similar to logarithms to base 10. A quick primer for those who didn't live in prehistoric times: in the olden days, when people had to do multiplications by hand, they used a tool called *logarithm tables* to do these multiplications. The logarithm tables basically converted the multiplication problem to an addition problem.

Here is the principle on which the logarithm tables worked. These tables allowed you to, for a given number x, find the approximate value of r such that $10^r = x$. This value of r is called $\log_{10} x$. Then, if you had to multiply x and y, you first found $\log_{10} x$ and $\log_{10} y$. It turns out that $\log_{10}(xy) = \log_{10}(x) + \log_{10} y$, because if $10^r = x$ and $10^s = y$, then $10^{r+s} = xy$ by properties of exponents. Thus, to find xy, we find $\log_{10} x$ and $\log_{10}(y)$ and add them. Then, there are antilogarithm tables, that allow us to find the antilogarithm of this sum that we have computed (or basically, raise 10 to the power of that number).

The principle of logarithm tables was later converted to a *mechanical device* called the *slide rule*. How many people have used slide rules? What a slide rule does is use a *logarithmic scale*, i.e., it places numbers on a scale in such a way that the distance between the positions of two numbers is determined by their quotient. So, on a logarithmic scale, the distance between 1 and 10 is the same as the distance between 10 and 100, and also the same as the distance between 0.01 and 0.1. The distance between 3 and 7 is the same as the distance between 30 and 70. (If you're interested in pictures of slide rules, do a Google image search. I haven't included any picture here because of potential copyright considerations).

A slide rule comprises two logarithmic scales (using the same calibration) but one of them can slide against each other. We can use the sliding scale to add lengths along the scale, but since the scale is logarithmic, this ends up multiplying the numbers. You may have heard about how people with an abacus can often do simple calculations faster than people with a calculator. It turns out that people with a slide rule can usually do multiplications faster than people with a calculator.[1]

Logarithmic scales are used in many measurements. Here are some examples:

(1) The **Richter magnitude scale** measures the intensity of earthquakes. It is calibrated logarithmically to base 10. An earthquake one point higher on the Richter scale is ten times as intense.
(2) The **pH scale** in chemistry is a logarithmic scale to measure the concentration of the H^+ (more precisely, H_3O^+) ions. It is a negative logarithmic scale to base 10. An increase in the pH value by 1 corresponds to a decrease in the hydronium ion concentration to 1/10 of its original value.
(3) The **decibel scale**, used for sound levels and other level measurements, is a logarithmic scale where an increase in 10 points along the scale corresponds to a ten-fold increase in amplitude. Thus, $20dB$ is ten times as loud as $10dB$.

The natural logarithm function can be thought of as creating a logarithmic scale on the positive reals. But the question we are concerned with is: what precisely is this scale? How does this compare with the usual logarithm to base 10? Our hunch is that there should be a number e such that $\ln(x)$ is the value r such that $x = e^r$. What must this number e be?

2. The back and forth of things: logarithm and exponential

2.1. **In search of e.** If such a number e exists, then it must be the unique number x satisfying $\ln(x) = 1$. Further, since ln is an increasing function, we can try locating e between two consecutive integers by determining $\ln 2$, $\ln 3$, and so on. Actually, we can be more clever.

We can begin by trying to compute $\ln 2$. We could do this using upper and lower sums. We could also do it by noting that $\ln x = \int_1^x dt/t$, and must be located between the antiderivatives of $x^{-1/2}$ and $x^{-3/2}$. We did this approximation a few weeks ago and found that $\ln 2$ is located between 0.58 and 0.83. Further approximations using either this method or upper and lower sums for partitions yields than $\ln 2$ is between 0.69 and 0.70. We will assume $\ln 2 \approx 0.7$ for calculations.

[1]On the other hand, the calculator is a lot more versatile than the slide rule, and is probably faster for computing roots, multiplying long sequences of numbers, or combinations of multiplication and addition.

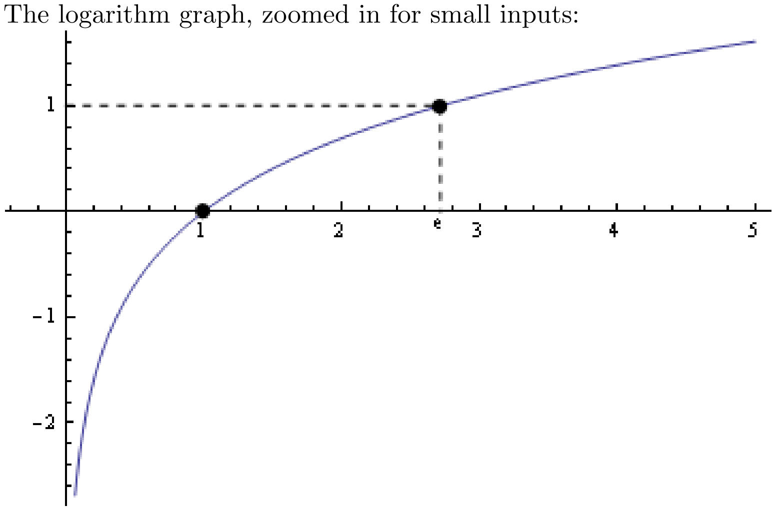
Now that we know $\ln 2$, we do not need to do any more messy work with upper and lower sums. We know that $2 < 2\sqrt{2} < 3$. We also have that $\ln(2\sqrt{2}) = (3/2)\ln 2 \approx 1.05$, which is bigger than 1. Thus, $\ln 3 > 1$, so $2 < e < 3$. In fact, $2 < e < 2\sqrt{2} \approx 2.83$. We can also calculate, for instance, that the cuberoot of 2 is about 1.26. Thus, $\ln(2.52)$ is approximately $(4/3)(\ln 2)$ which is approximately 0.93. Thus, we get that e should be bigger than 2.52. A similar process of successive approximations yields that the value of e is approximately 2.718281828. You should know e to at least three decimal places: 2.718.

Clearly, since $\ln(e) = 1$, we have that $\ln(e^{p/q}) = p/q$ for integers p and q, with $q \neq 0$. It is not clear a priori what we would mean by the notation e^r for irrational numbers r, but whatever we may mean, it should be the case that $\ln(e^r) = r$. In other words, the function $x \mapsto e^x$ must be the inverse function of the one-to-one function ln. The function $x \mapsto e^x$ is also called the exponentiation function, and sometimes denoted exp. By the way, the letter e could be thought of as standing for *exponentiation*, but historically it is believed to be named after Leonhard Euler, a prolific mathematician who studied the properties both of the number e and the exponentiation function.

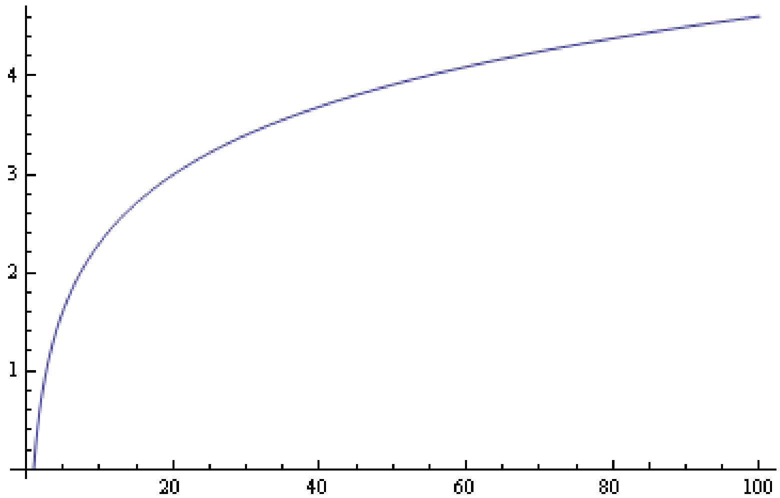
2.2. **Logarithms and exponents: some rules.** Here are some of the rules:

(1) ln is a one-to-one function from $(0, \infty)$ to $\mathbb{R}$ and exp is a one-to-one function from $\mathbb{R}$ to $(0, \infty)$. The two functions are inverse functions of each other.
(2) ln converts products to sums and exp converts sums to products. In other words, $\ln(xy) = \ln(x) + \ln(y)$ and $\exp(x + y) = \exp(x)\exp(y)$.
(3) $\ln(1) = 0$ and $\exp(0) = 1$.
(4) $\ln(1/x) = -\ln x$ and $\exp(-x) = 1/\exp(x)$.
(5) $\ln(x^r) = r \ln x$ and $\exp(rx) = (\exp(x))^r$.
(6) Both exp and ln are continuous and increasing functions.
(7) exp has the x-axis as a horizontal asymptote as $x \to -\infty$, while ln has the y-axis as a vertical asymptote as $x \to 0$.
(8) exp is concave up and ln is concave down.

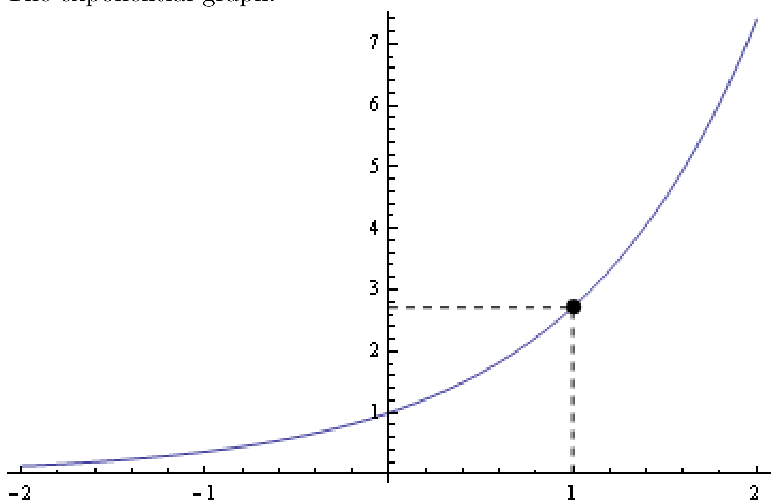
Here are the graphs:
The logarithm graph, zoomed in for small inputs:



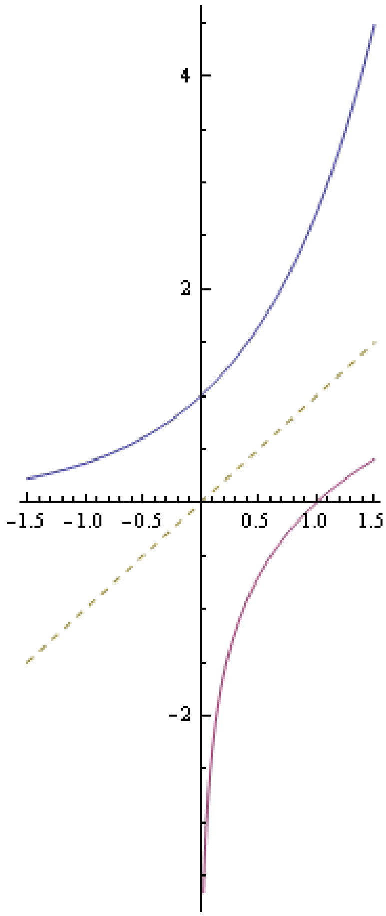
The logarithm graph, zoomed out:



The exponential graph:



Here are the logarithm and exponential graphs together, so that we can see that the graphs are reflections of each other about the $y = x$ line:



2.3. **Numerical shennanigans.** In his autobiographical book, Richard Feynmann discusses how he impressed a bunch of mathematicians by being able to calculate natural logarithms of many numbers to one or two decimal places. However, what he did was hardly impressive. It turns out that remembering the natural logarithms of a few numbers allows us to compute them approximately for many numbers.

For instance, it is useful to remember that $\ln(2) \approx 0.6931$, $\ln(3) \approx 1.0986$, $\ln(10) \approx 2.3026$, and $\ln(7) \approx 1.9459$. We can now calculate the logarithm values for most integers. How? Using the fact that logarithms translate multiplication to addition. Thus, $\ln(4) = 2\ln(2) \approx 1.3862$, while $\ln(5) = \ln(10) - \ln(2) \approx 1.6095$. In fact, we can readily calculate the natural logarithm of any positive integer all of whose prime factors are among 2, 3, 5, and 7. What about $\ln(11)$? While we cannot calculate this precisely, we can calculate $\ln(10)$ and $\ln(12)$ and thus obtain reasonable upper and lower bounds for $\ln(11)$. Even better, we know that $\ln(120) < 2\ln(11) < \ln(125)$, and since we can calculate both $\ln(120)$ and $\ln(125)$, we get a pretty small range for $\ln(11)$.

In fact, Feynman was able to impress physicists by doing calculations that essentially relied on only two facts: $\ln(2) \approx 0.7$ and $\ln(10) \approx 2.3$.

If we are able to quickly calculate natural logarithms, a happy corollary of that is that we can quickly integrate dx/x on intervals.

2.4. **Domain and range issues.** For domain computations in the past, we used the following basic guidelines:

(1) Things in the denominator must be nonzero.
(2) Things with squareroots or even roots must be nonnegative.
(3) Things with squareroots or even roots in the denominator must be positive.

We now add two more criteria:

(4) Things under logarithm must be positive.
(5) Things under logarithm of the absolute value must be nonzero.

2.5. **Logarithm of the absolute value.** The natural logarithm function is defined only for positive reals. However, we can extend it to a function on all reals by taking the absolute value first, i.e., we look at the function $x \mapsto \ln(|x|)$. This is an even function and its graph is obtained by takingthe graph of the logarithm function and adding its mirror image about the y-axis

It turns out that the derivative of $\ln(|x|)$ is $1/x$. In particular, we see that $\ln(|x|)$ serves as an antiderivative of $1/x$ for *all nonzero* x. This is an improvement on $\ln(x)$, which worked only for positive x. However, we should be careful because the domain of $1/x$ as well as of $\ln|x|$ excludes zero. Hence, the behavior on the positive and negative side are totally independent of each other. We shall return to this point in a later lecure.

3. Formulas for derivatives and integrals

3.1. **Derivative and integral formulas for logarithms.** The main formula that we have, which follows from our definition of natural logarithm, is the following:

$$\frac{d}{dx}(\ln x) = \frac{1}{x}$$

This is the formula for $x > 0$. A more general version, for $x \neq 0$, is:

$$\frac{d}{dx}(\ln|x|) = \frac{1}{x}$$

The corresponding antiderivative formula for $x > 0$ is:

$$\int \frac{dx}{x} = \ln(x) + C$$

In general, the antiderivative formula is:

$$\int \frac{dx}{x} = \ln(|x|) + C$$

However, it should be remembered that this formula is valid only when we are working with x either in $(0, \infty)$ or in $(-\infty, 0)$, i.e., we cannot use the formula to cross between the interval $(0, \infty)$ and $(-\infty, 0)$. In fact, if we have a function $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ such that $f'(x) = 1/x$, then we can guarantee that $f(x) - \ln(|x|)$ is constant on $x > 0$ and is constant on $x < 0$. However, these constants may differ. The behavior on the $(0, \infty)$ connected component does not in any way constrain the behavior on the $(-\infty, 0)$ connected component.

3.2. **The exponential and its derivative and integral formulas.** Recall that exp is the inverse of the ln function. Thus, we can use the rule for differentiating the inverse function to find $\exp'$. We have:

$$\exp'(x) = \frac{1}{\ln'(\exp x)} = \frac{1}{\frac{1}{\exp(x)}} = \exp(x)$$

Thus, we have the remarkable property that the exponential function, i.e., the function $x \mapsto e^x$, is its own derivative. Another way of thinking about this is that the rate of growth of the exponential function is *equal* to its value. Note that this also implies that the exponential function is infinitely differentiable and all higher derivatives equal the same function.

We rewrite the above in Liebniz notation:

$$\frac{d}{dx}(e^x) = e^x$$

We also note the corresponding statement for indefinite integration:

$$\int e^x \, dx = e^x + C$$

3.3. **Some corollaries.** Using the above, we obtain the following identities for the logarithm and exponent:

$$\begin{aligned}
\frac{d}{dx}(\ln(kx)) &= \frac{1}{x} \\
\frac{d}{dx}(\ln(x^r)) &= \frac{r}{x} \\
\frac{d}{dx}(e^{mx}) &= me^{mx} \\
\int e^{mx}\,dx &= \frac{1}{m}e^{mx} + C
\end{aligned}$$

Each of these identities can be derived in two ways: either by using the properties of logarithms and exponents on the inside and then differentiating, or by first differentiating and then simplifying. For instance, for the second identity, we can either simplify $\ln(x^r)$ as $r\ln(x)$ first and then pull the constant r out before differentiating, or we can use the chain rule to obtain $(1/x^r)\cdot rx^{r-1}$. It is gratifying to know that the answers we obtain both ways are the same.

4. Application to indefinite and definite integration

4.1. **The u-substitution: a textbook example.** We begin with an easy example:

$$\int_{\sqrt{n}}^{n} \frac{1}{x\ln x}\,dx$$

Here, n is an integer greater than 1.

Believe it or not, this integral actually came up in some asymptotic approximations I was doing some time ago to figure out whether some numbers have large prime divisors! Let us first look at the indefinite integral. The substitution $u = \ln x$ gives us:

$$\int \frac{1}{x\ln x} = \int \frac{du}{u} = \ln(u) = \ln(\ln x) + C$$

Note that we do not need to put absolute values here because on the interval of integration, ln is positive. Now, we can evaluate between limits:

$$[\ln(\ln x)]_{\sqrt{n}}^{n} = \ln(\ln n) - \ln(\ln\sqrt{n}) = \ln\left[\frac{\ln n}{\ln\sqrt{n}}\right] = \ln\left[\frac{\ln n}{(1/2)\ln n}\right] = \ln 2$$

So, the answer is $\ln 2$, which, as we computed earlier, is approximate 0.693. Apparently, this is the rough heuristic argument for why about 69.3% of the numbers have a prime divisor greater than their squareroot.

4.2. **Numerator as derivative of denominator.** The gist of this logarithmic substitution can be captured by the formula:

$$\int \frac{g'(x)}{g(x)}\,dx = \ln|g(x)| + C$$

The proof of this proceeds via setting $u = g(x)$. Thus, the general idea when using logarithmic substitutions is to try to obtain the numerator as the derivative of the denominator. For instance, consider the integral:

$$\int \frac{x}{x^2+1}\,dx$$

Here, the derivative of the denominator is $2x$, so we adjust by a factor of 2 to obtain:

$$\frac{1}{2}\int \frac{2x}{x^2+1}\,dx = \frac{1}{2}\ln(|x^2+1|) + C$$

Note that in this case, since x^2+1 is always positive, the absolute value can be dropped and we get $\frac{1}{2})(x^2+1) + C$. Further, this antiderivative is valid over all reals.

4.3. **Trigonometric integrals involving logarithms.** Recall so far that we have seen the antiderivatives of sin, cos, $\sec^2$, $\sec \cdot \tan$, $\csc^2$, and $\csc \cdot \cot$. We also used these, along with trigonometric identities, to compute antiderivatives for $\sin^2$, $\cos^2$, $\tan^2$, and $\cot^2$. All these results were obtained as corollaries of the differentiation formulas.

We now try to obtain a formula to integrate cot. The key idea is to note that:

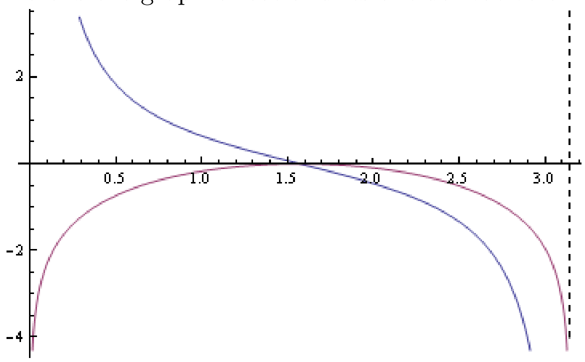
$$\cot x = \frac{\cos x}{\sin x}$$

Since cos is the derivative of sin, this matches up with the general pattern that we just discussed, and we obtain that the antiderivative of cot is $\ln \circ |\sin|$. In other words:

$$\int \cot x \, dx = \ln|\sin x| + C$$

Note that cot is undefined at multiples of π, and so any integration of this sort is valid only if the entire interval of integration lies strictly between two consecutive multiples of π. It is also instructive to graph the antiderivative of cot.

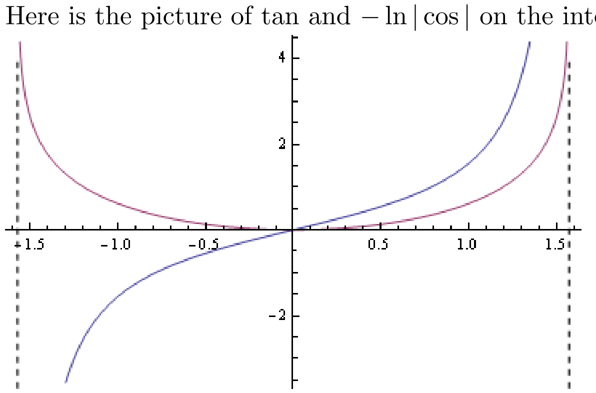
Here is the graph of cot and its antiderivative on the interval $(0, \pi)$, where both are defined:



Similarly, we obtain:

$$\int \tan x \, dx = -\ln|\cos x| + C = \ln|\sec x| + C$$

Here is the picture of tan and $-\ln|\cos|$ on the interval $(-\pi/2, \pi/2)$, where both are defined:



Note that those two expressions are the same because cos and sec are reciprocals of each other.

Let us look at a somewhat harder integral: the integral of the secant function:

$$\int \sec x \, dx = \int \frac{\sec x(\sec x + \tan x)}{\sec x + \tan x} \, dx = \int \frac{\sec^2 x + \sec x \tan x}{\tan x + \sec x} \, dx$$

The numerator is the derivative of the denominator, and we obtain:

$$\int \sec x \, dx = \ln|\sec x + \tan x| + C$$

In a similar vein, we obtain that:

$$\int \csc x \, dx = \ln|\csc x - \cot x| + C$$

4.4. **Domain and range issues.** When doing indefinite integration, it is often best to forget about issues of domain and range and just let the algebraic manipulations flow. However, to interpret the results at the end, it is important to look at the domain and range issues. Ideally, the antiderivative should be defined and should make sense on all intervals where the function itself is continuous. Further, if there are points where the function is continuous but the *expression obtained for the antiderivative* is not defined, we should try to obtain the limit at that point.

4.5. **An application: integrating the cube of the tangent function.** Let us look at an application of the above:

$$\int \tan^3 x \, dx$$

Before we proceed, it is worth remarking how different integration is from differentiation. For differentiation, there was just the formula for differentiating sin and cos, and everything else followed using the product rule and quotient rule. We still memorized more, but that was mainly to speed things up, not out of necessity. With integration, on the other hand, we need to have a whole bag of *ad hoc* tricks that we try one after the other.

Let us look at this integral. The key thing to do here is to break down $\tan^3 x = \tan x \cdot \tan^2 x$. Next, we use $\tan^2 x = \sec^2 x - 1$, and we have:

$$\int \tan x \sec^2 x \, dx - \int \tan x \, dx$$

The first integral can be quickly calculated using the chain rule or u-substitution, since $\sec^2 x$ is the derivative of $\tan x$. The second integral comes from our formula, and we get:

$$\frac{\tan^2 x}{2} + \ln|\cos x| + C$$

4.6. **A fancier formula.** Here is a formula that uses the chain rule twice:

$$\int \frac{g'(x) f(\ln|g(x)|)}{g(x)} \, dx = \int f(u) \, du$$

where $u = \ln(|g(x)|)$. For instance:

$$\int \frac{2x(\ln(x^2+1))^3}{x^2+1} \, dx = \int u^3 \, du$$

where $u = \ln(x^2+1)$. This further simplifies to:

$$\frac{1}{4}[\ln(x^2+1)]^4 + C$$

5. More tricks and techniques

5.1. **Logarithmic differentiation.** Logarithmic derivatives are both a conceptual and a computational tool. Currently, we focus on the computational aspects. The idea is to use the same formula that we obtained earlier, but in reverse:

$$\frac{d}{dx} \ln(|g(x)|) = \frac{g'(x)}{g(x)}$$

Rearranging the terms yields:

$$g'(x) = g(x) \frac{d}{dx} \ln(|g(x)|)$$

If g is a product of functions $g_1, g_2, \dots, g_n$, then $\ln|g(x)| = \ln|g_1(x)| + \cdots + \ln|g_n(x)|$, and we get:

$$g'(x) = g(x)\left[\frac{g_1'(x)}{g_1(x)} + \frac{g_2'(x)}{g_2(x)} + \cdots + \frac{g_n'(x)}{g_n(x)}\right]$$

Note that this expression is not *really* new and did not *really* require logarithms. You can in fact convince yourself that it is just a reformulation of the product rule. When $g = g_1g_2$, for instance, this says that:

$$g' = g\left[\frac{g_1'}{g_1} + \frac{g_2'}{g_2}\right]$$

Substituting $g = g_1g_2$, this simplifies to the usual product rule. However, the logarithmic formulation has some conceptual advantages.

For instance, suppose $g(x) := x(x-1)(x-2)$. Then, we immediately obtain that:

$$\frac{g'(x)}{g(x)} = \frac{1}{x} + \frac{1}{x-1} + \frac{1}{x-2}$$

Further, of we have $g(x) = g_1(x)^{a_1}g_2(x)^{a_2}\dots g_n(x)^{a_n}$, we obtain that:

$$\frac{g'(x)}{g(x)} = \frac{a_1g_1'(x)}{g_1(x)} + \frac{a_2g_2'(x)}{g_2(x)} + \cdots + \frac{a_ng_n'(x)}{g_n(x)}$$

So, if $g(x) = x^3(x-1)^4(x-2)^5$, we obtain that:

$$\frac{g'(x)}{g(x)} = \frac{3}{x} + \frac{4}{x-1} + \frac{5}{x-2}$$

5.2. **Exponentiation tricks.** We have already seen basic integration and differentiation identities for the exponentiation function. There are some ways of combining these identities with the chain rule. I note some special cases here.

(1) For any function f, the derivative of $f(x)e^x$ is $(f(x) + f'(x))e^x$. Thus, the integral of $g(x)e^x$ is $f(x)e^x + C$ where $f + f' = g$.
(2) (1) is particularly useful when integrating polynomial function times e^x. This is because we can use linear algebra to find, for a given polynomial g, the unique polynomial f such that $f + f' = g$.
(3) The integral $\int f(e^x)e^x\,dx$ is $\int f(u)\,du$ where $u = e^x$.
(4) The integral $e^{f(x)}f'(x)\,dx$ is $e^{f(x)} + C$.

Let us consider an example to illustrate this. Consider the function:

$$F(x) := (x^2 + 5x + 1)e^x$$

The derivative of this, by the product rule, turn out to be e^x times the sum of x^2+5x+1 and its derivative, giving:

$$F'(x) = (x^2 + 7x + 6)e^x$$

Note that this is a new polynomial times e^x. An intersting question would be how we could reverse this procedure, i.e., given $g(x)e^x$ where g is a polynomial, how do we find a polynomial f such that the derivaitve of $f(x)e^x$ is $g(x)e^x$? By the product rule, we obtain that:

$$g(x) = f(x) + f'(x)$$

Thus, we need to find the coefficients of f. Let us do this in our concrete case where $g(x) = x^2 + 7x + 6$.

We know that the degree of f' is strictly smaller than the degree of f, so $f + f'$ has the same degree and same leading coefficient as f. In this case, this forces f to be a quadratic polynomial of the form x^2+mx+n. We then get $f'(x) = 2x + m$, and we obtain that:

$$f(x) + f'(x) = x^2 + (m+2)x + (m+n)$$

Since we are given that $g(x) = x^2 + 7x + 6$, we can match coefficients and obtain:

$$m + 2 = 7, \qquad m + n = 6$$

Solving, we get $m = 5$, and $n = 1$, and we get $f(x) = x^2 + 5x + 1$, recovering our original polynomial.

Although the *specific procedure involving comparing coefficients* does require that we are *dealing with polynomials*, the general idea remains true in a broader sense: integrating $\int g(x)e^x$ is equivalent to finding a function f such that $f + f' = g$. In some cases, it is easier to think of it as an integration problem, and in others, it is easier to think of it in terms of the differental equation $f + f' = g$. The idea that these two apparently different computations measure the same thing is extremely important.

EXPONENTIATION WITH ARBITRARY BASES, EXPONENTS

MATH 152, SECTION 55 (VIPUL NAIK)

Corresponding material in the book: Section 7.5.

What students should definitely get: The definition of a^b, where $a > 0$ and b is real. The definition of logarithm to positive base. The method of differentiating functions where the exponent (or base of logarithm) itself is variable. Key properties of exponents and logarithms.

Executive summary

Words ...

(1) For $a > 0$ and b real, we define $a^b := \exp(b \ln a)$. This coincides with the usual definition when b is rational.
(2) All the laws of exponents that we are familiar with for integer and rational exponents continue to hold. In particular, $a^0 = 1$, $a^{b+c} = a^b \cdot a^c$, $a^1 = a$, and $a^{bc} = (a^b)^c$.
(3) The exponentiation function is continuous in the exponent variable. In particular, for a fixed value of $a > 0$, the function $x \mapsto a^x$ is continuous. When $a \neq 1$, it is also one-to-one with domain $\mathbb{R}$ and range $(0, \infty)$, with inverse function $y \mapsto (\ln y)/(\ln a)$, which is also written as $\log_a(y)$. In the case $a > 1$, it is an increasing function, and in the case $a < 1$, it is a decreasing function.
(4) The exponentiation function is also continuous in the base variable. In particular, for a fixed value of b, the function $x \mapsto x^b$ is continuous. When $b \neq 0$, it is a one-to-one function with domain and range both $(0, \infty)$, and the inverse function is $y \mapsto y^{1/b}$. In case $b > 0$, the function is increasing, and in case $b < 0$, the function is decreasing.
(5) Actually, we can say something stronger about a^b – it is *jointly* continuous in both variables. This is hard to describe formally here, but what it approximately means is that if f and g are both continuous functions, and f takes positive values only, then $x \mapsto [f(x)]^{g(x)}$ is also continuous.
(6) The derivative of the function $[f(x)]^{g(x)}$ is $[f(x)]^{g(x)}$ times the derivative of its logarithm, which is $g(x)\ln(f(x))$. We can further simplify this to obtain the formula:

$$\frac{d}{dx}\left([f(x)]^{g(x)}\right) = [f(x)]^{g(x)}\left[\frac{g(x)f'(x)}{f(x)} + g'(x)\ln(f(x))\right]$$

(7) Special cases worth noting: the derivative of $(f(x))^r$ is $r(f(x))^{r-1}f'(x)$ and the derivative of $a^{g(x)}$ is $a^{g(x)}g'(x)\ln a$.
(8) Even further special cases: the derivative of x^r is rx^{r-1} and the derivative of a^x is $a^x \ln a$.
(9) The antiderivative of x^r is $x^{r+1}/(r+1) + C$ (for $r \neq -1$) and $\ln|x| + C$ for $r = -1$. The antiderivative of a^x is $a^x/(\ln a) + C$ for $a \neq 1$ and $x + C$ for $a = 1$.
(10) The logarithm $\log_a(b)$ is defined as $(\ln b)/(\ln a)$. This is called the logarithm of b to base a. Note that this is defined when a and b are both positive <u>and $a \neq 1$</u>. This satisfies a bunch of identities, most of which are direct consequences of identities for the natural logarithm. In particular, $\log_a(bc) = \log_a(b) + \log_a(c)$, $\log_a(b)\log_b(c) = \log_a(c)$, $\log_a(1) = 0$, $\log_a(a) = 1$, $\log_a(a^r) = r$, $\log_a(b) \cdot \log_b(a) = 1$ and so on.

Actions...

(1) We can use the formulas here to differentiate expressions of the form $f(x)^{g(x)}$, and even to differentiate longer exponent towers (such as x^{x^x} and 2^{2^x}).
(2) To solve an integration problem with exponents, it may be most prudent to rewrite a^b as $\exp(b \ln a)$ and work from there onward using the rules mastered earlier. Similarly, when dealing with relative logarithms, it may be most prudent to convert all expressions in terms of natural logairthms and then use the rules mastered earlier.

1. Review and definitions

1.1. Exponents: what we know. Let us consider the expression a^b, with a positive. So far, we have made sense of this expression in the following cases:

(1) b is a positive integer: In this case, a^b is defined as the product of a with itself b times. This definition is fairly general; in fact, it makes sense even when $a < 0$.

(2) b is an integer: If b is positive, we use (1). If $b = 0$, we define a^b as 1, and if $b < 0$, we define a^b as $1/a^{|b|}$.

(3) $b = p/q$ is rational, p, q integers, $q > 0$: In this case, a^b is defined as the unique positive real number c such that $c^q = a^p$. The existence of such a c was not proved rigorously, but it essentially follows by an application of the intermediate value theorem. We proceed as follows: we show that the function $x \mapsto x^q$ is less than a^p for some positive x and greater than a^p for some positive x, and hence, by the intermediate value theorem, it must equal a for some positive x. The uniqueness of this is guaranteed by the fact that the function $x \mapsto x^q$ is increasing.

The notion of rational exponent has the added advantage that when the denominator is odd, it can be extended to the negative numbers as well. Also, note that when $b > 0$, we define $0^b = 0$.

So far, exponents satisfy some laws, namely:

$$\begin{aligned} a^0 &= 1 \\ a^{b+c} &= a^b \cdot a^c \\ a^{-b} &= \frac{1}{a^b} \\ a^{bc} &= (a^b)^c \end{aligned}$$

For the rest of this document, where we study arbitrary real exponents b, we restrict ourselves to the situation where the base a of exponentiation is positive.

1.2. And here's how mathematicians would think about it. We're not mathematicians, but since we're doing mathematics, it might help to think about the way mathematicians would view this. A mathematician would begin by defining positive exponents: things like a^b where a is a positive real and b is a positive integer. Then, the mathematician would observe that $a^{b+c} = a^b \cdot a^c$ and $a^{bc} = (a^b)^c$. The mathematician would then ask: is there a way of extending the definition to encompass more values of b while preserving these two laws of exponents? Further, is the way more or less unique or are there multiple different extensions?

It turns out that there is a way, and it is unique, and it is exactly the way I mentioned above. In other words, if we want exponentiation to behave such that $a^{b+c} = a^b \cdot a^c$ and if we have it defined the usual way for positive integers b, we are forced to define it the usual way for all integers b. Further, we are forced to define it the way we have defined it for rational numbers b. The rules constrain us.

1.3. Real exponents: continuity the realistic constraint. We now move to the situation involving real exponents. Unfortunately, the laws of exponents are not enough to force us to a specific definition of a^b for a positive real and b real. However, the laws of exponents, along with *continuity in both a and b* do turn out to be enough to force a specific definition of a^b. To see this, note that we already have defined a^b for b rational, and the rationals are *dense* in the reals, so to figure out the answer a^b for a real value of b, we take rationals c closer and closer to b and consider the limit of a^c.

For instance, to figure out $2^{\sqrt{2}}$, we look at 2, $2^{1.4}$, $2^{1.41}$, and so on. Each of these is well-defined, because in each case, the exponent is a rational number. Thus, $2^{1.4}$ is the unique positive solution to $x^5 = 2^7$, and $2^{1.41}$ is the unique positive solution to $x^{100} = 2^{141}$.

However, we would ideally like a clearer description that does not involve this approximation procedure. It turns out that the exponentiation function (obtained as the inverse function to the natural logarithm function) works out.

We define:

$$a^b := \exp(b \ln a) = e^{b \ln a}$$

Note that I use the exp notation because I want to emphasize that exp is just the inverse function to ln; it does not have an *a priori* meaning as exponentation. Note that this definition coincides with the usual definition for positive integer exponents, because $a^b = a \cdot a \cdot \cdots \cdot a$ b times, and thus we get:

$$\ln(a^b) = b \ln a$$

Applying exp to both sides gives:

$$a^b = \exp(b \ln a)$$

Similarly, we can show that the definition $a^b := \exp(b \ln a)$ coincides with the usual definition we have for negative integer exponents and for all rational exponents. Thus, this new definition extends the old definition.

Next, we can use the properties of exponents and logarithms to verify:

(1) With this new definition, the general laws for exponents listed above continue to hold.
(2) Under this new definition of exponent, a^b is continuous in each variable. In other words, for any fixed a, it is a continuous function of b, and for a fixed b, it is a continuous function of a. Actually, something stronger holds; it is jointly continuous in the two variables. However, joint continuity is a concept that is based on ideas of multivariable calculus, and hence beyond our scope.
(3) For a fixed $a \neq 1$, the function $x \mapsto a^x$ is a one-to-one function from $\mathbb{R}$ to $(0, \infty)$. When $a > 1$ the function is increasing, and when $a < 1$, the function is decreasing.
(4) For a fixed $b \neq 0$, the function is a one-to-one function from $(0, \infty)$ to $(0, \infty)$. When $b > 0$, it is an increasing function, and when $b < 0$, it is a decreasing function.

2. More perspective: exponents, logarithms, and radicals

2.1. **Addition, multiplication, and exponentiation.**

The *inverse operation* corresponding to addition is subtraction, and the *inverse operation* corresponding to multiplication is division. What is the inverse operation corresponding to exponentiation? The answer turns out to be tricky, because there are a couple of nice things about addition and multiplication that are no longer true for exponentiation:

(1) Addition and multiplication are both *commutative*: We have the remarkable fact that $a + b = b + a$ and $ab = ba$ for any a and b. On the other hand, exponentiation is not commutative. In fact, as we might see some time later, for every $a \in (0, \infty)$, there exist at most two values of $b \in (0, \infty)$ such that $a^b = b^a$, and one of those two values is a. For instance, the only numbers b for which $2^b = b^2$ are $b = 2$ and $b = 4$.
(2) Addition and multiplication are both *associative*: We have the remarkable fact that $a + (b + c) = (a + b) + c$ and $a(bc) = (ab)c$. On the other hand, exponentiation is not *associative*, i.e., it is not true in general that $a^{(b^c)} = (a^b)^c$. This is because $(a^b)^c = a^{bc}$, and so the equality would give that $b^c = bc$, which is very rare.

The noncommutativity of exponentiation would mean that there are two notions of inverse operation: a left inverse operation and a right inverse operation. The nonassociativity would mean that these inverse operations behave very differently from subtraction and division, and the analogy cannot be stretched too far.

2.2. **Radicals and logarithms.**

There are two kinds of inverse operations to the a^b operation. The first is to find solutions x to the equation:

$$x^b = c$$

Such a solution is a b^{th} root or b^{th} radical of c, and is given by:

$$x = c^{1/b}$$

It is also denoted as:

$$x = \sqrt[b]{c}$$

Note that as per our general discussion of the a^b function, we see that this is well-defined and unique if $b \neq 0$.

The other kind of inverse operation we may want is to solve the equation:

$$a^x = c$$

To solve this, we need to use the definition:

$$\exp(x \ln a) = c$$

Taking ln both sides, we obtain:

$$x \ln a = \ln c$$

Thus, we get that if $a \neq 1$:

$$x = \frac{\ln c}{\ln a}$$

Note that the uniqueness of x corresponds to the fact, observed earlier, that for fixed a, the map $x \mapsto a^x$ is one-to-one.

This solution x is also written as:

$$x = \log_a c$$

This is often read as *logarithm of c to base a*. The map $c \mapsto \log_a(c)$ is called *taking logarithms to base a*. As shown above, it is equivalent to taking natural logarithms and dividing by $\ln a$.

The upshot is that we define:

$$\log_a(c) := \frac{\ln c}{\ln a}$$

where a, c are both positive and $a \neq 1$. In particular, logarithms cannot be taken to base 1. Taking logarithms to base 1 is like division by zero, a forbidden operation.

2.3. **Properties of logarithms.** We have the following notable properties of logarithms, where we assume that all elements appearing in the base of the logarithm are positive and not equal to 1, and all elements whose logarithm is being taken are positive:

$$\begin{aligned}
\log_a(bc) &= \log_a(b) + \log_a(c) \\
\log_a(b^c) &= c \log_a(b) \\
\log_a(1/b) &= -\log_a(b) \\
\log_a(1) &= 0 \\
\log_a(a) &= 1 \\
\log_a(a^r) &= r \\
\log_{1/a}(b) &= -\log_a(b) \\
\log_a(b)\log_b(c) &= \log_a(c) \\
\log_b(a) &= \frac{1}{\log_a(b)}
\end{aligned}$$

All of these follow from the definition and the corresponding properties of ln, which follow from its definition as the antiderivative of the reciprocal function.

2.4. **Absolute and relative: natural bases and the bases for their naturality.** We can think of logarithm to a given base as measuring a *relative logarithm*. The natural logarithm is the logarithm to base e, which is the *natural logarithm*, in that the base e is the natural choice for a base. This behavior of logarithms is very similar to, for instance, the behavior of refractive indices for pairs of media through which light travels. For any pair of media, we can define a refractive index of the pair, but the *natural base* with respect to which we measure refractive index is vacuum. Natural logarithms play the role of vacuum: the natural base choice.

There are two other common choices of base for logarithms. One is base 10, which has no sound mathematical reason. The reason for taking logarithms to base 10 is because it is easy to compute the logarithm of any number by writing it in scientific notation using a table of logarithm values for numbers from 1 to 10. This allows us to use logarithms to base 10 as a convenient tool for multiplication, a convenience that seems to have been rendered moot in recent times with the proliferation of calculators.

The second natural choice of base for logarithms, which we will talk about next time, is logarithms to base 2. These come up for three reasons, listed below. We will return to some of these reasons in more detail when we study exponential growth and decay next quarter.

(1) Halving and doubling are operations to which humans relate easily. We measure and record the *half-life* of radioactive substances, talk of the time it takes for a country to *double* its GDP, and routinely hear campaign rhetoric and promotional NGO material that talks of *doubling* and *halving* arbitrary indicators. The reason is not that 2 has any special mathematical or real-world significance (or plausibility, in the case of politicians and NGOs) but rather, that it is easy for people to (believe they) understand. It's a lot less exciting to make a campaign promise to multiply the number of tax breaks or subsidies or scholarships by e, even though $e > 2$.

(2) The second reason is perhaps more legitimate. In computer science algorithms, it is customary to use *divide-and-conquer strategies* that work by breaking a problem up into two roughly equal subproblems, and solving both of them separately. The amount of time and resources needed to solve problems using such strategies typically involves a logarithm to base 2, since that is the number of times you need to keep dividing the problem into two equal parts until you get to problems of size 1. Thus, logarithms to base 2 frequently pop up in measuring the time and space requirements of algorithms. Similarly, in psychology and the study of human cognition and information processing, logarithms to base 2 play a role if we hypothesize that humans perform complex tasks by using divide-and-conquer strategies.

(3) The third reason has to do with biology, more specifically with the reproduction strategies of some unicellular organisms. These organisms divide into two organisms. This form of asexual reproduction is termed *binary fission*. Other reproduction strategies or behaviors may also be associated with logarithms to base 2 or 3, because of the discrete nature of numbers of offspring and number of parents.

One way of thinking about this is that logarithms to base 2 are more natural when working with *finite, discrete problems* while logarithms to base e are more common when dealing with *continuous processes*.

3. Applications to differentiation and integration

3.1. **Differentiating functions with variables in the exponent.** We are now in a position to discuss the general procedure for differentiation the function $f(x)^{g(x)}$, where f is a positive-valued function and g is a real-valued function.

To differentiate, note that:

$$f(x)^{g(x)} = \exp(g(x)\ln(f(x)))$$

We use logarithmic differentation and simplify to get:

$$(f^g)'(x) = (f^g)(x)\frac{d}{dx}[g(x)\ln(f(x))]$$

This can further be simplified using the product rule:

$$(f^g)'(x) = (f(x))^{g(x)} \left[g'(x)\ln(f(x)) + \frac{g(x)f'(x)}{f(x)} \right]$$

Two special cases are worth noting:

(1) The case where $g(x)$ is a constant function with value r. In this case, the derivative just becomes $rf(x)^{r-1}f'(x)$. An even further special case is where $g(x) := r$ and $f(x) := x$, in which case we obtain rx^{r-1}. This is the familiar rule for differentiating power functions that we saw, but now, we have established this rule for *all real exponents*, not just for the rational ones.

(2) The case where $f(x)$ is a constant function with value a. In this case, the derivative is $a^{g(x)}g'(x)\ln a$. In the further special case where $g(x) = x$, we obtain $a^x \ln a$. In other words, the derivative of a^x with respect to x is $a^x \ln a$.

3.2. **Key special case formula summary.**

$$\begin{aligned}
\frac{d}{dx}(x^r) &= rx^{r-1} \\
\frac{d}{dx}(a^x) &= a^x \ln a \\
\int x^r \, dx &= \left\{ \begin{array}{ll} x^{r+1}/(r+1) + C, & r \neq -1 \\ \ln|x| + C, & r = -1 \end{array} \right. \\
\int a^x \, dx &= \left\{ \begin{array}{ll} \frac{a^x}{\ln a} + C, & a \neq 1 \\ x + C, & a = 1 \end{array} \right.
\end{aligned}$$

Note: We don't need to put the $|x|$ in the ln antiderivative if the exponent is irrational because x^r isn't even defined for irrational exponents, so $\ln x + C$ is a valid answer for irrational exponents.

3.3. **Differentiating logarithms where both pieces are functions.** Consider:

$$\frac{d}{dx}\left[\log_{f(x)}(g(x))\right]$$

To carry out this differentiate, we first rewrite the logarithm as a quotient of natural logarithms:

$$\frac{d}{dx}\left[\frac{\ln(g(x))}{\ln(f(x))}\right]$$

Note that the base of the logarithm has its ln in the denominator.

We now use the quotient rule and get:

$$\frac{\ln(f(x))g'(x)/g(x) - \ln(g(x))f'(x)/f(x)}{(\ln(f(x)))^2}$$

4. Fun miscellanea

4.1. **Dimension and sense of proportion.** [This is optional – will cover only if I have time.]

When you double the lengths, the areas become four times their original value, and the volumes become eight times their original value. More generally, when you scale lengths by a factor of λ, the areas get scaled by a factor of λ^2, and the volumes gets scaled by a factor of λ^3.

The *exponent* of 2 occurs because area is two-dimensional and the exponent of 3 occurs because volume is three-dimensional.

Suppose there is a certain physical quantity such that, when we scale lengths by a factor of $\lambda \neq 1$, that quantity scales by a factor of μ. What is the dimension of that quantity? It is the value d such that $\lambda^d = \mu$. With our new understanding of logarithms, we can write:

$$d := \log_\lambda(\mu) = \frac{\ln \mu}{\ln \lambda}$$

When we put it this bleakly, it does not seem a foregone conclusion that d must be a positive integer. In fact, there is a whole range of physical objects that have positive measure in *fractal dimension*, i.e., there are quantities that we associate with them whose dimension is not an integer. For instance, there are certain sets such as Cantor sets that we design in such a way that when we triple the lengths, the size of those objects *doubles*. Thus, the dimension of such an object is:

$$\frac{\ln 2}{\ln 3} \approx \frac{0.7}{1.1} \approx 0.63$$

Similarly, there are sets with the property that when you double lengths they increase by a factor of three times. The dimension of such sets is:

$$\frac{\ln 3}{\ln 2} \approx \frac{1.1}{0.7} \approx 1.57$$