STUDIO CALCULUS

ADAM BOUCHER

These notes are intended to familiarize the student with calculus in a functional, applied context. Definitions and theorems are given in a rigorous framework, and problems and procedures are illustrated in the context of realistic problems rather than restricted to the 'toy problems' of more conventional courses. The order of topic presentation has been altered from the standard pedagogical order in order to emphasize the connections between differentiation and integration and to force the student to continually revisit both operations in multiple contexts of increasing complexity. Many topics and procedures are introduced before their rigorous justification in order to expose students to applications as early as possible. Once the utility of the concepts is well understood, we proceed to study the definitions from a more rigorous perspective, and build a sound theoretical foundation for further study in analysis.

1. Introductory Remarks: Notation and Pedagogy

In mathematics notation is power. As you pass from the study of elementary algebra and geometry to intermediate and advanced mathematics the role of notation changes. In your early study of mathematics your teachers presented you with formal rules and algorithms which allowed you to manipulate mathematical equations and expressions. These rules describe how quantities change when we perform arithmetic operations, and they also tell us that certain operations do not change the truth of statements of equality. When learning about these rules and procedures, notation provides a bookkeeping system. Each quantity or number within the problem at hand has some place in the notation. By following the rules, one can manipulate the notation into myriad different forms. By following directions carefully and mimicking the actions of your peers and instructors one can be wildly successful at this kind of mathematics with absolutely **no** understanding.

Similarly, you can learn to drive a car without any knowledge of what happens within the car's engine. Provided that the car continues to run, you might never need to open the hood. However if the car breaks down, runs out of gas, has a flat tire or any of a hundred other possible mechanical failures, then your ability to drive the car becomes completely irrelevant until the car gets fixed.

The study of mathematics at the intermediate and advanced level requires that we learn about the internal workings of mathematical machinery, and this requires **understanding**. Many of the concepts, definitions and theorems that we will discuss may not seem to have any direct connection to the real world, however my goal is not to simply teach you to drive the car. Ideally, you will be able to reconstruct the car from scratch and build it lighter, faster and more efficient than it started. Driving the car is only a side effect. As you pass from the elementary study of mathematical symbol manipulation to applied mathematics notation plays a different role. Mathematical notation becomes an compressed representation of complex ideas, and hand computations become a symbolic expression of the flow of abstract ideas. Unlike elementary algebra, not every important detail will be explicitly coded into the notation. You cannot understand calculus by surface level symbol manipulation, so you must concentrate on internalizing each definition and theorem that we discuss. Calculus is a required prerequisite of any serious technical education so you should learn the material for the long haul, not try to cram for the test.

We will use a variety of different notations in this class, these notations have been chosen to familiarize you with the most common notations used in mathematics today. Always remember that the notation is a representation of an idea, if you don't know the underlying idea, you shouldn't use the notation. Many of the notations we use have been designed to agree with your intuition and this agreement can lead you into a false sense of security about how well you know and understand the notation.

I will emphasize procedures which are technically correct **and** which generalize correctly when you see them in more advanced contexts. Many of the 'instructional aids' given in textbooks, by your peers, and in tutorials on the internet are rife with technical errors, these sources generally focus on an extremely local view of correctness, and it seems any set of nonsense transformations which leads to the correct answer on a particular problem is a reasonable 'explanation.' My techniques and methods may seem more involved, but I assure you that when you learn these procedures properly they don't come back to bite you on a test two years down the road.

In light of the long view I take on your intellectual progress, I always ask that you come to see me first for any clarification or difficulties with the material. If you are content that you have understood everything that I have told you about a particular topic, feel free to seek others' opinions and ideas, but it might be wise to double check any differences with me before committing them to memory.

2. Functions and function notation

A function is a mathematical object which formalizes the idea of a mapping or a transformation from one set of objects into another. The set of input objects is called the domain and the set of output objects is called the co-domain (target set is also commonly used). A formal study of the abstract properties of functions is usually conducted in an introduction to proof course, or in a first course in mathematical analysis. We proceed from a semi-formal definition of a function, but actively work with a more intuitive conception.

Definition:
$f : D \to R$ is said to be a **function** if and only if f maps every $x \in D$ to a *unique* $y \in R$. This definition can be made more or less formal as the situation requires. For our purposes this definition has two salient points which may be useful or important to know in certain contexts.

- The first point is that the function does something to **everything** in the domain.
- The second point is when you fix an input to the function, the output is always the same. In this sense the behavior of mathematical functions is repeatable, so basic mathematical functions cannot handle uncertain situations directly.

A classical function is completely specified by the way that it connects the inputs from the domain to the outputs of the target space. Thus we might have situations where we specify a function by creating a table of inputs and their respective outputs. Because of the first point, we will need one line in our table for every element in the domain, and because of the second point we know that our table only needs to have two columns, one for the inputs and one for the outputs.

In the case of large, or infinite domains, tables are inefficient, so in simple cases we might represent a function with a formula. From a theoretical perspective it's important to distinguish the function is the mapping, while the formula is a representation of the function; it tells us what the function does. We can think about functions without having recipes or formulae for their operations, and think is one of the more important applications of functions, by adopting notation for a function without knowledge of a formula, we may be able uncover that formula through symbolic manipulation alone. This outcome is rare, but possible when enough simplifying assumptions are adopted.

Consider the following three examples:

Example 1:
Imagine that you are trying to send a letter via first class mail. You bring the package to the post-office and based upon the weight of the letter you will be given a single cost that you must pay to send the letter. Since each letter (the input) gets exactly one price (the output) this fulfills the definition of a mathematical function.

Example 2:
Now consider the action of an old soda machine. You observe that sometimes you put a dollar into the machine and you get your soda. Sometimes you put a dollar into the machine and nothing happens. This would not satisfy the definition of a mathematical function because the same input (one dollar) can yield very different outputs.

Example 3: On a more mathematical note, consider the act of doubling. Imagine that you have a black box which copies anything you put into it and returns two of the same thing. This is also an example of a function because each argument is given a single output. You know what will happen for any viable input.

Example 1 gives us an example of a function which we might not want to represent with a formula, since the postal scale will have lots of different prices based upon the different weights of letters, this formula would be rather long and complex. If we want to work with this function, but we do not want to lug the whole formula around during our calculations, it is very common to use function notation.

Suppose for example that the price for first class postage followed the following formula:

$$P(x) = \begin{cases} .45 & & x \leq 6oz. \\ .60 & 6oz < & x \leq 8oz. \\ .75 & 8oz < & x \leq 10oz. \\ 1.00 & 10oz < & x \leq 13oz \end{cases}$$

Rather than writing this messy piecewise defined function in each step, we simply write $P(x)$, and carry the shorthand symbolic representation along with our computations.

In the case of example 3, we have a very simple function, which can conveniently be represented by a formula, in this case we might write:

$$f(x) = 2x$$

This tells us that the function notation $f(x)$ is exactly the same as multiplying x by a factor of 2. In simple situations like this, it may be easier to use the formula directly, than to use the abstract function notation. You should be careful to distinguish function notation $f(x)$ where x is the argument or input to the function, f, with regular multiplication where a quantity f is multiplied by another quantity x. You should be able to distinguish between the two cases in contex. Just note whether f is being used as a function or a quantity.

Functions are an extremely broad class of objects, and for the most part we will only be concerned with relatively nice functions, which in addition to a possible formulaic representation, will also have a graphical representation. I work under the assumption that you are familiar with two dimensional cartesian coordinates and you are capable of plotting the graphs of simple functions given their formulaic representation. (Again, the graph is not the function it is a representation of the function, when we talk of the graph we are talking about a picture that illuminates certain aspects of the function.) With calculus we will investigate how to extract exact information from symbolic representations that might be obvious from a graphical representation, and we will see how to use the symbolic representations to create qualitatively accurate graphical representations, without brute force computational assistance.

A few frequently asked questions about functions which may or may not have been answered correctly by previous teachers.

(1) *How do I find the domain of a function?*
This is a common type of question to see when first introducing functions. The ability to correctly specify the domain when given a formulaic representation of a function is functionally useless, in the sense that once the domain is specified exactly you don't actually possess any new or interesting information. Functionally, the important skill is to be able to analyze a formula and to find values of the argument at which the formula is undefined (These are points outside the domain). These holes are of the upmost importance when computing limits, derivatives and integrals, and they can lead to a variety of different behaviors. In elementary courses, one typically needs to look for division by zero, negative arguments to square-root functions (or other even roots), and zero or negative arguments to logarithm functions.

(2) *How do I find the range of a function?*
This is a difficult question in general, and one for which there is no general solution procedure. Typically, to answer questions of this form in elementary courses, one draws the graph of the function and uses the graph along with careful technical argumentation to make claims about the range. Well known functions: polynomials, rational functions, trigonometric functions, exponential and logarithmic functions all have well-known ranges, and these ranges often play a role in technical proofs.

(3) *What is a function, really?*
When one defines functions formally in an analysis course, one uses set theory and defines a function as a particular set. So at the most fundamental level a function is a set. The elements of the set defining a function, f are of the form $(x, f(x))$. So a function is a set whose elements describe a relation between the domain and the co-domain. The set theoretic definition gives a concrete handle on what a function really **is**, and is useful for constructing proofs about how functions behave. From a more applied perspective it is most helpful to conceive of a function as a transformation where the domain elements are turned into target elements. This conception keeps one in the correct frame of mind when applying functions to real world problems.

2.1. **Compositions of Functions.** Now that you have been briefly introduced to functions and basic function notation, we will investigate how to mix and match functions. A composition of two functions is a combination of both of their operations in a sequence. We sometimes denote composition with an open circle, $\circ$

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

In order to determine the composition of two functions, we give the final output of the first function $g(x)$ as input to the second function. We then perform the second function on this input value as normal.

Example:
Consider $f(x) = x^2 + 1$ and $g(x) = 3x - 2$, let us compute the composition of these two functions on a fixed argument $f \circ g(2)$.

First we compute $g(2)$.

$$g(2) = 3(2) - 2 = 6 - 2 = 4$$

We now use $g(2) = 4$ as the input for f.

$$f(g(2)) = f(4) = 4^2 + 1 = 17$$

Combining these two results we get $f \circ g(2) = 17$.

Sometimes you might want to use the composition of two functions more than once, making these sequential computations very tedious. In this case, we would simply substitute the whole function $g(x)$ into the function f to obtain a complicated expression for $f \circ g(x)$. We could then simply that expression as much as possible in order to define a new function.

$$f(g(x)) = (g(x))^2 + 1$$
$$f(g(x)) = (3x - 2)^2 + 1$$

If this expression was satisfactory, we could stop here. Otherwise:

$$f(g(x)) = 9x^2 - 12x + 4 + 1$$
$$f(g(x)) = 9x^2 - 12x + 5$$

Remark: Note that in the composition $f(g(x))$ we work from the inside to the outside. we do **not** work left to right.

2.2. **Inverse Functions.** Sometimes we want to "undo" the operation of a function. An inverse function satisfies this need, however, inverses do not always exist. An inverse function can only be found for functions which are 1 to 1, that is all of the elements of the domain get mapped to different elements of the range. In other words the elements of the range are never re-used.

Example 1:

The simplest example of a function without an inverse is a constant function. A constant function basically ignores the input and always returns the same thing. Since you lose all of the information you put into the function, there is no way to backtrack and determine what went into the function in the first place.

In order for an inverse to exist one must be always be able to distinguish the input that produces each output. If multiple inputs produce the same output, then the function cannot be inverted uniquely.

Example 2:

Consider the functions x^2 and x^3. The function x^3 is uniquely invertible for any choice of output (i.e. the cube root function is well defined for any real number.). The function x^2 also has an inverse, but this function is only well defined when we restrict the domain. Consider the following calculations:

$$f(x) = x^2, \quad f(2) = 2^2 = 4, \quad f(-2) = (-2)^2 = 4$$

When you square a number it loses its sign, so if we invert x^2 we do not know whether the input was positive or negative. We **define** the square root function:

$$g(x) = \sqrt{x}$$

as the **positive** number whose square is equal to x. This function **is** the inverse function for all positive inputs, but it is not the inverse function for the negative inputs.

$$\forall x \in \mathcal{R}^+ \cup \{0\}, \quad f^{-1}(x) = g(x) = \sqrt{x}$$

(In class we will look at this example graphically and discuss the implications for several important applications.)

This example raises certain issues for us. We must be careful to note the domain and range of functions under consideration, because by appropriately restricting the domain we may be able to define an inverse function. From a practical, problem solving perspective the existence or non-existence of a inverse function is usually only of indirect importance. At the level of problem solving, one either attempts to invert a function directly by algebraic manipulation or tries to find a solution indirectly by numerical approximation. Since algebraic manipulation only works for the simplest formulaic representations, the knowledge about the existence and behavior of inverse functions can be used to inform and improve the performance of numerical approximations, or to help determine when solutions do not exist.

3. Informal Introduction to Limits

While functions are the primary mathematical objects studied in calculus, the limit operation is the foundation on which calculus is built. The current rigorous definition of the limit was developed by Cauchy over a century after Leibniz and Newton developed the fundamental techniques of calculus in their theories of infinitesimals, and fluxions respectively. Let's repeat that last sentence. The definition of a limit we now use to **define** the operations of calculus was developed over a century **after** the operations of calculus were introduced. This is very important: Mathematics can be **used** effectively without having a rigorous theoretical foundation, however the rigorous foundation is the only way

to determine exactly where certain techniques will be valid or invalid.
This class is going to outline the concepts and techniques of calculus from a functional perspective before retracing the same concepts from a more precise, rigorous perspective. This is a pedagogical **choice** that we are making in this class. You will first investigate **how** calculus is used, then once we have developed a working knowledge of the subject, we will be prepared to investigate **why** it works and how the different concepts and definitions fit together. In this section we will outline the definition of a limit in informal terms, and use this informal definition to help build intuition about the connections between the graphical and formulaic representations of smooth functions. We will connect our intuitive understanding of continuity with the informal definition of the limit.

The limiting operation is based upon a very simple notion, which turns out to be very subtle when we try to pin it down. A limit operation has two working parts. When we specify a limit we need to specify two things, a function, and a limiting argument. When we take a limit we are trying to extract information about the function, and the information tells us about the behavior of the function around the limiting argument. We notate the limit operation in the following way:

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

Here the $f(x)$ is the function we are analyzing, and c is the limiting argument. x is the argument of the function, and it will appear in any formulaic representation of the function. A complete understanding of the scope of the limit operation requires dedicated study of the sets and spaces which define both the domain and the codomain of the function f. For our purposes these sets are almost always real numbers, so for this introduction we will restrict ourselves completely to this set.

Informally speaking, when we take the limit of a function around a limiting argument, c we are trying to learn about how the values of the function around the limiting argument relate to the value of the function at the limiting argument. The limit operation allows us to bridge the gap between the individual values of a function and the values of the function on intervals and sets. The limit operation is usually the first operation that you study which has the potential to fail, and this makes it difficult for many students.

When you perform arithmetic on the real numbers you can perform arbitrary addition and multiplication operations, and the operations are always well-defined. You never have to worry about the sum of two numbers not existing, or the product of two numbers having two possible values. (We have briefly discussed some of the complications that arise with the square root function, but we have a systematic way of handling the function and a definite set of rules which allow us to handle every circumstance which might arise.) Under normal circumstances, and circumstances which may be important for applications, the limit of a function may fail to exist, and that failure to exist is perfectly okay.

When dealing with the limit operation several different outcomes are possible. We can have limits that exist and take on regular, numerical values. We can have limits that tend toward infinity or negative infinity, or we can have limits which fail to exist altogether for several different reasons. The most frustrating part about working with limits, is that different problems may require different techniques to help evaluate the limit, so it is possible to reach a dead end with certain techniques without finding a definite answer. This is **okay**. As we first delve into the limiting operation, it is more important that you know your limitations, than to get into habits where you jump to conclusions from incomplete information. As the semester progresses, we will learn a variety of techniques which will eventually allow us to handle pretty much any limit we can write down, but as we begin, I want you to focus on making correct logical deductions from the information at hand rather than simply trying to memorize a set of possible answers.

Informal Definition: *We say that a function, f, has the limit L at the point $x = c$, notated by:*

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

if and only if the function values $f(x)$ **must** *get close to the value L whenever the argument x gets close to the value of c.*

Alt. Definition: *We say that a function, f, has the limit L at the point $x = c$, notated by:*

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

if and only if when the values of x get close to the value c, we can assert that the values of the function, $f(x)$ also get close to the value of L.

(**Remark:** As we will see later, the concept of 'must get close to' can be expressed in terms of quantitative mathematical expressions, but we do not delve into those yet.)

(**Remark**: Often students want to pin down the meaning of 'close.' The definition actually forces us to leave the idea of distance or 'closeness' unspecified, so the definition allows us to force the function values to be **as close as needed** to the limiting value.)

3.1. **Procedural Handling of Limits.** As a user of mathematics, the ability to evaluate limits will be absolutely central to your ability to transition from the definitions of the operations in calculus to the computational procedures we use to compute derivatives and integrals. Toward this end, we first introduce a procedural way of evaluating limits. This procedural handling is designed to slow you down and keep you honest with the limiting operation, so that when we revisit the material in a rigorous context your intuition will match the rigorous results, rather than being a mixture of ad-hoc rules and erroneous generalizations based upon an incorrect grasp of the limit.

In practice, when you are faced with a limit operation, the problem is usually to find the value of L when the function f, and the limiting argument c are both given. In elementary calculus the function f is usually given an explicit formulaic representation, and c is either a finite real number, or positive or negative infinity.

Because 95% of the mathematical functions you have worked with are continuous and well behaved, you may have incorrectly assumed that the limit operation is just function evaluation, that what you **do** when you evaluate a limit is simply to plug c into the formula for f and work out any necessary arithmetic. From a rigorous perspective, employing this strategy means that you have **completely missed the boat**. So before we say any more about limits, let's define continuity and use this to correctly handle some limits.

Definition: *A function f is* **continuous** *at the point c if and only if:*

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

This definition extends to continuity on an interval (a, b) or $[a, b]$ if the function is continuous at every single point within those respective intervals, and a function is continuous (without qualification) if it is continuous at every point in it's domain.

Thus the concept of continuity captures the simplistic strategy of function evaluation for limit evaluation: **If a function is continuous at the point c, then you may evaluate the limit simply by evaluating the function.**

When given a formulaic representation of a function to evaluate, trying direct evaluation is a good first step, because almost all of the elementary functions we use (polynomials, rational functions, trigonometric functions, exponential and logarithmic functions and their algebraic combinations) are continuous almost everywhere. (The use of almost everywhere is actually technical term so while it might sound like I'm being vague, I'm actually being precise in a mathematical sense). BUT you must understand that this 'technique' is not technically correct. In all cases, the actual value of L is totally independent of the value of $f(c)$, it just happens that we spend a great deal of our time working with examples where these values coincide.

In many cases of practical interest, the formula for f which is given will not be well defined at the point, $x = c$. If the formula for $f(x)$ is well defined for $x \neq c$, then you can perform any algebraic manipulations or simplifications you want on the formula provided you do not change its value. If some set of these manipulations allows you to evaluate the formula at $x = c$, then the value of the limit is usually the value of the simplified formula. If you cannot work the formula into a evaluable form, then you must stop, you have failed to evaluate the limit analytically. At this point, you might try approximating the value numerically by plugging in values of x which are very close to c. Without some type of external analytical information, this strategy does not give **any** definite information about the true value of the limit, or if the limit even exists.

As we progress, we will learn several more advanced techniques for finding limits analytically. Some of these techniques will be based upon direct rigorous proofs, while others will be based upon heuristic handling of different families of elementary functions.

At this point we state without proof that:

- All polynomials are continuous everywhere.
- All rational functions are continuous everywhere where the denominator is non-zero.
- Cosine and Sine are both continuous everywhere, the derived functions (Tangent etc.) are continuous everywhere the denominator is non zero.
- The exponential function, e^x is continuous, and the natural logarithm is continuous on $(0, \infty)$.

We also state that sums, products, powers, and compositions of continuous functions are continuous (on suitably restricted domains), and that quotients of continuous functions are continuous provided the denominator is non-zero.

The proofs of these facts are important, but as our aim is to work with calculus, we do not belabor these fundamental results at this time. We will spend more time later in the semester working through certain classical results and examples when we treat this material from a rigorous standpoint.

3.2. **Indeterminate Forms.** When we try to evaluate limits using direct substitution, we often encounter algebraic expressions which do not have a fixed value. These are called **indeterminate forms**. Eventually we will use calculus to help us handle these forms, but at this stage you must be able to recognize these forms, and you must understand that they are **never** a final answer. A short list of possible indeterminate forms is given below:

- $$\frac{0}{0},$$
- $$\frac{\pm\infty}{\pm\infty},$$
- $$\infty - \infty$$
- $$1^{\infty},$$

- $$\infty^0,$$

While one can create examples where each of these forms produces different results, we provide a set of example formulas for only the first indeterminate form showing that **any** limit is a possible conclusion when one obtains the indeterminate form from direct substitution.

Examples: Note that by direct substitution each of these formulae takes on the indeterminate form $\frac{0}{0}$. The correct limiting values are given without proof:

$$\lim_{x \to 0} \frac{e^x - 1}{x} = 1, \quad \lim_{x \to 0} \frac{\cos(x) - 1}{x} = 0, \quad \lim_{x \to 0} \frac{\ln(x+1)}{x^3} = \infty$$

These examples show us that when we get the form $\frac{0}{0}$, it tells us **nothing** about the value of the limit. More analysis is required. A similar set of examples can be generated for each of the indeterminate forms. I do not want to hear, or read any nonsensical arguments like "Anything minus itself is zero so infinity minus infinity must also be zero." This is total rubbish. Infinity is a concept not a fixed real number. We have a symbol to represent the concept, but the rules regarding the manipulation of finite numbers do not apply to the symbol for infinity.

The following forms are not indeterminate, and the correct conclusion for the limit is given:

- $$\frac{c}{\infty} \to 0, \forall c \in \mathbb{R}$$
- $$\frac{\pm\infty}{c} \to \pm\infty, \forall c \in \mathbb{R}^+$$
- $$\frac{\pm\infty}{c} \to \mp\infty, \forall c \in \mathbb{R}^-$$

3.3. **Examples of finite limits that exist.** We begin with some trivial examples, here the underlying function is continuous at the evaluating point, so evaluating the function is a successful strategy for evaluating the limit.

(1)
$$\lim_{x \to 3} x^2 = f(3) = 3^2 = 9.$$

(2)
$$\lim_{x \to \pi} \cos(2x) = f(\pi) = \cos(2\pi) = 1.$$

(3)
$$\lim_{x \to 2} \frac{x}{x+3} = f(2) = \frac{2}{2+3} = \frac{2}{5}.$$

3.4. **Finite Limits that exist, but require manipulation.** Next we consider some examples where direct evaluation fails. We then proceed to perform some algebraic manipulation to find a form where evaluation is possible, sometimes this works and sometimes it fails.

(1)

$$\lim_{x\to 3}\frac{x-3}{x^2-9}=\frac{0}{0},\qquad \text{indeterminate form.}$$

$$\lim_{x\to 3}\frac{x-3}{x^2-9}=\lim_{x\to 3}\frac{x-3}{(x+3)(x-3)}=\lim_{x\to 3}\frac{1}{(x+3)}=\frac{1}{6}.$$

(In this case algebraic simplification allowed us to evaluate.)

(2)

$$\lim_{x\to\frac{\pi}{2}}\frac{1}{\tan^2(x)+1}=?\qquad \tan(\pi/2)\ \textit{is undefined.}$$

$$\lim_{x\to\frac{\pi}{2}}\frac{1}{\tan^2(x)+1}=\lim_{x\to\frac{\pi}{2}}\frac{1}{\sec^2(x)}=\lim_{x\to\frac{\pi}{2}}\cos^2(x)=0.$$

(In this case a trigonometric identity allowed us to evaluate.)

(3)

$$\lim_{x\to 0}\frac{e^x-1}{x}=\frac{0}{0},\qquad \text{indeterminate form.}$$

(In this case, no simple algebraic manipulation allows us to handle the limit, we **cannot obtain the answer** without more sophisticated techniques.)

(4)

$$\lim_{x\to 0}\frac{e^x-1}{e^{2x}-1}=\frac{0}{0},\qquad \text{indeterminate form.}$$

$$\lim_{x\to 0}\frac{e^x-1}{e^{2x}-1}=\lim_{x\to 0}\frac{e^x-1}{(e^x)^2-1}=\lim_{x\to 0}\frac{e^x-1}{(e^x-1)(e^x+1)}=\lim_{x\to 0}\frac{1}{e^x+1}=\frac{1}{2}.$$

(In this case algebraic simplification allowed us to evaluate.)

3.5. **Discontinuities and limits that don't exist.** We have already seen that when a function is continuous at a point c, the limit exists and is equal to the function value at c. Thus, a function must be discontinuous in order for a limit to fail to exist. In order to better describe functions, we tend to classify discontinuities in three ways. In order of severity, when a function is discontinuous at a point c we say that the function has:

- a **removable** discontinuity. This happens when the function has a well defined limit at c, but the function is either undefined at c, or the value of the function doesn't agree with the value of the limit.

The function, f, defined by the formula:

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

has a removable discontinuity at the point $x = 2$. The given formula is undefined at that point, but the limit $x \to 2$ is well defined and takes the value $\frac{1}{3}$. This discontinuity can be removed by changing the function at a single point.

- a **jump** discontinuity. Here the function has well defined, one-sided limits, but the two one-sided limits have different values. Then the function "jumps" across that gap and is thus discontinuous. A jump discontinuity cannot be 'fixed' without changing the function on some small interval.

  A function exhibiting this type of discontinuity would be :

  $$f(x) = \frac{|x|}{x}$$

  At zero this function jumps from -1 to 1 as x increases.

- an **infinite** discontinuity. Infinite discontinuities occur when one or both sides of a limit go to either $\pm\infty$.
  An example of this type of discontinuity occurs in:

  $$g(x) = \frac{1}{(x-5)^2}$$

  As $x \to 5$ the denominator of this function becomes arbitrarily small, making the whole function grow without bound. This is an infinite discontinuity.

In terms of the limit operation, one should note that the limit is well defined in the case of the removable discontinuity and the one-sided limits are well defined in terms of the jump discontinuity, the limits fail to exist, for infinite discontinuities, but we might use $\pm\infty$ to denote the limit if this answer gives us relevant information for the problem at hand. This observation tells us that the limit operation allows a more delicate and sensitive analysis of a function than direct function evaluation. With the limit operation, we can distinguish different kinds of discontinuity, and in certain cases, we can use the limit operation to actually repair a function which has defects at individual points.

(**Remark:** In addition to having unbounded positive or negative growth, limits can also fail to exist due to oscillation. If a function undergoes wild oscillation which does not decay when approaching a limiting argument, then the limit can also fail to exist. This type of behavior is usually de-emphasized in elementary classes, but becomes important in a more detailed study of sequences of functions. A classical example of an function with wild oscillations is $\sin(\frac{1}{x})$. This function does not have a limit as $x \to 0$.)

3.6. **Theoretical Results for limits and continuous functions.** In this section we will cover two theoretical results which have important long range implications. The first result is called the squeeze theorem, or the sandwich theorem. While this theorem will appear rather obtuse on a first (and probably second) reading, the theorem exhibits a very important strategy which is often implemented in more advanced material. The idea is that if you can control the values of a complex function with an upper bound and a lower bound, then the possible values for the limits of the complex function are controlled in the same way. As a result, if you can sandwich a function between functions with known behavior, then you can often make conclusions about the behavior of the sandwiched function without direct evaluation.

Suppose that the function g is always bounded below by the function, f, in some interval f. ($f(x) \leq g(x) \forall x \in (a, b)$). Suppose also that the function g is always bounded above by the function, h, in the same interval. ($g(x) \leq h(x) \forall x \in (a, b)$). We can combine these two suppositions into an inequality sandwich:

$$f(x) \leq g(x) \leq h(x), \quad \forall x \in (a, b)$$

The Sandwich theorem tells us one way we can use this construction to help evaluate limits.

Sandwich Theorem: *If the chain of inequalities $f(x) \leq g(x) \leq h(x)$ holds in some interval (a, b) containing the point c, and if:*

$$\lim_{x \to c} f(x) = L, \qquad \text{and } \lim_{x \to c} h(x) = L.$$

Then we can assert $g(x) \to L$ as $x \to c$.

Proof:
Since the limit operation is order preserving, for any $c \in (a, b)$, we apply the same limit to each of the inequalities in our assumed ordering:

$$\lim_{x \to c} f(x) \leq \lim_{x \to c} g(x) \leq \lim_{x \to c} h(x).$$

As we assume the limiting values of $f(x)$ and $h(x)$ are known, we have:

$$L \leq \lim_{x \to c} g(x) \leq L.$$

Since this limit is bounded above and below by the same number, we can immediately assert the conclusion of the theorem.
Q.E.D.

This theorem can be hard to grasp, not because it is a particularly deep or complex idea, but since the implementation of this idea can be very subtle. To **use** the theorem, one must select $f(x)$ and $h(x)$ especially for the purpose of bounding and controlling $g(x)$ near c. If these functions are not chosen properly, then the hypotheses of the theorem will not hold and the conclusion of the theorem will not apply.

Classically we apply this theorem to evaluate the following limit:

$$\lim_{x\to 0} \frac{\sin(x)}{x}.$$

Note first that direct evaluation gives us an indeterminate form, and that no algebraic manipulation can help us evaluate this formula at zero. So we need to get around direct evaluation in some way or another. To use the squeeze theorem, we need to get both upper and lower bounds to this function which approach the same value. We use a non-standard argument to first show you how the theorem can be used without selecting the functions perfectly.We look at the right and left sided limits separately. When $x > 0$ you can check graphically that the following chain of equalities is true:

$$-x \le \sin(x) \le x$$

This inequality is true for each $x > 0$, so we can manipulate the chain algebraically. Dividing each term by x and taking the right sided limit we have:

$$\lim_{x\to 0^+} -1 \le \lim_{x\to 0^+} \frac{sin(x)}{x} \le \lim_{x\to 0^+} 1.$$

Thus we obtain the result that:

$$-1 \le \lim_{x\to 0^+} \frac{\sin(x)}{x} \le 1.$$

When $x < 0$ we have the reverse inequality so the same algebraic manipulations with the left hand limit give us the result:

$$-1 \le \lim_{x\to 0^-} \frac{\sin(x)}{x} \le 1.$$

Putting these results together, we have now established that both the left and right sided limits are bounded in the interval $[-1, 1]$. This gives us a **bound** on the actual value of the limit, if it exists. Next we proceed with the standard construction for proving the existence of this limit using the sandwich theorem.

This time we use the chain of inequalities given by:

$$\sin(x) \le x \le \tan(x)$$

(This chain of inequalities is generated by looking at the areas of an inscribed triangle, a sector, and an extended triangle, each contained in the previous figure, the derivation is not intuitive.) Manipulating this chain of inequalities to make the function $\frac{\sin(x)}{x}$ appear in the middle term we get:

$$1 \le \frac{x}{\sin(x)} \le \frac{1}{\cos(x)}$$

Reciprocating flips each inequality, so we obtain:

$$\cos(x) \le \frac{\sin(x)}{x} \le 1$$

(You can verify this part graphically if you wish).
Now taking limits and using the sandwich theorem we find:

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

Next we move on to an amazing theorem with highly under appreciated consequences. In terms of the scope of hard problems which you can actually solve using this theorem, it ranks among the most powerful ideas in mathematics.

Intermediate Value Theorem: *Let f be a continuous function on the interval $[a, b]$, for each y in the interval $[f(a), f(b)]$ there exists at least one $c \in [a, b]$ such that $f(c) = y$.*

(A rigorous proof of this theorem requires the formal definition of the limit, so we omit the proof now, but we will revisit the theorem later in more detail.)

Consequences of the intermediate value theorem:

(1) The graphs of continuous functions are connected. This topological result is often given as an intuitive definition of continuity (i.e. a continuous function is one that can be graphed in a single stroke without taking the pencil from the page.)
(2) The theorem can be used as a basis for an algorithm for solving equations. The bisection algorithm is given below.

3.7. **Bisection Algorithm.** Suppose we want to find a value of x which solves the algebraic equation: $f(x) = y$. (f is a given continuous function, and y is a given number.)

Given an x_{low} and an x_{hi} satisfying the following inequalities:

$$f(x_{low}) < y, \quad f(x_{hi}) > y$$

We can proceed to approximate the solution to the equation to within an **arbitrary** degree of accuracy.

Appealing to the intermediate value theorem, we assert the existence of some c satisfying $f(c) = y$. Furthermore $x_{low} \le c \le x_{hi}$. Using the midpoint:

$$x_{mid} = \frac{x_{hi} + x_{low}}{2}. \tag{1}$$

We now have an approximation of c with error no larger than the radius of the interval (x_{low}, x_{hi}). By checking the value of $f(x_{mid})$, we can update either x_{low} or x_{hi} and narrow the possible range of values for the solution. Repeat this process until the error for the approximation x_{mid} is within the desired tolerance. If by dumb luck we ever find that $f(x_{mid}) = y$, then we can terminate the algorithm with the exact solution.

The powerful feature of this algorithm is that f is only assumed to be continuous. Initializing this algorithm could be difficult, but once you find suitable starting values for x_{low} and x_{high}, the algorithm is fool proof.

4. Supplementary Examples regarding continuity and limits

4.1. **Continuity Analysis.** In order to analyze the continuity of an arbitrary function, one must be able to identify possible points of discontinuity and then classify those points by computing limits.
If a given function is some combination of elementary functions (including, algebraic, exponential, logarithmic and trigonometric) these points are usually quite easy to spot.

If the function consists of sums or differences of elementary functions, one should only need to check for points where any of the terms would be outside it's domain (i.e. $\tan(x)$ has infinitely many values of $x$which are outside it's domain, $\frac{-\pi}{2}$, $\frac{\pi}{2}$ etc.).

If the function has any terms with quotients of elementary functions, one will need to identify any points where the denominators could be equal to zero. Each of these points should then be analyzed by taking limits.

If the function is defined piecewise, the junctions between different pieces of the function should be checked for continuity.
Example 1:

Consider the function:

$$h(x) = \begin{cases} \frac{1}{x^2+4} & x < 0 \\ \\ \frac{x^2+1}{4} & x \geq 0 \end{cases}$$

This function is piecewise defined so we need to be concerned with the junction ($x = 0$ in this case) as well as the individual pieces.

We first check the junction. We compare the left and right hand limits:

$$\lim_{x \to 0^-} h(x) = \lim_{x \to 0^-} \frac{1}{x^2 + 4} = \frac{1}{0 + 4} = \frac{1}{4}$$

$$\lim_{x \to 0^+} h(x) = \lim_{x \to 0^+} \frac{x^2 + 1}{4} = \frac{0 + 1}{4} = \frac{1}{4}$$

Since both one sided limits are equal, the limit of the function must also exist. Evaluating $h(0)$ we find this limit is equal to the value of the function at $x = 0$. This means the function is continuous across the junction.

$$\lim_{x \to 0} h(x) = \frac{1}{4} = h(0)$$

Now we look at each individual piece of the function and search for any other possible points of discontinuity.

The first piece cannot have any discontinuities because both the numerator and denominator are polynomials and the denominator, $x^2 + 4$, is never zero.
The second piece is similarly a polynomial, and is continuous everywhere.

We conclude that $h(x)$ is a continuous function.

Example 2:

$$f(x) = x^2 \cot(x)$$

If we re-write this function by replacing $\cot(x)$ with $\frac{\cos(x)}{\sin(x)}$ we find that we have many potential problems, as the denominator, $\sin(x)$, is equal to zero for $x = n\pi$ for any integer, n.

Let us consider one of these points, and then try to generalize our findings to the other points.

$$\lim_{x \to \pi} x^2 \frac{\cos(x)}{\sin(x)}$$

$\sin(x)$ turns out to have different signs on either side of this limit, so we consider the left and right limits separately.

$$\lim_{x \to \pi^+} x^2 \frac{\cos(x)}{\sin(x)}$$

Direct substitution yields division by zero:

$$(\pi)^2 \frac{\cos(\pi)}{\sin(\pi)} = \pi^2 \frac{-1}{-0}$$

Since the numerator is heading towards a finite-non-zero value, we conclude there is no behavior which will cancel the growth in the function due to the shrinking of the denominator.
Since the numerator is negative and the denominator is heading to zero from the negative side, we conclude that the whole function must head to $+\infty$.

If we consider the other limit we find:

$$\lim_{x \to \pi^-} x^2 \frac{\cos(x)}{\sin(x)}$$

Direct substitution again yields division by zero, but now the denominator approaches zero from the positive side, combining this with the negative numerator we find this limit goes to $-\infty$.

Combining these two pieces of information we conclude the limit does not exist and the function has an infinite discontinuity at $x = \pi$.

An identical analysis will work for any potential discontinuity away from the origin. At $x = 0$ we need a different analysis:

$$\lim_{x\to 0} x^2 \frac{\cos(x)}{\sin(x)}$$

Here, trying direct substitution we find an indeterminate form:

$$0^2 \frac{1}{0} = \frac{0}{0}$$

(Remember, this is an abuse of notation, this substitution is not strictly valid)

Re arranging the function we find:

$$\lim_{x\to 0} x \frac{\cos(x)}{\frac{\sin(x)}{x}}$$

Now in the denominator we have formed a function whose limit we know. We can now use the properties of limits to substitute into this configuration and find an actual limiting value.

$$\text{since } \lim_{x\to 0} \frac{\sin(x)}{x} = 1$$
$$\lim_{x\to 0} x \frac{\cos(x)}{\frac{\sin(x)}{x}} = 0\frac{1}{1} = 0$$

In the limit we know that the ratio $\frac{\sin(x)}{x}$ tends to one, so we substitute into the remaining function with this knowledge and we find, this limit actually exists:

$$\lim_{x\to 0} x^2 \frac{\cos(x)}{\sin(x)} = \lim_{x\to 0} x \frac{\cos(x)}{\frac{\sin(x)}{x}} = 0$$

Since the limit exists, but the function is undefined at this point (x=0 is outside the domain of $\cot(x)$) this is termed a removable discontinuity.

Summarizing our findings:

The function $x^2 \cot(x)$ has infinite discontinuities for $x = n\pi$, n a non-zero integer. The function also has a removable discontinuity at $x = 0$.

Example 3:

Consider the following function:

$$g(x) = \frac{x-4}{x^2 - 2x - 8}$$

Since this function contains a quotient we need to analyze any points where the denominator goes to zero.

Factoring the denominator we find:

$$x^2 - 2x - 8 = (x + 2)(x - 4)$$

Thus the denominator goes to zero at $x = -2$ and $x = 4$. We will analyze these points separately:

$$\lim_{x \to -2} \frac{x - 4}{x^2 - 2x - 8}$$

Substitution yields division by zero, but the numerator heads to a finite value so we conclude the whole function must become unbounded and we will have an infinite discontinuity.

Considering $x = 4$ we find by substitution:

$$\lim_{x \to 4} \frac{x - 4}{x^2 - 2x - 8} = \frac{0}{0}$$

Algebraic simplification reveals that the function actually has a limit:

$$\lim_{x \to 4} \frac{x - 4}{x^2 - 2x - 8} = \lim_{x \to 4} \frac{x - 4}{(x + 2)(x - 4)}$$

$$\lim_{x \to -2} \frac{1}{x + 2} = \frac{1}{6}$$

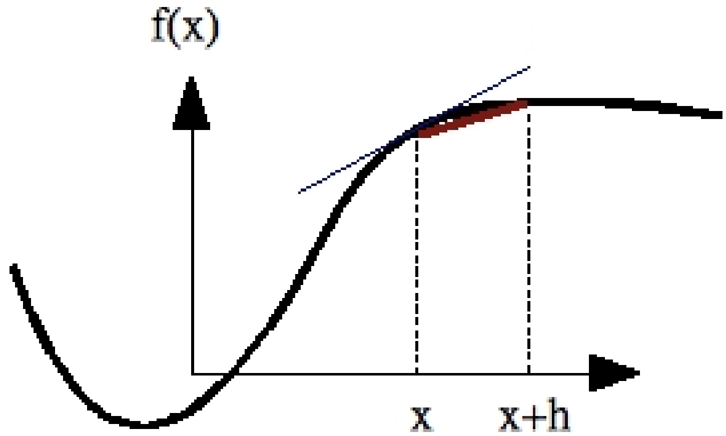
Since the limit exists, we determine that at $x = 4$ the function has a removable discontinuity.

Summarizing these results: We have an infinite discontinuity at $x = -2$, and a removable discontinuity at $x = 4$.

5. Basic Differentiation

In this section, we will cover the definition of the derivative, and the fundamental rules of differentiation, while many of the concepts we cover generalize to all of the elementary functions, we will restrict our attention to polynomials and rational functions to speed our coverage of the material. Later in the semester we will revisit these ideas with the trigonometric functions and exponential and logarithmic functions respectively.

The limit operation gives us a powerful way to generalize the idea of slope from straight lines to smooth curves. If we are interested in measuring the slope of a curve at a single point, we can always draw a line between **two** different points on the curve to approximate the slope, however if we want to measure the slope at a single point, our familiar 'rise over run' formula becomes undefined. The limit operation offers us a powerful way to get out of this problem. If we use the limit operation, we have the option of algebraically manipulating our slope formula before shrinking the distance between the two points on our curve.



When the curve is smooth enough, this manipulation allows us to actually determine a value for the slope at a single point.

Suppose f is a smooth function, then an **approximation** of the slope at the point x could be found by finding the slope of the secant line connecting the points $(x, f(x))$ and $(x+h, f(x+h))$. Using the familiar formula for slope we get the following estimate for the slope at x.

$$\frac{f(x+h)-f(x)}{x+h-x}$$

A graphical representation of this approximation is included in figure 1. The function is drawn in black, the secant line approximation is drawn in red and the exact slope we are trying to approximate is drawn in blue.

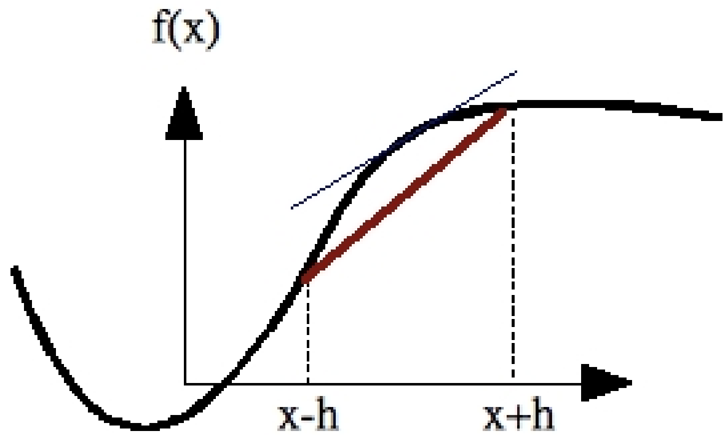
This formula makes sense for any $h \neq 0$, but is undefined when $h = 0$. By taking $\lim h \to 0$, we are asking the question, 'What happens to the slope of the secant lines when the arguments of the function get closer and closer together?'

When this question has a sensible answer (i.e. the limit actually exists) we call the function f differentiable at the point x, and we define the derivative of f at the point x by the value of the limit.

Several nice things to point out. First, the denominator can always be simplified, so we use the simplified version as the definition of the derviative.

$$\frac{df}{dx} \equiv \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$$

Second, if this limit has any hope of existing, then we need to get an indeterminate form when we try to plug $h = 0$ into the formula. This means that $f(x+h) - f(x)$ must go to zero as $h \to 0$, this gives the well known result that differentiability implies continuity, or



that continuity is a **necessary** condition for differentiability.

This definition gives us the basic mathematical structure which we use to deduce and prove the different properties of the derivative. (Returning to the car analogy this definition behaves like a blueprint for the engine of our car. We won't use it when we are out driving, but we will use it to figure out where all the parts go, and to figure out how to move the parts around to achieve interesting effects.)

Technical Remark: This definition is not unique, there are many other difference quotients which can be used to define the derivative which are equivalent in the limit $h \to 0$. One example which is commonly used in numerical computing is the centered difference quotient given by:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x-h)}{2h}$$

This difference quotient uses information from the left and the right of the point of interest to approximate the derivative. This gives slightly better approximations and reduces numerical error. For computing derivatives several different rules are important. We prove one rule for each of the different algebraic operations regularly performed with functions. Each of the important rules is given below using both Leibniz notation and primed function notation.

Assume that f and g are differentiable at x, and c is a constant. Then each of the following rules holds.
For sums and scalar multiples of differentiable functions we have:

$$\frac{d}{dx}\left[f(x) + cg(x)\right] = \frac{df}{dx} + c\frac{dg}{dx} = f'(x) + cg'(x)$$

For multiplication of functions we have the **product rule**:

$$\frac{d}{dx}[f(x)g(x)] = \frac{df}{dx}g(x) + f\frac{dg}{dx} = f'(x)g(x) + f(x)g'(x)$$

For compositions of functions we have the **chain rule**:

$$\frac{d}{dx}\left[f(g(x))\right] = \frac{df}{dg}\frac{dg}{dx} = f'(g(x))g'(x)$$

The power rule is another important rule which can be deduced from the product rule or proven independently (as done in class)

$$\frac{d}{dx}\left[x^n\right] = nx^{n-1}, \qquad n \in \mathbb{R}$$

(**Remark:** There is nothing special about the appearance of x as the independent variable in these formulae, the same structural rules apply when other letters are used to denote the independent variable (e.g. $\frac{d}{dt}\left[t^n\right] = nt^{n-1}$).)

5.1. **Geometric Interpretation of the Derivative.** As the derivative is defined through a generalization of the concept of slope or rate of change, the information it provides can be given a direct geometric interpretation. The values of the derivative do not necessarily correspond to the values of the function, but the values of the derivative do correspond to the shape of the function. For example, on intervals where the first derivative is positive the underlying function must increase. One intervals where the first derivative is negative, the underlying function must decrease.

In applications, the single most important observation about the derivative is that at **every** local peak and valley of a smooth function the derivative is zero. We can see this fact by drawing a smooth curve with several bumps and noting that at each peak and valley the slope of the local tangent line is horizontal. Horizontal lines have zero slope, so the derivative must also be zero at these points.

This observation allows us to trade one mathematical problem for another one. If we are interested in finding the maximum or minimum values of a function, we can instead try to find the roots of the first derivative. Since it is very likely that the derivative has fewer roots than the function has potential values this new problem may be easier to solve than to try to maximize or minimize the values of the function directly.

5.2. **Differential equations.** When we know a function or are given a function directly, computing the derivative (when it exists) and using the derivative to find maxima and minima is an interesting calculation, but derivatives can be applied in more subtle ways to help us understand both natural and man-made systems.

In practice if we are interested in studying and measuring a particular quantity it is often **much** easier to create a list of factors we think contribute to altering that quantity, rather than try to guess a function which correctly mimics the actual behavior of the quantity of interest.

If we consider the problem of throwing a tennis ball, we might be interested in the question 'where will the ball go?'

This is a complicated, multifaceted question whose answer depends strongly on many interrelated factors including: which direction we throw the ball, how hard we throw the ball, our location (presumably on the Earth), the local geography and current wind patterns, and probably to a lesser degree, gravitational fluctuations due to the location of the moon in orbit around the sun. Plus a whole host of other possible contributing factors.

If we try to mentally create a function which can handle all these factors, then we may as well order out for a straight-jacket, a padded cell,a and a sippy cup, because we will quickly drive ourselves mad. But we can create a model situation which includes the factors that we 'think' are important, and we can try to work out the details of that model in order to find a function which is compatible with that particular picture of the universe (or at least compatible with our cartoon of that particular tiny corner of the universe). Then, provided the physical laws of the universe don't go through a u-turn, we may be able to re-use the function we find any time we are in the same 'tennis-ball' throwing situation. We may find our function is 'good' or 'bad' at predicting the future depending on how well our model captures the physical world.

Amazingly, this process actually works, and very often we can make predictions which are practically useful with a very small collection of different models. In addition to making direct predictions, these models can also help us obtain a more coherent picture of the world around us, making initially complex and frightening phenomena such as hurricanes and tornadoes more understandable and more or less predictable. If anything mathematical modeling will make you more humble when you realize that we can only write down exact solutions to the simplest, most cartoonish models of the world. With our current computational resources we have made a great deal of headway in our ability to model, simulate and predict more complex situations, but the real limits of our ability are probably much closer than you realize. (i.e. most of the effects you see in movies are flowery approximations with a great deal of artistic license and very little predictive power.)

6. Basic Integration

In addition to the differentiation operation, calculus is also concerned with a second operation called **integration**. While the derivative gives us information about change at a single point, the integral operation gives us information about the area underneath curved regions.

The definition of the integral requires a great deal more technical specification than the definition of the derivative, however even without an understanding of the technical specifications, one can obtain a thorough conceptual understanding of the meaning of the integral.

Conceptually, we define the value of the integral in terms of the problem that we want it to solve. Given a function, $f(x)$, we use the notation:

$$\int_a^b f(x)dx$$

to describe the exact value of the area between the graph of the function and the x-axis on the interval (a, b). To fully specify this notation we must define both the function, $f(x)$ and the limits of integration a and b. Without all of these quantities fully specified, we cannot pin down the exact area.

With this conceptual idea defined, we could use graphical representations of functions to help us make rough quantitative judgements about the size of various areas, but the concept doesn't really help us pin down the **values** of the integral in specific cases. To determine the value of a specific integral we must try to use what we know about area to help us discover what we don't know. Just like using the slope of secant lines to approximate the value of derivatives, we might try to approximate the value of an integral using areas that we can compute exactly. If we can create a systematic way to improve our area approximation, then we might try to take a limit and see whether our approximations actually converge or settle down on particular values.

Starting from the simplest idea possible, we might try to approximate the area under a curve using a collection of rectangles. If we choose the rectangles intelligently, by looking at the area under the function, then we can expect our approximation to at least be 'close' to the area we are trying to approximate.

(**Technical Remark:** The primary difficulty with rigorously defining the integral operation, is that we have a great deal of freedom in choosing our rectangles, and we must be able to show that the final result we get for the area is independent of the particular collection of rectangles. If our notion of area under the curve really makes sense, then there must be one value for the integral, the integral cannot take different values depending on how we cut up the area.)

To keep things simple we might choose rectangles with equal widths, but with heights that are chosen to 'match' the function. Using this kind of a structure is a particular kind of Riemann Sum. We approximate the value of the integral, by adding the areas of all of the rectangles.

$$\int_a^b f(x)dx \approx \sum_{i \in P} f(x_i^*)\Delta x_i$$

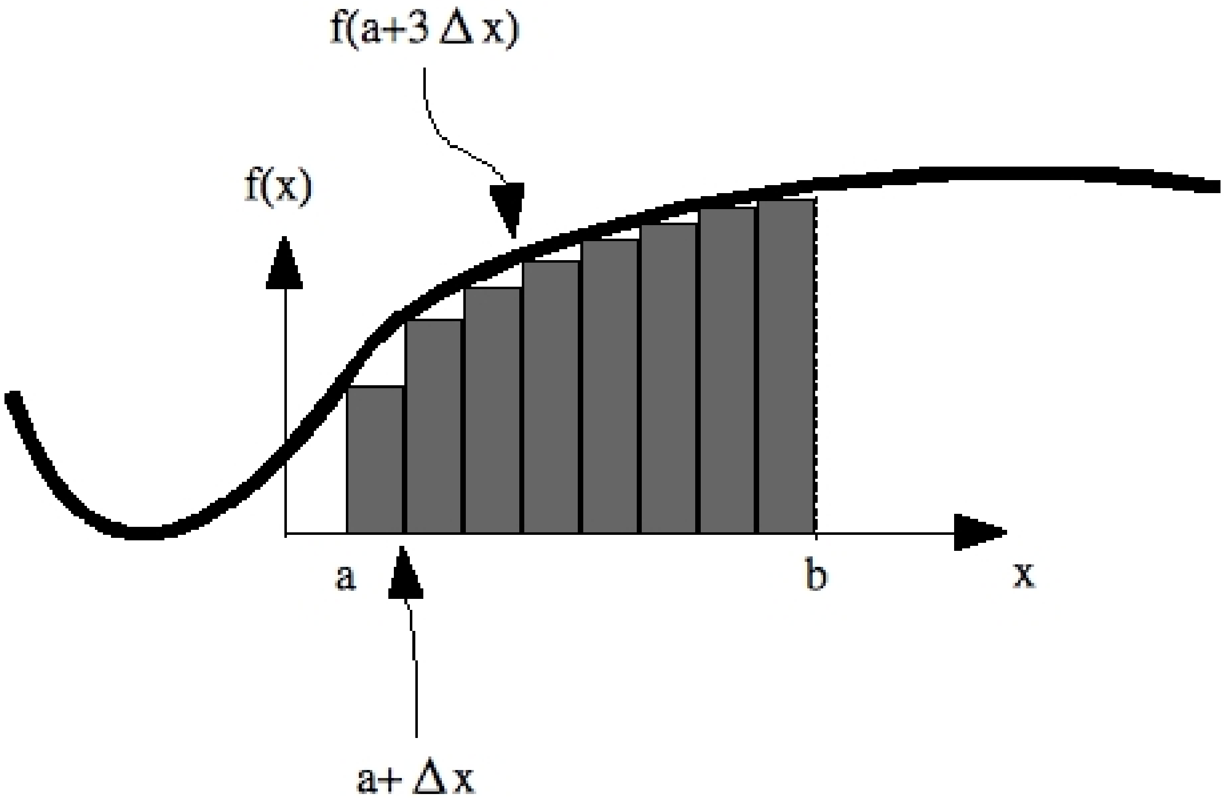


FIGURE 1. The left point rule with $n = 8$.

If we use rectangles which are all equal widths, then Δx_i must be constant. By choosing the function value in different ways we obtain different approximations of the integral.

Suppose we decide to approximate the area under the function f, using n rectangles of equal width, and we choose the height of each rectangle by using the value of the function at the left side of each rectangle. We can see that such an approximation gives a rough approximation of the area underneath the curve.

With this particular construction, the exact formula for our approximation would be given by:

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} f(a+(i-1)\Delta x)\Delta x, \qquad \Delta x = \frac{b-a}{n}$$

Alternately, we might try to approximate the area under f by using the middle of each rectangle, rather than the left endpoint. Graphically this looks slightly different. Using this approximation scheme the exact formula for our approximation with n rectangles would

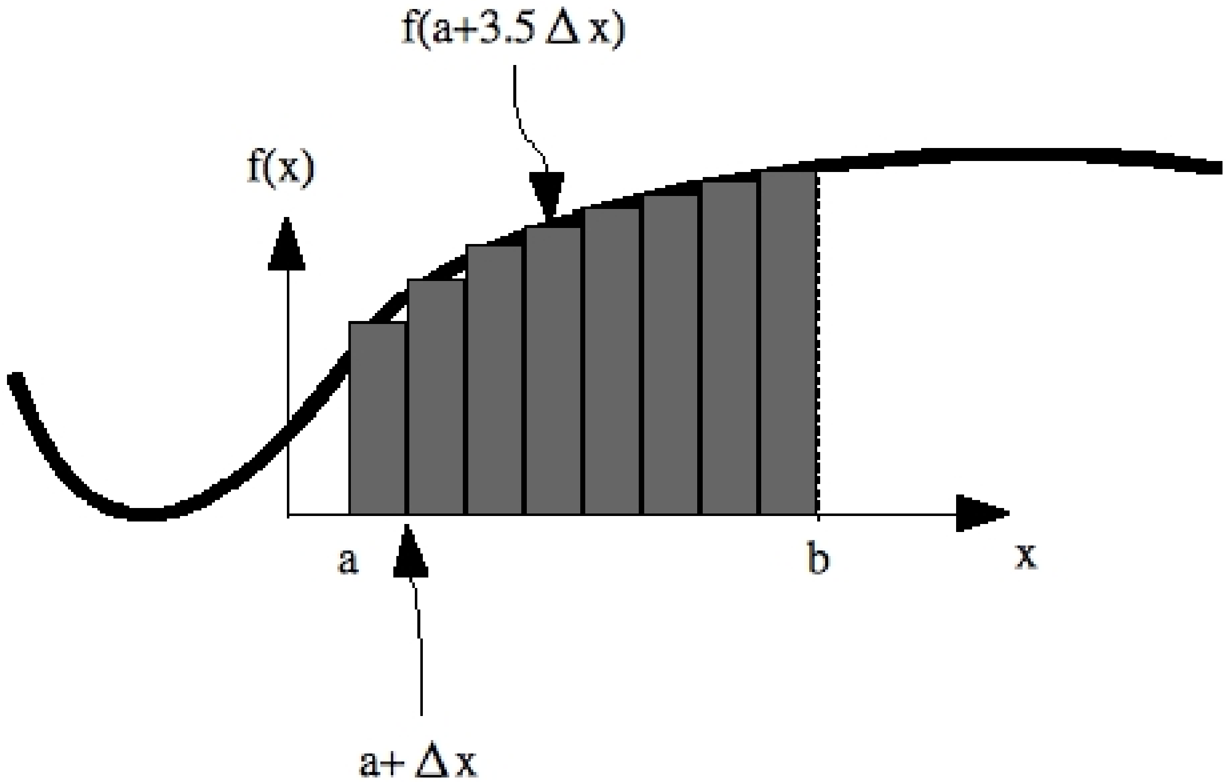


FIGURE 2. The midpoint rule with $n = 8$.

be given by:

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} f\left(a + (i-1)\Delta x + \frac{\Delta x}{2}\right)\Delta x, \qquad \Delta x = \frac{b-a}{n}$$

If we wish to find the exact value of the integral, then we can try to systematically improve this approximation by evaluating the area using more and more rectangles. When this process works (i.e. the limit $n \to \infty$ is well defined), then we have a computational method for approximating and **evaluating** the exact value of integrals.

$$\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{i=1}^{n} f\left(a + (i-1)\Delta x + \frac{\Delta x}{2}\right)\Delta x, \qquad \Delta x = \frac{b-a}{n}$$

In practice evaluating a limit involving a summation is very difficult, so we will seldom use this definition directly when evaluating integrals by hand. In cases where the value of the integral is needed, but an analytic result is not available, this definition acts as the foundation for many numerical approximation schemes.

Example:
The following example shows you exactly what is required to use the definition to determine the exact value of a particular integral.

Consider trying to evaluate:

$$\int_{-1}^{1}(x+1)^2dx$$

Using the right-point rule (not explicitly computed above) the value of this integral is given by:

$$\int_a^b f(x)dx = \lim_{n\to\infty}\sum_{i=1}^{n} f(a+i\Delta x)\Delta x$$

Substituting the given function and limits of integration into our approximation formula we have:

$$\int_{-1}^{1}(x+1)^2dx = \lim_{n\to\infty}\sum_{i=1}^{n}(-1+\frac{2}{n}i)^2\frac{2}{n}$$

Expanding the square and simplifying the resulting expression yields several different summations.

$$\int_{-1}^{1}(x+1)^2dx = \lim_{n\to\infty}\sum_{i=1}^{n}\left(1+\frac{4}{n}i+\frac{4}{n^2}i^2\right)\frac{2}{n}$$
$$\int_{-1}^{1}(x+1)^2dx = \lim_{n\to\infty}\sum_{i=1}^{n}\frac{2}{n}+\frac{8}{n^2}i+\frac{8}{n^3}i^2$$
$$\int_{-1}^{1}(x+1)^2dx = \lim_{n\to\infty}\sum_{i=1}^{n}\frac{2}{n}+\sum_{i=1}^{n}\frac{8}{n^2}i+\sum_{i=1}^{n}\frac{8}{n^3}i^2$$
$$\int_{-1}^{1}(x+1)^2dx = \lim_{n\to\infty}\frac{2}{n}\sum_{i=1}^{n}1+\frac{8}{n^2}\sum_{i=1}^{n}i+\frac{8}{n^3}\sum_{i=1}^{n}i^2$$

Using summation identities we can replace these summations with closed form formulae depending upon n. These formulae are given without proof. (Interested students can investigate proof by induction to help derive and prove these identities for themselves.)

$$\sum_{i=1}^{n}1=n, \qquad \sum_{i=1}^{n}i=\frac{n(n+1)}{2}, \qquad \sum_{i=1}^{n}i^2=\frac{n(n+1)(2n+1)}{6}$$

Returning to the integral calculation, we now have a series of expressions involving n for which the limit may be evaluated.

$$\int_{-1}^{1}(x+1)^2dx = \lim_{n\to\infty} \frac{2}{n}n + \frac{8}{n^2}\frac{n(n+1)}{2} + \frac{8}{n^3}\frac{n(n+1)(2n+1)}{6}$$
$$\int_{-1}^{1}(x+1)^2dx = \lim_{n\to\infty} 2 + 4\frac{n(n+1)}{n^2} + \frac{4}{3}\frac{n(n+1)(2n+1)}{n^3}$$

Taking the limit $n \to \infty$ we obtain:

$$\int_{-1}^{1}(x+1)^2dx = 2 + 4 + \frac{8}{3}$$
$$\int_{-1}^{1}(x+1)^2dx = \frac{26}{3}$$

As you can see, this is a rather lengthy calculation, and this was only for a quadratic function. While it is possible to derive summation identities to help with higher degree polynomials, we really would like a more efficient way to calculate these exact values, preferably a way which doesn't require complicated summation manipulation. The answer solution to this particular little problem comes from a surprising direction.

7. The Fundamental Theorem

In order to get our hands on an efficient way to compute definite integrals we need to be able to make a connection between the integrand and a formula for the value of the integral. The Riemann sum definition does not provide us with such a connection. It turns out that the connection is obtained by trying to differentiate an integral.

Suppose we define the following function:

$$g(x) \equiv \int_{a}^{x} f(t)dt$$

Here the value of g represents the area underneath the function f, on the interval (a, x). Thinking of this value as a function of x allows us to ask questions like: How quickly does the area under f change as we widen or narrow the interval (a, x)?

To answer this question we must be able to take the derivative of $g(x)$. We use the definition of the derivative to start the calculation.

$$\frac{d}{dx}[g(x)] = \lim_{h\to 0}\frac{g(x+h) - g(x)}{h}$$

Now using the definition of g we can fill in the details of the difference quotient.

$$\frac{d}{dx}[g(x)] = \lim_{h\to 0}\frac{\int_{a}^{x+h} f(t)dt - \int_{a}^{x} f(t)dt}{h}$$

If we consider the difference between these two areas geometrically, we can see that their difference must be the area under the sliver of function on the interval $(x, x+h)$. Writing this area as an integral we have:

$$\frac{d}{dx}[g(x)] = \lim_{h\to 0} \frac{\int_x^{x+h} f(t)dt}{h}$$

To proceed further we must find some way to evaluate this limit. The sandwich theorem provides us with a viable avenue.

Rather than deal with the integral of f, we can bound the integral above and below as follows:

$$\int_x^{x+h} \min_{s\in[x,x+h]} f(s)dt \leq \int_x^{x+h} f(t)dt \leq \int_x^{x+h} \max_{s\in[x,x+h]} f(s)dt$$

$$\min_{s\in[x,x+h]} f(s) \int_x^{x+h} dt \leq \int_x^{x+h} f(t)dt \leq \max_{s\in[x,x+h]} f(s) \int_x^{x+h} dt$$

$$\left[\min_{s\in[x,x+h]} f(s)\right] h \leq \int_x^{x+h} f(t)dt \leq \left[\max_{s\in[x,x+h]} f(s)\right] h$$

(i.e. the actual area under the function $f(t)$ on the interval $[x, x+h]$ is always bigger than the area of a rectangle whose height is the smallest value of f between x and $x+h$, and always smaller than the area of a rectangle whose height is the biggest value of f between x and $x+h$.

Dividing by h and letting $h \to 0$ in the above chain of inequalities, when f is a continuous function then the min and max of f must both approach $f(x)$. This yields:

$$f(x) \leq \lim_{h\to 0} \frac{\int_x^{x+h} f(t)dt}{h} \leq f(x)$$

Using this result in our previous computation we find:

$$\frac{d}{dx}[g(x)] = \lim_{h\to 0} \frac{\int_x^{x+h} f(t)dt}{h} = f(x).$$

This result is the fundamental theorem of calculus. If we remove any reference to g, we have just proven the following equality:

$$\frac{d}{dx}\int_a^x f(t)dt = f(x)$$

This result tells us that if we first perform a definite integration, then we follow that by differentiating by the upper limit of integration the result is exactly the function we started with evaluated at the upper limit of integration. This result is technically important, but is not useful when we are just trying to evaluate integrals. If we re-state this result in a form amenable to integral evaluation we find the following:

The statement

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

is true whenever $\frac{dF}{dx} = f$. This statement is helpful for the analytic evaluation of integrals, because it gives us a handle on the connection between a formula for F and the given formula for f. If we can find the function F whose derivative is the function we want to integrate, then we can evaluate any definite integral involving f.

Because of this formulaic connection, we also define a new notation for an indefinite integral:

$$\int f(t)dt.$$

We use this notation when we are interested in determining the anti-derivative of our integrand, without respect to any particular area. The indefinite integral is a one to many operation. When evaluating an indefinite integral, one typically adds a constant of integration to express the whole class of functions whose derivative is given by the integrand. Thus, to fully evaluate an indefinite integral one must find a function $F(t)$ satisfying the relationship $\frac{dF}{dt} = f$, then one writes:

$$\int f(t)dt = F(t) + C$$

C is arbitrary here, but may be specified by supplemental conditions.

8. The Logarithm and Exponential Functions

Now that we have defined both the differentiation and integration operations, we pause to define two functions which are very important for both physical applications and formal computations.

We define the natural logarithm to close a small technical hole in our ability to integrate. Using the fundamental theorem of calculus we find the following rule of integration:

$$\int x^n dx = \frac{x^{n+1}}{n+1} + C$$

this rule follows immediately from the power rule of differentiation and the fundamental theorem of calculus since the derivative of $\frac{x^{n+1}}{n+1}$ is exactly the integrand, x^n. We term this the **power rule for integration**. Using the linear properties of integration and this rule immediately allows us to integrate any polynomial function.

This rule has one small caveat, however. This rule does not tell us how to integrate the function x^{-1}, since the formula for the resulting anti-derivative would be ill-defined in that case (due to division by zero). Thus we have a gap in our ability to integrate powers of x.

Looking at a graph of the function $\frac{1}{x}$, we see the area under this curve is well-defined provided we avoid the infinite discontinuity at the origin. This leads us to define a function representing these areas. This new function will fill the gap in our ability to integrate the powers of x, and will have many interesting properties in it's own right. We define:

$$\ln(x) = \int_1^x \frac{1}{t} dt$$

The lower limit is chosen to be 1 by convention so that we keep away from the discontinuity at zero.

Using the fundamental theorem of calculus we can immediately differentiate this function:

$$\begin{aligned} \frac{d}{dx}\ln(x) &= \frac{d}{dx}\int_1^x \frac{1}{t} dt \\ \frac{d}{dx}\ln(x) &= \frac{1}{x} \end{aligned}$$

Using the definition and this result about differentiation allows us to deduce many interesting algebraic properties of this function. Using the definition along with the rules of differentiation we find:

$$\begin{aligned} \ln(xy) &= \ln(x) + \ln(y), \\ \ln(x^r) &= r\ln(x), \end{aligned}$$

Combining the previous two results and properties of exponents we can also deduce:

$$\ln\left(\frac{x}{y}\right) = \ln(x) - \ln(y)$$

Graphing this function and noting the derivative is strictly positive, we see that the natural logarithm is monotonically increasing and thus is globally one to one. Consequently, this function must possess an inverse function. We term the inverse function the exponential function and define it using the following relationship:

$$exp(x) = y \qquad \text{if and only if } \ln(y) = x$$

This definition ensures that:

$$\begin{aligned} exp(\ln(x)) &= x \\ \ln(exp(x)) &= x \end{aligned}$$

By using our knowledge of the derivative of $\ln(x)$, we can deduce a formula for the derivative of it's inverse function $exp(x)$. This is our first introduction to the technique of implicit differentiation. Suppose we wish to find a formula for the derivative of the exponential function. We use the following procedure:

Begin with an equation defining the exponential function as a new symbol.

$$y = exp(x).$$

Use the natural logarithm to replace this with an equivalent statement without any reference to the exponential.

$$\ln(y) = \ln(exp(x))$$
$$\ln(y) = x$$

Next differentiate both sides of this equation with respect to x.

$$\frac{d}{dx}\ln(y) = \frac{d}{dx}x,$$
$$\frac{1}{y}\frac{dy}{dx} = 1.$$

We have used the chain rule on the right hand side since y depends on x. This dependence forms a second 'layer' within $\ln(y)$, so we must use the chain rule to fully capture the change in this function. Finally, by solving the resulting equation for the quantity $\frac{dy}{dx}$ we will now possess a formula for the desired derivative.

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

Using our original definition of y, we find an explicit formula for the derivative of the exponential function.

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

Thus this function has the interesting property of being it's own derivative. Using more sophisticated mathematical techniques we can show that this function actually behaves like a single number raised to a variable power. The exponential function is often written as:

$$exp(x) = e^x,$$

recognizing the function has this algebraic structure, we can take advantage of the well known properties of exponents which generalize immediately to this function. (These properties can also be deduced by using the definition and differentiation properties.

$$e^{x+y} = e^x e^y$$
$$e^{xy} = (e^x)^y$$
$$e^{x-y} = \frac{e^x}{e^y}$$

9. Trigonometric Functions

The trigonometric functions are often a source of difficulty for students because of the sheer number of different individual facts which must be mastered to use the functions effectively in practice. In addition to memorizing the definitions of the four derived trig functions one usually memorizes the derivatives of each trigonometric function as well as a a whole plethora of identities and reductions which help facilitate various computations. In these notes we emphasize a minimal memorization strategy for dealing with the trigonometric functions. By memorizing the fundamental differentiation identities for cosine and sine and the standard pythagorean identity and having a competent mastery of the rules of differentiation we can differentiate any expression involving the derived trig functions. Using our knowledge of implicit differentiation we are able to derive the formulae for the derivatives of the inverse trig functions, again without any separate memorization effort.

We begin with a short presentation of two dimensional rotation matrices, which can be used to derive angle addition and subtraction identities. Consider the following 2×2 matrix:

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}.$$

For each value of θ this matrix maps points from $\mathbb{R}^2$ into $\mathbb{R}^2$. The net effect of this matrix is a clockwise rotation by θ radians. These rotation matrices are helpful in many circumstances in both physics and mechanics problems, and provide an easy means of rotating a coordinate system to a convenient angle. For our purposes, we only need to know that these matrices behave naturally under composition so that if we transform the plane by rotating by θ radians, and then we rotate again by θ radians, the final transformation would be identical with rotating the original plane by 2θ radians. Algebraically this gives a convenient way of deriving both angle addition and angle subtraction identities.

$$\begin{pmatrix} \cos(\theta+\phi) & \sin(\theta+\phi) \\ -\sin(\theta+\phi) & \cos(\theta+\phi) \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix}$$

Performing matrix multiplication we obtain a different representation of the sines and cosines of the argument $\theta + \phi$.

$$\begin{pmatrix} \cos(\theta+\phi) & \sin(\theta+\phi) \\ -\sin(\theta+\phi) & \cos(\theta+\phi) \end{pmatrix} = \begin{pmatrix} \cos\theta\cos\phi - \sin\theta\sin\phi & \cos\theta\sin\phi + \sin\theta\cos\phi \\ -\sin\theta\cos\phi - \cos\theta\sin\phi & -\sin\theta\sin\phi + \cos\theta\cos\phi \end{pmatrix}$$

For derivation purposes the important identities here are:

$$\cos(\theta+\phi) = \cos\theta\cos\phi - \sin\theta\sin\phi, \qquad \sin(\theta+\phi) = \cos\theta\sin\phi + \sin\theta\cos\phi$$

You can derive the angle subtraction identities form these by carefully replacing ϕ with $-\phi$ everywhere. Such 'on the fly' derivation techniques are very helpful as you mature in your mathematical studies because they free you from the tedium of memorizing long and complex formulae which are easy to confuse and mix up. By using these identities and some well known limits we can easily differentiate cosine and sine.

Using the definition of the derivative we compute the following:

$$\frac{d}{d\theta}\sin\theta = \lim_{h\to 0}\frac{\sin(\theta+h) - \sin\theta}{h}$$
$$\frac{d}{d\theta}\sin\theta = \lim_{h\to 0}\frac{\cos\theta\sin h + \sin\theta\cos h - \sin\theta}{h}$$

Since the limit operation only depends upon h, we group terms and pull the functions of θ outside the limiting operation:

$$\frac{d}{d\theta}\sin\theta = \cos\theta\left[\lim_{h\to 0}\frac{\sin h}{h}\right] + \sin\theta\left[\lim_{h\to 0}\frac{\cos h - 1}{h}\right].$$

Using the well known values of these limits we find:

$$\frac{d}{d\theta}\sin\theta = \cos\theta\,[1] + \sin\theta\,[0]\,,$$
$$\frac{d}{d\theta}\sin\theta = \cos\theta$$

Using the angle addition identity for $\cos(\theta + h)$ we can similarly derive the relationship:

$$\frac{d}{d\theta}\cos\theta = -\sin\theta$$

Using these identities along with the definitions of the derived trig functions ($\sec\theta$, $\csc\theta$, $\tan\theta$, $\cot\theta$). we can differentiate any algebraic expression involving the trig functions.

The inverse trigonometric functions are very helpful in practice for obtaining angles from side lengths, and in dynamic problems we often see inverse trig functions when we seek optimal angles for certain types of optimization problems. Below I will use implicit differentiation to derive an explicit formula for the derivative of the arctangent function.

Arctangent is defined by the following equivalence:

$$\arctan(x) = y \leftrightarrow x = \tan(y)$$

Applying implicit differentiation to the relationship on the right we obtain:

$$\frac{d}{dx}x = \frac{d}{dx}\tan y$$
$$1 = \sec^2 y\frac{dy}{dx}$$

Solving this equation for $\frac{dy}{dx}$ we obtain:

$$\frac{dy}{dx} = \frac{1}{\sec^2 y}$$

Using one of pythagorean identities to write this expression in terms of tangent function, we have:

$$\frac{dy}{dx} = \frac{1}{1+\tan^2 y}.$$

Finally, we note that $x = \tan y$ so this formula may be written:

$$\frac{dy}{dx} = \frac{1}{1+x^2}$$

Returning to the original definition of the arctangent function we now have:

$$\frac{d}{dx} \arctan x = \frac{1}{1+x^2}$$

The formulae for the $\arcsin(x)$ and $\arccos(x)$ are derived using a similar process (exactly the same idea, but you need to employ a slightly different trig identity).

10. Implicit Differentiation and Related Rates

Implicit differentiation has applications beyond differentiating inverse functions, implicit differentiation allows us to apply the differentiation operation to whole equations, and thus to derive new relationships connecting the derivatives of quantities of interest. Word problems which build upon this application are called related rates problems. While these problems may seem impractical and arbitrary, they are an important precursor to the study and manipulation of differential equations which can give us far more realistic and powerful models for analyzing the physical world.

11. Extreme values and Optimization

One of the most powerful and potentially financially regenerative applications of calculus is optimization. When one is interested analyzing smooth functions, calculus allows one to take the problem of finding the maximum or minimum of that function and reduce it to finding and classifying the roots of another function (the derivative). This reduction is very useful because when you are searching for maxima or minima, the values of those extreme values are not usually known beforehand, so you are left trying to find an unknown function value which is bigger (or smaller) than all other function values. When you use calculus, you are instead able to focus your attention on finding places where the derivative takes the value zero. Thus you trade a rather nebulous, non-local condition (the property of being an extreme value) and replace it with a well defined local condition which must hold at any extreme point.

As we study the process of optimization, we will gently sketch out the way the optimization process generalizes in more complex problems, and we will probe the boundaries of traditional calculus 1 material. Keep in mind that the ideas we explore here only form the core and the kernel of the techniques that are applied to complex, real world problems. We begin with a statement of the extreme value theorem, a very important result which gives important theoretical guarantees about the existence of maximum and minimum values.

Extreme Value Theorem:

Any function $f(x)$ which is continuous on the closed interval $[a, b]$ attains both a maximum and a minimum on that interval. Formally we can write this theorem as statement of existence.

If $f(x)$ is continuous on $[a,b] \subset \mathbb{R}$, then:

$$\exists x_{min}, x_{max} \in [a,b] \Big| \forall c \in [a,b] f(x_{min}) \leq f(c) \leq f(x_{max}).$$

The proof of this theorem is a little beyond our current understanding of limits, but the crux of the proof depends upon the topological properties of the closed interval $[a,b]$. In topological terms this interval is **compact**, which conceptually means that the domain is closed and bounded. Closure tells us that anything we can get 'close' to in the domain is actually inside the domain, and boundedness tells us that the domain is totally enclosed so that the continuous function cannot run off to infinity.

At this point the use of a closed interval vs. an open interval may seem very subtle, but it can have a drastic effect on the existence of maxima and minima. This theorem is powerful and useful for use since it provides a certain set of circumstances where maxima and minima are guaranteed to exist.

First and Second Derivative tests

Using information from the first and second derivatives of differentiable functions we can extract geometric information about the functions, and as a consequence we can use these derivatives to efficiently search for and find local extreme values.

The core idea which allows us to apply derivative information towards the search for extreme values, is the observation that the sign of the derivative coincides with increasing, decreasing properties of the underlying function. To summarize:

Let f be a differentiable function:

- If $\forall c \in (a,b) f'(c) > 0$, then f is increasing on (a,b). (If $f'(c) \geq 0$, then we can say f is non-decreasing.)
- If $\forall c \in (a,b) f'(c) < 0$, then f is decreasing on (a,b). (If $f'(c) \leq 0$, then we can say f is non-increasing.)

A function f possesses a **local maxima** at x_{max} whenever

$$\exists (a,b), \forall c \in (a,b) f(x_{max}) \geq f(c)$$

In English this statement says that x_{max} is a local maxima for the function f, whenever $f(x_{max})$ is bigger than or equal to everything in **some** interval around x_{max}. Local minima are defined similarly.

Combining these two ideas we find that in order to have a local maximum at x_{max} there must exist some interval to the left of x_{max} where the function is non-decreasing, and an interval to the right of x_{max} where the function is non-increasing. In order for a local minima to exist we must reverse the conditions. In both of these cases we find that At the local extreme value we find that both these conditions force $f'(x) = 0$. Thus an efficient way to find local extreme values is to search for points where $f'(x) = 0$. (In cases where

f is not differentiable everywhere, we must also extend our search to points where $f'(x)$ does not exist, since our analysis does not cover those points.

In order to emphasize the importance of these points we define the **critical points** of f to be the **set** of all values x_c for which either $f'(x_c) = 0$ or $f'(x_c) = \text{d.n.e.}$

In practice we search for the critical points of f, and then we conduct further analysis to determine the behavior of f surrounding each critical point.
Critical points may be local extreme values, inflection points, cusps, or points of discontinuity and we must analyze them individually to determine their behavior.

12. Optimization

One of the most powerful and versatile applications of calculus is optimization. The goal of an optimization problem is to find the 'best' solution to the problem at hand. When you formulate an optimization problem with mathematics different feasible, or admissible solutions to the problem are ranked by an **objective function**, and depending upon the context one may wish to either maximize or minimize the objective function.

When considering an optimization problem, calculus is an invaluable tool as it allows you to bypass the brute force approach of trying all possible solutions by using the geometric properties of the derivative to seek out important, or **critical** points of the objective function.

Most textbook optimization problems or max/min problems are usually amenable to a formulaic approach. One reads the problem description, determines the objective function (which may be a distance, or a geometric quantity such as area or volume), then one uses any additional constraints or restrictions to write the objective function in terms of a single variable or control. Once the objective function is written in terms of a single variable, one can use the derivative to seek out all of the critical points and hopefully uncover the desired optimum by using one of the critical points.

As a first example consider the following purely geometric optimization problem.
Of all rectangles with unit perimeter, which rectangle has the maximum area?

To formulate this problem we first note that the **area** is the quantity we wish to optimize. We then write down a formulaic representation of this quantity which is relevant to the admissible solutions. We recall the formula for area of a rectangle from elementary geometry:

$$A = LW$$

Next we use the restriction that the perimeter must be one unit long to make a second connection between the control variables of our rectangle. Using the given restriction we obtain:

$$2L + 2W = 1$$

Solving this equation for one of the variables, say L, we find:

$$L = \frac{1}{2} - W$$

Next we substitute this into the formula for our objective function, and seek critical points is terms of W.

$$A = \left(\frac{1}{2} - W\right) W$$

$$\frac{dA}{dW} = \frac{1}{2} - 2W$$

Setting the derivative equal to zero, we find that this particular problem has a unique critical point when $W = \frac{1}{4}$. If we consider the problem restrictions, we see that the restrictions on perimeter restrict the domain for W to the interval $[0, \frac{1}{2}]$. We know from the extreme value theorem that the area function must attain its max and its min on this interval, and that the max and min must occur at either the endpoints, or at a critical point.

Evaluating the area function at either endpoint of the domain, we find $A = 0$, while when we evaluate the area function at the critical point we find $A = \frac{1}{8}$, so this critical point must be the max.

As an example of a more complex mathematical optimization problem we'll work through a proof of the 'well-known' fact that the shortest distance between two points is a line. This is a rather unorthodox example for an introductory calculus course, but it provides us with an actual motivation for defining the arclength of a curve.

In order to formulate this problem in mathematical terms, we pose the following question:

Given two points in the plane, what is the shortest curve which connects both points?

This optimization problem is on a different level of abstraction than our standard max-min problems because the admissible functions are **curves**, not real values. In max min problems you try to find real valued input arguments which make a real valued function large or small, but here your job is to select a curve which has a shorter length than any other admissible curve.

In order to solve this problem we need to formalize what we mean by the length of a curve. Without an objective measure of curve length, we cannot compare the lengths of different curves. To define the exact length of a curve we use the same strategy we used when defining both derivatives and integrals. We use what we know to create an approximation to the length of our curve, then we investigate what happens if we try to systematically improve our approximation. We define the length of a line segment using the geometric distance formula derived from elementary geometry. If we connect two points on a curve f, we can calculate the geometric distance between those points to be given by:

$$dist((x_i, f(x_i)), (x_{i+1}, f(x_{i+1}))) = \sqrt{(x_{i+1} - x_i)^2 + (f(x_{i+1}) - f(x_i))^2}$$

if we partition the segment of the curve which connects the two given points in the plane, then we can add the individual line segments to find an approximation of the length of the curve.

$$L \approx \sum_{i=1}^{n} \sqrt{(x_{i+1} - x_i)^2 + (f(x_{i+1}) - f(x_i))^2}$$

Now if we consider finer and finer partitions of the curve, using more and more line segments, we would suspect that this approximation would get better and better, however we don't actually possess an operation that we can use to take this limit specifically for arclength...so we choose to transform this approximation into a more familiar form. By transforming this formula into the form of a Riemann sum, we know the resulting sum will converge to an integral operation when we refine the partition. To transform this sum into a riemann sum we must introduce a width (i.e. a $\Delta x_i = x_{i+1} - x_i$), so we multiply by a very clever form of 1, and manipulate the resulting expression.

$$\sum_{i=1}^{n} \sqrt{(x_{i+1} - x_i)^2 + (f(x_{i+1}) - f(x_i))^2} = \sum_{i=1}^{n} \sqrt{(x_{i+1} - x_i)^2 + (f(x_{i+1}) - f(x_i))^2}\frac{x_{i+1} - x_i}{x_{i+1} - x_i}$$

$$\sum_{i=1}^{n} \sqrt{(x_{i+1} - x_i)^2 + (f(x_{i+1}) - f(x_i))^2}\frac{x_{i+1} - x_i}{x_{i+1} - x_i} = \sum_{i=1}^{n} \sqrt{(\frac{x_{i+1} - x_i}{x_{i+1} - x_i})^2 + (\frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i})^2}(x_{i+1} - x_i)$$

Simplifying this formula yields:

$$\sum_{i=1}^{n} \sqrt{1 + (\frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i})^2}(x_{i+1} - x_i)$$

Now by taking the limit $n \to \infty$ (with the appropriate adjustment of the x_i), whenever f is at least differentiable this Riemann sum must converge to the following integral:

$$L = \int_{x_{min}}^{x_{max}} \sqrt{1 + \left(\frac{df}{dx}\right)^2} dx$$

This is our definition of the **arclength** of the function f on the interval (x_{min}, x_{max}). In terms of the shortest curve problem, this integral is our objective function. We wish to choose a function f which makes the value of this integral as small as possible.

We approach this problem by using a beautiful argument from the calculus of variations. This argument as a whole uses two pieces which we have not yet investigated in detail, so I will include a short remark at each of those steps.

The basic idea of the argument is to assume the existence of an optimal solution, which we will call $y(x)$, and to use our intuitive understanding of the increasing and decreasing properties of functions of a single variable to re-phrase the problem involving unknown functions, in terms of a new problem with fixed functions and a single unknown variable. This re-phrasing of the problem reduces the problem to a single variable calculus problem. Once we work through the single variable problem, we will have to use integration by parts and a variational lemma to obtain the final answer.

To begin, lets assume that $y(x)$ **is** the shortest curve connecting (x_{min}, y_{min}) to the point (x_{max}, y_{max}). This means that If we plug y into our arclength formula, the result is smaller than the arclength for **any** other curve $z(x)$.

$$L(y) \leq L(z) \forall z(x),$$
$$\int_{x_{min}}^{x_{max}} \sqrt{1 + \left(\frac{dy}{dx}\right)^2} dx \leq \int_{x_{min}}^{x_{max}} \sqrt{1 + \left(\frac{dz}{dx}\right)^2} dx$$

This is really messy, so now we propose to study the implications of this condition on a very special collection of functions. Let's take $y(x)$ and change it slightly (Technically, we are going to introduce a small perturbation to the values of y.) we will change it slightly by adding another function of x. In order to control our perturbation we also introduce a parameter, t. This parameter acts as a 'volume' control for our perturbation. Thus we are now interested in the behavior of the arclength formula when it acts on functions of the form $y(x) + t\eta(x)$.

$$J(y(x) + t\eta(x)) = \int_{x_{min}}^{x_{max}} \sqrt{1 + \left(\frac{d}{dx}\left[y(x) + t\eta(x)\right]\right)^2} dx$$

If we want the function $y+t\eta$ to be an admissible curve connecting the two given points, then we had better select the perturbation function in such a way that $\eta(x_{min}) = \eta(x_{max}) = 0$. Now let's fix y and η and ask how this integral depends on the volume parameter, t.

$$f(t) = J(y + t\eta)$$

Translating our assumptions about y being a minima of J into this new parameterization, we find that $t = 0$ should be a minimum of $f(t)$. This implies $\frac{df}{dt} = 0$ when $t = 0$. If we can successfully differentiate the arclength integral, then we will possess a new equation which must be satisfied by y.

$$
\begin{aligned}
f(t) &= \int_{x_{min}}^{x_{max}} \sqrt{1 + \left(\frac{d}{dx}\left[y(x) + t\eta(x)\right]\right)^2} dx \\
\frac{d}{dt} f(t) &= \frac{d}{dt} \int_{x_{min}}^{x_{max}} \sqrt{1 + \left(y'(x) + t\eta'(x)\right)^2} dx \\
\frac{d}{dt} f(t) &= \int_{x_{min}}^{x_{max}} \frac{d}{dt} \sqrt{1 + \left(y'(x) + t\eta'(x)\right)^2} dx \\
\frac{d}{dt} f(t) &= \int_{x_{min}}^{x_{max}} \frac{1}{2\sqrt{1 + \left(y' + t\eta'\right)^2}} 2\left(y' + t\eta'\right)\eta' dx \\
\frac{d}{dt} f(t) &= \int_{x_{min}}^{x_{max}} \frac{\left(y' + t\eta'\right)}{\sqrt{1 + \left(y' + t\eta'\right)^2}} \eta' dx
\end{aligned}
$$

At this point we recall that we are specifically interested in the situation $t = 0$, since this is an assumed minima of our objective function.

$$
\frac{d}{dt} f(t)|_{t=0} = \int_{x_{min}}^{x_{max}} \frac{y'}{\sqrt{1 + \left(y'\right)^2}} \eta' dx
$$

At this point we will perform a technique called integration by parts. This is a technique for manipulating integrals involving products of different functions, it can be used to compute the values of integrals exactly, and to transform integrals into convenient forms. In this case we are going to use integration by parts to replace η' with η. This transformation comes at a cost.

Integration by parts works by combining the fundamental theorem of calculus with the product rule for differentiation.

$$
\begin{aligned}
\frac{d}{dx}\left[U(x)V(x)\right] &= \frac{dU}{dx}V + U\frac{dV}{dx} \\
\int \frac{d}{dx}\left[U(x)V(x)\right] dx &= \int \frac{dU}{dx}V + U\frac{dV}{dx} dx \\
UV &= \int \frac{dU}{dx}V dx + \int U\frac{dV}{dx} dx \\
\int U\frac{dV}{dx} dx &= UV - \int \frac{dU}{dx}V dx
\end{aligned}
$$

Applying this identity to the problem at hand we obtain the following:

$$\int_{x_{min}}^{x_{max}} \frac{y'}{\sqrt{1+(y')^2}} \frac{d\eta}{dx} dx = \frac{y'}{\sqrt{1+(y')^2}} \eta \Big|_{x_{min}}^{x_{max}} - \int_{x_{min}}^{x_{max}} \frac{d}{dx} \left[\frac{y'}{\sqrt{1+(y')^2}} \right] \eta dx$$

Since we chose $\eta(x_{min}) = \eta(x_{max}) = 0$, so that our perturbed curve would still connect the given points, the boundary terms from this integration by parts are zero. This yields the following:

$$0 = -\int_{x_{min}}^{x_{max}} \frac{d}{dx} \left[\frac{y'}{\sqrt{1+(y')^2}} \right] \eta dx$$

Here we note that η was an arbitrary perturbation, so the calculations we just slogged through apply to any allowed choice of η. This means that the integral above is zero, **independent** of the particular choice of η. Since we could potentially select an η which was positive anywhere we want, the only this could happen is if the 'other stuff' in the integrand was zero. (There is a variational lemma that proves this particular argument whenever you can choose η from a sufficiently broad class of perturbation functions.)

Using this argument we deduce, that if y is a minima of our arclength functional, then the function $y(x)$ must satisfy the **differential equation**:

$$\frac{d}{dx} \left[\frac{y'}{\sqrt{1+(y')^2}} \right] = 0$$

This is a rather nasty looking non-linear differential equation, but we can handle this using a little logic (and some foresight since we know to expect a straight line solution...). First we can immediately integrate both sides of the equation:

$$\left[\frac{y'}{\sqrt{1+(y')^2}} \right] = C$$

Second we can ask when this particular configuration of y''s is constant. Graphing this function, we quickly deduce that the configuration:

$$Q(x) = \frac{x}{\sqrt{1+x^2}}$$

is monotonic and never takes the same value twice...thus if y' changes then this quantity is not constant...we conclude that in order for this equation to be satisfied y' must be constant. The only functions which have constant derivatives are straight lines. By selecting the unique line passing through the two given points, we have found the only possible minima for this variational problem.