Math 137 Physics Based Section

Marek Stastna

1 PROLOGUE

"Calculus is a machine of the mind." That's how my own University career began, in the fall of 1991, taking Math 147 with Frank Zorzitto as a professor. Much has changed since that day, mostly having to do with the ubiquity of computers and all the changes in thinking that goes along with that. One thing that has not changed is my firm belief in the importance of calculus as a truly awesome machine of the mind. Of course one doesn't build an F1 car to drive it to the corner store at 40 clicks, and in the same way it is through its application that calculus truly comes to life. This course and the accompanying set of notes is an attempt to put calculus into the context of its oldest field of application, namely physics. While this is a natural project to attempt, it is not without risks. As my PhD advisor, Kevin Lamb, once said to me in a moment of exasperation: "The field is 400 years old and we still don't know how to teach it."

While it is not true that we do not know how to teach calculus, it is certainly true that professors who teach the upper year courses in disciplines that have traditionally been called 'applied math' or 'theoretical physics' often find themselves frustrated with the background that students come in with. This often has less to do with the best efforts of the students involved, and more to do with the structure of the calculus sequence. Not really surprising when you consider that a desktop Mac G5 of the 2004 vintage dwarfs the power of a supercomputer of the 1978 vintage.

Your involvement in the development of the course is crucial. For one, it helps you develop a strong speaking voice; something that I feel should be part and parcel of any University experience, and that will benefit you in ways ranging from the purely practical to the intensely personal. I once asked Mike Watkins, who was the project head for the Mars rover programme at NASA, how he got so good at public speaking. He laughed and told me that they make him practice 10 hours a week in front of cameras that record every dumb gesture of his hands and wooden turn of phrase. The second reason you should be actively involved in pointing out parts of the course that require improvement, is that any course is ultimately about teaching students, and not pleasing the professor. This means the improvements you make keep the course evolving and this is not only positive, but perhaps even essential in a world that changes with light speed. At the very least, read the notes. I have tried to write them in a descriptive way and the level of discussion in them is the level at which the tests and exams will be set. Try to explain the ideas to yourself or your friends, and make it a

personal goal to gain confidence in the way you both write and speak about calculus as the term goes on.

Logistically this section is equivalent to other sections of Math 137. We share the textbook (though we will depend on the written notes more than other sections), the test dates, the material covered (albeit with differences in order), and the standard of testing (if I can give you one piece of advice; do not worry about grades during the term, with sufficient work they have a way of taking care of themselves), and the assignment schedule. We will have **different** tests, assignments, and final exam. There are several logical points in time to transfer to different sections of the calculus stream. The first is within two weeks of the term beginning. While I hope not many of you will exercise this option, you should not feel that I am in any way constraining you from doing so. Transferring later in the term is possible, though obviously requires some catch-up. A third logical point of transfer is at the start of Math 138. Conversely, if you have friends who are in other sections, please tell them about our way doing calculus. The bewildering choices university offers have a way of clouding the range of what is actually possible, and many people who you think could enjoy the way of doing calculus presented in these notes, actually would. As they say in showbiz, word of mouth is better than any positive review.

Finally, I must thank a few influences. Foremost, the influence of my mother's take on the teaching of mathematical analysis can be found everywhere in these notes, even if true mathematical rigour only makes the occasional appearance. Pino Tenti and my father, two stubborn defenders of physics, are responsible for any physical insight contained in these notes. John Wainwright's various course notes, though I never used them as a student, provided an inspiring economy of presentation. Kevin Lamb's influence goes far beyond the PhD dissertation I wrote under his guidance. Finally, Michael Spivak's Calculus text still serves in my mind as the greatest exposition of first year calculus available in the english language.

2 INTRODUCTION AND NUMBERS

Back when I was a UW mathie undergrad I used to get a big kick out of all the people who, upon finding out I was a math major, would say something like "You must really like numbers, eh" Most of my courses, if they had any numbers in them at all, used them sparingly and with a great deal of apology. I'm not sure if much has changed in this regard today. Still, numbers are at the core of what mathematics is, and if one wishes to couple mathematics with science , as we do, then it's really all about the numbers in the end. This introduction is in part a review of things you would want to know about numbers coming into a university calculus course. More importantly it is an introduction to the style of thinking I hope to get across to you throughout the course. We aim to **build up** a way to describe the world around us. Historically this description has found its best expression in the subject of physics, though today essentially all sciences benefit from some of the mathematics we are going to learn. You should note that this is a very different mode of thinking than the "decreed from on high" style of mathematics that you have likely been exposed to. This approach deserves some mention and a surprising amount of credit, and indeed we will speak about it later in the course, however for now we begin as if we had nothing but our wits about us.

One often hears "Math is hard", or "I hate math", yet "Numbers suck" is a much less common put down. Indeed numbers themselves go hand in hand with human development going back to well beyond recorded history. It doesn't take much imagination to think of our cave dwelling ancestors counting, or attempting to count, a vast herd of deer (one, two,... many). Counting involves numbers like 1, 2, 20 and so on. In today's mathematics these are called **natural** numbers. Probably the first true mathematical breakthrough was the discovery of the number zero to represent the absence of something. Once we have zero we can define adding, the first operation or **machine**. The machine will take in two natural numbers (say 3 and 2) and output one natural number (5 in the example). Today we denote the machine by two numbers separated by a + symbol with the output generally placed to the right of the symbol =. Of course a cave man adding the deer he or she caught today to those he or she caught yesterday would hardly use such a complicated idea. However, one thing worth noting that he or she would do, is **record** the number. This suggests numbers go hand in hand with writing and language, and indeed writing mathematical ideas down, often called **notation**, is a vital part of the act of learning and doing mathematics. Moreover, defining the "+ machine" carefully even one time, makes it possible to use it later with more confidence.

Why did we need the number 0? Well for any operation we should have the option of doing nothing. Moreover, a desirable property is **reversibility**, for example, I dress in the morning with the understanding that I can undress in the evening before going to bed. The caveman of the above paragraph hunts to get food that he or she eats, and in a more modern setting I do the same thing in the veggie aisle of the grocery store. So if we were to start with 5 and wish to get back to 3 we must take away two objects. This is the reverse of adding and is represented by the symbol − which operates much like the + symbol and follows the same

notation (write it out for yourself).

Now ask yourself what sort of natural numbers can I add together? The answer is of course, all of them and zero, too. Aha, but what about subtraction? Well, subtracting 3 from 5 gives 2, and 5 from 5 gives 0, but subtracting 5 from 3 does not give a natural number! What to do? Either the $-$ machine must come with restrictions (something like the second number is no bigger than the first), or we need to allow negative numbers. What are negative numbers? Well, you could say they are the results of the $-$ machine when the second number is bigger than the first (mathematicians pull this sort of stunt all the time). Alternatively you think of borrowing some lunch money (say five dollars for concreteness) from your best friend that you plan to pay back. This is an example of using mathematics to represent something in the real world. We hope to do a lot more of this as the course goes on, but the present example is a good start. You should ask yourself, "OK does this negative number thing make sense?" Well, adding -3 to 3 is the same as subtracting 3 from itself and hence gives 0. Adding -5 to -5 gives -10, but subtracting -3 from -5 requires a bit of thought. For positive numbers subtracting meant "taking away" so for negative numbers, which mean, we owe something, subtracting should mean we are getting something back. Thus subtracting a negative number is like adding a positive number of the same "size" as the negative number. In other words $-5-(-7) = -5+7 = 7-5 = 2$. I used a bunch of rules in the calculation (like $a+b = b+a$) and I don't want to make too much of a big deal of them, except to point out that just because I can do something on auto-pilot doesn't mean that the occasion won't arise when I should **question my assumptions** (in other words, was it OK to switch order?).

With a bit of time you can convince yourself that the set of **integers**, or all the natural numbers, zero, and all the natural numbers with a minus in front of them, makes both $+$ and $-$ well defined in the sense that given any two integers (call them "purple" and "cadillac", you don't need to use a, b, x or y all the time) can be fed into the $+$ machine and the $-$ machine to yield an integer in each case.

So that's what we got for our troubles. Proper definitions mean we can do something with complete impunity thereafter. In the case of adding and subtracting integers we also know that we can get back to where we started.

Incidentally we can define another machine called **absolute value**. This one takes in any integer and "takes away the minus" if there is one. In other words it tells you how far you are from zero, provided you don't care whether the number is positive or negative. The notation is $|3| = 3$ and $|-3| = 3$.

Now that we have the integers you can think of defining other useful operations. Adding a number together a certain number of times defines multiplication. On a computer, $4 \times 3 = 4+4+4 = 12$, however there is no guaranteed way to get back to where we started (try to come up with one). How do things go wrong? Well say you buy a pizza with three friends, where everyone wants to eat the same amount. Naturally, you divide the pizza into quarters and get your munch on. A quarter of a pizza, but a quarter of the integer 1 is meaningless

in our present discussion. So what do we do? Well we actually use the seemingly silly idea mentioned above: Define **fractions** as numbers that give back an integer when you multiply by the integer on the bottom, $4 \times \frac{1}{4} = 1$. Notice that the integer you get in the end does not need to be 1 (think of dividing two pizzas between three people, $3 \times \frac{2}{3} = 2$). Now, we'd better allow this new bunch of numbers to have negative integers on the top (something like 'If I owe my three best friends two pizzas, what do I owe each friend?'). If we consider this new set of numbers (the integers and all zero) than we must conclude that the new set is bigger than the set of integers (why is it bigger?), moreover both addition and multiplication are reversible (for example $4 \times 3 \times \frac{1}{3} = 4 \times \frac{3}{3} = 4 \times \frac{1}{1} = 4$ and $4 + 3 - 3 = 4$). The operation that undoes multiplication is called division and in general the operation that undoes what a given machine does is called the **inverse**.

So, a big success, right? Well hold on a minute, first of all we have not come up with rules of precedence, say what does $4 \times 2 + 1$ mean? Indeed you might want to use your knowledge of brackets and such to write out the machine form of $4 \times 2 + 1$ and $4 \times (2 + 1)$. Still, with some care the rules can be worked out. A bigger problem is that $\frac{1}{0}$ is tough to figure out; I mean one pizza split into zero pieces, what's that all about? In fact no useful way to deal with this problem has been found, so we make a new rule: **Division by zero is not defined.**

It turns out that the set of fractions has a name, it is called the **rational numbers**. If you think of a pizza again, it turns out that in principle you could split the pizza into as many pieces as you wish, and then put as many of these pieces together as you wanted. Thus the set of rational numbers is very large and can be used to accurately describe almost any object we wish to "turn into numbers".

To wrap up, we have done a brief walk through of numbers. Along the way we used numbers to do a variety of things (count, split pizza, and so on). The idea was, to accomplish a certain task, certain operations had to be defined (as carefully as we could), then we had figure out if an operation actually did what we wanted it to on **ALL** the numbers that we allowed as "input". On occasion we had to restrict what was allowed as input (for example no zero on the bottom for division). Finally we took some care in making the operations **reversible** (subtracting undoes adding and so on). This turned out to be quite complex (like for multiplication). In the final analysis, the test of a new concept is whether it is useful (useful itself has a broad range of meaning).

3 THE MEANING OF =

In the previous section we used the notation $\text{number}_1 + \text{number}_2$ = the sum of the two numbers. Here the symbol = was used merely as a way of indicating that to the left of the symbol we write the two things to be added and to the right we write the answer after the adding has been done. I say "things" and not numbers because we could just as easily write

something like

$$3x + 1 = y$$

for the two symbols x and y. You might recognize this expression as the equation for a line (of slope 3 with the y-intercept 1). However we have now used the symbol = in a much more powerful way. What the equation of a line says is that **given** a value of x you can **compute** the value of y. Moreover, **given** y a bit of rearranging lets you **compute** the value of x.

Now, lines are rather abstract things (can you recall the geometric definition of a line?) so perhaps we should consider something more concrete, say, Newton's Universal Law of Gravitation between two objects of mass m_1 and m_2 that are found a distance r apart

$$F = \frac{Gm_1m_2}{r^2} \tag{1}$$

where F is the force and G is a constant that must be determined from experiment. You have probably seen this formula in high school physics class (and I am actually taking a certain liberty when I write it the way I do, can you see how?). But what does it actually say?

Start with what it assumes you can measure. Well we could just weigh the two masses, right? Except that weight is not mass, so something is fishy here (resolve the apparent paradox). Even if we don't know the masses or the value of G we can change the distance between the two objects and measure the change in the force (you can look in your physics text to see if there are some clues as to how clever experimentalists find ways to do this). This should give us the r^2 term on the bottom of the right hand side. But why r^2 (which just means the distance multiplied by itself) and not something like $r^{2.2}$ (which we would have some trouble defining with what we know so far) or even $r^{2.0000002}$ (which is pretty much like r^2 except when r is either very big or very small)? As you can imagine, the exact value of the exponent is very important and there has been a big effort to determine it from experiment, with the upshot being that 2 is pretty much spot on.

Then there is the whole question of units. How do we measure mass, force, distance and so on? If we were to encounter aliens how could we convey the meaning of our units to them (can you think of a speed, whose value would be unchanged for aliens or humans?).

The above discussion surely "feels" very strange. Math is supposed to be full of certainty, and yet here we are having trouble getting at the exact meaning of a basic universal law! Of course, if we treat the law as a fact we can solve for r, or any other value of interest just by rearranging. However, the rearranging itself tells us nothing new about the physics of the situation. I don't mean to suggest that the rearranging cannot be helpful to an eventual solution, but only to remind you that it is the interpretation of the mathematics that contains the physics (or biology, chemistry or any other scientific field of interest).

Now "rearranging" is commonly known as solving. Sometimes this is easy and gives one answer, say solving,

$$3x + 7 = 16$$

to get

$$x = 3.$$

More often there is trickery to be used and more than one answer is possible, like solving

$$x^2 - 5x + 6 = 0$$

by factoring to get

$$(x - 3)(x - 2) = 0$$

and noticing that when two numbers multiply to give 0 then at least one is equal to zero, therefore both $x = 3$ or $x = 2$ are valid answers. Sometimes a solution can be shown not to exist, like

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

Here, our rules for multiplication tell us that any positive number times itself is positive and any negative number times itself is also positive, so no number squared could possibly equal -1.

As a general rule, only approximate solutions can be found (we'll see examples of this later), and this is usually done by using a computer. When this is the case an obvious thing we would like is the ability to make the "approximate" solution as close to the actual solution as we can. To make this notion precise will require a fair bit of work. Let's summarize

- The symbol = has two meanings
- Mathematically it means that it is possible to rearrange the expression and still give a valid relationship between the various symbols
- Physically, it means more
- For the physics "=" we need an experimental means to determine the various quantities to make the expression useful

Thus the way forward is clear. We need to increase the number of objects we know and to define as precisely as we can, operations on them, so that we can formulate descriptions of the world around us. When a sample description is set up, we can confront its predictions with measurement to find where the theory fails. This in turn leads to modified theories. Thus the process is not static, a point often missed by science textbooks which focus on science that is well understood enough to be put into book form. Note that this does not mean that science is a form of politics, in which different facts are true at different times. It is highly unlikely, for example, that the r^2 in the Universal Law of Gravitation will change to a $r^{2.1}$ anytime soon.

4 FUNCTIONS

Let's look back for a moment. In treating the + symbol as a machine we not only allowed the input of two numbers (say $3+8=11$) but symbols as well, for example $y=3x-5$ in the case of the equation for a line (Note the weirdness of math grammar, $3x-5=y$ means the same thing as $y=3x-5$, though in english 'eat fruit' and 'fruit eat' do not mean the same thing; unless you're Yoda that is). If we pick a value of x, say $x=2$ then the expression now gives $y=6-5=1$. Thus the expression has another meaning as a machine that takes in a single value for x and gives back a single value of y. This is undeniably useful, though it may help to look at it from another point of view. If we knew the constant G and the value of the two masses in question then the Universal Law of Gravitation reads

$$F(r)=\frac{number}{r^2}$$

and we may as well make our life easier by taking the number in the numerator to be 1. Then we have

$$F(r)=\frac{1}{r^2} \tag{2}$$

which is written in a highly suggestive notation. Again we are inputting one value (the distance between the two masses, r) and outputting one value, namely the force. For example if $r=2$ (we should have units, but we won't worry about that for now) then $F(r=2)=\frac{1}{4}$ (because $2^2=4$). Now think for a second, could we have more than one input? Well in some cases this seems sort of silly, say if we input the distance between the two masses and their colours, to make three inputs. Indeed, we do not expect that the force of gravity depends on colour (but if colour is due to electromagnetic waves, why don't we?). But if we had not specified the two masses then the Universal Law of Gravitation actually takes in **three** inputs, m_1, m_2 and r. The constant G is assumed **Universal** and thus does not change. In fact, if we considered particles with electric charge (like a positively charged proton and negatively charged electron) then unless we were careful (or lucky) then the *total* force is all we could expect to measure. Now we would have 5 inputs (check your physics text to find out why).

OK so much for inputs, what about outputs? Well let's say that we know the values of the masses and the value of G and then fix the value of the distance by moving the two masses to known positions (say one unit of length apart). How many outputs does the Universal Law of Gravitation give? Of course the answer is one. We are sure of this almost instinctively. It is part of our physical experience from very early on; my two year old pushes on a closed door and has full expectation that the door will not move. Indeed it would be very difficult to predict a whole lot of anything if the laws of physics (and chemistry and biology) yielded anything but a single output for each set of inputs. Note that an algorithm, on the other hand, can be quite useful and return multiple outputs.

An argument as useful as that given above must surely have captivated mathematicians. Indeed, for the purposes of our course we define a function as a machine (some might say

mapping or operator) which acts on a certain set of objects (called the **domain**) and **for each input the machine produces a single output**. The set of outputs is called the **range**). It is common to write the input as x and the output as $f(x)$, though it goes without saying that this is not necessary.

Example 1: $f(x) = x$ for any set of inputs just gives back the set of inputs and is sometimes referred to as the **identity mapping**.

Example 2: $f(x) = x^2$ for the domain of all integers returns some of the positive integers and 0. It does not return any negative integers and in fact does not return a lot of the positive ones either (can you list a few examples?)

In some sense the set of outputs for the second example is smaller than the set of inputs. You might be wondering if it is possible to have functions that have a larger set of outputs than inputs. This is impossible and you should go through the argument as to why this is so.

Example 3: $f(x) = x^2$ on all fractions between 0 and 2. It seems intuitive that the set of outputs is all the fractions between 0 and 4. However to show this with certainty would take a fair bit of work (try it).

Example 4: $f(x) = x^4$ on all integers. There isn't much new here, except that we could consider the original input to be x, feed it into the machine $g(x) = x^2$ and get the output $y = x^2$. Now take y as the input into $g(.)$ again to get the output $z = y^2 = x^4$. Thus $f(x) = x^4$ can be thought of as a **composition** or repeated application of the squaring function $f(x) = x^2$.

If you work on the challenge from **Example 3** you will use the idea of the square root from high school (for example $\sqrt{\frac{1}{4}} = \frac{1}{2}$). But this brings up two questions

1. What is $\sqrt{2}$? Is it a fraction? If it is not what kind of number is it?
2. It seems that $\sqrt{x^2} = x$ so thus the square root undoes the action of $f(x) = x^2$. Can we undo the action of any function?

Both questions are worth answering and we will pursue them at length later.

Now back to the physics. In the Universal Law of Gravitation, how do we get the output from the input? Simple right, you give me r and I multiply it by itself to get r^2 then I take the numerator (we took this to be 1) and divide this value by r^2 to get the final answer. For example if $r = \frac{3}{4}$ then $r^2 = \frac{9}{16}$ and the final answer is

$$F(r = \frac{3}{4}) = \frac{16}{9}.$$

So the news is good, we input a fraction and after some simple operations built on **multiplication** (which is just repeated addition) and **division** (which is more complicated but we've covered that already) we get our answer. Even better, the answer is a fraction as well. On second thought, it seems like a minor miracle to me that the granddaddy of all physical laws comes down to such simple operations!

Example 5 The idea of functions as machines which accept certain inputs and produce certain outputs can be used to organize much of the material of calculus (as we will hopefully see). However, its true glory comes when we consider more complicated physical situations where the number of inputs is more than one (the course MATH 237 considers this topic in detail). As a preview consider the well-known IDEAL GAS LAW,

$$PV = nRT$$

where P denotes the pressure, V the volume, T the temperature and n and R two physical constants which we assume are externally given (in a manual, for example). We can solve for any one of the three variables P, V and T, in terms of the other two. For example

$$P(V,T) = \frac{nRT}{V}.$$

Thus P is a machine that takes in **two** inputs and produces **one** output. For fixed V, P is directly proportional to T, so if we heat a container of fixed size the pressure of the gas inside increases. On the other hand if we fix the temperature, and decrease the size of the container (and hence V) we again raise the pressure.

5 BUILDING FUNCTIONS

So far we have defined what a function is and even built up a few examples. In fact you can convince yourself that all the functions we can build up as of now are of two types. The first form looks like $5x$, $6x^3 + 2x - 5$, $17x^{100} - 64x^{57} + x^5 - x + 1$ and so on. Functions like this are called polynomials and they can be written in the somewhat fancier (though no more revealing) sigma notation as

$$p(x) = \Sigma_{n=0}^{n=N} a_n x^n$$

where N gives you the highest power (1, 3 and 100 in the examples above) and 0 is the lowest power because of the rule $x^0 = 1$. The various a_ns are called coefficients. The notation $p(x)$ is chosen so that the letter p makes you think of, wait for it, polynomial.

Now you can convince yourself that if we have two polynomials, call them $p_1(x)$ and $p_2(x)$ then multiplying the two together is still a polynomial. The second type of function we can think of is called a rational function and takes the form of one polynomial divided by another polynomial, or

$$r(x) = \frac{p_{top}(x)}{p_{bottom}(x)}.$$

We need to be careful though because division has an exclusion rule, namely "can't divide by zero". So we can say that $r(x)$ should only allow for inputs those x values for which $p_{bottom}(x) \neq 0$. For example if

$$r_1(x) = \frac{x^2 + 2}{x^2 + x}$$

then $x = 0$ and $x = -1$ are not valid inputs, or put another way cannot be in the domain of $r(x)$.

Since we can take some very high powers it would seem we can reproduce almost any behaviour we measure in the lab, with a polynomial. We will come back to this idea that any function meeting certain mathematical requirements measured for a finite time (wait you can't measure for infinite time, so I guess that is any function) can be approximated by a polynomial.

In practice this may not be the best way to describe many different types of physical behaviour. Think of a heartbeat or a pendulum swinging. It seems intuitive that heartbeats and pendulums **repeat** themselves nearly exactly and thus the function to describe them should repeat as well. But polynomials, by definition can repeat themselves only a certain number of times (can you explain why?).

The whole concept of endless repetition or **periodicity** has captivated mankind for a long time. Indeed the circle is a profound symbol to many outside of mathematics. A circle has no end an no beginning. In that sense it is no different from a line. However, while a line goes out to forever (an ill defined idea at the best of times) a circle is right there in front of us; easy to draw (how would you do it without only a pencil and a string?) and measure. Indeed measuring a circle leads to the interesting observation that the distance around the circle, called the **circumference**, is a linear function of the **radius**,

$$C(r) = kr$$

for some special number k. What is the number k? Careful measurements indicate that $k = 2\pi$ (so $C(1) = 2\pi$ for example) where the number π, which is approximately 3.14159265, actually cannot be written as a fraction.

Leaving π alone for a minute, consider a circle of radius one centered at the origin of a right handed set of Cartesian coordinates. Kind of a mouthful, but really we just have two perpendicular axes where the x axis runs to the right and the y axis runs upward. In these coordinates a circle is written by the relation

$$x^2 + y^2 = 1^2.$$

Note that this is **NOT** a function (why not?). Now consider a little cart or bug that starts at the point $(1, 0)$ and begins moving upward along the circle. If we consider its coordinates as functions of the **distance travelled along the circle** then both the x and y coordinates (call them $c(\theta)$ and $s(\theta)$) repeat themselves after the cart has travelled a distance of 2π. I'm messing around a little bit here, because we usually call θ an **angle** (measured in *radians*)

and the two functions are $x = \cos(\theta)$ and $y = \sin(\theta)$. So angles are really a way to measure *distance*, something that you would not see if you used the Babylonian religious symbols called degrees.

Now since $x^2 + y^2 = 1$ we know that $(\sin(\theta))^2 + (\cos(\theta))^2 = 1$. This fact is certainly worth memorizing, but if one remembers where sin and cos come from there is no need to memorize anything.

The two periodic functions can be combined in various ways and indeed you will see this on your assignments. However for now we will define

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

This function will also be periodic, but it is also undefined whenever $\cos(\theta) = 0$ (at $\frac{\pi}{2}$ and $-\frac{\pi}{2}$ for example). The function gets very large in absolute value near these points.

Incidentally the proper definition of a periodic function goes something like this:
DEFINITION: A function $f(x)$ is periodic with period L if for any integer n and any valid input x, $f(x + nL) = f(x)$.

Thus we have further enriched our mathematical language with the ability to describe at least some periodic phenomena. Of course the nicely behaved sin and cosine functions may describe a pendulum pretty well, but can't possibly describe something as complex as a heartbeat.

Note however that there is nothing stopping us from linking our functions together in a string of inputs and outputs, something like

$$f(x) = p(\sin(x))$$

for a polynomial p, where we take x, feed it into the sine machine then take the output and feed that into a polynomial function, $p(.)$, to get another output. While it will take a couple of years before you see why, these polynomials made up of sines and cosines let you describe almost any measurement in a compact and succinct way. Here's a little sneak preview.

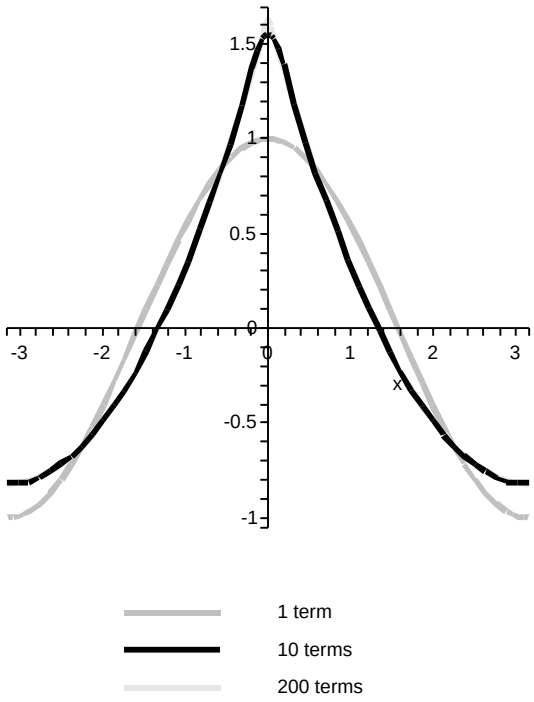
Consider $s_1(x) = \cos(x)$, $s_2 = s_1 + \frac{1}{4}\cos(2x)$, $s_3(x) = s_2 + \frac{1}{9}\cos(3x)$ and so on. Pretty simple pattern, right? You might want to check that the general formula can be written recursively as

$$s_n(x) = s_{n-1} + \frac{1}{n^2}\cos nx. \tag{3}$$

Of course we need to also know that $s_1(x) = \cos(x)$, but once we know this, we can build $s_n(x)$ for any integer n. You can also use sigma notation to get

$$s_n(x) = \sum_{k=1}^{k=n} \frac{1}{k^2}\cos(kx). \tag{4}$$

This would be a big waste of time if it wasn't for the result shown in the following picture.



Here I have plotted $s_1(x)$, $s_{10}(x)$ and $s_{200}(x)$ for $-\pi < x < \pi$. You can see that $s_1(x)$ is just $\cos x$, but that s_{10} and s_{200} are essentially the same. This means that if I had a function that was kind of pointy near $x = 0$ but round near $x = \pm\pi$ then instead of finding some new form for this function I could just use $s_{200}(x)$ as an **approximation** (actually most of the time I could get away with $s_{10}(x)$. All it takes is a recipe for choosing the numbers in front of $\cos(x)$, $\cos(2x)$, $\cos(3x)$ and so on.

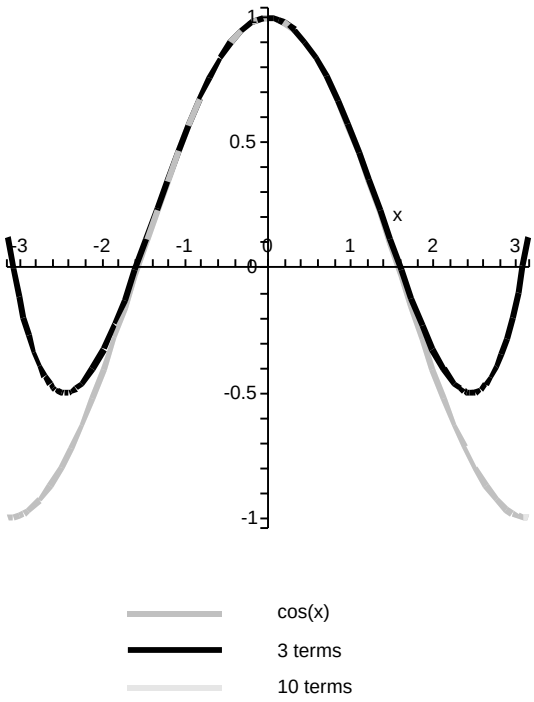
You might be thinking this whole business of $p(\cos(x))$ is too complicated, why not just use polynomials $p(x)$ to approximate? Toward this end consider the following example that you will be able to understand in full by the end of first year. I claim that for x pretty close to zero (can you make this more precise?) that I can approximate $\cos(x)$ by the polynomial

$$p_n(x) = \sum_{k=0}^{k=n} (-1)^k \frac{x^{2k}}{(2k)!} \tag{5}$$

where you should try to recall that $anything^0 = 1$ and that $k! = k \times (k-1) \times (k-2)...2 \times 1$ for all positive integers. Also, by default $0! = 1$. It might help to write a few terms out,

$$\begin{aligned} p_0(x) &= 1 \\ p_1(x) &= 1 - \frac{x^2}{2} \\ p_2(x) &= 1 - \frac{x^2}{2} + \frac{x^4}{24} \end{aligned} \tag{6}$$

and so on. You should notice that $p_0(x)$ has one term, $p_2(x)$ has three terms, and in general $p_n(x)$ has $n+1$ terms. OK onto the picture,



here I plot the target function $\cos x$ as well as the two approximations for $-\pi < x < \pi$. I label with the number of terms for the approximation. You can see that $p_2(x)$ is a bad approximation over all of $-\pi < x < \pi$. However if we only cared about $-1 < x < 1$ then it is actually quite a good approximation! Furthermore, $p_9(x)$ is a very good approximation on all of $-\pi < x < \pi$. So the question becomes:

If I know I want the approximation to be a polynomial, how do I find the numbers out front (the coefficients)?

The answer will have to wait a bit. Still the results are impressive. How so? Well, pick any natural phenomenon you want to look at, and imagine writing down a function to describe it. Given that we only know polynomials and sines and cosines, so far, this seems impossible. But now ask a more sophisticated question: "What do I really want to know, and how **precise** of a knowledge do I require?" Perhaps I don't care that neither the Earth or the Sun are perfectly spherical if I simply wish to compute the orbit of the Earth around the Sun. Perhaps an *approximation* may do the trick (we take both the Earth and Sun to be point masses with gravity the only force and off to Newton-land we go). Well, if an approximation will do the trick then there is no reason to search out all manner of exotic functions when an approximation can be achieved with polynomials or polynomials of sinusoids!

6 GRAPHING

In the last section I showed two graphs that, in my opinion, made the point of‘ “what is a good approximation?” rather well. Indeed experience suggests that the eye is a very powerful and even precise instrument. Thus, we should look at the relationship between our machines called inputs and their graphs. To draw graphs we employ the idea of Descartes (17th century, french) that the value of the input variable will be represented by the distance along one axis, and the value of the output will be represented by the distance along another axis that is perpendicular (mathematicians sometimes say orthogonal) to the first. In practice the input (or x axis) is usually taken to be horizontal, with increasing values to the right, while the output axis (or y axis) is usually taken to be vertical, with increasing values taken to point upward. You should note that both positive and negative values can be represented if we follow the convention that positive means distance to the right (up) and negative means distance to the left (down) for x (y).

To a pure mathematician a graph is a set of ordered pairs $(x, f(x))$ where each x is drawn from the domain of the function $f(.)$. We'll tend to be more visual than that and so we'll be concerned with the question:

Given a function $f(x)$ how do we draw its graph?

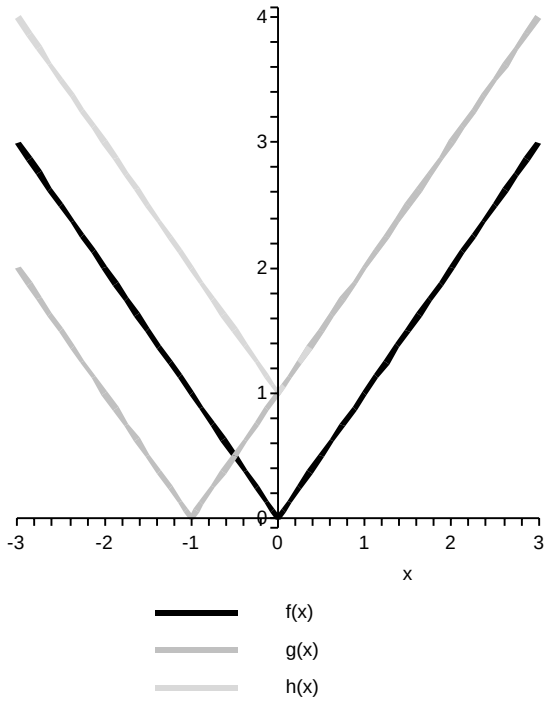
Following our philosophy of building up from what we know, we'll start with a line. Given the linear function $L(x) = 3x - 2$ (what does 'linear function' mean, anyway?) we recall that the slope of the line is given by the coefficient of x; three in this case, and when the input is $x = 0$ then the output is $L(0) = -2$; or that the y-intercept is -2. Now the graph is easy to draw.

Example 1: $f(x) = |x|$. This function is nearly a line, the only thing is that the slope can take two values since $f(x) = x$ when $x > 0$ and $f(x) = -x$ when $x < 0$. Of course, $f(0) = 0$. Now you can draw the function easily. The point $(0, 0)$ where the slope jumps is called a **cusp**.

Example 2: $h(x) = |x| + 1$. All I have done is to demand that we add one to the output. This means the **shape** of the graph is unchanged, but it does have to be translated (shifted, in everyday words). Noting $h(0) = 1$ we can draw the graph.

Example 3: $g(x) = |x+1|$. You can think of this function as a two step process, first take x as an input, add one to it to get $z = x+1$ then feed z into $f(.)$ (If I want to write the machine without a particular input I will use a period to fill the space where the input should go). You already know the graph of $f(.)$ so the only question is what effect the "preprocessing step" in which x gets mapped to (turned into, in everyday words) $x + 1$ has. Well, $f(0) = 0$ means $g(-1) = f(-1 + 1) = f(0) = 0$, so the point $(-1, 0)$ is the cusp. Also $f(z) = z$ when $z > 0$ means $g(x) = x+1$ when $x+1 > 0$ or $x > -1$. Similarly, $g(x) = -x-1$ when $x < -1$. This means the graph has the same shape as the graph of $f(x)$ but is shifted horizontally.

Here is a graph of all three functions.

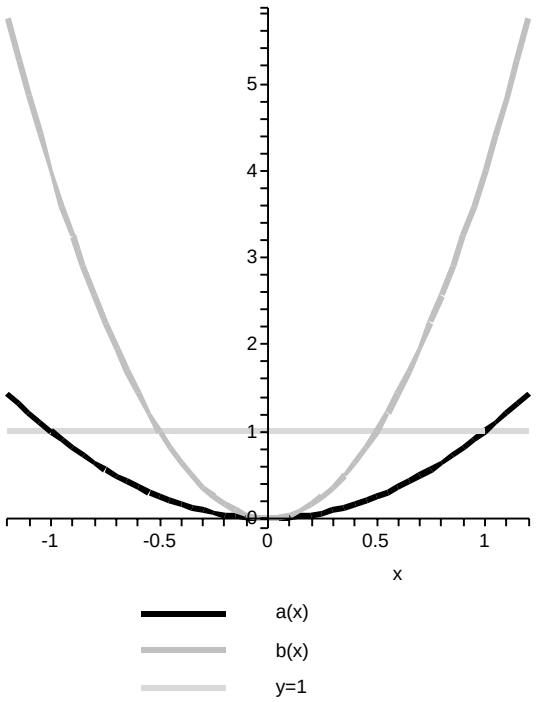


We used $f(x) = |x|$, a function whose graph we knew, but even if you didn't know the function and only knew its graph (say from a bunch of lab data you collected) then the above tells you that:

- **Given the graph of $f(x)$, the graph of $h(x) = f(x) + C_1$ is just the same graph shifted up by C_1**
- **Given the graph of $f(x)$, the graph of $g(x) = f(x + C_2)$ is just the same graph shifted to the *left* by C_2**

Example 4: $a(x) = -x^2$. Everyone knows this is a parabola opening downwards with the top found at $(0, 0)$. The real question is **how do you know it?**. Well one way is to do what the computer does; take lots and lots of input values and for each compute output values, plot each ordered pair and connect the dots freehand. When you do this (say you restrict yourself to $-2 < x < 2$) you'll find that out near the ends you are essentially drawing straight lines, however near $x = 0$ you are curving more and more. Furthermore, if you draw from left to right, then the curve you are drawing increases (goes up) until you hit $x = 0$ and thereafter decreases (goes down). **Is there a way to characterize both the increasing-decreasing and curviness properties of a function mathematically?** This will be one of our major goals in the first half of this course.

Example 5 $b(x) = 4x^2$. Notice, $b(x) = a(2x)$ since $2^2 = 4$. Surely the graphs must somehow be related? For $a(x)$ we know $x = \pm 1$ are special points since $a(\pm 1) = 1$. For $b(x) = a(2x)$ the input $x = \pm\frac{1}{2}$ will be "preprocessed" before being fed into the $a(.)$ machine by being multiplied by 2. This means the input into the $a(.)$ machine will be ± 1. You can follow this argument through for other points and you will conclude that the graph of $b(x)$ will be a squeezed together version of the graph of $a(x)$. Here is a picture:



Once again, we used something we knew, namely parabolas, but in general:

- **Given the graph of $f(x)$, the graph of $f(\alpha x)$ will be a horizontally squeezed (if $\alpha > 1$) or stretched (if $\alpha < 1$) version of the graph of $f(x)$.**

There are other rules, some of which you will get practice with on the assignment, however this should get you started when it comes to graphing.

7 INVERTIBILITY

If I gave you the equation of a line, say $y = 3x - 1$ and asked for you to rearrange to get x as a function of y you'd do something like

$$y = 3x - 1$$

$$y+1 = 3x \tag{7}$$
$$x = \frac{y+1}{3}. \tag{8}$$

In fact, you would probably point out to me that you could do it for any line, except for those one in which x never appears (horizontal lines). Then I would ask you, "what's so special about horizontal lines?" After a bit of thought you'd likely come back to me with something like: If you think of the equation as a recipe then the equation for a horizontal line just says y equals some number, **no matter** what x is, so there's no way you can think of assigning any **one** y to any **one** x.

The idea of rearranging a formula to get the y in terms of the x may seem simple, but in fact it can run rather deep. Let's reconsider the Universal Law of Gravitation which tells us the force of attraction between two masses (m_1 and m_2 separated by a distance r) with the bunch of constants in the numerator set to one, namely

$$F(r) = \frac{1}{r^2}.$$

Usually you think of measuring the distance, r, and using it to determine F, but what if you did measure F, what sort of information would this tell you? Mechanically you could just rearrange to get

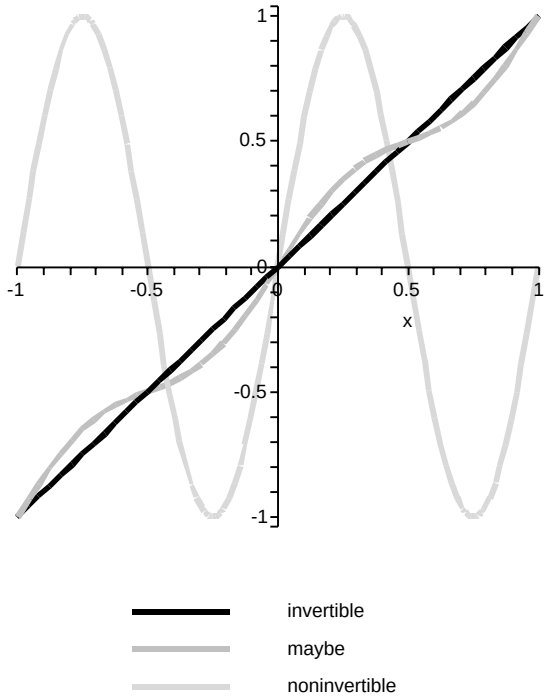
$$r^2 = \frac{1}{F}.$$

In fact you could even write $g(F) = r^2 = \frac{1}{F}$. You could subsequently let $h(g) = \sqrt{g}$ and claim, quite rightly, that since the square root is always positive, you have recovered the distance between the particles (you could even label it by the physics textbook r) having measured the force F. Much more interesting is the question "What couldn't I actually solve for?". For concreteness, let's say the positions of our particles can be represented by points in a set of right-handed Cartesian axes (sounds fancy but don't get fooled), then no generality is lost (think about what this means) if the first particle (mass one) is at $(0, 0)$ and the second (mass two) at $(0, r)$. Aha, you say, why not mass one at $(0, r)$ and mass two at $(0, 0)$, or how about mass one at $(0, 0)$ and mass two at $(0, -r)$. At this point I am, of course, forced to admit that **from a measurement of force I can make no conclusion about the actual position of the two particles, only the distance between the two**.

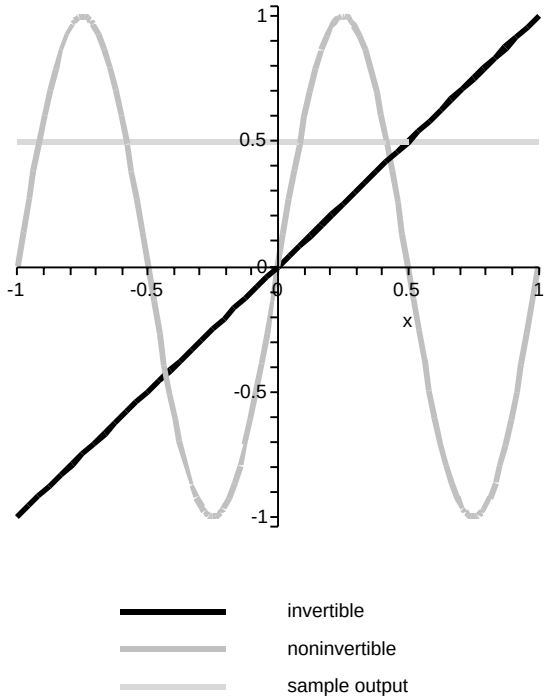
So now, what would a mathematician say about all of this?

1. If the set of inputs, r, into $F(r)$ is allowed to be both positive and negative then $F(r)$ cannot be inverted to give $r(F)$.
2. If r is restricted to be non-negative (which is natural for a physical quantity like distance) then $F(r)$ can be inverted to give $r = g(F)$.

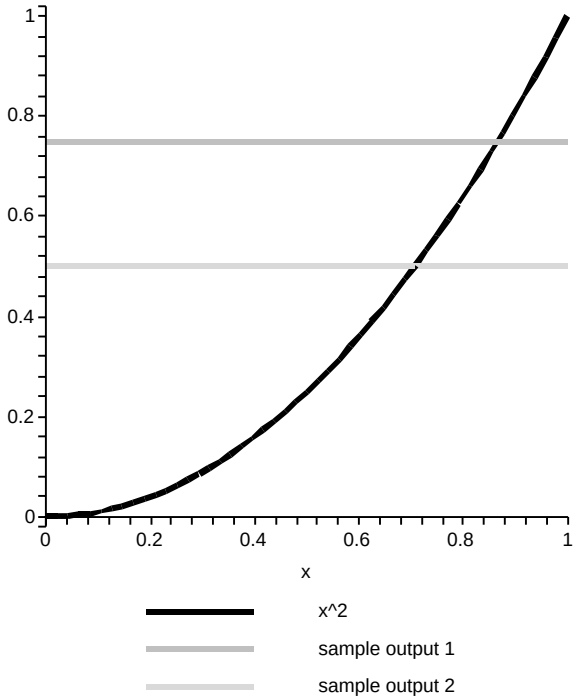
A couple of pictures might help, and here is the first:



The picture shows three functions. The one that is absolutely, for sure invertible is $f(x) = x$. Why do I know it is invertible? Well, being the so-called **identity function** it doesn't actually do anything, so it is its own inverse (can any other function be its own inverse?). We'll leave a discussion of 'maybe' for later. The function that is not invertible is $\sin(2\pi x)$ (what effect does the 2π multiplier have?) which repeats itself twice on the interval $-1 < x < 1$. The reason you can't go back from the output back to the input is because if you fix a particular output (say $y = 0.5$) you wouldn't know which of the four possible inputs you should go back to. The second picture may help to make the point:



As you can see, given the output $y = 0.5$ the identity function readily yields the input as $x = 0.5$. Now let's consider a function we can invert but one for which the inverse is not so easy to calculate. If $f(x) = x^2$ and $x \leq 0$ then the inverse is easily written down as $x = \sqrt{(}y)$. Of course square roots are not always easy to calculate ($\sqrt{2}$ is not even a rational number), but in a pinch we could use the graph of $f(x) = x^2$. Here is the picture:



I plot $y = x^2$ for $0 < x < 1$ as well as two horizontal lines at $y = 0.5$ and $y = 0.75$. These fix the outputs for $f(x)$. But this is the same as saying as, they fix the **input** for the inverse, or $g(y)$. To get an estimate we would just draw a perpendicular line from the points where the horizontal line of our choice and the graph meet. Where this new **vertical** line meets the x-axis gives the value of the inverse. Eye-balling it I guess $\sqrt{0.5} \approx 0.7$ which is not that far from the value your calculator will give.

Finally, what makes it possible to invert? It is the fact that in some cases not only is there one output produced for one input (the definition of a function), but that each output is produced only once. A function which produces each member of the range exactly once is called **1-1**.

If a function is 1-1 then it is invertible. Though there is no guarantee that a formula for the inverse can be written down.

If a function $f(x)$ is invertible then its inverse is commonly written as $f^{-1}(.)$ and we know that:

1. $f^{-1}(f(x)) = x$
2. $f(f^{-1}(y)) = y$

for any x in the domain of f and y in the range of f.

Example 1: $f(x) = x^2$ on $x \leq 0$. The inverse exists because for larger inputs $f(x)$ produces larger values (this property makes $f(.)$ an **increasing function**) and this in turn guarantess that $f(.)$ is 1-1. Moreover we can write down the inverse, as $f^{-1}(y) = -\sqrt{y}$ and if need be, resort to using the graph to compute it.

Example 2: $g(x) = -5x + 3$ all real numbers x. This is a linear function with a negative slope, and is thus decreasing and 1-1. The inverse thus exists and in this case can be written down in terms of elementary operations as $x = g^{-1}(y) = \frac{3-y}{5}$.

Example 3: $h(x) = \sin x$ for a subset of the real numbers for which this function is 1-1. Using the circle definition, one choice is $-\frac{\pi}{2} \leq x \leq \frac{\pi}{2}$ (can you find another?). For this domain the sine function is invertible and can formally be written down as $h^{-1}(y) = \arcsin y$. Notice that the domain of the inverse is the range of the original, in this case $-1 \leq y \leq 1$.

Finally, we note that it would be very handy to have some simple test to make sure a function is increasing or decreasing, especially for messy ones like the "maybe" in the first picture. For a preview of the answer, imagine making an approximation of a graph by a large number of line segments. On their own each line segment is 1-1 (and hence invertible) provided it is not horizontal, or provided that it does not have zero slope. As a further case notice that if some of the line segments have positive slope and some have negative slope the graph must reverse direction, and hence the function cannot be invertible.

8 RATE OF CHANGE *(NOT ON TEST)*

You probably learned in particle mechanics that provided the velocity is constant (recall that velocity is a *vector* and to avoid talking about vectors we will just line up our choice of x-axis with the velocity vector) then the displacement (again a vector) is just $\vec{d} = \vec{v}t$. For one dimensional motion we can skip the vector symbols and just write $d = vt$, where velocity is now just a scalar. Now consider a situation in which the velocity is held at one value, call it v_1, for a time t_1, then changed instantly (is this realistic?) to a value v_2 that is held for a time t_2. What is the total distance travelled?

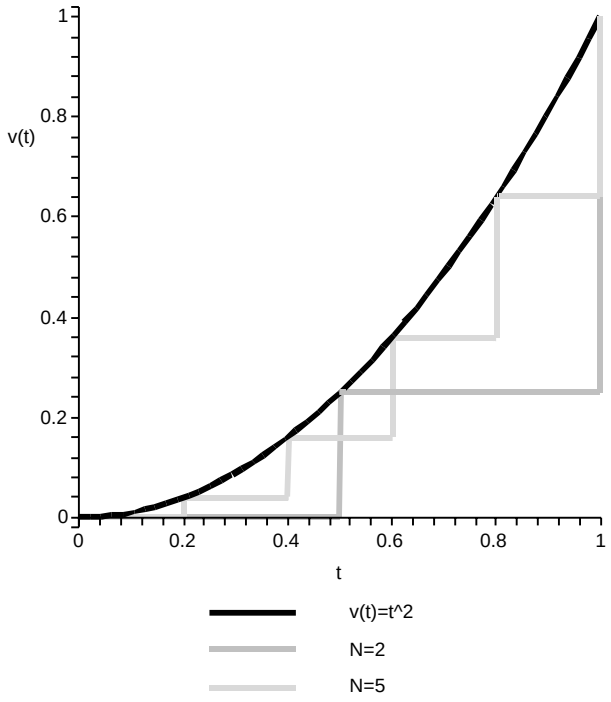
For the first time segment it is just $d_1 = v_1 t_1$ and for the second just $d_2 = v_2 t_2$, hence in total, $d = d_1 + d_2$ which we will write in a funny, but suggestive, way as

$$d = \sum_{k=1}^{k=N} v_k t_k, \tag{9}$$

where for the simple example $N = 2$. The formula (9) however suggests, why stop at $N = 2$. Indeed we could think of splitting up our travel into lots of little segments during which the velocity is constant. What would we accomplish by this?

First of all, as I hinted at above, it is probably true that the velocity can't just change

instantly (no 0 to 60 in 0 seconds). If the total travel time is fixed (let's call it T) and the length of all the segments is equal, then the length of a segment is T/N. Now the velocity can possibly change instantly $N-1$ times. This may not be realistic, but at least with more segments a smooth curve can be "approximated" better. Consider the following picture:



Here the velocity is given by the quadratic function of time $v(t) = t^2$, $T = 1$ and I show you the approximations with $N = 2$ and $N = 5$. Consider $N = 2$ first. To get v_1 I evaluate $v(t)$ at $t = 0$, and to get v_2, I evaluate $v(t)$ at $t = \frac{T}{N} = 0.5$. Thus $v_1 = 0$ and $v_2 = 0.25$. The estimate of the distance travelled is thus $d = 0.25 \times 0.5 = 0.125$.

With $N = 5$ the time segments have a length of $\frac{T}{N} = 0.2$, and I follow the same algorithm, evaluating $v(t)$ at $t = 0$, 0.2, 0.4, 0.6 and 0.8. The estimate is now $d = (0.2^2 + 0.4^2 + 0.6^2 + 0.8^2) * 0.2$ (how come I could factor out the time segment length?), or $d = 0.24$. This is much closer to the actual distance travelled (trust me for now) of $d = \frac{1}{3}$.

With $N = 10$ we get the even better estimate of $d = 0.285$, and with $N = 100$ we improve to $d = 0.32835$ and with $N = 1000$ to $d = 0.3328335$, or two digit accuracy.

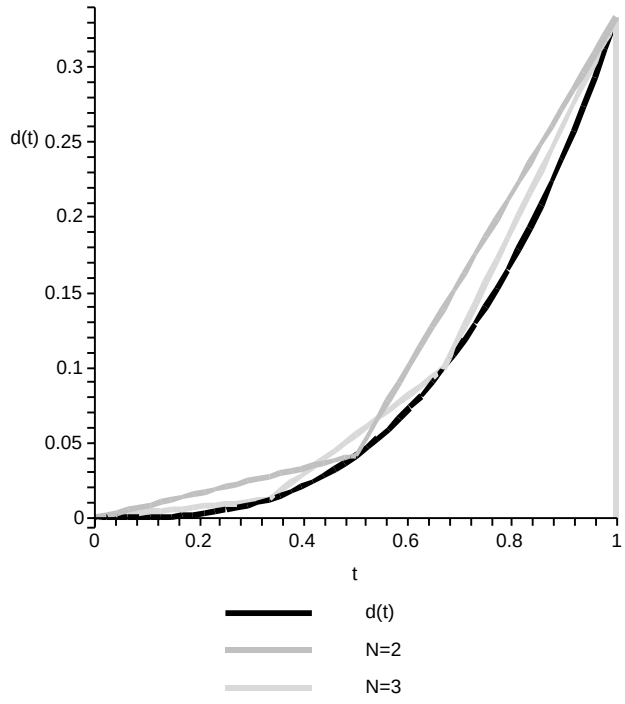
This is the second time we have taken a process which can be refined as many times as we wish and found that the answer we arrive at in the end "tends" to a unique value as the process is refined.

Let us summarize. We know from highschool that the formula

$$v = \frac{\Delta d}{\Delta t} \tag{10}$$

where Δ or "Delta" denotes "change in". As a recipe between three numbers this is OK; compute the change in distance, record the change in time, divide the two and there is your velocity. Our work above suggests that there is a better and more general way to think of this relation. This "better way" allows for the relation to be true not just once, but all the time (though the numbers can change). Moreover, the "better way" has a way to compute an approximation that is as accurate as we would like.

This is probably still a bit unclear, so let's look at it from the opposite point of view. Let's say we know the displacement as a function of time and ask ourselves how we could approximate the velocity. For concreteness, we'll use $d(t) = t^3/3$ with $T = 1$, and now, for a given N, we will again look at time segments of length $\frac{T}{N}$. We will now define the velocity in each segment by the distance travelled at the end of the time segment, minus the distance travelled at the beginning of the time segment (the δd) and then divide by the length of the time segment. Here is a picture:



From the picture we see that the velocity is just the **slope** of each time segment. Moreover the extent to which the line segments follow the curve increases as N increases. Even with $N = 3$ the shape of the $d(t)$ is reflected rather well by the line segments.

It is thus clear that we must somehow get a handle on this whole "limiting" or approximating

process. Once we do (and it will be a bit of a mathematical battle) we can come right back to simple mechanics and hopefully add some solid algebra to the geometric intuition we have built up.

Math 137 Physics Based Section

Marek Stastna

9 LIMITS

The limiting process and the concept of limit form the backbone of much of what is to follow. In their essence they are purely mathematical concepts, however the wealth of applications that they make possible makes them well worth studying in detail.

Example 1 Consider the identity function $f(x) = x$ on the set of inputs $0 \leq x < 1$. Notice that at the left end point, $x = 0$ is **included** in the set of possible inputs, while at the right end point $x = 1$ is **not included** in the set of possible inputs. We have chosen things to be this way in this example, but there will be instances in which the input of interest is disallowed (think of a rational function in which the denominator is not allowed to equal zero).

Now let's consider what happens as we get closer and closer to $x = 0$. The identity function makes the output just the same as the input so as the input gets closer to zero, the output gets closer to zero. Indeed we can go right to zero, and taking $x = 0$ as the input we get the output $f(0) = 0$. You might be saying to yourself that we could have skipped the whole business of inputting smaller and smaller values and gone right to inputting the actual value $x = 0$. To see why this isn't true consider what happens near the right end point.

Here the input $x = 1$ is prohibited (remember we chose it to be that way) so we can only get closer and closer to $x = 1$. We could, for example take the sequence of inputs 0.9, 0.99, 0.999 and so on. The resulting sequence of outputs is, obviously, 0.9, 0.99, 0.999. Aha, you say we are tending towards an output of 1. At the risk of being pedantic, let's ask, "What do I mean by that?". Look at the difference between the **candidate value** (1 in this case) and the outputs. The differences are, 0.1, 0.01, 0.001. Now what about the inputs? Since $f(x)$ is the identity function the differences are the same! So if I **demand** that the output is a certain distance from 1, or closer, then all I have to do is ensure that the input is the same distance or closer from the limiting input value of 1.

Example 2 Consider $f(x) = 2x$ on $0 < x < 1$ and again ask what the limit of the function is as we get closer to 1. If we again consider the sequence of inputs that get closer and closer to 1, such as 0.9, 0.99, 0.999, 0.9999 then the sequence of outputs is 1.8, 1.98, 1.998, 1.9998 and so on. Clearly the limiting value is 2 (you could even cheat and check that $f(1) = 2$ even though on a test you wouldn't want to do that since $x = 1$ is not an allowable input). More importantly let's say we want to guarantee the output to be within 0.002 of the limiting

value of 2 then we would have to restrict the input to be 0.999 or larger, or within 0.001 of the limiting input value of 1.

You can think of this as a quality control problem at a factory. Given that a bolt must have a radius $R \pm 0.001$ how must accurately must the machine that makes the bolt be callibrated.

DEFINITION We say the function $f(x)$ has the limit L as x tends to a if **for any desired constraint on the output** (call it $\epsilon > 0$) we can **guarantee** that $f(x)$ lies a distance ϵ or less from L, e.g. $|f(x) - L| < \epsilon$, by restricting the input to lie a certain distance from $x = a$, or $|x - a| < \delta$.

In the above I used the fact that the absolute value function measures the distance between numbers. Note that for one particular value of ϵ (say 0.001 we get one value of δ. For a smaller value of ϵ we get a different value of δ. The key however is that we can make the process work **for any** $\epsilon > 0$.

NOTATION We write $\lim_{x \longrightarrow a} f(x) = L$ if the limit exists. Often we say a function tends to the value L when $\lim_{x \longrightarrow a} f(x) = L$.

The concept of limit is our first confrontation with the concept of infinity. The infinity in this case is the idea of making sure the output is infinitely close to the limiting value by restricting the input appropriately. Of course infinity is more commonly associated with the idea of large things. Arguing similarly to the above we see that infinity is not a number, but the concept of getting arbitrarily large (you give me a number I can go bigger, you give me a bigger one still, I get bigger again).

One hardly ever uses the definition to show limits exist. In practice there are a variety of tricks to compute limits. Here's just one.

Example 3 Consider

$$f(x) = \frac{3x^2 + 5x + 3}{x^2 + 4}$$

Does the limit as x gets arbitrarily large exist? If it does, what is it?

Consider first the simpler rational function

$$g(x) = \frac{1}{x^n}$$

where n is a natural number. Here the numerator is fixed to be 1 and as the input x gets larger we can see that the denominator is at least as large as the input. This means we are dividing 1 by larger and larger values and thus $g(x)$ tends to 0 as x gets arbitrarily large. If I gave you a value for n then you could use the definition of limit to show that the limit is 0 from the definition. Finally what if the 1 in the numerator was a 31 or a 513451? Would anything change? Not really, since the input is allowed to get arbitrarily large we would still eventually be dividing by a number so big the result would be very close to 0.

OK back to $f(x)$, let's do a nothing operation by dividing the numerator and denominator by the same thing. The trick is to choose that something as the highest power in the denominator. Here is the result:

$$f(x) = \frac{3 + \frac{5}{x} + \frac{3}{x^2}}{1 + \frac{4}{x^2}}.$$

Now look at the various pieces. As x gets arbitrarily large the constants 3 and 1 don't change, but all the other terms get very close to 0. In fact they get arbitrarily close to 0. This means

$$\lim_{x \longrightarrow \infty} f(x) = \frac{3 + 0 + 0}{1 + 0}$$

or

$$\lim_{x \longrightarrow \infty} f(x) = 3.$$

You should notice how I structured my argument. I showed something simple first,

$$\lim_{x \longrightarrow \infty} \frac{1}{x^n} = 0$$

for any natural number n. I then extended it to

$$\lim_{x \longrightarrow \infty} \frac{A}{x^n} = 0$$

for any rational number A. With this fact in hand I rewrote $f(x)$ so it was made up of pieces that were either constants or of the form A/x^n. Then I could find the limit of $f(x)$ by putting together the limits of the pieces.

This means I used certain properties of limits (check your textbook for a full list) which allow me to do arithmetic on limits (the limit of a sum is the sum of individual limits and so on). There is one key feature of all these properties, they only work if the limits of all the pieces and the original limit exist.

Example 4 Consider $\lim_{x \longrightarrow \infty} x$. Here the fact that the input can get arbitrarily large and the function of interest is the identity function together imply that the output can get arbitrarily large. In this case we say the limit **does not exist** or that the function **grows without bound**.

Example 5 There is a technique to evaluate certain limits that has a very broad range of application. It is called the **Squeeze Theorem** and we will illustrate it by considering $g(x) = x^2 \sin 1/x$ as $x \longrightarrow 0$. The limit is troublesome because $1/x$ grows without bound in magnitude as $x \longrightarrow 0$ and this means that $\sin(1/x)$ oscillates faster and faster as x gets smaller and smaller. Incidentally $\sin(1/x)$ is one of those examples that despite its unphysical nature, tends to pop up an awful lot in math courses. So you might want to spend a minute or two looking at its behaviour yourself.

Back to the Squeeze Theorem. Even though $\sin(1/x)$ oscillates quickly when x is small, it is still a sine function and that means that it is no smaller than -1 and no larger than 1, i.e. $|\sin(1/x)| \leq 1$. We will use this fact to construct two "fences" around $g(x)$. The first will be given by the upward opening parabola $h_1(x) = x^2$ and the second by the downward opening parabola $h_2(x) = -x^2$. Mathematically we write this as

$$-x^2 \leq x^2 \sin(1/x) \leq x^2.$$

Notice that the above is true *regardless of what the input is*. In particular it is true for $x \longrightarrow 0$. When x tends to zero both $h_1(x)$ and $h_2(x)$ tend to zero as well. Thus $g(x)$, which is trapped in between, has no choice but to tend to zero as well and we thus conclude that

$$\lim_{x \longrightarrow 0} x^2 \sin\left(\frac{1}{x}\right) = 0 \tag{1}$$

The Squeeze Theorem will pop up here and there throughout the course and again in future courses. You can think of it as a new trick to try if you are given a problem you don't know how to solve right away.

10 THE DERIVATIVE

We could talk about limits for another month. However, I want to get back to the matter of rates of change and the topic of mechanics. Recall that when for motion in a straight line the velocity is the rate of change of distance, provided the velocity is constant. Of course, as anyone who has stomped on the accelerator pedal knows, it's when the velocity isn't constant that things get interesting. So let's start with the simple mechanics relation distance=velocity×time. Now let's assume we have a valid stop watch and some Cartesian axes with the motion taken along the x-axis. This allows us to use the following definitions:

- $v(t)$ represents the **velocity** as a function of **time**
- $x(t)$ represents the **position** (this is another way to say distance travelled) as a function of **time**

and ask, "What is the relation between $x(t)$ and $v(t)$?" If we look at two instants in time we can write t_0 and $t_0 + \Delta t$ to give them mathematical names. The position at the two instants is given by $x(t_0)$ and $x(t_0 + \Delta t)$ and the distance travelled would thus be $x(t_0 + \Delta t) - x(t)$, which we could write as the change in position Δx. If we next divide by the difference in time, Δt (here the choice of notation becomes clearer) we would get

$$G(t_0, \Delta t) = \frac{\Delta x}{\Delta t}. \tag{2}$$

Notice two things. First, all I did was assign the name $G(t_0, \Delta t)$ to the ratio, in reality, for any one t_0 and time increment Δt, $G(t_0, \Delta t)$ is a **number**. Of course if you think of changing the input time or the time increment then you come to the second point, namely G as a function, has **two** inputs. You could think of these as the two instants in time, but I claim the better way to do it is to think of it as the original time t_0 and the increment Δt. My motivation is the fact that I am not happy about having two inputs (partly because I know $v(t)$ has only one and I am eventually hoping to relate $v(t)$ and $x(t)$). How to get rid of one input? Well, consider which of the two is troublesome. Changing t_0 means a different time, and that's really not such a big deal, in fact we should be able to change the input time. But what if I wish to choose $\Delta t = 0.1$ seconds while someone else would prefer $\Delta t = 0.0001$ seconds. Which is the better choice? It seems kind of arbitrary, but perhaps the limiting process can help. What if we consider the limit of time increments tending to 0? If this limit exists then on the left hand side we get some function $G(t_0)$ which equals a limit, namely

$$G(t_0) = \lim_{\Delta t \longrightarrow 0} \frac{\Delta x}{\Delta t}.$$

Of course if we fix t_0 to take a value, say $t = 1$, then we expect a number from the limit.

OK, fine, can we calculate any of these limits? Let's start with a very simple case:
Example 1 Say $x(t) = 2t$ for $0 \leq 0$. What is G(1)? From above

$$G(1) = \lim_{\Delta t \longrightarrow 0} \frac{\Delta x}{\Delta t}$$

provided the limit exists (this statement is the 'small print' from contest forms, but for mathematicians). To make progress let's write out Δx to get

$$G(1) = \lim_{\Delta t \longrightarrow 0} \frac{x(1 + \Delta t) - x(1)}{\Delta t}.$$

Next we use $x(t) = 2t$ to get

$$G(1) = \lim_{\Delta t \longrightarrow 0} \frac{2(1 + \Delta t) - 2(1)}{\Delta t},$$

or

$$G(1) = \lim_{\Delta t \longrightarrow 0} \frac{2\Delta t}{\Delta t}.$$

Now the power of the limit comes to the forefront. Remember we are not actually ever going to set $\Delta t = 0$. We are only going to get arbitrarily close. This means we can cancel the $\Delta t = 0$ from the numerator and the denominator to get

$$G(1) = \lim_{\Delta t \longrightarrow 0} \frac{2}{1}.$$

Of course this gives $G(1) = 2$.

It thus appears we are successful in applying the limit.

Example 2 What about $G(t_0)$ where $t_0 > 0$? We could try doing the same thing as above. Using $x(t_0) = 2t_0$ gets us

$$G(t) = \lim_{\Delta t \longrightarrow 0} \frac{2(t_0 + \Delta t) - 2t_0}{\Delta t}$$

and the simple algebra again gives $G(t) = 2$.

But hold on a minute, when velocity is constant then

$$v = \frac{d}{t}$$

and from the above for $x(t) = 2t$ we get $G(t)$ =constant. Thus for a linear function of position $G(t)$, which is a constant, agrees with the velocity. Moreover since

$$G(t_0) = \lim_{\Delta t \longrightarrow 0} \frac{\Delta x}{\Delta t}.$$

where the fraction on the right hand side translated into english reads "the limit of the distance travelled divided by the time taken to travel it". Thus we have the following

DEFINITION The velocity, $v(t)$, is defined as

$$v(t) = \lim_{\Delta t \longrightarrow 0} \frac{x(t + \Delta t) - x(t)}{\Delta t} \tag{3}$$

provided that this limit exists. Moreover, the velocity is interpreted as the **rate of change of position with time**.

Example 3 Say $x(t) = t^2$ and $0 \leq t$. What is $v(t)$? Working from the definition gives

$$v(t) = \lim_{\Delta t \longrightarrow 0} \frac{x(t + \Delta t) - x(t)}{\Delta t}$$

or using $x(t) = t^2$

$$v(t) = \lim_{\Delta t \longrightarrow 0} \frac{(t + \Delta t)^2 - t^2}{\Delta t}.$$

Now expand $(t + \Delta t)^2$ to get

$$v(t) = \lim_{\Delta t \longrightarrow 0} \frac{t^2 + 2t\Delta t + (\Delta t)^2 - t^2}{\Delta t}$$

or

$$v(t) = \lim_{\Delta t \longrightarrow 0} \frac{2t\Delta t + (\Delta t)^2}{\Delta t}.$$

Finally, dividing out by the Δt gives

$$v(t) = \lim_{\Delta t \longrightarrow 0} (2t + \Delta t).$$

Now $2t$ does not depend on Δt and the second term is just Δt itself and hence tends to zero. Thus

$$v(t) = 2t.$$

We have shown that a quadratic displacement implies a velocity that is non-constant and is linear as a function of time.

You have just computed your first non-trivial derivative (for some reason mathematicians do not like 'rate of change'). Notice that from a graphing point of view the rate of change (or derivative) generalizes the idea of 'slope' (rise over run) to functions whose graphs are not lines (more on this later). Also notice how the machinery of the limit was required to carry out the calculation, but how it remained in the background compared with the physics.

11 THE MEANING OF DERIVATIVES

While theoretical advances do not often gain notoriety on their own (with notable exceptions such as the quantization of energy, and the Lorentz transformation of Special Relativity), theoretical advances that save work are celebrated. For example, in the field of computing, the Fast Fourier Transform, dates back to Gauss for its trickiness, but it was not until it sped up computations by an order of magnitude or more that it got its own acronym (FFT; a friend actually chose her son's names so his initials were FFT) and became a standard topic to teach in the undergraduate curriculum.

Though we will talk about this in more detail in the next section, let's proceed as we did for velocity in the previous section and write

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

where $f'(x)$ denotes **the rate of change of a function with respect to its input**.

So what is the derivative, or rate of change, good for? Well, let's look at the simple parabola $f(x) = x^2$. We know from our high school education that this parabola opens upward and that its minimum occurs at the point $(0, 0)$. From the previous section (just change t to x and $x(t)$ to $f(x)$) we know that the rate of change is given by $f'(x) = 2x$. This is a linear function and clearly $f'(0) = 0$, $f'(x) > 0$ when $x > 0$, and $f'(x) < 0$ when $x < 0$. A simple check of the graph will reveal that the function $f(x) = x^2$ is increasing when $f'(x) > 0$ and decreasing $f'(x) < 0$, with $f'(0) = 0$ occurring at the minimum. This is no accident.

Indeed, **provided we can find the derivative (this is not a strong restriction, though worth being careful about) then $f(x)$ is increasing (decreasing) at an input point x provided $f'(x) > 0$ ($f'(x) < 0$).**

You can look at the definition of the derivative, or rate of change, to see why this is true. Consider a small, positive h (say $h = 0.0001$) and pick $x = 5$, just so we don't have too many variables running around. The expression on the right hand side now says, take what f gives when you input a number just a bit bigger than 5 subtract off $f(5)$ and divide the result by the difference in inputs. Let's call the outcome of the calculation M and notice that M should be pretty close to $f'(5)$ (otherwise the limit would not exist). Let's say $f'(5) > 0$ and thus $M > 0$ and a simple rearrangement gives

$$f(5 + h) = f(5) + hM.$$

Because both h and M are positive we conclude that $f(5 + h) > f(5)$.

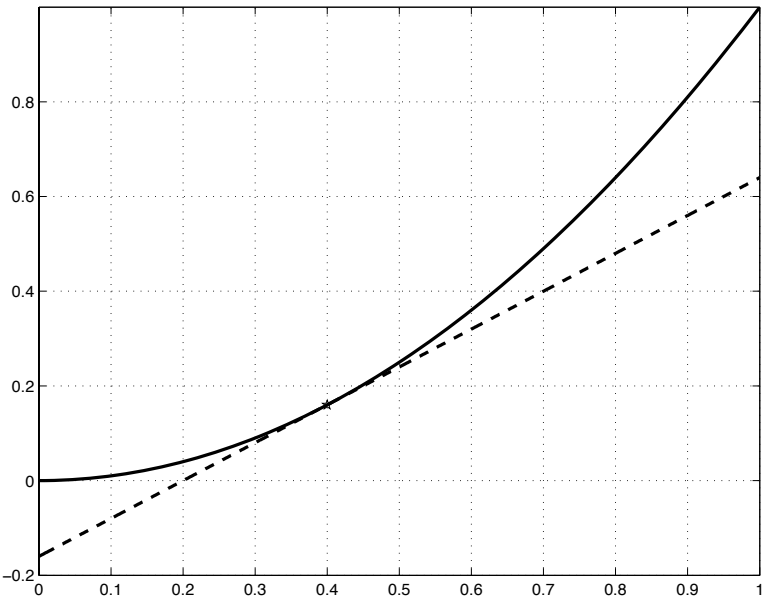
Now to conclude that $f(.)$ is an increasing function at $x = 5$ we need to also check $h < 0$, or inputs just a bit smaller than 5. Arguing similarly to the above we find (because $h < 0$ I use the absolute value to clearly make the negative numbers show up, it is a common trick, though may look weird on first sight)

$$f(5 + h) = f(5 - |h|) = f(5) - |h|N$$

where again N is near $f'(5)$ and if $f'(5) > 0$ then for small enough $|h|$ $N > 0$. Thus the output just to the left of 5 on the number line will be smaller than the output at 5. Thus $f'(5) >$ guarantees that the function $f(.)$ is increasing at 5.

Example 1 Consider $f(x) = x^4 - x^2 + 5$. Either some grinding, though possible, limit calculations, or the power rule of derivatives (see your text or the next section for details) give us $f'(x) = 4x^3 - 2x$ and now we can conclude, among other possibilities that at $x = 0.1$ $f(x)$ is decreasing because $f'(0.1) < 0$ (check it) while at $x = 1$ $f'(1) > 0$ (check this, too) means that $f(.)$ is increasing at $x = 1$. Finally we could ask where all the weird points at which $f'(x) = 0$ are. In this case $f'(x) = 2x(2x^2 - 1)$ and so $x = 0$ and $x = \pm 1/\sqrt{2}$ give the three answers.

To put a picture to the words, recall that the slope of a line is the rise over the run, and thus the derivative, or rate of change, is just a way to generalize the idea of "slope" to functions that are not lines. Of course if I give you just a slope, you can come up with many different lines. Thankfully, since we are computing the rate of change at a particular input (say $x = x_0$), it makes perfect sense to demand that the line pass through the point $(x_0, f(x_0))$ and has the slope $f'(x_0)$ (ideally you would want simpler ways of computing $f'(x_0)$ than going from the limit, and indeed we will discuss them shortly). This special line is called the **tangent line** and in the following picture I give an example of one with $f(x) = x^2$ and $x_0 = 0.4$.



Notice how the line touches the curve at the point $(0.4, 0.4^2)$ and how near this point it is very close to the curve itself (even though a line and a parabola are very different objects geometrically).

In summary, derivatives, or rates of change, let us:

- Determine if a function is locally increasing ($f'(x) > 0$) or decreasing ($f'(x) < 0$).
- Construct a tangent line to the graph of $f(x)$ passing through the point $(x_0, f(x_0))$. The tangent line is a good approximation to the original function near x_0 (though how near is unclear).

12 MACHINES ON FUNCTIONS

The rate of change, or derivative, that we discussed in the previous two sections is a particularly useful example of a machine that takes a function as input and produces a (possibly different) function as output. This is really no different than thinking of $f(x) = \sin(x^2)$ as a two part function. The first, call it $g(.)$, takes in x as input and produces x^2 as output. The second, call it $h(.)$ takes in g as input and produces $\sin(g)$ as output. Since $g(x)$ is a function, the function h acts on a function! Of course for $f(x) = \sin(x^2)$, it makes more sense of thinking about the action of $f(.)$ one input at a time. However, as we showed, velocity is the rate of change of position, and here it is quite useful to think of velocity, the function of time $v(t)$, to be given by the action of a certain machine (the derivative) on the function of time, $x(t)$, which represents position.

Indeed this idea is particularly useful if we can avoid calculating the derivative, or rate of change from the limit. Let's consider polynomial functions first. To make progress we need to agree on some notation.

NOTATION If, at a point t_0, the rate of change, or derivative, of a function $f(t)$ can be found from the limit we will write for the rate of change either

$$\frac{df}{dt}(t_0)$$

or $f'(t_0)$. If the rate of change could be computed from the limit at all input points we write

$$\frac{df}{dt}$$

or $f'(t)$. Thus for example for $x(t) = t^2$ we have $f'(t) = 2t$ and $f'(1) = 2$.

Example 1 It is easy to show using the binomial theorem that for any natural number n, $f(t) = t^n$ has a rate of change for all t and that $f'(t) = nt^{n-1}$. This is often called the **power rule of derivatives**.

Example 2 If we take a linear combination of two functions $f(t)$ and $g(t)$, or in other words, we define $h(t) = a_1 f(t) + a_2 g(t)$ for two numbers a_1 and a_2, then provided $f'(t)$ and $g'(t)$ can be computed then

$$h'(t) = \lim_{\Delta t \longrightarrow 0} \frac{h(t + \Delta t) - h(t)}{\Delta t}$$

by definition. Of course we can just substitute for $h(.)$ to get

$$h'(t) = \lim_{\Delta t \longrightarrow 0} \frac{a_1 f(t + \Delta t) + a_2 g(t + \Delta t) - a_1 f(t) - a_2 g(t)}{\Delta t}$$

which can be regrouped an split up as

$$h'(t) = \lim_{\Delta t \longrightarrow 0} \frac{a_1 f(t + \Delta t) - a_1 f(t)}{\Delta t} + \lim_{\Delta t \longrightarrow 0} \frac{a_2 g(t + \Delta t) - a_2 g(t)}{\Delta t}$$

where it's OK to split the limit into two limits provided each of the two new limits exists. Similarly the a_1 and a_2 can be brought out front to yield

$$h'(t) = a_1 \lim_{\Delta t \longrightarrow 0} \frac{f(t + \Delta t) - f(t)}{\Delta t} + a_2 \lim_{\Delta t \longrightarrow 0} \frac{g(t + \Delta t) - g(t)}{\Delta t}.$$

Now the two limits just give $f'(t)$ and $g'(t)$ and so finally we have the linearity rule of derivatives, or rates of change, $h'(t) = a_1 f'(t) = a_2 g'(t)$.

Example 2 is interesting, because, first of all it shows us the utility of a fairly dull mathematical argument. Mainly, once shown one time the mathematical fact is ours to use at will. Second it tells us that we can compute rates of change of polynomials.

Example 3 Let's consider $x(t) = t^5 + 3t^3 - 1$, to find $v(t)$ we must find $x'(t)$, to do this think of writing

$$\frac{d}{dt}(t^5 + 3t^3 - 1) = \frac{dt^5}{dt} + 3\frac{dt^3}{dt} - \frac{d1}{dt}$$

where we think of the machine $\frac{d}{dt}$ acting on the polynomial $x(t) = t^5 + 3t^3 - 1$. Using the linearity of the derivative we get the right hand side, and the three pieces can now be evaluated from Example 1, namely $v(t) = x'(t) = 5t^4 + 9t^2$. You should confirm using the limit that the rate of change of any constant is zero (this is obvious from the term "rate of change").

There are more complicated rules for products of functions and we will come back to those later. For now I want to get back to our simple mechanics problems. We defined the velocity at an instant in time as the rate of change of position. It thus makes sense to define **acceleration**, which we will write $a(t)$, as the rate of change of velocity. This means $v(t) = x'(t)$ and $a(t) = v'(t)$. Notice, if we think of $v(t)$ as the output from feeding the position, $x(t)$, as input into the derivative machine, then we can think of the acceleration as the output from a two step process in which we feed the position into the derivative machine to get velocity and then feed velocity into the derivative machine to get the acceleration.

NOTATION Repeated application of the derivative machine is written like

$$a(t) = \frac{d^2x}{dt^2}$$

or just $a(t) = x''(t)$.

Now, it is a cornerstone of classical physics (really all physics) that under an appropriate choice of coordinate system the relation governing a particle's motion is **Newton's Second Law**, namely Force=mass×acceleration. In general both force and acceleration are *vectors*, but for the present we assume the motion is in one dimension and write $F = ma$. On the surface of it this is a formula, much like the Universal law of gravitation. However, imagine that we are not so interested in the acceleration, but the position of our particle. Then provided we can measure the particle mass and have a known the functional form for the force, then we have

$$\frac{d^2x}{dt^2} = \frac{F(t)}{m}. \tag{4}$$

This is a strange sort of equation because it is not a straight up recipe like the Universal law of Gravitation. We can rewrite it as

$$\frac{dv}{dt} = \frac{F(t)}{m}. \tag{5}$$

which tells us that the rate of change of the velocity is given by the force (which is a known function of time) divided by the mass (which is assumed to be measured). An equation like this is called a **differential equation**. I guess **rate of change equation** did not roll off the

tongue quite as well (actually, notation is a serious matter and indeed Newton was so bad at choosing notation that literally noone writes mechanics the way newton did). Given what little we know, you may be surprised to find that we can solve many differential equations.

Example 4 Consider $F(t) = 0$, or no force. Newton's second law then tells us that the acceleration, or the rate of change of velocity with time, is zero. Thus velocity is constant, say $v(t) = v_0$. What about position? This is a bit trickier. From the power rule with $n = 1$, or from the limit, we see that if $x(t) = v_0 t$ then $x'(t) = v_0$. OK, reasonable, right? We have this derivative machine, so we can make it work in reverse, at least sometimes, by working backwards. **But** hold on a minute, we know that the derivative of a constant is zero so in fact I could have added any one number to $x(t)$! Let's write that as

$$x(t) = v_0 t + x_0 \tag{6}$$

and ask ourselves if this makes sense. We know our particle is moving in the absence of forces, moreover let's say at $t = 0$ we know the particle is found at the position $x = 3$ and is moving with the velocity $v = 2$. Our process of starting with $F = ma$, and reversing the derivative machine gives us $v(t) = v_0$ and if $v(t = 0) = 2$ then $v_0 = 2$ and $v(t) = 2$. So far so good. Now consider the position, $x(t)$. As we showed above $x(t) = v_0 t + x_0 = 2t + x_0$, using $v(t) = 2$. Now we measured that $x(t = 0) = 3$ and thus $x_0 = 3$. Thus:

Given that $F(t) = 0$ **and** $x(0) = 3$ **and** $v(0) = 2$ **The differential equation known as Newton's Second Law gives us** $x(t) = 2t + 3$**.**

Try working out other sets of initial conditions for yourself to get a bit of practice.

To conclude, each time we reversed the derivative machine we had to keep an extra **arbitrary constant**. With the appropriate measurement this constant could be uniquely determined. Thus the mathematics takes care of the functional form, but it is the physics that tells us the values of any constants. A proper scientific model, of course, requires both mathematics and physics.

13 AN EXAMPLE AND SOME NUMERICS

We now consider the most basic problem in particle mechanics, namely that of a projectile (say a ball), thrown upwards under the action of the Earth's gravity. It is not really necessary to throw the ball straight up, but it will let us avoid talking about vectors and this seems reasonable for now. To make progress on any problem we require a choice of coordinates. You can think of this as a choice of measuring sticks, though it is somewhat deeper than that. For now we will assume that all motion is along the x axis which points upward. Moreover since we don't plan to throw the ball too high, the Universal Law of Gravitation can be simplified simply to $F = -mg$ where m is the mass of the ball and g is the acceleration due to gravity (usually about $9.81 m/s^2$). It is worth noting that assuming the ball doesn't fly

too high also allows us to neglect the earth's rotation, a very important simplification that other problems (say weather prediction) do not permit.

OK, so we should start with Newton's second law,

$$-mg = ma$$

or using rates of changes

$$x''(t) = -g$$

where I have divided out by the mass. How interesting, the path of the ball does not depend on its mass!

We have the governing differential equation, now what to do with it? Perhaps it might be easier to work with the velocity first:

$$v'(t) = -g.$$

Thus the rate of change of velocity is a constant. We've done this before and conclude that the velocity must thus be a linear function of time, which we write,

$$v(t) = -gt + C$$

where C is a constant, but we can't really say what its value is, yet.

To get at C consider what we want to specify about the ball. One obvious choice is "How hard we throw it up". This is a bit ambiguous, and we could reasonably think, "well I'm pretty sure I can control how fast the ball leaves my hand." That means, if we start our stopwatch just as the ball leaves our hand then $v(0) = v_0$ where v_0 is a known value. We thus conclude

$$v(t) = v_0 - gt \tag{7}$$

Great, now nwhat about the position? Well we do know that the velocity is, by definition, the rate of change of position, so

$$x'(t) = v_0 - gt.$$

Using the power rule of derivatives we guess that

$$x(t) = v_0 t - \frac{1}{2}gt^2 + D$$

where D is an arbitrary constant. Now, we know the ball is in our hand just at the time of release, and while this is not necessary, it does make some sense to take that point to be the origin. This would mean that $x(0) = 0$ and hence the position of the ball is given by

$$x(t) = v_0 t - \frac{1}{2}gt^2. \tag{8}$$

These formulae are likely familiar ones form high school, though I hope you agree that we have arrived at them in a more reasonable way using the calculus. What information can we get from them?

Consider the question: "Given v_0 how high will the ball fly?" This is another way of saying, "At what time does the ball stop moving upward?" Putting it in the latter manner means we look at $v(t)$ and try to find where it is zero. Since $v(t)$ is a linear function there is precisely one time where this is true, and if we call it t^*, then

$$t^* = \frac{v_0}{g}.$$

To find the height the ball reaches (call it H) we just use t^* in the formula (8) to get, after a bit of algebra,

$$H = \frac{1}{2}\frac{v_0^2}{g}.$$

There is one more thing worth noticing. Namely, the maximum height reached is, obviously, the maximum of the function $x(t)$ for all $0 \leq t$. At this maximum the velocity must be zero (otherwise the ball would be going up and would thus reach a higher height). Now using the definition of $v(t)$ we conclude that **at the maximum of $x(t)$ the rate of change of $x(t)$ is zero.**

What about the velocity? Intuitively we know that since gravity acts downward, the velocity can never be larger than what it is immediately upon the ball's release, or $v(0) = v_0$. Computing the rate of change of velocity gives us $-g$, the *downward* gravitational acceleration, which is never zero. So what gives? Well the point $t = 0$ is special because it is at this time that we start our stopwatch and our model makes no claim as to what happened for $t < 0$. The point $t = 0$ is thus a **boundary point** and we conclude that **maxima (and reasonably, minima) can occur at boundary points as well**.

So there you have a simple problem done in its full glory. However, the forcing need not always be so simple, and indeed the governing equation itself may be more complex (as it is for the wind blowing past the Math building). For hundreds of years this was a very serious restriction on what could be done with the calculus. However, computers have largely done away with this problem, because they allow us to calculate accurate approximations to solutions of differential equations. Let's look at Newton's second law written for the velocity,

$$\frac{dv}{dt} = \frac{F(t)}{m}$$

where $F(t)$ is known, but we can't think of how to reverse the derivative machine to write a formula for $v(t)$. We will also assume that $v(0) = v_0$, a known value. Now let's return to the definition of derivative, namely

$$\frac{dv}{dt} = \lim_{\Delta t \longrightarrow 0} \frac{v(t + \Delta t) - v(t)}{\Delta t}.$$

A finite precision machine like the computer cannot be expected to compute the limit on the right hand side, but perhaps we don't want it to! Perhaps we can get away with only

finding $v(\Delta t)$, $v(2\Delta t)$, $v(3\Delta t)$, and so on, where $\Delta t = 10^{-4}$ or something small like that. In that case we could write

$$\frac{v(t+\Delta t) - v(t)}{\Delta t} \approx \frac{F(t)}{m}$$

and since $v(t+\Delta t)$ is the velocity at a time later than t, we rearrange into recipe form to get what is called an update equation

$$v(t+\Delta t) = v(t) + \Delta t \frac{F(t)}{m}. \tag{9}$$

You can imagine using this equation by hand, but better still you can imagine a computer doing thousands of steps in mere seconds. Indeed this is the power of numerical methods, though we won't have much more to say about numerical methods in this course.

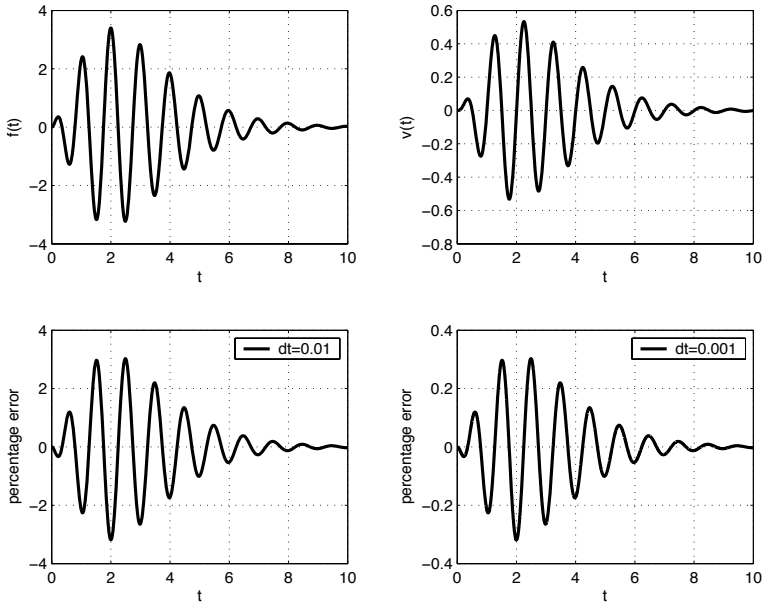
Example 1 Consider a point particle of mass m, given by the somewhat complicated force

$$f(t) = t^2 \exp(-t) \sin(2\pi t).$$

Now all three parts of $f(t)$ are reasonably familiar, though in this course I have been somewhat coy about using the exponential (not for long, though). At $t = 0$, $f(0) = 0$, while for $t > 0$ the sine part will be oscillatory while the other two parts will compete in making the amplitude of the sine shrink or grow. You could try sketching $f(t)$ for practice. Now with some moderate effort and first year behind you, you could reverse the derivative machine to solve

$$\frac{dv}{dt} = f(t)$$

to get a formula for $v(t)$. But why not try our numerical method! It took me about four lines of Matlab code and exactly one for loop to produce the data for the following picture:



In the upper left panel I show the forcing function for $0 \leq t \leq 10$. In the upper right panel i show the numerical solution with $\Delta t = 0.01$. In the bottom left panel I show the difference between the numerical solution and the "exact" solution one would get by reversing the derivative machine by hand. I scale things to show the *percentage error* and you can see that the numerical method error is no more than 4%. In the bottom right panel I increase the accuracy of the numerical method by shrinking Δt to equal 0.001. You can see that the error shrinks by a factor of ten or so (often one would say the error shrinks by an order of magnitude). Thus numerical methods are almost ridiculously easy to use (on simple equations anyway). Indeed the practice of science has been utterly changed by the fact that computing is so readily available to solve problems the great scientists of the past could only speculate about.

There is one more interesting thing about the way we approximated the limit with a very small, but not arbitrarily small, Δt. If this approximation works we can write

$$v'(t) \approx \frac{v(t + \Delta t) - v(t)}{\Delta t}$$

and rearranging we say

$$v(t + \Delta t) \approx v(t) + \Delta t v'(t). \tag{10}$$

Now let's say for concreteness that $t = 1$, $v(1) = 2$ and $v'(1) = -1$ then (10) reads

$$v(1 + \Delta t) \approx 2 - \Delta t$$

which tells us that when Δt is small, or saying it another way, when t is near 1, then the actual function $v(t)$ can be approximated by a line. Moreover this line is specified so that at $t = 1$ it gives the same output as $v(.)$ and the slope of the line is given by $v'(1)$, the rate of change at the input point $t = 1$. While we don't know how small Δt has to be for the approximation to be good, it is a remarkable fact nevertheless that provided we can find the rate of change at a point, then locally the function can be approximated by a line (or linear function) no matter how complex the original function was.

14 THE CHAIN RULE

As impressive as the mechanics applications of the rate of change, or derivative, concept are, they only scratch the surface. In this section we open up a whole world, both of applications and pure mathematics, by looking at the so-called **Chain Rule**. The Chain Rule is a fundamental building, and is often placed on a high pedestal as a result. However, it is, at root, a very simple idea.

Consider a measuring device moving along a time dependent path $x(t)$ and measuring a quantity $C(x)$ which varies with space, but is assumed not to vary in time. An example of this is a weather balloon, or airplane, that measures the concentration of stable trace gases in

the atmosphere. A more exciting, and exotic one is the GRACE II satellite which measures the variations in the gravitational field of the Earth (the interested can ponder for a few moments how such a thing could be done) due to things like underground aquifers (whose size and rate of change is important for city planning). Of course the assumption of C being a function of one variable only is overly simplistic, still it is a good place to start.

OK, what do we have? Well we have a single input, namely t, for time, being fed into the machine called $x(.)$ which is assumed to be given by a recipe (like $x(t) = t^2$, say) and that outputs the position $x(t)$. Next we have a machine $C(.)$, representing concentration of a quantity of interest (ozone, nitrous oxide, etc.) that requires the position as input. We ask: **What is the rate of change of $C(x(t))$ with time?**

We could make a diagram of inputs and outputs like

$$t \longrightarrow x(.) \longrightarrow x(t) \longrightarrow C(.) \longrightarrow C(x(t))$$

and even write the definition of limit as

$$\frac{dC(x(t))}{dt} = \lim_{\Delta t \longrightarrow 0} \frac{C(x(t+\Delta t)) - C(x(t))}{\Delta t}$$

provide the limit exists. Again, if we wanted an approximation then we could just take Δt very small and evaluate the right hand side (see assignment). However, long before calculators, people wanted a better expression for the right hand side. There are two ways to get it.

The first argues that, physically, we should **only be able to calculate the rate of change of a machine with respect to its input**. This means that the rate of change of $C(x(t))$ with respect to time must be done in two parts. First we calculate the rate of change of $C(.)$ with respect to its true input, namely position, and then we multiply by the rate of change of position with time.

Does this make sense? Let's say $C(x) = x$, then the above rule says that

$$\frac{dC(x(t))}{dt} = \frac{dC}{dx}\frac{dx}{dt}$$

and since $C(x) = x$

$$\frac{dC}{dx} = 1$$

and

$$\frac{dC(x(t))}{dt} = v(t)$$

the velocity. This makes perfect sense because in this case $C(.)$ was just the identity function and any change in C was due to how fast the measuring device was moving. Notice that if $C(x)$ =constant, then

$$\frac{dC}{dx} = 0$$

and

$$\frac{dC(x(t))}{dt} = 0,$$

or no measured change, regardless of how fast the measuring device is moving.

Alright, this appears to make sense, how to get it from the limit? Use the trick of multiplying by one to make the limit into the product of two limits. Here is the series of steps:

$$\frac{dC(x(t))}{dt} = \lim_{\Delta t \longrightarrow 0} \frac{C(x(t+\Delta t)) - C(x(t))}{\Delta t}$$

becomes

$$\frac{dC(x(t))}{dt} = \lim_{\Delta t \longrightarrow 0} \left[\left(\frac{C(x(t+\Delta t)) - C(x(t))}{x(t+\Delta t) - x(t)} \right) \left(\frac{x(t+\Delta t) - x(t)}{\Delta t} \right) \right]$$

now split the limit up into two limits (remember this is OK only if both exist)

$$\frac{dC(x(t))}{dt} = \lim_{\Delta t \longrightarrow 0} \left(\frac{C(x(t+\Delta t)) - C(x(t))}{x(t+\Delta t) - x(t)} \right) \lim_{\Delta t \longrightarrow 0} \left(\frac{x(t+\Delta t) - x(t)}{\Delta t} \right)$$

and rewrite as

$$\frac{dC(x(t))}{dt} = \lim_{\Delta t \longrightarrow 0} \left(\frac{C(x+\Delta x) - C(x)}{\Delta x} \right) \lim_{\Delta t \longrightarrow 0} \left(\frac{x(t+\Delta t) - x(t)}{\Delta t} \right).$$

Finally, use the definition of the derivative and you get back the result we proposed on physical grounds. Even if you like following through mathematical arguments, the one part that may puzzle you is how I just changed $C(x(t+\Delta t))$ to $C(x+\Delta x)$. The answer is, that since neither Δt or Δx are actually given values in the limiting process it is OK to relabel (say $x(t+\Delta t) = 5$ and $x(t) = 3$, then you would think nothing of writing $x(t+\Delta t) = 3+2 = x(t) + \Delta x$ where $\Delta x = 2$). Indeed some of the trickiest mathematical arguments are often just fancy versions of the idea of relabeling.

Example 1 $C(x) = 5 + \sin(x)$ and $x(t) = t^2$. Then

$$\frac{dC}{dx} = \cos(x)$$

or

$$\frac{dC}{dx} = \cos(t^2)$$

once we use $x(t) = t^2$. Furthermore,

$$\frac{dx}{dt} = 2t$$

and thus the rate of change of $C(x(t))$ with respect to time is

$$\frac{dC(x(t))}{dt} = 2t\cos(t^2)$$

where I placed the $2t$ term out in front of the cosine to agree with what is commonly done in books.

Example 2 Find the rate of change of $C(x(t))$ at the times $t = 1$ and $t = 10$. On the surface this is very easy, just evaluate $2t\cos(t^2)$ at the prescribed times. However I want to do it in pieces to prove a point. Remember that

$$\frac{dC(x(t))}{dt} = \frac{dC}{dx}\frac{dx}{dt}$$

and that the derivative is really a machine that takes functions as input and produces other functions as output. Above we have two functions $C'(x)$ and $x'(t)$ where the apostrophe or 'prime' denotes rate of change with respect to the input (pure mathematicians would say, "with respect to the argument"). **But the input is different for the two functions!** Thus to evaluate at $t = 1$ we evaluate $x'(t) = 2t$ at $t = 1$ to get $x'(1) = 2$. Similarly at $t = 10$ we get $x'(10) = 20$. For $C'(x)$ we have to be more careful. First we evaluate $x(1) = 1$ and $x(10) = 100$. Then we recall that $C'(x) = \cos(x)$ and use this to evaluate $C'(1) = \cos(1)$ and $C'(100) = \cos(100)$.

Example 3 Often we use the chain rule without even thinking about it. Say I wanted you to find the rate of change of $f(x) = \cos(2\pi x)$. You know that the rate of change of cosine is the negative of sine, but you may or may not be able to guess how the multiple of 2π will work out. A useful way to get the right answer (and not just for beginners, I still force myself to do this in my own work), is to **define a new variable**, in this case $z = 2\pi x$ and a new function $g(z(x)) = f(x)$. Then to find $f'(x)$ I use the chain rule on $g(z(x))$. This gives

$$\frac{dg(z(x))}{dx} = \frac{dg}{dz}\frac{dz}{dx}$$

or using $g(z) = \cos(z)$ and $z = 2\pi x$

$$\frac{dg(z(x))}{dx} = -\sin(z)2\pi.$$

Finally we use the fact that

$$f'(x) = \frac{dg(z(x))}{dx}$$

to conclude that $f'(x) = -2\pi\sin(2\pi x)$.

Example 4 Upping the ante a bit on the above argument, we consider $f(x) = \sin(\sin(x))$. We let $g(z) = \sin(z)$ where $z(x) = \sin(x)$ and the chain rule then gives

$$\frac{dg(z(x))}{dx} = \frac{dg}{dz}\frac{dz}{dx}$$

or

$$\frac{dg(z(x))}{dx} = \cos(z)\cos(x)$$

and this allows us to conclude $f'(x) = \cos(\cos(x))\cos(x)$.

In conclusion, here are some hints to using the chain rule effectively.

- Write out the problem as a series of inputs and outputs
- Decide how many steps are being done
- Remember that rates of change can only be computed with respect to the input
- Follow through until the original input is reached. This produces a bunch of rates of change multiplied together
- If the result is being evaluated remember to evaluate each piece at the appropriate input.

Math 137 Physics Based Section

Marek Stastna

14 THE CHAIN RULE

As impressive as the mechanics applications of the rate of change, or derivative, concept are, they only scratch the surface. In this section we open up a whole world, both of applications and pure mathematics, by looking at the so-called **Chain Rule**. The Chain Rule is a fundamental building, and is often placed on a high pedestal as a result. However, it is, at root, a very simple idea.

Consider a measuring device moving along a time dependent path $x(t)$ and measuring a quantity $C(x)$ which varies with space, but is assumed not to vary in time. An example of this is a weather balloon, or airplane, that measures the concentration of stable trace gases in the atmosphere. A more exciting, and exotic one is the GRACE II satellite which measures the variations in the gravitational field of the Earth (the interested can ponder for a few moments how such a thing could be done) due to things like underground aquifers (whose size and rate of change is important for city planning). Of course the assumption of C being a function of one variable only is overly simplistic, still it is a good place to start.

OK, what do we have? Well we have a single input, namely t, for time, being fed into the machine called $x(.)$ which is assumed to be given by a recipe (like $x(t) = t^2$, say) and that outputs the position $x(t)$. Next we have a machine $C(.)$, representing concentration of a quantity of interest (ozone, nitrous oxide, etc.) that requires the position as input. We ask: **What is the rate of change of $C(x(t))$ with time?**

We could make a diagram of inputs and outputs like

$$t \longrightarrow x(.) \longrightarrow x(t) \longrightarrow C(.) \longrightarrow C(x(t))$$

and even write the definition of limit as

$$\frac{dC(x(t))}{dt} = \lim_{\Delta t \longrightarrow 0} \frac{C(x(t+\Delta t)) - C(x(t))}{\Delta t}$$

provide the limit exists. Again, if we wanted an approximation then we could just take Δt very small and evaluate the right hand side (see assignment). However, long before calculators, people wanted a better expression for the right hand side. There are two ways to get it.

The first argues that, physically, we should **only be able to calculate the rate of change of a machine with respect to its input**. This means that the rate of change of $C(x(t))$ with respect to time must be done in two parts. First we calculate the rate of change of $C(.)$ with respect to its true input, namely position, and then we multiply by the rate of change of position with time.

Does this make sense? Let's say $C(x) = x$, then the above rule says that

$$\frac{dC(x(t))}{dt} = \frac{dC}{dx}\frac{dx}{dt}$$

and since $C(x) = x$

$$\frac{dC}{dx} = 1$$

and

$$\frac{dC(x(t))}{dt} = v(t)$$

the velocity. This makes perfect sense because in this case $C(.)$ was just the identity function and any change in C was due to how fast the measuring device was moving. Notice that if $C(x)$ =constant, then

$$\frac{dC}{dx} = 0$$

and

$$\frac{dC(x(t))}{dt} = 0,$$

or no measured change, regardless of how fast the measuring device is moving.

Alright, this appears to make sense, how to get it from the limit? Use the trick of multiplying by one to make the limit into the product of two limits. Here is the series of steps:

$$\frac{dC(x(t))}{dt} = \lim_{\Delta t \longrightarrow 0} \frac{C(x(t+\Delta t)) - C(x(t))}{\Delta t}$$

becomes

$$\frac{dC(x(t))}{dt} = \lim_{\Delta t \longrightarrow 0} \left[\left(\frac{C(x(t+\Delta t)) - C(x(t))}{x(t+\Delta t) - x(t)} \right) \left(\frac{x(t+\Delta t) - x(t)}{\Delta t} \right) \right]$$

now split the limit up into two limits (remember this is OK only if both exist)

$$\frac{dC(x(t))}{dt} = \lim_{\Delta t \longrightarrow 0} \left(\frac{C(x(t+\Delta t)) - C(x(t))}{x(t+\Delta t) - x(t)} \right) \lim_{\Delta t \longrightarrow 0} \left(\frac{x(t+\Delta t) - x(t)}{\Delta t} \right)$$

and rewrite as

$$\frac{dC(x(t))}{dt} = \lim_{\Delta t \longrightarrow 0} \left(\frac{C(x+\Delta x) - C(x)}{\Delta x} \right) \lim_{\Delta t \longrightarrow 0} \left(\frac{x(t+\Delta t) - x(t)}{\Delta t} \right).$$

Finally, use the definition of the derivative and you get back the result we proposed on physical grounds. Even if you like following through mathematical arguments, the one part that may puzzle you is how I just changed $C(x(t+\Delta t))$ to $C(x+\Delta x)$. The answer is, that since neither Δt or Δx are actually given values in the limiting process it is OK to relabel (say $x(t+\Delta t)=5$ and $x(t)=3$, then you would think nothing of writing $x(t+\Delta t)=3+2=x(t)+\Delta x$ where $\Delta x=2$). Indeed some of the trickiest mathematical arguments are often just fancy versions of the idea of relabeling.

Example 1 $C(x)=5+\sin(x)$ and $x(t)=t^2$. Then

$$\frac{dC}{dx}=\cos(x)$$

or

$$\frac{dC}{dx}=\cos(t^2)$$

once we use $x(t)=t^2$. Furthermore,

$$\frac{dx}{dt}=2t$$

and thus the rate of change of $C(x(t))$ with respect to time is

$$\frac{dC(x(t))}{dt}=2t\cos(t^2)$$

where I placed the $2t$ term out in front of the cosine to agree with what is commonly done in books.

Example 2 Find the rate of change of $C(x(t))$ at the times $t=1$ and $t=10$. On the surface this is very easy, just evaluate $2t\cos(t^2)$ at the prescribed times. However I want to do it in pieces to prove a point. Remember that

$$\frac{dC(x(t))}{dt}=\frac{dC}{dx}\frac{dx}{dt}$$

and that the derivative is really a machine that takes functions as input and produces other functions as output. Above we have two functions $C'(x)$ and $x'(t)$ where the apostrophe or 'prime' denotes rate of change with respect to the input (pure mathematicians would say, "with respect to the argument"). **But the input is different for the two functions!** Thus to evaluate at $t=1$ we evaluate $x'(t)=2t$ at $t=1$ to get $x'(1)=2$. Similarly at $t=10$ we get $x'(10)=20$. For $C'(x)$ we have to be more careful. First we evaluate $x(1)=1$ and $x(10)=100$. Then we recall that $C'(x)=\cos(x)$ and use this to evaluate $C'(1)=\cos(1)$ and $C'(100)=\cos(100)$.

Example 3 Often we use the chain rule without even thinking about it. Say I wanted you to find the rate of change of $f(x)=\cos(2\pi x)$. You know that the rate of change of cosine is the negative of sine, but you may or may not be able to guess how the multiple of 2π will work out. A useful way to get the right answer (and not just for beginners, I still force

myself to do this in my own work), is to **define a new variable**, in this case $z = 2\pi x$ and a new function $g(z(x)) = f(x)$. Then to find $f'(x)$ I use the chain rule on $g(z(x))$. This gives

$$\frac{dg(z(x))}{dx} = \frac{dg}{dz}\frac{dz}{dx}$$

or using $g(z) = \cos(z)$ and $z = 2\pi x$

$$\frac{dg(z(x))}{dx} = -\sin(z)2\pi.$$

Finally we use the fact that

$$f'(x) = \frac{dg(z(x))}{dx}$$

to conclude that $f'(x) = -2\pi \sin(2\pi x)$.

Example 4 Upping the ante a bit on the above argument, we consider $f(x) = \sin(\sin(x))$. We let $g(z) = \sin(z)$ where $z(x) = \sin(x)$ and the chain rule then gives

$$\frac{dg(z(x))}{dx} = \frac{dg}{dz}\frac{dz}{dx}$$

or

$$\frac{dg(z(x))}{dx} = \cos(z)\cos(x)$$

and this allows us to conclude $f'(x) = \cos(\sin(x))\cos(x)$.

In conclusion, here are some hints to using the chain rule effectively.

- Write out the problem as a series of inputs and outputs
- Decide how many steps are being done
- Remember that rates of change can only be computed with respect to the input
- Follow through until the original input is reached. This produces a bunch of rates of change multiplied together
- If the result is being evaluated remember to evaluate each piece at the appropriate input.

15 APPLYING THE CHAIN RULE: DEs

The Chain Rule allows us to greatly expand the variety of differential equations. In the context of Newton's Second Law, all we need to do is notice that the force is the derivative

of some complicated function that can be solved by chain rule. Here's an example where we work backwards:

Example 1 Given the functional form of the velocity, $v(t) = \sin(t^2)$ we know that the acceleration is the rate of change of velocity with time. To apply the chain rule, let $b(t) = t^2$, then

$$a(t) = \frac{dv}{db}\frac{db}{dt}$$

by the Chain Rule. Evaluating gives

$$a(t) = \cos(t^2)2t$$

and Newton's Second Law finally gives $F(t) = ma(t)$ where m is the particle's mass (assumed constant), or

$$F(t) = 2mt\cos(t^2).$$

While this is an interesting result, it suggests that one should know the answer ahead of time! This isn't really how problems get solved (or is it?), so we need a way to rewrite the problem that will allow us to go beyond pure guess-work. What we want is a mnemonic device to help us remember as well as a means to test hypotheses (fancy language for guessing). Let's rewrite Newton's Second Law as

$$\frac{dv}{dt} = \frac{F(t)}{m}$$

and now let's allow ourselves the freedom to split up the rate of change as if it were a regular fraction. For approximations this is certainly valid (because Δt is just a small number), and indeed even in the limit $\Delta t \longrightarrow 0$ can be made mathematically rigorous. If we agree to accept this trick, then we write

$$dv = \frac{F(t)}{m}dt.$$

Now let's return to Example 1. There we found $F(t) = 2mt\cos(t^2)$ and so

$$dv = [\cos(t^2)](2t)dt$$

where I just added a bunch of unnecessary seeming brackets. But here's the trick: The cosine in the square brackets is a function not of t directly, but t^2, and we will define $b(t) = t^2$, like in Example 1. Of course, we know that the rate of change of a sine with respect to its argument is the cosine, but since b is a function of t, we must multiply by the rate of change of $b(t)$ with respect to time. Now, notice that what is in the round brackets is the rate of change of b with respect to time. This means that by defining $b(t) = t^2$ we rewrite the differential equation as

$$dv = \frac{d\sin b}{db}\frac{db}{dt}dt.$$

But if we allow ourselves the luxury of treating the rate of change like a fraction then, for mnemonic purposes anyway, we cancel the dt terms and are left with

$$dv = \frac{d\sin b}{db}db.$$

Moving the db back to the denominator on the left we get

$$\frac{dv}{db} = \frac{d \sin b}{db}$$

in which it is very easy to reverse the derivative machine to get

$$v(b) = \sin b + C$$

where C is a constant which we must get from the initial conditions. This is the answer to the problem given in terms of the new variable b. This is why the method is called **THE CHANGE OF VARIABLES**. To produce an answer in terms of the original, physical, variable t we just use $b(t) = t^2$ and for the case $v(0) = 0$ we get

$$v(t) = \sin (t^2)$$

as was originally assumed in Example 1.

So may be we don't need to know the answer ahead of time after all. Let's try another:

Example 2 Consider

$$F(t) = \frac{-6t^2 m}{(t^3 + 1)^2}$$

and $v(0) = 1$. Newton's Second Law reads

$$\frac{dv}{dt} = \frac{F(t)}{m}$$

or

$$dv = [\frac{1}{(t^3 + 1)^2}](-6t^2)dt$$

where I have written the example out using the square and round brackets according to what I am going to try for my change of variables. You might think that it is unfair for me to do that, since you may do not have the ten years of post-calculus experience I have. Sadly, change of variables problems are **inductive** in nature and you get better at spotting the patterns once you have done a few. To help you spot the pattern I will rewrite the differential equation one more time,

$$dv = [(t^3 + 1)^{-2}](-6t^2)dt.$$

In the square brackets we have a polynomial to the power of -2 while in the round brackets we have another polynomial (a monomial really). The rate of change with respect to time of $t^3 + 1$ is, using the power rule $3t^2$ which isn't quite what's in the round brackets, but at least the powers match up.

At this point we get our confidence up and say, let's try $b(t) = t^3 + 1$. Then the square brackets look like b^{-2} and using the power rule of derivatives we write this as

$$b^{-2} = -\frac{d(b^{-1})}{db}.$$

We also know that

$$\frac{db}{dt} = 3t^2$$

or

$$db = 3t^2 dt$$

so that

$$-6t^2 dt = -2db.$$

Putting it all together gives

$$dv = -\frac{d(b^{-1})}{db}(-2db)$$

and simplifying

$$\frac{dv}{db} = 2\frac{d(b^{-1})}{db}$$

which can be solved to give

$$v(b) = \frac{2}{b} + C.$$

Since we are interested in the physical variables we use $b(t) = t^3 + 1$ to get

$$v(t) = \frac{2}{t^3 + 1} + C$$

and finally to satisfy $v(0) = 1$ we must have

$$1 = \frac{2}{0 + 1} + C$$

so that $C = -1$ and

$$v(t) = \frac{2}{t^3 + 1} - 1.$$

Here it is step by step

1. make sure the derivative machine cannot be reversed by simpler methods
2. by guesswork, insight, or experience rewrite the equation using the square and round brackets (a good rule for now that you can drop later)
3. Define your new variable $b(t)$ and find its derivative
4. Use the derivative of $b(t)$ to replace dt and the round brackets with db
5. Rewrite the square brackets in terms of b and hopefully as a derivative
6. Solve the new easier problem

I am making no claims that this is easy. It can be frustrating, but its usefulness is difficult to deny, especially once you start using the patterns that others before you have dreamt up.

16 APPLYING THE CHAIN RULE: THE EXPONENTIAL

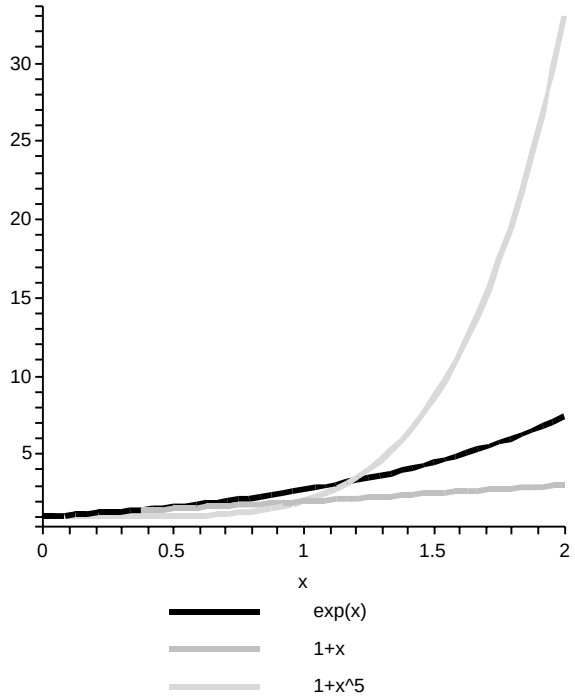
Near the start of the course we had a brief discussion about just what sorts of functions we can actually build from what we knew. We came up with polynomials and rational functions (fractions made up of polynomials) right away. Using the geometry of the circle we defined the sine and cosine functions (recall they measured the coordinates along the circle in terms of the the angle in radians, which is really a way to measure *distance*). However, this is where we left things. Now that we know about the derivative (or rate of change) machine that takes functions as input and produces functions as output, we can greatly expand the family of functions we have. We can do this by **defining a function as solution to a differential equation!**

You probably have some doubts about this enterprise at this point. In reality we will only consider one very simple case, but it will give us a very rich set of functions. Toward this end, consider the differential equation

$$\frac{df(x)}{dx} = f(x) \tag{1}$$

or in other words, the statement that the rate of change of $f(.)$ is actually given by $f(.)$ itself. From a practical point of view this is very reasonable. Consider a monomial like $g(x) = x^n$. A simple application of the power rule of derivatives gives $g'(x) = nx^{n-1}$. Now let's think about what happens when the input is large. For concreteness let's take $n = 5$. This means $g'(x) = 5x^4$ and so the rate of change is the input multiplied by itself four times and then multiplied by five. When the input is big (say a thousand) then $g(x)$, which is the input multiplied by itself five times, will give an output that is **much larger** than $g'(x)$. Thus $g'(x) \ll g(x)$ when x is large.

Thus the function for which the differential equation (1) is true grows much faster than any polynomial provided we make the input large enough. Here is a picture:



I picked the exponential and the two polynomials (one linear, or first order, and one fifth order) to pass through the point $(0, 1)$. For $0 < x < 1$ the three are reasonably close to one another, however you can see that as x grows beyond 1.5 or so, the exponential function takes off and grows much faster than the polynomials shown. I could have kept the agreement better for longer by choosing something like $1+x^{100}$, but no matter what I chose there would be some value of x beyond which the exponential wins out.

There is one small matter to take care of with defining the exponential as the solution to a differential equation and that is the arbitrary constant that results from putting the derivative machine in reverse, since

$$\frac{dA}{dx} = 0$$

for any constant A. We simply choose the exponential to have the value 1 at $x = 0$.

Now let's see how the chain rule comes in. We write the exponential like $\exp(x)$ and ask, "What kind of differential equation does $h(x) = \exp(kx)$ satisfy if k is a constant (like 3 or 42.6 or even π)?"

Let's define a new variable $z = kx$. Then $h(z) = \exp(z)$ and

$$\frac{dz}{dx} = k.$$

By the Chain rule,

$$\frac{dh}{dx} = \frac{dh}{dz}\frac{dz}{dx}$$

or

$$\frac{dh}{dx} = kh(x).$$

Thus the exponential isn't just one function, it is a whole family of functions, all of whom have the property that their rate of change is proportional to the function value.

Example 1 You've probably already seen exponentials introduced a different way. Let's see if we can derive the property $\exp(5x) = \exp(3x)\exp(2x)$. Start with the definition to state

$$\frac{d\exp(5x)}{dx} = 5\exp(5x).$$

next feed $\exp(3x)\exp(2x)$ into the derivative machine and apply the product rule to get

$$\frac{d}{dx}[\exp(3x)\exp(2x)] = \frac{d\exp(3x)}{dx}\exp(2x) + \exp(3x)\frac{d\exp(2x)}{dx}$$

now using the definition of the exponential on the right hand side, we find

$$\frac{d}{dx}[\exp(3x)\exp(2x)] = 3\exp(3x)\exp(2x) + \exp(3x)2\exp(2x)$$

or

$$\frac{d}{dx}[\exp(3x)\exp(2x)] = 5\exp(3x)\exp(2x).$$

But wait we, just showed that when the derivative machine is applied to $\exp(3x)\exp(2x)$ we get, as output, five times the original function. This is exactly what $\exp(5x)$ should give when it is used as input in the derivative machine. Thus $\exp(3x)\exp(2x) = \exp(5x)$, and indeed in general,

$$\exp(ax)\exp(bx) = \exp([a+b]x) \tag{2}$$

for any two real numbers a and b.

The hard core mathematician in you may protest that we don't really know that the solution to the differential equation (1) is unique (perhaps something like how some quadratic equations have multiple solutions). It is a simple trick to prove that this is the case. Alternatively you can take it as a given fact (indeed uniqueness of the solution can be proven for a very broad variety of differential equations).

Finally, you could ask, "What does the exponential give as output when negative values are input?" One way to answer this is to start at $(0, 1)$ and take tiny steps in the *negative* direction. Another is to let $h(x) = \exp(-x)$ and $s = -x$ to get the differential equation

$$\frac{dh}{ds} = -h(s)$$

where $h(s = 0) = h(x = 0) = 1$. From this we see that since $h(.)$ has a rate of change that is the negative of its value, starting at $(0, 1)$ will lead to a decrease in $h(.)$ as the inputs increase. However, since the decrease is proportional to the value output by $h(.)$ the decrease

is **always smaller than 1** and indeed the large the input, the smaller the rate of decrease. Indeed you can take as given that

$$\lim_{x \longrightarrow -\infty} \exp x = 0$$

but that you never actually hit zero.

In conclusion, for the full set of inputs $-\infty < x < \infty$ the output of the exponential can be split into two regions, one in which the function grows rapidly, and one in which it grows slowly.

17 THE LOGARITHM

Now that we have the exponential we should at least attempt to look for its inverse. Do we expect one to exist? Well, if $f(x) = \exp xk$ then $f'(x) = kf(x)$ and $f(0) = 1$. Thus we have that the rate of change of $f(x)$ is positive whenever $f(x)$ is positive, assuming $k > 0$. Starting with $f(0) = 1$ we find that the exponential is growing, or increasing, as x, the input, increases. If this is the case we expect $f(x)$ to be invertible. Moreover, recall that exponentials grew faster than any polynomial (once the input was large enough). This means that we expect the inverse to grow **very slowly**.

Let's be a bit more concrete. To simplify the algebra (this is no big deal, really) we take $k = 1$. The definition of inverse gives us

$$f^{-1}(f(x)) = x$$

and now we can apply the derivative machine (with respect to x) to both sides of the equation. You should note that you can always apply the derivative to the two sides of an equation **BUT** if the original equation has no solutions then any rate of change you work out from it is meaningless. For the present we know that the exponential has all real numbers as allowable inputs and its set of outputs are all positive real numbers. Hence both the function and the inverse can always be computed, so there is nothing to worry about.

Once we've settled the "Can we do this legally?" question, we note that upon applying the derivative machine to both sides we find that the right hand side just gives 1, but on the left we must use the chain rule. If we let $y = f(x)$ then the left hand side reads

$$\frac{d}{dx} f^{-1}(f(x)) = \frac{df^{-1}(y)}{dy} \frac{dy}{dx}.$$

However, from the definition of the exponential,

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

Thus, if we put all this together we get

$$\frac{df^{-1}(y)}{dy}y = 1$$

or

$$\frac{df^{-1}(y)}{dy} = \frac{1}{y}. \tag{3}$$

This is an interesting, though in hindsight not unexpected result. The inverse of the exponential has a rate of change that is proportional to the reciprocal of its value. So this function grows very fast when the input is between 0 and 1 and then the growth slows more and more, the larger the value becomes. The inverse is important enough to have its own name, the natural logarithm, and is generally written as $\ln(y)$.

The idea of differentiating the definition of the inverse using the derivative as a machine was a very good one (it's a good bet that this author didn't think of it). It can be done in general, like this

$$\frac{d}{dx}g^{-1}(g(x)) = \frac{dx}{dx}$$

or by Chain Rule, writing $G(y) = g^{-1}(g(x))$,

$$G'(g(x))g'(x) = 1.$$

Thus if we choose an input point, and compute $g(x)$, then feeding x into the rate of change of g and feeding $g(.)$ into the rate of change of $g^{-1}(.)$ and multiplying the two numbers will always give one.

Example 1 Consider $D = \ln a + \ln b$, then $\exp D = \exp(\ln a + \ln b)$, just by applying the exponential machine to both sides of the equation. However, we know that the exponential of a sum is just the product of the two exponentials. So,

$$\exp D = \exp(\ln a)\exp(\ln b).$$

But $\ln(.)$ is the inverse of $\exp(.)$ so

$$\exp D = ab$$

or on applying the ln machine to both sides and replacing D with its definition

$$\ln a + \ln b = \ln(ab).$$

Example 2 We are now in a position to define a^b for cases in which b is not a fraction (something strange like 2^π). Let

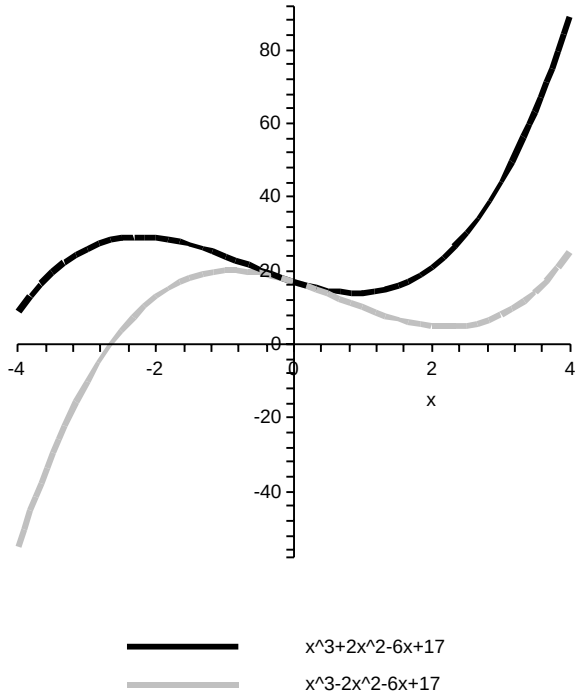
$$a^b = \exp(b\ln(a)).$$

You can readily check that for fractions the usual interpretation in the sense of repeated multiples and appropriate roots can be recovered.

Example 3 Recall that as x gets large and positive $\exp(x)$ grows without bound. However, as x gets large and negative, $\exp(x)$ tends to zero. This means the set of outputs of the exponential is bounded between zero and positive infinity. Hence the **domain** of the natural logarithm is given by $0 < y < \infty$ while the **range** is all the real numbers $(-\infty < \ln(y) < \infty)$.

18 SKETCHING USING DERIVATIVES

For very simple functions like parabolas, lines and even sine, cosine, the exponential and logarithm, the definitions of the functions tell us how to sketch their graphs. What I mean by this is that, given a reasonable amount of time you could do it from scratch. However even for moderately complex functions we can run into trouble quickly. Consider the graph of two very similar looking polynomials:



Can we sketch the graphs using what we know about derivatives and their implications of functions being increasing and decreasing?

Consider $p(x) = x^3 + 2x^2 - 6x + 17$. The derivative is easy to compute and reads $p'(x) = 3x^2 + 4x - 6$. The derivative is also polynomial, though a somewhat simpler one than $p(x)$. We can now ask at what inputs $p(x)$ is zero (I will call them x^*). We really want to know where $p'(x) > 0$ and $p(x)$ is increasing, but finding the places where $p'(x) = 0$ is a good start. You can use the quadratic formula to get $x^* = (1/3)(-2 \pm \sqrt{22})$, or on evaluating

$x^* \approx -2.23$, 0.897. Now look at what happens when $x > 0$ and large. Clearly the $3x^2$ term wins out and $p'(x) > 0$. Similarly, when $x < 0$ and large in magnitude the $3x^2$ term, which is positive even though $x < 0$, wins out and $p'(x) > 0$.

Now let's think about what it would take for $p'(.)$ to change sign. One possibility is that as a function, $p'(.)$ jumps all over the place, or in other words, is not continuous. From experience, you would likely argue that this is not true for polynomials, and so the only possibility is that $p'(.)$ changes sign by passing through a place where $p'(x) = 0$. For our case this means $p'(x)$ changes sign at the two values of x^*, or approximately at -2.23 and 0.897 (even though these values are approximate I will use them as the exact values for this example; I could have doctored the problem so as to get "nice" numbers, but I think this is misleading). What is more these are the **only** places where $p'(x)$ can change sign.

Putting it all together we conclude that $p(x)$ is increasing for $x < -2.23$, decreasing for $-2.23 < x < 0.897$ and increasing for all $x > 0.897$. As $x \longrightarrow \infty$ $p(x)$ grows like its term with the highest power, namely x^3. As $x \longrightarrow -\infty$ $p(x)$ grows in magnitude, but its value is negative, because its term with the highest power, namely x^3, is an odd power. It remains only to evaluate $p(x)$ at the two values of x^* and get $p(-2.23) \approx 29.24$ and $p(0.897) \approx 13.95$ and connect these points in the appropriate way. You should try it and convince yourself that you get the black curve in the above figure.

Once you've got that worked out consider the grey curve, namely the graph of $p_2(x) = x^3 - 2x^2 - 6x + 17$. With some calculator work you can convince yourself that $p'(x) = 0$ when $x \approx -0.897$ and $x \approx 2.23$. This should set some bells off, because it looks just like the previous example except with the signs backwards. So how will the two graphs be related (after all if we want an exact graph we can just feed the two into a program like Matlab)? Well, the term with the highest power is x^3 in both cases so the two must have the same behaviour as $|x| \longrightarrow \infty$. What is more, both have $p'(x) >$ for $|x|$ large, and have two points at which $p'(x) = 0$, and hence $p'(x) < 0$ in between and only in between these two points. At this point you can reexamine the graph I showed you at the start of the section and think, "OK, now I know why everything looks the way it does."

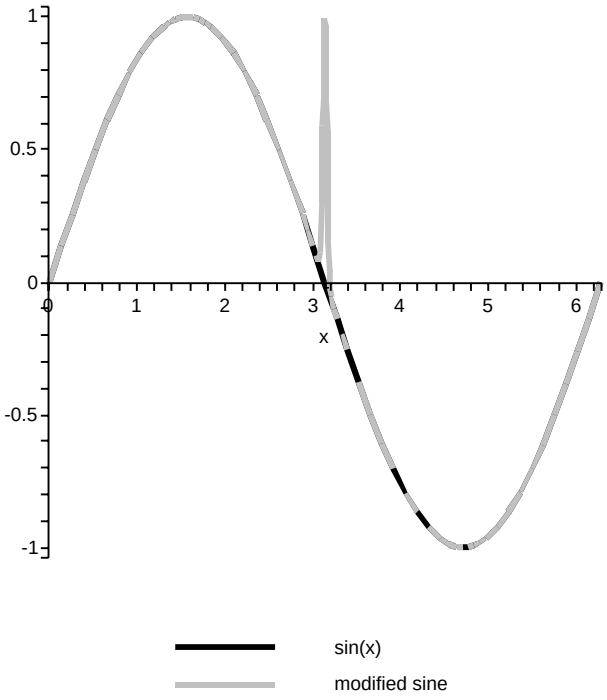
The points at which $p'(x) = 0$ are called **critical points**. They often given places where the function has a local maximum or minimum. Not always, though, just think of $f(x) = x^3$ so that $f'(x) = 3x^2$ and even though $f'(0) = 0$ and $x = 0$ is a critical point, we notice that $f'(x) > 0$ for all $x \neq 0$ and thus the function is increasing, and $x = 0$ is neither a maximum or minimum.

The reason for the designation "local" is that when we have more than one critical point, or when we restrict the set of inputs, the overall (or global) maximum and minimum, does not need to occur at the critical points. Indeed for our example this was the case.

Consider $p(x)$ restricted to the set of inputs $-10 \leq x \leq 10$. An examination of the graph indicates that the global maximum of 1157 and global minimum of -723 occur at $x = 10$ and $x = -10$ respectively. **Thus in a situation in which the inputs are restricted you**

must check any endpoints to successfully determine the global maximum and minimum.

As a little aside, note that is often said that the derivative is useful in situations like



where two functions are nearly identical over most of the domain, but vary wildly over a short distance. This is true, but it seems to me that it is better to think of using derivatives in sketching as part of an overall package of tools that let you do things that a computer may not be able to do (like determine symmetries, look at limiting behaviour and so on).

Finally, what about those nasty situations in which we get neither a maximum or minimum? A simple fix for the $f(x) = x^3$ example is to take the second derivative, i.e. $f''(x) = 6x$, from which we conclude that $f'(0) = f''(0) = 0$. You can contrast this with the two cases $g_1(x) = x^2$ and $g(x) = -x^2$ for which the point $x = 0$ is both a local and global minimum (for $g_1(x)$) or maximum (for $g_2(x)$). A quick calculation shows $x = 0$ is a critical point for all three functions, but notice $g_1''(x) = 2$ and $g_2''(x) - 2$. Thus we can conjecture (check for the proof in your text if you wish) that:

Second Derivative Test If $f'(x^*) = 0$ then $f''(x^*) > 0$ means that $(x*, f(x^*))$ is a local maximum point and $f''(x^*) < 0$ means $(x^*, f(x^*))$ is a local minimum point. If $f''(x^*) = 0$ then the critical point is called an inflection point and is neither a local maximum or minimum.

19 SOLVING EQUATIONS AND CONTINUITY

From your own mathematical education, you would probably agree with me in saying that a big part of the mathematical enterprise is solving equations. However, it is an often overlooked fact that most interesting problems do not have solutions that can be written down as formulae. So what do we do?

Well, one thing mathematics can tell us is when we should expect answers at all. Towards this end consider the analogy between the graph of a function (say $f(x)$) and a string. The string is held so that its shape matches that of the graph and with no loss of generality (mathematicians love these sorts of silly phrases) we can take the height of the table to represent the horizontal line $y = 0$. To solve the equation $f(x) = 0$ we look for all the places where the graph passes through the x-axis, or alternatively, the string crosses the level of the table. Of course if I don't want to find the answer, but only to know if one can be found I argue like this. If I find a piece of string below the table and another piece above the table, then somewhere in between I am guaranteed that the string passes through the level of the table. This then suggests that if I can find a so that $f(a) < 0$ and b so that $f(b) > 0$ then I am guaranteed the existence of an input that I call c so that $a < c < b$ and $f(c) = 0$. The only catch is that I need the function $f(x)$ to share the properties of the string.

This is worth pursuing a bit further. Let's say we have a light switch that has been off for all $t < 0$ and suddenly is switched on at $t = 0$. We measure the output from the light bulb and call the energy output $E(t)$. Then $E(t) = 0$ for $t \leq 0$ and $E(t) = E_0$ for $t > 0$ where E_0 is a constant value. We have made a number of simplifications and assumptions, but overall our intuition suggests $E(t)$, as we defined it, will have a reasonable amount of utility to it. Now ask yourself if the equation

$$E(t) = 0.25E_0$$

will have a solution. We can rewrite this equation to fit into the above discussion, by moving the number on the right hand side to the left hand side, $E(t) - 0.25E_0 = 0$. Because we have defined $E(t)$ to reflect the experience of throwing a light switch, it starts at an output of 0 and instantaneously moves to an output of E_0 without passing any of the outputs in between. Indeed if you graph $E(t)$ you will not be able to draw the graph without lifting your pencil, because of this jump. Finally if you take the limit as $t \longrightarrow 0$ you will find the limit does not exist because coming from $t > 0$ you get the limiting value of E_0, while coming from $t < 0$ you get the limiting value of 0.

Functions with jumps are called **discontinuous functions**. There is nothing wrong with them, and indeed they can prove quite useful in describing physical phenomena (like throwing a light switch). However, they have different properties than continuous functions.

Definition A function $f(x)$ is continuous at $x = a$ if

$$\lim_{x \longrightarrow a} f(x) = f(a).$$

Definition A function $f(x)$ is continuous, if it is continuous at all its possible input points. Though on intervals care must be taken at the end points.

The discussion with the string above can now be converted into the form of a mathematical theorem:
INTERMEDIATE FUNCTION THEOREM If $f(x)$ is continuous and $f(a) < 0$ and $f(b) > 0$ then there is at least one $a < c < b$ so that $f(c) = 0$.

Of course in the real world we are usually asked to actually find c. The theorem tells us *absolutely nothing* about how to do that. Though we can implicitly use the theorem to construct the following algorithm:

1. Start with $a < b$ so that $f(a) < 0$ and $f(b) > 0$.
2. Go half way in between to $d = 0.5(a + b)$ and check what sign $f(d)$ has
3. if $f(d) < 0$ then repeat the process with d and b
4. if $f(d) > 0$ then repeat the process with a and d
5. otherwise $f(d) = 0$ and you're done

This simple algorithm is called **bisection**, because it repeatedly divides the interval in half. In practice you never find the exact solution, but settle for the place at which $|f(x)| < \epsilon$ for an ϵ of your choosing (like 10^{-6} say). Notice that the assumption of continuity is implicit in the algorithm (try it on $E(t) - E_0$ to see how it fails).

A better algorithm is Newton's method which computes the tangent line at a point and uses the place where the tangent line crosses the x-axis to provide the next guess for where $f(x) = 0$.

As mentioned above functions with sudden jumps prove to be quite useful in modeling on/off type phenomena, like throwing a switch in a circuit. The most famous of these is the Heaviside step function (after the english polymath, Oliver Heaviside), defined as:

Definition The Heaviside step function is given by $H(t) = 0$ when $t < 0$ and $H(t) = 1$ when $t > 0$. The input point $t = 0$ is not terribly important (using what you know about limits you should be able to tell me why). I prefer to take $H(0) = 0$ though other authors favour $H(0) = 1/2$.

Finally we should note that the point of view we have taken on solving equations is pretty much the same one we took when we looked at the differential equations of Newtonian Mechanics. Namely, we are aware that we do not know all the "proper" mathematical theory, however, with what we do know we can make useful conclusions about broad sets of physical problems. Thus in the present we identified the concept of continuity without making a

terribly big deal about it all. We used continuity in solving $f(x) = 0$ by the process of bisection, but noted that discontinuous functions, like the Heaviside step function, for which the bisection algorithm and intermediate value theorem fail, can prove very useful in the description and modeling of physical situations.

20 GEOMETRIC INTERPRETATIONS

While we defined the rate of change, or derivative, according to the usual definition, we have largely used the derivative in the form of a machine acting on functions to produce other functions, i.e.

$$\frac{d}{dt}t^3 = 3t^2.$$

and in reverse

$$\left(\frac{d}{dt}\right)^{-1} 3t^2 = t^3 + C$$

where C is any constant value. However, we noted that the derivative is really a generalization of the slope of a line and even used the derivative to approximate functions near a point. The argument goes like this:

- Assume we know $f(a)$ and $f'(a)$
- we can construct a line that passes through $(a, f(a))$ and that has slope $f'(a)$. This line has the equation $y - f(a) = f'(a)(x - a)$.
- This line touches the graph of $f(x)$ at the point $(a, f(a))$ and provided $f(.)$ is well behaved (technically this is called **differentiable**) the line will be close to the graph of $f(x)$ near $x = a$.

The problem with the above is that we do not know the error we are making, and an approximation without an error estimate is something like saying the radius of the Earth is 100m, beacause I don't care about the error. We will learn about how to estimate the error later in the calculus sequence. For now we note that the limit does rescue us as $x \longrightarrow a$.

Thus the derivative has an operational as well as a geometrical definition. Does the reverse of the derivative have a similar interpretation?

Let us return to mechanics. Recall that the velocity $v(t)$ is the rate of change of position $x(t)$ with respect to time, or $v(t) = x'(t)$. This means that if I know my position at $t = 0$, the time I turn on my stopwatch, then I can reverse the argument and write

$$\left(\frac{d}{dt}\right)^{-1} v(t) = x(t) + x(0). \tag{4}$$

Now let's say that I want to know the position at one time, say $t = T$. If I can't get an exact solution, then I wish to construct approximations to $x(t)$ and use the limiting process to evaluate them in the limit. Moreover I know the answer has to be given by (4) if out previous work is to be mathematically sound.

OK, let's divide the time interval between $t = 0$ and $t = T$ into sub-intervals. You can think of having ten parts, a hundred parts, whatever you wish. I will label the length of the sub-interval by Δt and I will ask myself: "What distance do I travel in this subinterval?"

Almost immediately I note that I cannot answer this question, because it is just rephrasing the original problem I wish to solve. However what I could say is that I can come up with **bounds** on the distance travelled in the sub-interval. If I assume the velocity does not jump in the sub-interval (or that $v(t)$ is continuous) then I will have exactly one velocity which is the largest attained in the interval, and one velocity which is the smallest attained in the interval. In mathematical language then

$$\min v(t) \leq v(t) \leq \max v(t) \text{ on the sub-interval.}$$

It is worth noting that this is like those tricks where you multiply the top and bottom of a fraction by the same thing; we are not saying anything new, just writing it in a way that suits us better. Now the distance travelled in the interval, which I will call Δx, can be bounded by taking the velocity to be either the maximum (upper bound) or minimum (lower bound) for all times in the interval. Again mathematically,

$$\Delta t \min v(t) \leq \Delta x \leq \Delta t \max v(t) \text{ on the sub-interval.}$$

Now to get the total distance travelled we would just add over the intervals. To write this mathematically let's define some terms. Say we have N intervals, so that $\Delta t = T/N$. Then for the ith interval we define v_i^+ to be the maximum and v_i^- to be the minimum velocity (note that we can have negative velocities). We will also take $x(0) = 0$. This is no big deal because in the end we can just add any non-zero value to $x(t)$ for cases in which $x(0)$ is not zero. The bounds read:

$$\Delta t \sum_{i=1}^{N} v_i^- \leq x(t) \leq \Delta t \sum_{i=1}^{N} v_i^+.$$

Now, we could certainly use the above bounds to *approximate* $x(t)$, and we could get some very good approximations without too much trouble (see assignment). However we could also recall that the squeeze theorem tells us that if the limit of the lower bound equals that of the upper bound, then the limit of whatever is trapped in between must equal the same value. So if we call the lower and upper bound sums S^- and S^+, respectively, then

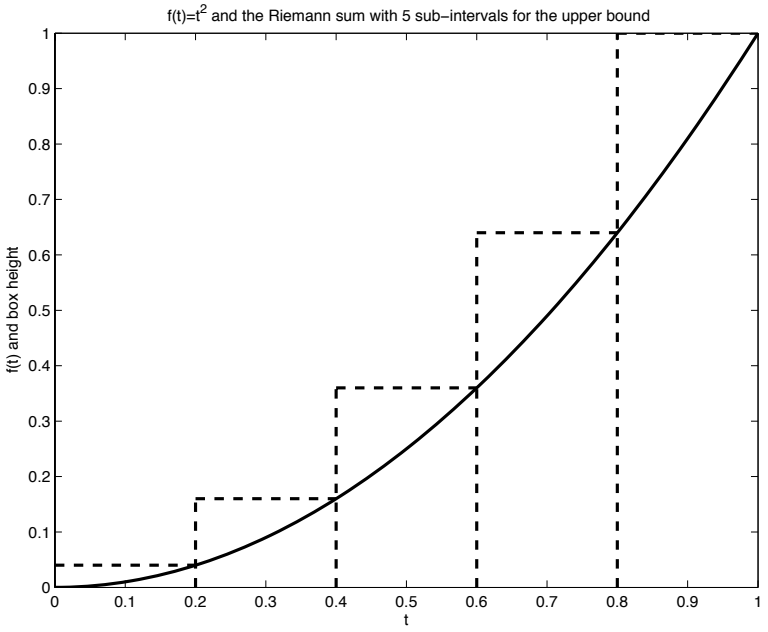
$$\begin{aligned} x(t) &= \lim_{\Delta t \longrightarrow 0} S^- \\ x(t) &= \lim_{\Delta t \longrightarrow 0} S^+ \end{aligned} \tag{5}$$

provided

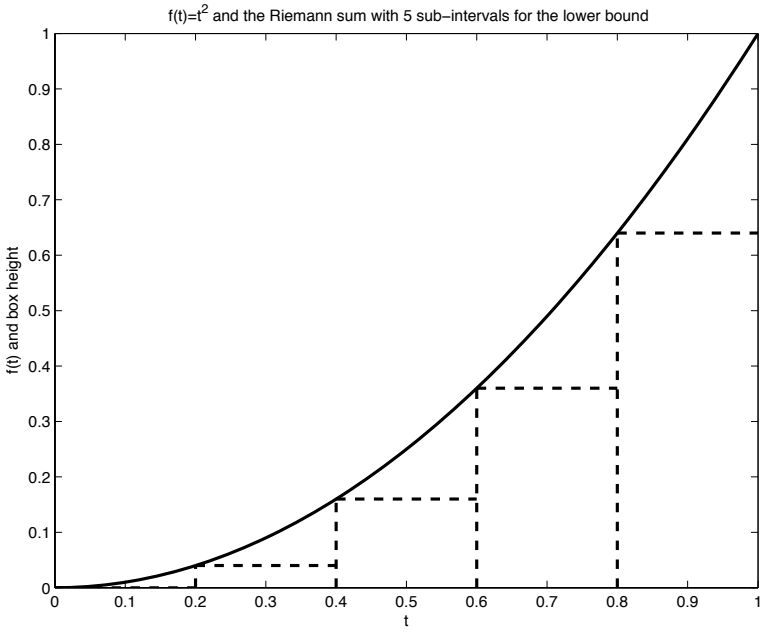
$$\lim_{\Delta t \longrightarrow 0} S^- = \lim_{\Delta t \longrightarrow 0} S^+.$$

OK, that seems fine and dandy for a pure mathematician (incidentally the upper and lower sumes are called **Riemann sums** after a very famous mathematician), but where is the geometric interpretation we have been looking for?

Let's return to the bounds on the sub-interval. Here you can think of graphing $v(t)$ and drawing two boxes. Both will have the bottom at $x = 0$ but one will have the top at the horizontal line $y = v^-$ and the other at $y = v^+$. The area of the bigger box is $\Delta t \times v^+$ and the area of the smaller box is $\Delta t \times v^-$. Here are a couple of pictures. First the larger boxes:



then the smaller boxes:



From the pictures, and writing out the sum we see that the bounds on $x(t)$ are given by the sum of the areas of the bigger and smaller boxes for each sub-interval. Thus we can interpret $x(t)$ as **the area below the curve** $v(t)$ **and above the t-axis**. Moreover since $x(t)$ is given by the inverse of the derivative machine we can interpret the inverse of the derivative machine as giving the area beneath the graph of the function which we are feeding into the inverse derivative machine.

Of course there is a much more complicated way of saying all this. We call the inverse of the derivative machine the **indefinite integral** and the geometrical interpretation the **definite integral**. We write the former as

$$\int 3t^2 dt = t^3 + C$$

where C is an arbitrary constant. The latter is written as

$$\int_0^T v(t)dt = x(T) - x(0).$$

and if the two limits agree we call the function $v(t)$ **integrable**.

The relation between the two "types" integral (which mechanics has already shown us are really part of the same package) is called

THE FUNDAMENTAL THEOREM OF CALCULUS (FTC): Provided that $f(t)$ is continuous on the interval $[a, b]$ then

$$\int_a^b f(t)dt = F(b) - F(a)$$

where we know that $F'(t) = f(t)$.

Again, this has that strange feel of something important and difficult, but in fact it only restates what we figured out already. Namely, when we can reverse the derivative operation (like for polynomials, exponentials, sinusoids and so on) then we can use the results of the inverse derivative operation to compute the area under the graph of the input function. In mechanical terms this means we can get the distance travelled if we "integrate" the velocity. Of course, if one is willing to learn the various tricks of analytical integration, FTC suggests you will never need to do another Riemann sum again. This is an unfortunate misrepresentation of reality since many interesting and useful integrals cannot be done analytically and must be tackled by approximation methods, which amounts to better versions of the Riemann sum upper and lower bounds implemented on a computer.

In practice it is a good idea to say thing like "by FTC" on assignments and tests when working out an integral, at least for this first term.

21 APPLICATIONS OF THE INTEGRAL

With the definition of the definite integral in hand and the FTC to make the actual calculation of integrals tractable, we can come up with a variety of applications of the integral. Many are of the "add something up" variety, and indeed the definite integral is a natural extension of the concept of a sum:

Example 1 Consider a tube of length L meters. In this tube a chemical is distributed as a function of space. If we place the left end of the tube at the origin, we have that the concentration per unit length is given by the function $C(x)$. Let's say we wish to find the C_T, total amount of the chemical. We could start by splitting the tube into N pieces that were

$$\Delta x = \frac{L}{N}$$

If c_n^+ (c_n^-) is the maximum (minimum) concentration in the nth interval then

$$\sum_{n=1}^{N} c_n^- \Delta x \leq C_T \leq \sum_{n=1} c_n^+ \Delta x.$$

This inequality is nice to have, but cumbersome. If we now assume that $C(x)$ is continuous and take the limit as $N \longrightarrow \infty$ or $\Delta x \longrightarrow 0$, then

$$C_T = \int_0^L C(x)dx.$$

For example consider $L = 1$ and $C(x) = x^2$. Then

$$C_T = \int_0^1 x^2 dx = F(1) - F(0)$$

by FTC, where $F'(x) = x^2$. Of course since $C(x)$ is just a quadratic, we know that $F(x) = x^3/3$ and so $C_T = 1/3$.

Example 2 A slightly more sophisticated use of the integral would consider the same physical set-up as Example 1, but now aim to find the *average* concentration. If we were to have three sub-intervals of equal length then an estimate of the average value would look like

$$C_{av} \approx \frac{1}{3}(C_1 + C_2 + C_3)$$

where C_1 is a representative concentration for sub-interval 1, C_2 is a representative concentration for sub-interval 2, and so on. For N equal subintervals we have

$$C_{av} \approx \frac{1}{N}(\sum_{n=1}^{N} C_n)$$

but we know

$$\Delta x = \frac{L}{N}$$

so we can rewrite the estimate as

$$C_{av} \approx \frac{1}{L}\Delta x(\sum_{n=1}^{N} C_n).$$

In the limit of Δx going to zero we get

$$C_{av} = \frac{1}{L}\int_0^L C(x)dx.$$

Take $L = 2$ and consider $C(x) = x$. Some simple algebra gives

$$C_{av} = \frac{1}{2}(\frac{2^2}{2} - 0) = 1$$

a result you can confirm either from the graph or by arguing based on symmetry.

Example 3 A more mundane application is the area between two curves (say the graphs of $f(x)$ and $g(x)$). This can be found in your text. First you need to solve for points of intersection (call them a and b with $a < b$) and as discussed previously, for this to be possible analytically the problem pretty much has to be doctored. Once you have the points of intersection you estimate the total area by a sum of the areas of rectangles of length Δx and height $f(x) - g(x)$ when $f(x) > g(x)$ and $g(x) - f(x)$ when $f(x) < g(x)$. In the limit $\Delta x \longrightarrow 0$ and assuming for now that $f(x) > g(x)$ for our interval

$$A = \int_a^b [(f(x) - g(x)]dx.$$

A far more profound example of applications of the definite integral, at least for fans of classical mechanics, is the concept of work. As I will ask you to show on one of your

assignments, **Work** is defined as the change in total energy. Usually work is either performed by the system on the external world (a motor uses combustion to do mechanical work) or performed by the external world on the system (I use my hand to stop a moving ball). If you have seen Work in high school it was probably given in the mysterious looking formula

$$W = F\Delta l.$$

where F is the force and Δl is the distance moved while the force is acting. The formula is only true if the force does not change with x. When F does change with x we could think of splitting up the total displacement into sub-interval Δx and the total work ,W, into a sum of pieces of the form

$$W_n = F_n \Delta x$$

where F_n is a typical value of the force for the nth interval. In the limit of $\Delta x \longrightarrow 0$ we have

$$W = \int_{x_1}^{x_2} F(x)dx.$$

This is a somewhat more substantial result than the high school formula. The result would get even stronger if we considered a situations in which the particle moved in space with the position given by the vector

$$\vec{x}(t) = (x(t), y(t), z(t))$$

then $\Delta\vec{x}$ is a vector as well, as is the force and since energy is a scalar, we need a way to build the scalar quantity Work out of two vector quantities $\vec{F}(\vec{x})$ and $\vec{x}(t)$. Physically this is done by noting that it is only the part of the force in the direction of $\Delta\vec{x}$ that is changing the energy, and hence doing work. Of course, playing with vectors is a messy business and you'll have to wait until AM 231 before you get to do calculus on vectors (though you will see Work a few times in the introductory physics courses). Relating the Work, as we wrote it, to changes in energy is easy via the Chain Rule and one version will appear on your assignment.

22 A BIT ON RIGOUR

Typos and such aside, I have tried to keep these notes fairly easy to read. This is hardly the rule in mathematical writing. While the nature of mathematics makes its communication difficult it is hardly true that it is impossible to pitch mathematics at a reasonable level. So what gives?

I think that a part of the story has to do with the fact that many of the equations in mathematics have proven notoriously difficult to solve. Think for example that Fermat's Last Theorem states simply that for $n > 2$ there is no triplet of natural numbers (a, b, c) so that

$$a^n + b^n = c^n.$$

It looks pretty harmless, yet it has led to centuries of effort and a final resolution only in terms of mathematics that very, very few people can comprehend. Given that certain simple mathematical facts require an almost inexpressible machinery to explain, it is thus no wonder that mathematicians are an exacting lot.

The proper term for proper mathematical expression is "rigour". We have encountered rigour in this course, for example when we defined the derivative at a point x_0 as

$$f'(x_0) = \lim_{h \longrightarrow 0} \frac{f(x_0 + h) - f(x_0)}{(x_0 + h) - x_0}$$

provided that the limit exists. We did not, however, use rigorous arguments very much. So let's return to the definition of a limit and do one rigorous argument:

Example 1 Consider $f(x) = x^2$. We wish to prove that the limit as $x \longrightarrow 0$ of $f(x)$ is zero. We begin by recalling the definition of the limit, namely, for any $\epsilon > 0$ we must be able to find $\delta > 0$ so that $|x - 0| < \delta$ ensures that $|x^2 - 0| < \epsilon$. By a stroke of inspiration we hypothesize that $\delta = \sqrt{\epsilon}$ will work. If $|x| < \delta = \sqrt{\epsilon}$ then $|x|^2 < \delta^2 = \epsilon$. Furthermore since x^2 is never negative $|x|^2 = |x^2|$. Hence

$$|x^2 - 0| < \epsilon$$

provided $|x - 0| < \delta = \sqrt{\epsilon}$ and the proof is finished.

Not so bad, really, though a bit tiresome. Why bother? Well, historically the manner of writing mathematical proofs used above dates to Friedrich Gauss (1777-1855). Gauss, perhaps the greatest western mind of the modern era, revolutionized too many fields of mathematics and the physical sciences to list. However, a persistent complaint of his was the sloppy nature of the mathematical writings of the day. Often theology and meta-physics was intertwined with mathematics and Gauss felt this should not be. In his own writing he was meticulous to the point of being difficult to follow, though few could argue with his assertion:
"I mean the word proof not in the sense of the lawyers, who set two half proofs equal to one whole proof, but in the sense of the mathematician, where 1/2 proof = 0 and it is demanded for proof that every doubt becomes impossible."
It is one of the testaments to Gauss' genius that this statement is now accepted worldwide as the basis for mathematics.

Still, you must surely ask yourself whether the world of 2005 and the world of the early 1800s have much in common. Indeed, mathematical rigour takes away as much as it gives, in the sense that very few of us are actually predisposed to accept that our thoughts should be organized in this manner. What is more, that other great product of the Enlightenment, the scientific method, appears to stand in direct opposition to Gauss' rigour. For if we are constantly searching to invalidate the latest, and best, scientific theory, then we have not arrived at Gauss' destination and our argument is not a formal proof.

It is interesting that despite this apparent opposition mathematics and science benefit greatly from one another. The derivative and its formal definition allow us to write differential equations for mechanics, and these in turn allow us to predict the motion of billiard balls, projectiles, and even to some extent molecules. Conversely the failure to describe the complicated phenomena of Nature using the few mathematical functions we know leads to the mathematical project of finding approximations of arbitrary accuracy made up of polynomials (Taylor series) or sinusoids (Fourier series).

Often rigorous results are striking in their symmetry and beauty. Here is one:

THE MEAN VALUE THEOREM (MVT) If the function $f(.)$ is differentiable for $a < x < b$ and continuou at the two end points $x = a$ and $x = b$ then somewhere in (a, b) say at c we have

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

or put another way, at $x = c$ the rate of change of $f(.)$ matches the slope of the line connecting the two points $(a, f(a))$ and $(b, f(b))$.

The MVT can be used to bring rigour to the statements $f'(x) > 0$ implies $f(.)$ is increasing at x, among other uses. Still, you should note that intuitively it is an obvious fact. The assumptions tell us that $f(.)$ needs to be a 'reasonable' function and the conclusions then follow naturally. For example, say $f(a) = f(b) = 0$ then for 'reasonable' functions MVT tells us the obvious fact that somewhere between a and b the function has a critical point $(f'(c) = 0)$.

Let's play one more game with the MVT. Consider a function $g(x)$ that is continuous on $a \leq x \leq b$. From the FTC we know that if $G'(x) = g(x)$ then

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

Moreover we know that $G(x)$ is differentiable and continuous at $x = a$ and $x = b$ (otherwise the integral of $g(.)$ would not exist). So we can apply the MVT to $G(.)$ and guarantee ourselves the existence of a c somewhere between a and b so that

$$G'(c) = \frac{G(b) - G(a)}{b - a}.$$

But $G'(c) = g(c)$ and

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

and so we have actually guaranteed ourselves a c so that

$$g(c) = \frac{1}{b - a} \int_a^b g(x)dx$$

or that there is at least one c so that $g(c)$ is equal to the **average of $g(.)$ over the interval** $[a, b]$. Again intuitively obvious, but nevertheless pretty.

So what's the take home message? I think at the very least a practicing applied mathematician or theoretical physicist owes mathematical rigour respect. It allows us to do much of what we do, while rarely getting in the way. Moreover for some areas of study, like high energy physics, the deductive approach of pure mathematics whereby we start with axioms and derive all logical consequences is actually how the science (some would claim it cannot be science, but that is another story) is carried out. Still it should be noted that much of applied mathematics is not rigorous, not because of any internal failing, but because constructing approximate descriptions of Nature is what it is all about.

I end this section by returning to the construction of the real numbers. I mentioned in the very first section of these notes that numbers like $\sqrt{2}$ and π cannot be expressed as fractions. That is, even though we can get very close to these numbers using fractions, we cannot get "right there". Yet the machinery of the calculus requires, to be truly complete, a way to construct the real numbers. While being far too advanced for our discussions, I note that the construction due to Dedekind, as presented in Michael Spivak's "Calculus", for example, post dates the active use of the calculus by at least 150 years! So perhaps one can get quite far without all out rigour.

23 OPTIMIZATION

If you asked a hundred mathematics graduates about what first year calculus was all about, the answer you might get (right after "pain in the rear end") is "Word Problems". Nothing is quite as good at putting fear into the eyes of students as word problems. I, in some sense an eternal student, fully agree. Word problems tend to be doctored and artificial exercises which are tough to learn anything from. So, having properly vented we will proceed to solve some (and provide, hopefully, helpful commentary as we go along).

Example 1 Consider a length of rope L. Find the rectangle with the largest area that can be constructed with the length of rope.
Let's label the area of the rectangle A and now let's call the length of the rectangle x and the height h. Because we know that we only have a total length of rope L, and because we figure it might be a good idea to use all the rope, we have $L = 2x + 2h$. This allows us to write

$$h = \frac{L}{2} - x.$$

This allows us to write the area as a function of x only, i.e.

$$A(x) = x\left(\frac{L}{2} - x\right).$$

What is the valid set of inputs? Well certainly $x \geq 0$ and so $x = 0$ is one boundary point. Arguing by symmetry then gives that the other end-point must have $h = 0$ and hence $x = \frac{L}{2}$. For both end-points a ready calculation shows $A = 0$.
Next we find the first and second derivatives of $A(x)$, namely $A'(x) = -2x + L/2$ and $A''(x) = -2$. We thus see that the point $x = L/4$ is a critical point and from the second derivative that it is a local maximum.
Because $x = L/4$ is the only critical point and because the end-points both give $A = 0$, $A(L/4) = L^2/16$ gives the global maximum. Finally from the formula for h we find that the global maximum occurs when $h = x$ or for a square.

Your textbook has many other sample problems for you to try if you wish. Here we want to talk about what we learned from doing the above example:
Check the endpoints!
This means you have to formulate the problem in such a way as to identify what the end-points are. You also have to be clear about what all the points that you have to check are. Certainly actual endpoints need to be checked, but also places where the derivative does not exist, as in the following
Example 2 The payoff from a particular (and crooked) financial derivative is given by twice the absolute difference between the present value of a stock and a set threshold E. Find the minimum pay-out (and hence explain why the derivative is crooked).
Let's call the pay-out P and the value of the stock v. According to the above we have

$$P(v) = 2|v - E|.$$

It seems reasonable to expect that the value of the stock cannot be negative, hence $v \geq 0$ and $v = 0$ is a boundary point. We have been given no bound on the value of the stock and hence $v = 0$ is the only boundary point. Now we would like to apply the derivative tests. This means we must consider cases:
Case 1 $v < E$

$$P(v) = 2|v - E| = 2E - 2v$$

so $P'(v) = -2$ and $P''(v) = 0$. Clearly there are no critical points.
Case 2 $v > E$

$$P(v) = 2|v - E| = 2v - 2E$$

so $P'(v) = 2$ and $P''(v) = 0$. And again there are no critical points.
What is there left to check? Of course it is the point $v = E$ where the derivative is not defined. At this point $P(E) = |E - E| = 0$ and comparing to $P(0) = 2|0 - E| = 2E$ we conclude that the global minimum occurs at $v = E$ and has a value of 0. Of course a financial product that has no possibility of making money is either crooked or ineptly constructed

OK, what about the matter of the first versus the second derivative test? We have already talked about the fact that solving the equation $f'(x) = 0$ for x may not always be possible analytically. Let's say

$$f(x) = \frac{x^3 - 16x^2 + 7x - 5}{x^3 + 5x^2 - 3x + 1}$$

then $f'(x)$ will be bad enough, never mind $f''(x)$. In such a case (which will not appear on the final exam) you could battle through an get the first derivative, find the critical points (perhaps approximately) and then just compare the values delivered by the function at each critical point and all the end-points plus the places where the derivative is not defined. The largest value will be the global maximum and the smallest will be the global minimum. **There is no need to invoke the second derivative test in this case.**

Finally, there is a famous trick to remember. Let's say you are trying to minimize the distance between a curve and a point. You may recall that this will involve minimizing

$$d(x) = \sqrt{(x - x_0)^2 + (y(x) - y_0)^2}$$

but differentiating the square root will be messy so why not consider

$$s(x) = d(x)^2$$

instead? Since distance is either positive or zero, minimizing the distance squared is the same as minimizing the distance.

Example 3 Find the minimum distance between the line $y(x) = x + 3$ and the point $(2, 1)$. As per the hint write

$$s(x) = (x - 2)^2 + (x + 3 - 1)^2 = x^2 - 4x + 4 + x^2 + 4x + 4 = 2x^2 + 8$$

and then $s'(x) = 4x$ and $x = 0$ is the only critical point. The second derivative test (why not it's easy in this case) gives $s''(x) = 4 > 0$ so the point $x = 0$ is a local minimum. What about end points? Since the line extends out to infinity in either direction there are none to check. Finally, $s'(x)$ can be found for all x so there are no places where the derivative is undefined. Hence $s(0) = 8$ is the global minimum. The minimum distance is thus $d(0) = 2\sqrt{2}$. You might want to confirm this result from geometry by making a sketch.