Discrete and Continuous Calculus: The Essentials

Author: R. Scott McIntire

Date: January 19, 2009

EMail: scottrsm@gmail.com

I

Preface

If you're anything like me you hate to have to read large tomes devoted to *any* subject, even ones you like. When I see a large volume on a mathematical subject, it is even worse; I find even opening the book depressing. I'm often interested in the essential ideas of a subject and I don't want to wade through an encyclopedic reference. My experience from watching others - as well as myself - is that essential ideas get lost in the endless details of such a reference. Ironically, you can see this in the way people *forget* subject material they've learned. This is not so strange if you think about it. If you've ever had the experience of leaving a town after living there a while I think you'll see what I mean. You'll forget the names of streets after a bit. But, most likely, you will forget the more insignificant streets and places first. If you were to go back to this town, you could probably pick up the information you lost fairly easily. You could do this even if the town had grown or changed a fair amount. If instead, you were to memorize a list of street names with disconnected snap shots of locations I think you would have a different experience remembering/forgetting them. You are just as likely to forget a major road or landmark as a minor one. Of course, if certain roads were bold-ed in the list and you were told that they were more important than the others, this would help you remember them. What helps is to arrange the information in a hierarchy. In this way we prioritize the more important bits of information. The idea of hierarchy occurs in other academic areas as well. If you read a novel there will be major and minor themes in the text. These themes help you remember and connect the novel to other works. For instance, the notion of death and rebirth, or free will versus fate. You may forget much of the details of the novel after you have read it but the themes that the book is dealing with help you remember the plot. While you will forget subject material in this book, I would hope that you will remember the larger themes. If exposed to calculus after reading this text, you will (hopefully) be able to recall the essential nature of the subject.

You may be saying "Well of course the subject should be organized in a hierarchy – every text book is arranged that way." This is certainly true, but in academic subjects, and mathematics in particular, there is a tradition of creating a set of supporting concepts which build to a major result. This process (or series of such) leads to a crescendo with a powerful/amazing result proclaimed. My observation, watching students is that they are disconnected with the material. The feeling one gets with mathematical instruction is that education is done *to you*, rather than one being an active participant in the exploration of the ideas one generates when attempting to solve a problem.

Consequently, the material becomes a linear sequence of small sections that, step by step, obscure the bigger picture of the subject. I saw this process when I was a young boy also. My mother would try to teach me how to balance a check book. It was a step by step process yet I found I couldn't remember the steps if I hadn't done them recently. The problem was that I didn't see the larger purpose, it was all a jumble of steps that didn't make sense to me. Later on, I understood the purpose and could remember what to do. I was also more *adaptable* to change. If the procedure needed to be tweaked, I could make small changes; I wasn't locked into doing the steps exactly the way I was told. My observation is that although books are organized by the hierarchy of chapters, sections, and so forth as any book is, they often fail to impart a useful *conceptual hierarchy* to the student. This is because text books are trying to squeeze a great deal of detailed information between pages. Students typically only see a blur of sections. Regardless of the book's hierarchy, their perception is often of a linear collection of sections; each as important as the last or the next.

Today, mathematics texts are covering more and more ground at shallower and shallower depths. The argument that is often used is "They need to know about such and such before they take course such and such."† I much prefer someone who has an understanding of the basics and can use this base to solve problems than someone who has memorized a long list of information for a test. To do this, I think you need to understand how to think about your subject. This is the intent of this book, to get you to see the how and the why of the *essential* concepts of calculus. This book is not meant to be encyclopedic; there are already a host of such references.

To this end, we will introduce concepts through motivating examples. This means that the concepts may not appear in the order that you would see them in other calculus books, but it is hoped that you will see the point of the various definitions, theorems, and computational rules that arise. This is not to say that other texts don't use motivating examples. It is the case here that the examples will be of primary importance in creating and understanding the hierarchy of concepts of calculus. Major themes will be addressed straight away. In confronting these major themes we will see the need for a number of supporting ideas. Rather than build up to a major theme by investigating related minor themes, the minor themes will be seen as ancillary to our main track. This is what is meant when we say that the concepts may appear out of order; in fact, most of these supporting ideas will be found in the appendices.

In this regard, this text may be considered a form of Test Driven Development. TDD is an idea from the software engineering world that attempts to produce software by starting with tests first and then writing code to ensure that the tests work. In a similar way, we will start with problems we wish to solve and then see what mathematical machinery is needed to solve them. We might call this Problem Driven Development (PDD).

Another organizing principle is the reoccurring theme of discrete versus continuous. Here we develop a discrete as well as the traditional continuous calculus. This comparison is not done in a traditional course on calculus, but I think you will see that the discrete calculus also solves important problems and reinforces concepts from the continuous calculus. The discrete version has the advantage of not having to deal with the technical challenges that the continuous case forces us to consider.

We start by looking at problems dealing with the differencing and the summing of discrete data. In the process we find that differencing and summing are essentially inverses of one another. Next, we look at an analog of differencing and summing applied to continuous data; what are called differentiation and integration and find a similar relationship between them.

I expect the reader to have a good understanding of algebra and algebraic notation. In particular, we take xy to mean x times y; if there is confusion we will explicitly write $x * y$. Although trigonometric functions are used in one of the sections, it is not necessary for an understanding of the rest of the material. Finally, I hope that your experience reading this work gives you an understanding of calculus that will last when you leave this book.

R. Scott McIntire

May 2009

---

† Strangely, physics courses are often six months to a year ahead of the mathematics they use.

Contents

Chapter 1: Difference Calculus

Motivational Problem for Differencing

Suppose there are several cars suspected in the use of an armed robbery, all of a particular make and model. A high speed chase ensued soon after the robbery by the police and the car. The car used had no license plates at the time of the crime; however, it's now believed that the getaway car has legitimate plates. There are not enough visual clues to distinguish the getaway car from the group of suspected cars. The police would like to be able to give enough evidence to a judge in order to get a search warrant for the owner of one of the cars. The one thing the police have going for them is that each car has a sensor which records the odometer reading every 1/10 of a second. The odometer data is stored on an internal disk in feet. They want to examine the data in order to find out if there were high speeds occurring in any of these cars during the time of the getaway.

Each car would have a very long list of numbers representing odometer readings – meaning distance. How should we analyze the data to find velocity? Well, from physics we know the average velocity over a time interval is the distance traveled divided by the time. So, the average velocity over each time interval of 1/10th of a second is found by taking the difference in odometer readings over that interval and then dividing by 1/10. Before discussing any more details, it would be convenient to have a notation to describe long lists of numbers in order to talk about operating on them.

Sequences

We'll define a *sequence* of numbers to be a list of numbers. It is more than a collection of numbers; it is an ordered collection. The distinction is the same as when one considers an ordered list of tasks to perform, a "to do" list, as opposed to an unordered collection of tasks. For instance, the list $\{1, 20, 3, 5, 1, 7\}$ is a sequence of numbers as is the set of positive integers listed in increasing order; or the list of odd numbers listed in increasing order. Since the sequence values are ordered we can talk about the 5^{th} or the 10^{th} element of a sequence (provided there are that many elements). The 5^{th} positive number is of course 5; the 10^{th} odd number is 19. In order to talk about sequences in more generality we use the notation $\{x_n\}_{n=1}^{N}$ to represent the sequence: $x_1, x_2, x_3, \ldots, x_N$; where the x values are numbers and the three dots, ..., mean the intermediate values, in this case the sequence values between x_3 and x_N. Similarly, the notation $\{x_n\}_{n=0}^{N}$ is used to represent the sequence: $x_0, x_1, x_2, \ldots, x_N$. All that changed here is the choice of how to start off the indexing. More generally, $\{x_n\}_{n=a}^{N}$ (a is an integer) is used to represent the sequence: $x_a, x_{a+1}, x_{a+2}, \ldots, x_N$. If there is no value listed after the dots then we mean that there is no end to the sequence – like the sequence of odd numbers. In this case one would write: $\{x_n\}_{n=1}^{\infty}$.† The 5^{th} and 10^{th} values of such a sequence is x_5 and x_{10} respectively.

Sequences as Functions

Sequences can be thought of as functions on the integers, they take an integer as input and return a number. The functions that you are more familiar with, "continuous" functions like

† The symbol ∞ is meant to represent infinity.

$f(x) = x^2$, take a number and return a number. For both sequences and continuous functions there is a distinction to be made between the function and the formula – if any – that defines it. See the appendix "Functions, Formulas, Graphs, and Notation" for a discussion of this. An example of a sequence that can be described by a simple formula is the even numbers. This sequence is $\{2, 3, 6, 8, \ldots\}$. If we call this sequence x, then we can write the sequence values succinctly as $x_n = 2n$. In this case there is a formula that allows us to describe any element of the sequence simply. In cases where there is such a formula we will write the sequence using the formula. In the case of the even numbers we would write $\{2n\}_{n=1}^{\infty}$. Consider now the sequence of primes: $\{2, 3, 5, 7, 11, 13, 17, \ldots\}$. Calling this sequence p, we would like to write $p_n =$ some-nice-formula-in n. However, there is no such nice formula. The sequence exists, but there is no formula for it.

Differencing of Sequences

We know from the introduction that we need to "difference" the distance data to find the velocity. We define the difference symbol Δ (Greek letter Delta which corresponds to D – for difference), to operate on sequences by differencing successive elements of a sequence. Given a sequence, $\{x_n\}_{n=1}^{\infty}$, we define Δx_n by:

$$\Delta x_n = x_{n+1} - x_n \quad \text{for } n \geq 1$$

When using Δ on a sequence, x, we may replace x_n by its formula – if it exists. For instance, if $x_n = 5n + 1$, then we might write Δx_n as $\Delta(5n + 1)$. For this sequence, $\Delta x_n = x_{n+1} - x_n = (5(n+1)+1) - (5n+1) = 5$. If p is the sequence of primes then what is Δp_n? Well, there is no formula for p_n so we can't say – meaning we can't write down a formula for this. However, given a particular n, say 7, $\Delta p_7 = p_8 - p_7$. Since the 8^{th} prime is 19 and the 7^{th} prime is 17, we have $\Delta p_7 = 19 - 17 = 2$.

We can think of Δ as taking a sequence as input and returning a new sequence as the output. If we are given a finite sequence, then one must be careful at the end point. For a finite sequence, $\{x_1, x_2, \ldots, x_N\}$, $\{x_n\}_{n=1}^{N}$, in our notation, Δ is defined to act on this sequence by

$$\Delta x_n = x_{n+1} - x_n \quad \text{for } n \in [1, N-1]$$

In other words, the differencing is only defined up to $N - 1$. That is, up to where the differencing makes sense.

Example 1.1: If our sequence is $\{1, 3, 7, 15, 22\}$, then applying the differencing operation Δ to this sequence yields a new sequence: $\{2, 4, 8, 7\}$.

Notice that this sequence has one fewer element than the original.

Example 1.2: If $\{x_n\}_{n=1}^{\infty}$ is the list of odd numbers listed in increasing order, then after applying Δ to the sequence we obtain a new sequence: $\{2, 2, 2, 2, 2, \ldots\}$.

Sample Difference Formulas

Here are a few difference formulas for sequences of successive powers. Let $x_n = n$, $y_n = n^2$, and $z_n = n^3$, then

$$\begin{aligned}
\Delta x_n &= \Delta n = (n+1) - n = 1 \\
\Delta y_n &= \Delta n^2 = (n+1)^2 - n^2 = n^2 + 2n + 1 - n^2 = 2n + 1 \\
\Delta z_n &= \Delta n^3 = (n+1)^3 - n^3 = n^3 + 3n^2 + 3n + 1 - n^3 = 3n^2 + 3n + 1
\end{aligned}$$

Notice that the right hand side of the above formulas is a polynomial of one less power than the left hand side.

Back to finding the maximum velocity

We can express the average velocity(in feet per second) of the car over the n^{th} interval $[n/10, (n+1)/10]$ as: $v_n = \Delta x_n/10$. So, we have a new sequence, $\{v_n\}$ which represents an approximation to the velocity. More generally, if h is the length of the time interval, then the sequence of approximate velocities may be computed as:

$$v_n = \Delta x_n/h$$

To find the maximum velocity we could write a program which walks the sequence, $\{v_n\}$, looking for the maximum. At each point one compares the maximum value found so far with the current data point. If the data point has a larger value, we change the maximum value to this value. Such a program might take an array (representing the sequence of velocities) and return the maximum value. We offer some code written in the programming language equivalent of Esperanto‡.

```
find-max(v) {
  // The current maximum - so far
  vmax := v[0]
  // The number of values to examine
  N := length(v)

  // Loop over the rest of the values and compare...
   for i in [1,N]  do

     // Update current maximum if bigger value found
     if (vmax < v[i])  then
        vmax := v[i]
     end if
   end for

  // Return the maximum
   return vmax
}
```

Notice the work involved in finding the maximum velocity involves comparing N values.

After running this program on the data from the cars the police crime lab is able to identify a car which hit speeds of 80 mph. Unfortunately, the manufacturer informs them that while the dash-board odometer reading the owner sees is correct, the *historical* odometer data stored on disk is somewhat inaccurate. This "historical" odometer data was designed as a tracking system; originally intended to defeat attempts by used car dealers to illegally modify the odometer readings. It was soon realized that this data could also be used to help service departments do a better job maintaining the car. However, before this idea was implemented, the legal department of the auto manufacturer became worried about potential privacy law suits stemming from this data. They argued that no one wants their every move recorded. Sales and Marketing heard about the odometer tracking and were concerned that even if

‡ We use the characters // to denote the beginning of a comment line.

there were no lawsuits the tracking system could be a liability to sales; it would depend on how customers and consumer groups reacted.

The engineering department had an ingenious solution to these problems. To address the privacy concerns they would alter the data so that it was a bit fuzzy. The data would still be useful to their service departments who would only be using the data for general driving patterns and yet be able to determine velocity and acceleration over short intervals. There was a secondary benefit to modifying the data: They could embed a small signal in the modified data which would act as a finger print; thus making it easier to spot if someone had tampered with the recorded data. This was enough to address the concerns of the Legal and Sales departments.

This left the police with a problem though. They asked the auto manufacturer if they could give some information about their process of "fuzzying up" their data. After all, all the police wanted to know was if one of the cars went over 80 mph on the day of the robbery. The auto maker said that the details of their process was a Trade Secret and they wouldn't release them. However, they would say roughly what the process was. At time step n the actual odometer value, x_n, is modified by multiplying it by a factor which depends on n. Let us call this factor f_n, then the value stored on the disk is $d_n = x_n f_n$. Although the auto maker would not tell what f_n was, they would say that it varied in a range from 0.5 to 2.0 back to 0.5. It varied in this way over a period greater than $10,000,000$ time steps. This meant the period was over 10 days of continuous driving.

What the police have recorded on disk was not the odometer reading, x_n, but $d_n = x_n f_n$. They needed a way to relate the differencing of $x_n f_n$ to the differencing of x_n (Then it is a simple matter to get the velocity – divide by h).

The difference of the disk data is $\Delta d_n = \Delta(x_n f_n) = x_{n+1} f_{n+1} - x_n f_n$. We would like to relate this difference to the difference of the actual distance values: Δx_n. This is accomplished by adding and subtracting the term $x_n f_{n+1}$.

$$\begin{aligned}
\Delta(x_n f_n) &= x_{n+1} f_{n+1} - x_n f_n \\
&= x_{n+1} f_{n+1} - x_n f_{n+1} + x_n f_{n+1} - x_n f_n \\
&= (x_{n+1} - x_n) f_{n+1} + x_n (f_{n+1} - f_n) \\
&= \Delta x_n \, f_{n+1} + x_n \Delta f_n
\end{aligned}$$

Or,

$$\Delta(x_n f_n) = \Delta x_n f_{n+1} + x_n \Delta f_n$$

This gives us the relationship we were after. Since $x_n f_n = f_n x_n$ we also have

$$\Delta(x_n f_n) = \Delta(f_n x_n) = \Delta f_n x_{n+1} + f_n \Delta x_n$$

Let us formalize what we have so far as a theorem. We state the result with different names for the sequences.

Theorem(Discrete Product Rule). *If $\{x_n\}_{n=0}^{\infty}$ and $\{y_n\}_{n=0}^{\infty}$ are sequences then the difference of their product is related to the individual differences by:*

$$\Delta(x_n y_n) = \Delta x_n y_{n+1} + x_n \Delta y_n$$

Or,

$$\Delta(x_n y_n) = x_{n+1} \Delta y_n + \Delta x_n y_n$$

Getting back to our problem, we wish to find Δx_n from $\Delta(x_n f_n)$(the difference of the actual recorded odometer values), and we have the relation: $\Delta(f_n x_n) = \Delta(x_n f_n) = \Delta x_n f_{n+1} + x_n \Delta f_n$. Ideally, we would like to solve this equation for Δx_n. Something like:

$$\Delta x_n f_{n+1} = \Delta(x_n f_n) - x_n \Delta f_n$$

So that

$$\Delta x_n = \big(\Delta(x_n f_n) - x_n \Delta f_n\big)/f_{n+1}$$

Unfortunately, the formula for Δx_n involves x_n, f_n, and Δf_n which are all unknown. But we are not necessarily looking for an exact solution; we would be happy with a statement which said $v_n = \Delta x_n/h \geq 80$ for some value of n. This is actually more incriminating than saying $v_n = 80$.

Since f changes over a very long scale in comparison with x, it is essentially constant for the purposes of our calculations. That is, for all intents and purposes over the short time span of the robbery f_n can be treated as a constant equal to some value, say F. If a sequence, f_n, is constant then $\Delta f_n = 0$. Therefore, the last equation becomes†

$$\Delta x_n \approx \Delta(x_n f_n)/F$$

We still don't know F. Here it is a simple matter for the police to run the car for a short distance and compare the odometer tracking data with the actual distance traveled. Since f_n is virtually constant since the robbery (a few hours before) they are able to run an experiment to find F. They find that it was approximately 0.8 for the suspect car during the robbery, while it is close to 2 for the others. This means their maximum velocity calculation must be corrected by dividing by 0.8 for the suspect car. In this case $v_n = \Delta x_n/h \approx \Delta(x_n f_n)/(Fh) \approx 80/0.8$ mph $= 100$ mph. The police now have enough evidence to get a search warrant for the premises of the owner of this car.

Exercise 1.1: Find a similar formula expressing the difference of the product $x_n y_n z_n$ in terms of the differences: Δx_n, Δy_n, and Δz_n.

† The symbol $\approx$ means approximately equal.

Chapter 2: Summation Calculus

Telescoping Sums

In the last chapter we had a sequence of distance data, $\{x_n\}_{n=0}^{N}$. If x_0 is the odometer reading at the beginning of the day and x_N is the reading at the end of the day, then the distance traveled that day is $x_N - x_0$. Notice also that the distance traveled over the i^{th} interval (our time intervals were every $1/10^{\text{th}}$ of a second) is $x_{i+1} - x_i$. The distance traveled for the whole day should be just the sum of all the distances traveled over each small interval. In mathematical terms, this would be written:

$$\overbrace{(x_1 - x_0)}^{\text{first interval}} + \overbrace{(x_2 - x_1)}^{\text{second interval}} + \overbrace{(x_3 - x_2)}^{\text{third interval}} + \cdots + \overbrace{(x_N - x_{N-1})}^{\text{Nth interval}} = \overbrace{(x_N - x_0)}^{\text{total distance}}$$

It seems like we have taken something obvious and produced a deep formula. If you look at the left hand side of the formula you will see that after adding the first and second distances the x_1's cancel. If we then add the third distance the x_2's cancel. And so on, all the intermediate x's cancel except the first and the last. It isn't that mysterious after all. We say that such a sum is a telescoping sum; the sum acts like a collapsible telescope.

Fundamental Theorem of Difference Calculus

Previously, we introduced a notation to describe sequences and their differences; now, since we are talking about sums, it is convenient to use a notation for summation. The symbols $\sum_{n=1}^{N} x_n$, will mean the sum: $x_1 + x_2 + \cdots + x_N$. In our new notation the above telescoping sum becomes:

$$\overbrace{\sum_{n=0}^{N-1}}^{\text{Discrete Sum}} \overbrace{\Delta x_n}^{\text{Telescoping Difference}} = \overbrace{x_N - x_0}^{\text{Difference at Ends}}$$

Note: The Greek letter $\sum$ (sigma) corresponds to the letter S – for sum.

Although the above formula seemed rather obvious from the way we derived it, we will refer to this formula as *The Fundamental Theorem of Difference Calculus*, or FTDC. The reason why it is important is that it relates something that is computationally hard to do, sum N things, with something that is computationally easy, subtract *two* things. This is very useful if we have a sum of a large number of things and then recognize that we are summing the differences of a sequence. If we make this connection, the computation becomes easy. Suppose for instance, that someone asks you to sum a sequence of the form: $\sum_{n=0}^{N-1} v_n$. If you can find another sequence, $\{x_n\}_{n=0}^{N}$ such that $\Delta x_n = v_n$, then you can sum this series easily by the Fundamental Theorem of Difference Calculus since $\sum_{n=0}^{N-1} v_n = \sum_{n=0}^{N-1} \Delta x_n = x_N - x_0$. The tricky part, of course, is finding the sequence $\{x_n\}_{n=0}^{N}$ whose difference is $\{v_n\}_{n=0}^{N-1}$.

In later chapters we will introduce a notion of "infinitesimal" differences and "continuous" sums which will have a similar relationship. The corresponding theorem is the "Fundamental Theorem of Calculus". It is the corner stone of calculus and it is a generalization of the simple observation above.

Linearity of Summation and Differencing

We write down a basic property of both the summing and differencing operations, $\sum$, and Δ.

1. $\sum_{n=1}^{N}(a\,x_n + b\,y_n) = a\sum_{n=1}^{N} x_n + b\sum_{n=1}^{N} y_n$.
2. $\Delta(a\,x_n + b\,y_n) = a\,\Delta x_n + b\,\Delta y_n$.

From this it follows that $\sum_{n=1}^{N} ax_n = a\sum_{n=1}^{N} x_n$ and $\Delta(ax_n) = a\Delta x_n$. This can be seen by taking $y_n = 0$. These properties say that summation and differencing distribute through (that is, pass through) sums and constant factors. We say that these operations are *linear*. This allows us to more easily compute sums and differences of sequences in terms of their component parts.[†]

Applications of the Fundamental Theorem

As an example of how this might be useful consider the problem of summing the integers from 1 to N: $\sum_{n=1}^{N} n$. It would be nice if we could find a simple formula for this sum in terms of N. If we could recognize this sum as a sum of telescoping differences then we could easily find the answer. That is, if we could find a sequence $\{x_n\}$ such that $\Delta x_n = n$, then $\sum_{n=1}^{N} n = \sum_{n=1}^{N} \Delta x_n = x_{N+1} - x_1$. We call x_n an *anti-difference* of the sequence $\{n\}_{n=1}^{N}$. But, can we find such a sequence $\{x_n\}$? Well, from the last chapter we know that $\Delta n^2 = 2n+1$. We also know that $\Delta n = 1$. Therefore, $\Delta(n^2 - n) = \Delta n^2 - \Delta n = (2n+1) - 1 = 2n$.[‡]

Finally, if we divide by 2 we have $\Delta(n^2 - n)/2 = n$.[*] Therefore, if $x_n = (n^2 - n)/2$ then $\sum_{n=1}^{N} n = \sum_{n=1}^{N} \Delta x_n = x_{N+1} - x_1 = ((N+1)^2 - (N+1))/2 - (1^2 - 1)/2 = N(N+1)/2$. Summarizing:

$$\sum_{n=1}^{N} n = N(N+1)/2$$

So the sum of the first 1000 numbers is $(1000 * (1000+1))/2 = 500,500$. One can extend this technique to finding the sum of squares of the first N integers. The difference of the sequence n^3 is a polynomial of degree 2. So n^3 is an imperfect anti-difference in that its difference will produce n^2 but also create terms in n as well as a constant term. But, as above, if we add combinations of n^2 and n to n^3 we may rid ourselves of these extraneous terms. That is, we would try to find an anti-difference of n^2 as a linear combination of n^3, n^2, and n. Once we have this anti-difference we can easily find the formula for the sum of n^2. Let us flesh this out. First, find a linear combination of n, n^2, and n^3 whose difference is n^2. That is, find c_1, c_2, and c_3 such that

$$\Delta(c_1 n + c_2 n^2 + c_3 n^3) = n^2$$

Since Δ is linear it distributes over the sum and through multiplication by the c's. That is,

$$\Delta(c_1 n + c_2 n^2 + c_3 n^3) = c_1 \Delta n + c_2 \Delta n^2 + c_3 \Delta n^3 = c_1 + c_2(2\,n+1) + c_3(3\,n^2 + 3\,n + 1)$$

As a polynomial in n this is

$$(c_1 + c_2 + c_3) + (2\,c_2 + 3\,c_3)n + 3\,c_3 n^2$$

† If a is 1 and b is -1 the resulting linear combination becomes a difference.
‡ Using the linearity of Δ.
* Linearity of Δ again.

We need to find c_1, c_2, and c_3 so that this is n^2. For this to happen the first two terms should be 0 and the last coefficient should be 1. This gives us three equations:

$$\begin{aligned} c_1 + c_2 + c_3 &= 0 \\ 2\,c_2 + 3\,c_3 &= 0 \\ 3\,c_3 &= 1 \end{aligned}$$

Starting at the last equation and working backwards we have $c_3 = 1/3$, $c_2 = -3\,c_3/2 = -1/2$, and $c_1 = -(c_2 + c_3) = 1/6$. So, an anti-difference is $1/6\,n - 1/2\,n^2 + 1/3\,n^3$. Now we can compute the sum using the FTDC:

$$\begin{aligned} \sum_{n=1}^{N} n^2 &= \sum_{n=1}^{N} \Delta(\frac{1}{6}n - \frac{1}{2}n^2 + \frac{1}{3}n^3) \\ &= (\frac{1}{6}(N+1) - \frac{1}{2}(N+1)^2 + \frac{1}{3}(N+1)^3) - (\frac{1}{6} * 1 - \frac{1}{2} * 1^2 + \frac{1}{3} * 1^3) \\ &= (2\,(N+1)^3 - 3\,(N+1)^2 + (N+1))/6 - 0 \\ &= (N+1)\big[2\,(N+1)^2 - 3\,(N+1) + 1\big]/6 \\ &= (N+1)\big[(N+1) - 1)(2\,(N+1) - 1\big]/6 \\ &= (N+1)(N)(2\,N+1)/6 \end{aligned}$$

Therefore, the formula is $\sum_{n=1}^{N} n^2 = N(N+1)(2\,N+1)/6$.

Summation By Parts

Powerful applications of the FTDC result from equations that have a difference on one side. By summing both sides of the equation, the FTDC implies that the difference side is easy to compute. This means that there is a simple formula for the summation of the other side of the equation. We do have such a formula, it is the Product Rule for differences in the last chapter. It states that:

$$\Delta(x_n y_n) = \Delta x_n y_{n+1} + x_n \Delta y_n$$

Summing both sides of this equation from 0 to N gives us:

$$\sum_{n=0}^{N} \Delta(x_n y_n) = \sum_{n=0}^{N} \Delta x_n y_{n+1} + \sum_{n=0}^{N} x_n \Delta y_n$$

By the FTDC the left hand side is just $x_{N+1} y_{N+1} - x_0 y_0$. On the right hand side we have two complicated sums; moving one of them to the other side of the equation gives us a way to compare two seemingly unrelated sums.

$$\sum_{n=0}^{N} x_n \Delta y_n = \big(x_{N+1} y_{N+1} - x_0 y_0\big) - \sum_{n=0}^{N} \Delta x_n y_{n+1}$$

This formula is referred to as “Summation by Parts”. It relates one sum to another. If one sum is known, we can find the other from this formula.

Applications of Summation By Parts

Let's try to apply summation by parts to the previous example of summing the first N integers. To use "Summation by Parts" we have to map our problem onto this formula. We can do this if we let $x_n = n$ and $\Delta y_n = 1$. We need to find a y_n such that $\Delta y_n = 1$; but, we already know that the sequence $y_n = n$ works. Using the Summation by Parts formula gives:†

$$\sum_{n=1}^{N} \overbrace{n * 1}^{x_n \Delta y_n} = \overbrace{(N+1) * (N+1)}^{x_{N+1} y_{N+1}} - \overbrace{1 * 1}^{x_0 y_0} - \sum_{n=1}^{N} \overbrace{1 * (n+1)}^{\Delta x_n y_{n+1}}$$

We're trying to use this formula to relate our complicated sum to a simpler sum. Instead, we end up with a term that is very similar to what we started with – the last term of the above equation. This last term can be broken into a piece that is *exactly* like the left hand side of the equation. Unfortunately, it seems like we are going backwards. We have

$$\sum_{n=1}^{N} \overbrace{n * 1}^{x_n \Delta y_n} = \overbrace{(N+1) * (N+1)}^{x_{N+1} y_{N+1}} - \overbrace{1 * 1}^{x_0 y_0} - \sum_{n=1}^{N} \overbrace{n * 1}^{x_n \Delta y_n} - \sum_{n=1}^{N} 1$$

But, just as you would when solving an equation for "x" in algebra, you collect all expressions of the unknown to one side and isolate it; everything on the other side is the solution. Our unknown is $\sum_{n=1}^{N} n * 1$; that is, $\sum_{n=1}^{N} n$. So, collecting our unknown on one side and placing the rest of the expressions on the other yields:

$$2 \sum_{n=1}^{N} n * 1 = (N+1) * (N+1) - 1 * 1 - \sum_{n=1}^{N} 1 = N^2 + 2N + 1 - 1 - N = N^2 + N$$

Solving for the sum we have:

$$\sum_{n=1}^{N} n = N(N+1)/2$$

Example 2.1: Find a formula for the sum: $\sum_{n=1}^{N} n^2$.

We can use the summation by parts formula again by letting $x_n = n^2$ and finding a y_n such that $\Delta y_n = 1$. $y_n = n$ works as before. Therefore,

$$\sum_{n=1}^{N} n^2 * 1 = (N+1)^2 (N+1) - 1^2 * 1 - \sum_{n=1}^{N} \Delta n^2 (n+1)$$

This simplifies to:

$$\sum_{n=1}^{N} n^2 * 1 = (N+1)^3 - 1 - \sum_{n=1}^{N} (2n+1)(n+1)$$

Or,

$$\sum_{n=1}^{N} n^2 * 1 = (N+1)^3 - 1 - \left(\sum_{n=1}^{N} 2n^2 + \sum_{n=1}^{N} 3n + \sum_{n=1}^{N} 1 \right)$$

† For clarity we will occasionally write "$a * b$" instead of the conventional "$a\,b$" to indicate the multiplication of a and b.

We have now, as before, an equation for the unknown value $\sum_{n=1}^{N} n^2$. Solving for this value gives:

$$3\sum_{n=1}^{N} n^2 = (N+1)^3 - 1 - 3\sum_{n=1}^{N} n - \sum_{n=1}^{N} 1$$

We already know how to sum $\sum_{n=1}^{N} n$, so we replace it with the formula we just worked out.

$$3\sum_{n=1}^{N} n^2 = (N+1)^3 - 1 - 3N(N+1)/2 - N$$

Collecting terms on the right yields

$$3\sum_{n=1}^{N} n^2 = N^3 + 3N^2 + 3N + 1 - 1 - \frac{3}{2}N^2 - \frac{3}{2}N - N$$

This becomes

$$3\sum_{n=1}^{N} n^2 = N\left(N^2 + \frac{3}{2}N + \frac{1}{2}\right) = N(N+1)(N+\frac{1}{2}) = N(N+1)(2N+1)/2$$

Finally, we have

$$\sum_{n=1}^{N} n^2 = N(N+1)(2N+1)/6$$

Exercise 2.1: Find a formula for the sum: $\sum_{n=1}^{N} n^3$.

Summary

The fundamental theorem of difference calculus can be stated succinctly: the summation operation is the inverse of the difference operation. This can be seen from the FTDC: $\sum_{n=0}^{N-1} \Delta x_n = x_N - x_0$, and writing it more crudely as $\sum \Delta x = x$. What makes this result useful is if for a given sum, $\sum_{n=0}^{N-1} v_n$, one can recognize v_n as the difference of another sequence x_n, then this sum is $\sum_{n=0}^{N-1} \Delta x_n$ and by the FTDC this is just $x_N - x_0$.

The process of recognition is helped greatly if there is a well developed stock of difference formulas. Then when confronted with summing a sequence, we may be able to recognize, or more easily deduce, a sequence which is an *anti-difference* and then easily compute the sum. Actually, finding differencing and anti-differencing formulas for sequences is much like multiplying and dividing numbers. Differencing is easier – like multiplying, while finding anti-differences is harder – like division. To do long division it helps to be fluent with multiplication, because trial and error multiplications are necessary. The same is true for anti-differencing, it helps first to be fluent with difference formulas for a number of sequences. To this end, one would

1. Develop difference formulas for a base collection of sequences.
2. Develop difference formulas for sequences formed in various ways from the base.

This program has been partially carried out in the last chapter. There, we had difference formulas for a few sample sequences: $x_n = c$, $x_n = n$, $x_n = n^2$, $x_n = n^3$. We also had a way of finding formulas for sequences formed by linear combinations and products of other sequences. This helped us above as we tried to find summation formulas for $x_n = n$ and $x_n = n^2$. We will see in the next two chapters this strategy carried out in more detail. It turns out that this process often yields simpler formulas when dealing with the continuous analogs of sequences – continuous functions.

Chapter 3: Differential Calculus

A Motivating Problem for Differential Calculus

Let me be clear that there is not only one motivating problem for this subject. But I think the problem we introduce illuminates the parade of definitions you see in standard calculus books dealing with differential calculus. You would in the course of study of a standard book come across such a problem only after you had built up a certain amount of mathematical machinery. As mentioned in the preface, I prefer to see the problem first and see what motivated the machinery. It helps me, and I hope it helps you, get a better perspective of the subject.

Problem. *Suppose one is to make a storage box from a flat square sheet of metal 12 inches by 12 inches. The box is made by by doing the following:*

1. *Cut along the dotted lines; that is, cut out the corner pieces.*
2. *Fold the remaining four "flaps" up.*

Now, the resulting "box" is a container without a lid. See the figure below.

Figure 3.1: Metal Sheet

Question: What should the "flap distance" 'x' be in order to make a box with maximum volume?

First, it seems clear that the volume is a function of the parameter x. What is this function? The volume of a box is the area of the base times the height. The area of the base is $(12-2x)^2$ while the height of the box is x. So, the volume function (in cubic inches) can be expressed as[†] $V(x) = x(12-2x)^2$. It is not clear from looking at this formula what the best choice of x should be. We could try many values of x and see which gives the largest value of V; it's clear from the diagram that x must be in the range $[0,6]$ (in inches). This is a valid approach, and though we may not get an exact best answer it could give us a value close to the best. For instance, we could plug in the input values $0,1,2,\ldots,6$, in the function V seeing which gave the biggest answer. But how would we know if we missed any input value to try? We could increase the sampling by plugging in the input values $0, 1/2, 1, 1\,1/2, \ldots, 6$ into V. Again, comparing the values of V after plugging in each of these inputs would perhaps give a better answer. But to find the best solution we would have to plug in all the numbers in the interval $[0,6]$ into the function V to find the best solution. The problem here is that there are an infinite number of numbers in the interval, $[0,6]$!

One of the difficulties with this problem is one of visualization. We can visualize a box and its construction but the function V, or for that matter any function, is a fairly abstract notion. It is not a number, it is a thing that takes a number as an input and produces a number as an output. One way to make a function more tangible is to draw its *graph*. When we do this we get back to things that we can see and touch. By drawing the graph of V, we make V more concrete, and we can see another way to find an input, x, which maximizes V. See the figure below.

Figure 3.2: Box volume as a function of flap distance.

If you place a level[‡] against the graph at any point, the level is said to be *tangent* to the graph at that point. The line that passes through the level is called a *tangent line*. If the level is placed at the high point of the graph, the level lies flat. That is, the tangent line at that point is flat. See the figure below.

† See the appendix "Functions, Formulas, Graphs, and Notation".
‡ That is, something with a straight edge like a level or a ruler.

Figure 3.3: Graph of a function and sample tangent lines.

So, here is a strategy for finding the maximum: If someone gives us a graph of this function, place a level at some point on the graph and move it along the graph. Notice that at a bump in the graph – where potentially there is a maximum or minimum value – the level is flat. If we mark the point or points where this occurs you have a set of input candidates for the maximum. You need only compare the value of the function to maximize at these points and at the end points of the interval over which you are searching for a maximum. This, in general, should be a small number of points. See the figures below.

Figure 3.4: Find the maximum of $f(x)$ on the interval [0,4]

Figure 3.5: Find the maximum of $f(x)$ on the interval $[0,\ 4.2]$. Although this is the same function as the last figure, the change in the interval leads to a different solution.

The great thing about this method is that if we can identify these points we have reduced the comparison to a handful (hopefully) of points.

The problem with this method is that it is not much different from the last method. To draw a good approximation to the graph of this function over the interval of interest, $[0, 6]$, you need to sample the function at a large number of points on the interval $[0, 6]$. The more you sample, the better the approximation to the graph. So, the work is not really much different. To make matters worse, for some applications the interval could be huge. To draw the graph would be a great deal of work.

Suppose we had a way to numerically measure the "steepness" of the level applied at any point of the graph *WITHOUT HAVING TO GRAPH IT.* In addition, this "steepness measure" had a numeric value of 0 when the level was flat. In effect, we would have a new function, which I will call V' such that $V'(x)$ is the "steepness" of V at x. If someone were to hand us this function given V, we could find an x which maximizes V by examining *only* points in the set, $\{\, x \mid V'(x) = 0 \,\}$ (the places where V is flat), and the end points. To find a best x, we would plug in each of these inputs into V – hopefully a small number of inputs – and find a maximizing x. Notice again that instead of having to sift through an infinite number of input values we would only have to check a small handful.

But all this is predicated on easily finding the magical function V' given V. It turns out that the "measure of steepness" that we will introduce can be computed as $V'(x) = 12(12-8x+x^2)$. Now we need only find where the steepness is 0. But this is a quadratic and can easily be solved. In this case, $\{\, x \mid V'(x) = 0 \,\} = \{2, 6\}$. If we add the end points of the interval $[0, 6]$, then the set of points to check becomes, $\{0, 2, 6\}$. Plugging either 0, or 6 into V gives 0. If we plug in 2 into V we get 128. Therefore, $x = 2$ gives us the biggest value for V and is therefore the solution to our problem.

Again the magic is that rather than compare an infinite number of input values we end up looking at only a handful of candidate inputs that have the potential to produce the maximum value of V provided:

1. We are able to easily find the steepness of the graph of the function to maximize as a

function of the input, x. That is, we are able to find a steepness function that has a nice formula.

2. We are able to find the places where the steepness function is zero. That boils down to setting the steepness function to zero and solving for inputs. Then comparing the function to maximize on only these inputs (and the end points of the interval).

Measure of Steepness

How do we measure steepness of these tangent lines, or lines in general? Road builders have a similar problem when describing the steepness of their roads. The standard way to describe road steepness is to refer to the road's grade. This is the amount of "rise" of the road divided by its (horizontal) "run"[*]. The measure is *independent of the particular "run" length*. This can be seem from the graph below using simple geometry. This "rise" over "run" number is called the slope of a straight line. It has the property that it is zero when the line is flat.

Figure 3.6: The grade gives the same value regardless of the amount of run.

So, our definition of steepness of a line is the slope of the line – see the appendix "Straight Line Functions". Therefore, the steepness at a point on the graph of a function f at x is the slope of the line tangent to the graph of f at x. Let's formalize this.

Definition Steepness. *The steepness of a function f at x is the slope of the line tangent to the graph of f at x (provided the tangent exists).*

Remember, we need to move from the geometric view of steepness (a tangent line) to an algebraic/mechanical one so we can compute efficiently. That is, we need to avoid having to graph our function in order to solve the maximization problem. What we have now done, in effect, is represent our *geometric* notion of a tangent line in an *algebraic* way as a single number. This is important enough to state formally.

[*] This ratio is actually multiplied by 100, but this distracts from the main idea.

Theorem(Straight Line Representation). *A straight line passing through the origin can be represented by a single number - its slope.*

proof: From the appendix on straight line functions we see that a line with slope m passing through the origin is the graph of the *linear function* $f(x) = mx$. Thus m completely characterizes such a line.

The big question now is: How do we compute this measure of steepness for a given function?

Computing the Slope: Approximating Tangents

So, we know how to measure the steepness of a straight line - it is its slope. Although it's clear what we mean by tangent when we physically put a level against a graph, it is not so clear how to compute it algebraically in terms of the original function.

One way to get a handle on this idea of tangent is to work with approximate tangents to a function at a point, $(x, f(x))$ on the graph. In the figure below you see that the line which passes through the points $(x, f(x))$ and $(x+h, f(x+h))$ is approximately the same as the tangent line. The approximation gets better the closer h is to 0.

Figure 3.7: Tangent and approximating tangent.

Therefore, if we can find the slope of the approximating tangents then we will get an approximation of the slope of the tangent line. Computing the slope of the approximating tangent line is easy, we just need a "rise" and a "run". But the two points on the approximate line provide us with this. The vertical "rise" between the two points is $f(x+h) - f(x)$ while the horizontal "run" is h (we use the variable h to indicate that it is a *horizontal* increment).

Therefore, the slope of the approximating tangent line is†

$$\frac{f(x+h)-f(x)}{h} \tag{1}$$

So, if the approximation to the slope gets better and better the closer h is to 0, then it must be best when h is 0. However, if we plug 0 in for h in the equation above we get $0/0$ – oops! – $0/0$ is not defined. What we need to do is track the values of (1) as h gets smaller and smaller and see where these values "tend".

For instance, given the sequence of numbers $1, 1/2, 1/3, \ldots$, the 10^{th} element in the sequence is $1/10$, the 103^{rd} element in the sequence is $1/103$. In general, the n^{th} element of the sequence is $1/n$.

Question: Where is the sequence tending? Put another way, what is the *limit* of the sequence?

Answer: It seems clear that although this sequence never reaches 0 it is heading there.

Before continuing, let us introduce a notational convenience. To make it easier to talk about sequences we will use the notation $\{1/n\}_{n=1}^{\infty}$ to refer to the sequence $1, 1/2, 1/3, \ldots$; and more generally, we use the notation from chapter 1 to refer to a sequence of values: $\{x_n\}_{n=1}^{\infty} = \{x_1, x_2, \ldots\}$.

Now, referring to the sequence, $1, 1/2, 1/3, \ldots$ ($\{1/n\}_{n=1}^{\infty}$ in our notation) we say that its limit is 0. This leads to a general operational definition of a limit.

Definition (Intuitive Limit). *We say a sequence $\{x_n\}_{n=1}^{\infty}$ has the limit a if the sequence "tends to" a as n "tends to" ∞. We write this with the following notation: $\lim_{n\to\infty} x_n = a$. We also say that x_n converges to a.*

We leave for the time being what exactly "tends to" means. When Calculus was originally developed this notion was not rigorously understood. It was defined much the same way beauty is defined in many people's minds: "I can't define it, but I know it when I see it." For a precise definition of limit see the appendix "Infinite Sequences and Limits".

Now, going back to the approximating slopes we have our first major definition.

Definition (Slope of Tangent Line). *The slope of the tangent line at a point $(x, f(x))$ of the function f exists if the limit*

$$\lim_{n\to\infty} \frac{f(x+h_n)-f(x)}{h_n}$$

exists for every sequence $\{h_n\}_{n=1}^{\infty}$ such that h_n goes to 0 with $h_n \neq 0$ for any n; and, *every one of these limits is identical. In this case, the slope of the tangent line is this unique limiting value.*

That is, if *every* sequence of approximating slopes tends to the same value, then the slope of the tangent line exists and is this value.

† Some books will use Δx rather than h to indicate a small "run". The approximating slope is then written as $\frac{f(x+\Delta x)-f(x)}{\Delta x}$.

Note: This limit may not exist. There are many examples of functions where the limit does not exist at a particular point or a collection of points. Graphically this means the tangent line doesn't exist or is not unique – how can this be? We list a few examples of functions with points on their graphs where the tangent line does not exist.

Figure 3.8: This function does not have a tangent line at $(5, f(5))$.

Figure 3.9: The function $\mathrm{Osc}(x) = \begin{cases} x\sin(1/x) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$ has no tangent at $(0, \mathrm{Osc}(0))$.

Figure 3.10: No unique tangent line at $(0, f(0)) = (0, 0)$. Two possible tangents are drawn with dashed lines.

Definition (Derivative). *The slope of the tangent line (when it exists) at a point $(x, f(x))$ is called the derivative of f at x and is denoted $f'(x)$. We take f' to mean the function whose value at any x is the derivative of f at x. The domain of this new function will be, in general, a subset of the domain of f.‡ We call f' the derivative of f. We say that we differentiate f to get f'.*

We will also write $(f(x))'$ to denote $f'(x)$ when f is represented by an expression, or if it is more readable. For instance, $(g(x)h(x))'$ means the derivative of the function represented by the expression of g times h evaluated at x.

Recap

Recall that in our example problems we needed a function f' which would give the steepness of the tangent line at each point $(x, f(x))$ (now called the derivative). The idea was that f' would be easy to compute from f. From the last section, it seems things are getting more complicated. In the process of trying to find the slopes of the tangent lines, we introduced a fairly complicated notion of limit. Further, it seems we are back to having the same problem we had with sampling and graphing the function to find a tangent slope. In those methods one needed to sample the function at many points. In fact, for a completely accurate result all these methods require one sample f at an infinite number of points to find the slope of f at just one point x! Worse, since we want the slope of the tangent line not just at a given point x, but at all points, things look grim.

The good news is that while the limiting process of finding tangent slopes (the derivative) is an infinite process, it turns out that for a large number of nice functions, f, one can write down a formula for f' at *any* point without doing an infinite amount of sampling.

Derivatives of Straight Line Functions

We start with straight line functions first because the tangent line at any point of a straight line function is the straight line function itself. Therefore, we don't have to go through any

‡ *See the appendix "Functions, Formulas, Graphs and Notation".*

complicated limiting process.

The simplest such straight line functions are the constant functions. A constant function is one which has only one output for all inputs. Such a function has the following formula:

$$f(x) = c$$

What is the slope of the tangent line at a given point x. The graph of this function is a flat line whose height is c. Its tangent line is the line itself. We know a flat line has 0 as its slope. So, what is the value of the derivative function, f' for each x? It is another constant function which is identically zero. That is, $f'(x) = 0$. Again, this says that the slope of the tangent line (derivative) at $(x, f(x))$ is just the slope of the graph of f, which has slope 0.

What about finding the derivative function for a general straight line function? See the appendix "Straight Line Functions" for more information about the formulas for straight line functions. In this case, the function has the formula: $f(x) = m\,x + b$. As in the constant function case the tangent line to each point $(x, f(x))$ is just the graph of f. Therefore, the slope of the tangent line (derivative) is just the slope of the graph of f which is m. So, $f'(x) = m$.

In each case we avoided the limiting process because the graph of the functions *were already* the tangents to the function. This is not so for more complicated functions.

Computing a Non-Trivial Derivative

From algebra class, the next functions studied after straight line functions are the quadratic functions. Let us choose the simplest quadratic, $f(x) = x^2$. Unlike straight line functions, the graph of this function is not the same as the tangent line at any point. To compute the derivative at a point x, we will have to use the approximating tangent lines to compute the answer. Let $\{h_n\}_{n=1}^{\infty}$ be a sequence of numbers which tend to 0 and have the property that $h_n \neq 0$ for all n. Let us compute, for a given n, the slope of the nth approximating tangent. This is given by $\frac{f(x+h_n)-f(x)}{h_n}$ which is

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

Expanding the expression $(x + h_n)^2$ and simplifying yields:

$$\frac{2xh_n + h_n^2}{h_n}$$

Since h_n is not 0, we may cancel an h_n from the top and the bottom of the last expression. This yields

$$2x + h_n$$

What does this tend to as n tends to infinity? The limit is $2x$ since h_n goes to zero. Notice, that this is true for *every* such sequence $\{h_n\}_{n=1}^{\infty}$ that goes to zero and $h_n \neq 0$. In other words, we've found the derivative (slope of the tangent line) at each point x. The formula is $f'(x) = 2x$. *Notice that in one calculation not only have we found a formula for the derivative at one particular point, but for all points!*

A Function without a Derivative

As an example of a function that is not differentiable at a point, consider the absolute value function, $f(x) = |x|$, from the previous figure. The figure demonstrates that there is not a unique tangent at $x = 0$. We now show that the derivative does not exist when $x = 0$ according to our definition. Let $h_n = 1/n$ and try to take the limit of $\frac{f(0+h_n)-f(0)}{h_n}$. This is $\lim_{n\to\infty}(1/n-0)/(1/n)$, since $h_n > 0$; which is $\lim_{n\to\infty} 1 = 1$. Now let's try a different sequence converging to zero. Let $h_n = -1/n$, then $\lim_{n\to\infty} \frac{f(0+h_n)-f(0)}{h_n} = (-1/n-0)/(1/n)$, since $h_n < 0$; this is $\lim_{n\to\infty} -1 = -1$. Since different sequences converging to 0 provide different limiting slopes, the derivative does not exist for this f at 0.

In fact, one can show that if w is any sequence with $w_n > 0$ such that w_n goes to 0 as n goes to infinity then $\lim_{n\to\infty} \frac{f(x+w_n)-f(x)}{w_n} = 1$. And, if z is any sequence with $z_n < 0$ such that z_n goes to 0 as n goes to infinity then $\lim_{n\to\infty} \frac{f(x+z_n)-f(x)}{z_n} = -1$. For the derivative to exist at a given point we need the approximating slopes formed from *any* sequence h (with $h_n \neq 0$ such that h_n goes to 0 as n goes to infinity) to have the same limiting value. It is not enough to have it true for one sequence or a family of sequences, but for *all* sequences with the requisite properties.

However, we did find the derivative of $f(x) = x^2$; so, one function down, how many more to go? Well, there are an infinite number of functions and this was one of the easy ones. This may mean that given a function which has nice algebraic properties, we can do manipulations like we did above and get a formula for the derivative at all points of the function. This is indeed the case for many functions; the algebra just gets messier for more complicated functions. Can we do better than to do this limiting process again and again?

If we look at our function $f(x) = x^2$ with derivative function $f'(x) = 2x$ and then look at the related function $g(x) = x^2 + c$, where c is some constant, can we determine the derivative function, g', from our knowledge of f and its derivative f' without going through the limiting process? Yes, the graph of the function g is just the graph of the function f shifted up by the amount c. That is, g has the same shape as f and so the slope of its tangent lines must be the same as f. We've just argued that $g'(x) = f'(x) = 2x$ without having to go through the limiting process. What about other functions which are simple modifications of f, can we find their derivatives so easily? For instance, consider the function $h(x) = 2x^2$. Can we relate the derivative h' to f' without going through a limiting process? The answer again is yes.

Differential Calculus: Computational Strategy

So, we see differential calculus comes down to finding the derivative function, f', from a function f. Our plan to find the derivative of a function is: Compute the derivative of a base set of functions; we only have to do this once. Next, find formulas for computing the derivatives of functions composed in various ways from these base functions. The base functions and their compositions will be the only functions for which nice formulas will be found for their derivatives. That is to say, we are giving up on finding a formula for the derivative of an arbitrary function. This is not so bad since, hopefully, the pool of functions which we can differentiate will be big enough for the applications we consider. If not, then we will have to compute the derivative of a function the hard way; that is, use the limit process directly.

A Base – a Family of Functions and their Derivatives

We already saw how to find the derivative of the function $f(x) = x^2$ using the limit of tangent line approximations. We present formulas for derivatives of the family of *power* functions, $f(x) = x^n$ but omit the details.

Theorem(Derivative of Power Functions). *If f is the power function $f(x) = x^n$ with n an integer, then the derivative, f', is given by*

$$f'(x) = nx^{n-1}$$

provided $x \neq 0$ when $n < 0$.

Notice that when n is zero the function is $f(x) = x^0 = 1$ and its derivative is $f'(x) = 0$. For a proof of this theorem see the applications section of the appendix on "Mathematical Induction".

Exercise 3.1: Show this result in the case when $n = 3$. That is, derive the derivative of the function $f(x) = x^3$ using the definition of the derivative.

Linearity of Differentiation

A common way of constructing new functions from others is through what is called a *linear combination*. Given functions $f_1, f_2, \ldots, f_n$ a new function f is a linear combination of the functions $\{f_i\}_{i=1}^n$ with constants $\{c_i\}_{i=1}^n$ if

$$f(x) = c_1 f_1(x) + c_2 f_2(x) + \cdots + c_n f_n(x)$$

As an example consider $f_i(x) = x^i$, for $i \in [0, 1, 2]$. Then the function

$$p(x) = 2 + 3x + 5x^2$$

is a linear combination of the functions $f_0(x) = 1$, $f_1(x) = x$, and $f_2(x) = x^2$. The coefficients of the combination (the c's) are 2, 3, and 5.

We are able to form a much larger collection of functions from a set of base functions in this way. And, it turns out one can easily compute the derivative of the new function in terms of the derivatives of its component functions.

Theorem(Linearity of Differentiation). *If the functions f_1' and f_2' exist and $f(x) = c_1 f_1(x) + c_2 f_2(x)$, then the derivative, f', of f exists and may be computed as*

$$f'(x) = c_1 f_1'(x) + c_2 f_2'(x)$$

Note that this means that $(cf)' = cf'$. This follows by letting $f_2(x) = 0$ in the above theorem. This result may be extended to an arbitrary linear combination of functions.

Corollary(Linearity of Differentiation). *If the functions: f_1', f_2', …, f_n' exist and $f(x) = c_1 f_1(x) + c_2 f_2(x) + \cdots + c_n f_n(x)$, then the derivative, f', of f exists and may be computed as*

$$f'(x) = c_1 f_1'(x) + c_2 f_2'(x) + \cdots + c_n f_n'(x)$$

This follows from the previous theorem and mathematical induction.† The theorem and corollary say that the operation of differentiation is linear. This is directly analogous to the linearity result for the difference operation, Δ, of chapter 1. In effect, differentiation (like differencing) passes through constants and sums.

We show this for the case when f is the sum of two functions; that is $f(x) = c_1\, f_1(x)+c_2\, f_2(x)$. Let h be a sequence with the property that $h_n \neq 0$ and h_n tends to 0 as n tends to infinity. This sequence allows us to form approximating tangent lines whose slopes at $(x, f(x))$ are $\frac{f(x+h_n)-f(x)}{h_n}$. This is related to approximating tangent slopes of f_1 and f_2 as follows:

$$\begin{aligned} \frac{f(x+h_n)-f(x)}{h_n} &= \frac{\big(c_1\, f_1(x+h_n)+c_2\, f_2(x+h_n)\big)-\big(c_1\, f_1(x)+c_2\, f_2(x)\big)}{h_n} \\ &= c_1\left(\frac{f_1(x+h_n)-f_1(x)}{h_n}\right)+c_2\left(\frac{f_2(x+h_n)-f_2(x)}{h_n}\right) \end{aligned}$$

Taking the limit of both sides of this equation as h_n goes to 0 gives‡ $\lim_{n\to\infty}\left(\frac{f(x+h_n)-f(x)}{h_n}\right) = c_1\, f_1'(x)+c_2\, f_2'(x)$. Since this limit exists and is the same for *all* such sequences h the derivative of f at x exists and is this value. That is, $f'(x) = c_1\, f_1'(x)+c_2\, f_2'(x)$.

Example 3.1: Let's use this theorem to find the derivative of the polynomial $p(x) = 2+3x+5x^2$.

The theorem says that if we can *recognize* p has a linear combination of functions for which we know how to compute the derivative, then we can easily compute p'. We see that $p(x) = 2f_1(x) + 3f_2(x) + 5f_3(x)$, with $f_1(x) = 1$, $f_2(x) = x$, and $f_3(x) = x^2$. Therefore, $p'(x) = 2f_1'(x) + 3f_2'(x) + 5f_3'(x) = 2*0+3*1+5*2x = 3+10x$.

The linearity theorem combined with the rule for differentiating powers allows us to compute the derivative of any polynomial!

Theorem(Polynomial Differentiation). *Let $p(x) = c_0+c_1x+c_2x^2+\cdots+c_nx^n$, ($n$ a non-negative integer) then*

$$p'(x) = c_1 + 2c_2x + 3c_3x^2 + \cdots + nc_nx^{n-1}$$

This result follows easily because of the linearity of differentiation; we need only know how to find the derivative of each term. But each term is just a power function whose derivative we know.

Armed with our ability to compute derivatives for our base family – the polynomials, let's go back and solve the original problem we posed about the metal box container.

Application Revisited

Consider again the original problem of the metal box container. The problem was to maximize the volume of a box made from a piece of sheet metal. The volume of the box was a function of the "flap distance" x. A formula for the volume was $V(x) = x(12-2x)^2$.

† See the appendix "Mathematical Induction".

‡ Passing the limits through constant multipliers is discussed in the section "Computational Strategy for Limits" in the appendix "Infinite Sequences and Limits".

We will solve this again using our knowledge of how to compute derivatives. We proceed by finding all solutions to $V'(x) = 0$ combined with the end points of our search and then determine a value which maximizes V.

Now, V is a polynomial since, after expanding the formula for V we have: $V(x) = x(144 - 48x + 4x^2) = 144x - 48x^2 + 4x^3$. But we know now how to compute V' when V is a polynomial. According to the last theorem it is $V'(x) = 144 - 2 * 48x + 3 * 4x^2$; or, $V'(x) = 144 - 96x + 12x^2$. Recall the point of having V': *By finding those inputs, x, such that $V'(x) = 0$ we have a list of places where V is flat and therefore a potential maximum.* Solving the equation, $V'(x) = 0$, for x implies that x satisfies $144 - 96x + 12x^2 = 0$. Or, $12 - 8x + x^2 = 0$. Again, the two solutions are 2 and 6. To these values we merge the set of end points, $\{0, 6\}$, to give the candidate input set, $\{0, 2, 6\}$. Plugging in 0 and 6 yields 0. Plugging in 2 produces 128. Therefore, $x = 2$ is the input that gives the maximum value of V, the same result as before.

One may argue that with the advent of computers it might be easier to find this solution by brute force. This is a reasonable argument. However, what if someone asked to maximize the volume that could be produced from square sheets of metal other than 12 inches? One could resolve this problem using Calculus with a parameter representing the size of the square. The resulting formula would allow you to solve an entire *family* of problems. This is something the brute force method cannot do.

Enlarging the Class of Differentiable Formulas

For our derivative technique to be useful for more complicated maximization problems we need to enlarge the class of functions, f, for which we can find a formula for f'. As we know from our strategy we need to do two things: Enlarge the set of base functions for which we know how to compute the derivative; and, find ways to combine these functions in a way that we can determine the derivative from the component functions. How else can we form more complex functions, f, from simpler functions, g, and h? Here are three ways:

1. $f(x) = g(x)h(x)$
2. $f(x) = g(x)/h(x)$
3. $f(x) = g(h(x))$

It turns out one can find the derivative of functions formed in the ways enumerated above in terms of the derivatives of the functions used to compose them. Here are the three corresponding theorems for computing their derivatives. We also restate the theorem on linear combinations for completeness.

Theorem(Linearity Rule). *Let $f(x) = c_1 g_1(x) + c_2 g_2(x) + \cdots + c_n g_n(x)$ then*

$$f'(x) = c_1 g_1'(x) + c_2 g_2'(x) + \cdots + c_n g_n'(x)$$

Theorem(Product Rule). *Let $f(x) = g(x)h(x)$ then*

$$f'(x) = g'(x)h(x) + g(x)h'(x)$$

Theorem(Quotient Rule). *Let $f(x) = g(x)/h(x)$ then*

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

Theorem(Chain Rule). *Let* $f(x) = g(h(x))$ *then*

$$f'(x) = g'(h(x))h'(x)$$

We derive the product rule as it is similar to the product rule in the Discrete Calculus. Let h be a sequence with $h_n \neq 0$ such that h_n goes to 0 as n goes to infinity. Given that $f(x) = g(x)h(x)$, consider the slope of an approximating tangent to f at x based on the sequence h: $\frac{f(x+h_n)-f(x)}{h_n}$. There is a naming problem here. We have been using a sequence named h in our approximating slopes but the function f is composed of a function with the name h. What do we do? Well, we could press on and not let it bother us because a subscripted h is an element of the *sequence* h while $h(x)$ means the *function* h evaluated at x. Or, to be more clear we could change the name of either the sequence or the function h. We will change the name of the sequence. There is nothing special in the name of this sequence. All that is important is that its elements are not 0 and that the sequence tends to 0. So, let's rename the sequence h to z. An approximating tangent slope of f at x is $\frac{f(x+z_n)-f(x)}{z_n}$. We now try to relate this to the approximating tangent slopes of g and h.†

$$\begin{aligned}
\frac{f(x+z_n)-f(x)}{z_n} &= \frac{g(x+z_n)h(x+z_n)-g(x)h(x)}{z_n} \\
&= \frac{g(x+z_n)h(x+z_n)-g(x)h(x+z_n)+g(x)h(x+z_n)-g(x)h(x)}{z_n} \\
&= \left(\frac{g(x+z_n)-g(x)}{z_n}\right)h(x+z_n)+g(x)\left(\frac{h(x+z_n)-h(x)}{z_n}\right)
\end{aligned}$$

Taking the limit as n tends to infinity of the equation above gives: $\lim_{n\to\infty}\left(\frac{f(x+z_n)-f(x)}{z_n}\right) = g'(x)\lim_{n\to\infty} h(x+z_n) + g(x)h'(x)$. We are assuming that the limit of $h(x+z_n)$ exists and are using the properties of the limiting operation as described in the section "Computational Strategy for Limits" in the appendix, "Infinite Sequences and Limits". It turns out that a differentiable function is a continuous function and this means‡ that $\lim_{n\to\infty} h(x+z_n) = h(\lim_{n\to\infty}(x+z_n)) = h(x)$. Finally, since this limit exists for *all* such sequences z, the derivative of f exists and is equal to this value. That is, $f'(x) = g'(x)h(x) + g(x)h'(x)$.

So, given a function f, we can find its derivative if we can recognize it as being either:

1. One of the base functions for which we know the derivative. For us, this is the set of all polynomials and power functions.
2. Composed (in one of the ways above) of functions for which one knows the derivatives.

Keep in mind, there may be many ways to go about computing derivatives. It's really a kind of pattern recognition.

Example 3.2: Suppose $f(x) = (1+x+5x^2)(1-x^2)$. Find $f'(x)$.

We first check to see if this function is one of our base functions. Currently, we only recognize the set of polynomials as our base set. So, this is not a base function. Next, we check to see

† The term $g(x)h(x+z_n)$ is subtracted and added in an analogous way to the derivation of the product rule in the Discrete Calculus.

‡ Also discussed in the section "Computational Strategy for Limits" in the appendix "Infinite Sequences and Limits".

of f can be decomposed into one of the above patterns. Yes, f can be written as the product of two functions, g an h, where $g(x) = (1 + x + 5x^2)$ and $h(x) = 1 - x^2$. Do we know how to compute the derivative of g or h? Yes, each of these are polynomials and we have a formula for computing their derivatives. So, using the product rule we have

$$\begin{aligned} f'(x) &= g'(x)h(x) + g(x)h'(x) \\ &= (1 + x + 5x^2)'(1 - x^2) + (1 + x + 5x^2)(1 - x^2)' \\ &= (0 + 1 + 2 * 5x)(1 - x^2) + (1 + x + 5x^2)(0 - 2x) \\ &= (1 + 10x)(1 - x^2) - 2x(1 + x + 5x^2) \end{aligned}$$

Of course, another way to compute the derivative would be to multiply out the product of g and h. This would give us a polynomial and we could use the formula for differentiating polynomials.

Example 3.3: Let $f(x) = 1 + x + x^2 - 2x$. Find $f'(x)$.

We can find the derivative in one of two ways. First, f is a polynomial isn't it? Well, no, not in the form our theorem on polynomial differentiation requires. Our theorems refer to polynomials as being a linear combination of *unique* power functions. In the case of f, the "x" term appears twice; first as x and then later as $-2x$. It is easy enough to turn this into a "canonical" polynomial, just add like powers. f becomes: $f(x) = 1 - x + x^2$. Now we can apply the theorem for computing derivatives of polynomials giving us: $f'(x) = -1 + 2x$. Another way to compute this is to use the linearity theorem which just distributes the derivative through sums and products with constants. It says $f'(x) = 1' + x' + (x^2)' - 2x'$. Applying the rule for differentiating powers and constants gives us $f'(x) = 0 + 1 + 2x - 2 = -1 + 2x$. This agrees with our previous computation.

These rules are usually easy to apply. The only one which is harder to recognize is the chain rule. Here is an example of how it is used. Let $f(x) = (1 + x^2)^7$. What is the derivative of f? Again, in this case we could raise $(1 + x^2)$ to the 7^{th} power to give us a polynomial and then write down its derivative. We can also recognize f as a composition of the function $g(x) = x^7$ with the function $h(x) = 1 + x^2$. That is, $f(x) = g(h(x))$. Applying the chain rule for the composition of two functions gives

$$\begin{aligned} f'(x) &= g'(h(x))h'(x) \\ &= 7(1 + x^2)^6(1 + x^2)' \\ &= 7(1 + x^2)^6(0 + 2x) \\ &= 7(1 + x^2)^6 2x \\ &= 14x(1 + x^2)^6 \end{aligned}$$

Notice that $g'(x) = 7x^6$ and therefore, $g'(h(x)) = 7h(x)^6$. Since $h(x) = (1 + x^2)$ it follows that $g'(h(x)) = 7(1 + x^2)^6$.

Enlarging the Base

So far, our collection of functions for which we know the derivative is the collection of power functions, $f(x) = x^n$, for integer n, and all polynomials. We now enlarge this base collection to include formulas for computing the derivatives of logarithmic functions (and their inverses) as well as trigonometric functions. The next subsections help to explain the number e that you may have heard of as well as the angle measure called radians. This section is not really needed to understand the essential aspects of Calculus; you may safely skip the next two subsections if these functions don't interest you. You will, however, need to know about the log and exponential functions for the chapter on differential equations.

Derivatives of Log and Exponential Functions

Let's try to compute the derivative at a given point x of the log function base b, $\log_b$. The logarithm only makes sense for positive inputs, so through out our computations, x is assumed positive. The derivative of this function at x is the limit – if it exists – of $\lim_{n\to\infty}\big(\log_b(x+h_n)-\log(x)\big)/h_n$. Here $\{h_n\}_{n=1}^{\infty}$ is a sequence which converges to 0 with the property that $h_n \neq 0$ for all n. This may be written as $\big(\log_b\big(x(1+h_n/x)\big)-\log_b(x)\big)/h_n$. Using the property of logs, this may be written: $\big(\log_b(x)+\log_b(1+h_n/x)-\log_b(x)\big)/h_n$; or, $\log_b(1+h_n/x)/h_n$. Using the property of logs one more time we may write this as $\log_b(1+h_n/x)^{1/h_n}$; or, $\log_b\left[(1+h_n/x)^{x/h_n}\right]^{1/x}$. We need to find the limit of this as n goes to 0. Since the logarithm and the power function $z(s)=s^{1/x}$ are continuous for $x>0$, we may pass the limit through these functions‡ yielding $\log_b\left[\lim_{n\to\infty}(1+h_n/x)^{x/h_n}\right]^{1/x}$. It turns out that the limit inside the log (regardless of $x>0$) is the same as the limit: $\lim_{n\to\infty}(1+1/n)^n$. This limit exists and its limiting value is called e which is approximately 2.718281828459045. So the limit becomes $\log_b\big(e^{1/x}\big)=(1/x)\log_b(e)$. Therefore, we have the formula:

$$\big(\log_b(x)\big)' = (1/x)\log_b(e)$$

If we choose the base of the logarithm to be e, then

$$\big(\log_e(x)\big)' = (1/x)\log_e(e) = 1/x$$

The function, ln, is the name used to denote $\log_e$.

We can use this knowledge to compute the derivative of ln's inverse, the function e^x. By definition, $\ln(e^x)=x$, taking the derivative of this equation yields $\big(\ln(e^x)\big)'=x'=1$. Or,

$$1 = \big(\ln(e^x)\big)' = \overbrace{\tfrac{1}{e^x}}^{h'(g(x))}\ \overbrace{(e^x)'}^{g'(x)}$$

The last equality follows from the chain rule with $h(x)=\ln(x)$ and $g(x)=e^x$. Rewriting, this becomes:

$$(e^x)' = e^x$$

We repeat this procedure for b^x, where b is any positive number. Taking the derivative of the identity $\log_b(b^x)=x$, gives $\big(\log_b(b^x)\big)'=x'=1$. Or,

$$1 = \big(\log_b(b^x)\big)' = \overbrace{\tfrac{1}{b^x}\log_b(e)}^{h'(g(x))}\ \overbrace{(b^x)'}^{g'(x)}$$

‡ A continuous function, f, has the following property: If $\lim\limits_{n\to\infty} x_n = x$ then $\lim\limits_{n\to\infty} f(x_n) = f(\lim\limits_{n\to\infty} x_n) = f(x)$. We say in this case that the limit passes through f. See again, appendix C for a discussion of continuous functions.

The last equality follows from the chain rule with $h(x) = \log_b(x)$ and $g(x) = b^x$. Rewriting, this becomes: $(b^x)' = \frac{1}{\log_b(e)} b^x$. Since, $1/\log_b(e) = \log_e(b) = \ln(b)$, we may write this last expression as

$$(b^x)' = \ln(b) b^x$$

We finish this section with a formula for computing the derivative of the function $e^{a\,x}$. This is really the composition of functions $h(x) = e^x$ and $g(x) = a\,x$; that is, $e^{a\,x} = h(g(x))$. By the chain rule the derivative is $h'(g(x))\,g'(x)$. Since $h'(x) = (e^x)' = e^x$ and $g'(x) = a$ it follows that $(e^{a\,x})' = e^{a\,x} a = a\,e^{a\,x}$. This formula will be of use when we get to the chapter on differential equations. We list our main results:†

1. $\ln'(x) = \frac{1}{x}$

2. $\left(e^x\right)' = e^x$

3. $\left(e^{a\,x}\right)' = a\,e^{a\,x}$

Derivatives of Trig Functions

Let us attempt to compute the derivative of the sine function, sin, at a point x. Again, let $\{h_n\}_{n=1}^{\infty}$ be a sequence which converges to 0 with the property that $h_n \neq 0$ for any n. For any n, the approximating slope to the tangent line is

$$\begin{aligned} \frac{\sin(x + h_n) - \sin(x)}{h_n} &= \frac{\overbrace{\sin(x)\cos(h_n) + \cos(x)\sin(h_n)}^{\sin(a+b) = \sin(a)\cos(b) + \cos(a)\sin(b)} - \sin(x)}{h_n} \\ &= \sin(x)\frac{\cos(h_n) - 1}{h_n} + \cos(x)\frac{\sin(h_n)}{h_n} \end{aligned}$$

We need to be able to find the limits (if they exist) of $\frac{\sin(h_n)}{h_n}$ and $\frac{\cos(h_n)-1}{h_n}$. We will have to do some real work to compute the limit of $\frac{\sin(h_n)}{h_n}$; but once we do, the other limit, $\frac{\cos(h_n)-1}{h_n}$ will follow and, therefore, we will be able to compute the derivative of the sine function. Before we make a precise argument for the limit of $\frac{\sin(h_n)}{h_n}$, note that in the figure below that when θ is small, the length of the segment BC and the length of the arc DC seem to be very close to the same value. In fact, one might conclude that in the limit their ratio $\frac{\sin(\theta)}{\theta}$ is 1. This is what we show now.

Examining the figure and comparing the areas of the triangular region ABC to the sector ADC to the triangle ADE we find that

$$\frac{\overbrace{\cos(\theta)\sin(\theta)}^{\text{Area of triangle ABC}}}{2} \leq \frac{\overbrace{\theta}^{\text{Area of sector ADC}}}{2} \leq \frac{\overbrace{1 * \tan(\theta)}^{\text{Area of triangle ADE}}}{2}$$

† The domain of $\ln'$ is the same as ln which is $\{x \,|\, x > 0\}$.

Figure 3.11: Comparison of triangle ABC, sector ADC, and triangle ADE.

Here, we are assuming θ is positive, measured in radians, and is less than $\pi/2$. The area of each triangle is found using the formula, base * height / 2, while the area of the sector is determined by the formula $\theta r^2/2$. In our case, $r = 1$, giving $\theta/2$; again, provided θ is measured in radians. We briefly describe radian measure for those who have not seen it before.

The radian measure of an angle is the arc length of the angular sector formed by this angle on the unit circle. That is, if a sector has angle θ in radians, then the corresponding arc (on the unit circle) has length θ (see the figure above). The arc length of a complete circle (360 degrees) is 2π; therefore, there are 2π radians in 360 degrees. This means that there are $\frac{360}{2\pi} = \frac{180}{\pi}$ (which is approximately 57) degrees per radian. Also, the area of the unit circle is π; so, the area of a sector of θ radians (on the unit circle) is a fraction of the area of the full circle π. What fraction? Well, the sector of θ radians has an arc length of θ while the arc length of a complete circle is 2π. Since the area of a complete circle is π, the area of the sector must be $\frac{\theta}{2\pi}$ of π. So, the area of the sector is $\theta/2$. One can use other angular measures (such as degrees or grads) but the formulas for the area of a sector have messy constants; and, in turn, the derivatives of the trig functions using these angular measures also have messy constants. Note that this is not unlike the derivative of the logarithmic functions; any choice of base other than e leads to a messier formula.

Continuing with our calculations, the first inequality gives the relation:

$$\frac{\sin(\theta)}{\theta} \leq \frac{1}{\cos(\theta)}$$

while the second inequality yields the relation:

$$\cos(\theta) \leq \frac{\sin(\theta)}{\theta}$$

Combining these gives:

$$\cos(\theta) \leq \frac{\sin(\theta)}{\theta} \leq \frac{1}{\cos(\theta)}$$

A similar argument shows that the relation holds when θ is negative and larger than $-\pi/2$. We may replace θ with the n^{th} element of our sequence h giving:

$$\cos(h_n) \leq \frac{\sin(h_n)}{h_n} \leq \frac{1}{\cos(h_n)}$$

Now, since the limit of the first and third terms exists and is 1, it is the case that the middle term has the same limit.† Therefore,

$$\lim_{n\to\infty} \frac{\sin(h_n)}{h_n} = 1$$

We can use this result to compute the other limit of interest, $(\cos(h_n) - 1)/h_n$. First note that

$$\frac{\cos(h_n)-1}{h_n} = \frac{(\cos(h_n)-1)(\cos(h_n)+1)}{h_n(\cos(h_n)+1)} = \frac{\cos^2 h_n - 1}{h(\cos(h_n)+1)}$$

This may be written

$$\frac{-\sin^2(h_n)}{h_n(\cos(h_n)+1)} = \frac{\sin^2(h_n)}{h_n^2}\frac{h_n}{\cos(h_n)+1}$$

Since the last expression is a product of expressions which have limits, we may write:‡

$$\lim_{n\to\infty} \frac{\cos(h_n)-1}{h_n} = \lim_{n\to\infty} \frac{\sin^2(h_n)}{h_n^2} \lim_{n\to\infty} \frac{h_n}{\cos(h_n)+1} = 1^2 * (0/2) = 1 * 0 = 0$$

Applying these limits to calculate the derivative of sine we have:

$$\begin{aligned}
\lim_{n\to\infty} \big(\sin(x+h_n) - \sin(x)\big)/h_n &= \sin(x) \lim_{n\to\infty} \frac{(\cos(h_n)-1)}{h_n} + \cos(x) \lim_{n\to\infty} \frac{\sin(h_n)}{h_n} \\
&= \sin(x) * 0 + \cos(x) * 1 \\
&= \cos(x)
\end{aligned}$$

Therefore, we have shown that $\sin'(x) = \cos(x)$.

The derivative of the cosine function can be done now without much effort.

$$\begin{aligned}
\frac{\cos(x+h_n) - \cos(x)}{h_n} &= \frac{\overbrace{\cos(x)\cos(h_n) - \sin(x)\sin(h_n)}^{\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b)} - \cos(x))}{h_n} \\
&= \cos(x)\frac{\cos(h_n) - 1}{h_n} - \sin(x)\frac{\sin(h_n)}{h_n}
\end{aligned}$$

Taking the limit we have $\cos'(x) = \cos(x) * 0 - \sin(x) * 1 = -\sin(x)$.

The other trig functions can be described in terms of sin and cos so we may find their derivatives through the derivatives of sin and cos. We list the derivatives of these trig functions and restate the results for sin and cos:

1. $\sin'(x) = \cos(x)$
2. $\cos'(x) = -\sin(x)$
3. $\tan'(x) = \sec^2(x)$
4. $\sec'(x) = \sec(x)\tan(x)$
5. $\csc'(x) = -\csc(x)\cot(x)$
6. $\cot'(x) = -\csc^2(x)$

† This follows from the Squeeze Limit Theorem from the appendix "Infinite Sequences and Limits".

‡ This follows from the Product of Limits Theorem in the appendix "Infinite Sequences and Limits".

To see how these formulas are derived, we compute the derivative of the tangent function, tan. Since tan is the ratio of sin to cos we can use the quotient rule to compute its derivative. The quotient rule says $(f/g)' = (f'g - f\,g')/g^2$. In our case, $f = \sin$ and $g = \cos$. Plugging these into the quotient rule gives

$$\begin{aligned}
\tan'(x) &= \big(\sin(x)/\cos(x)\big)' \\
&= \big(\sin'(x)\cos(x) - \sin(x)\cos'(x)\big)/\cos^2(x) \\
&= \big(\cos^2(x) + \sin^2(x)\big)/\cos^2(x) \\
&= 1/\cos^2(x) \\
&= \sec^2(x)
\end{aligned}$$

This is not valid for $x = \frac{\pi}{2}$; in fact, it is not valid for any x of the form $x = k\,\frac{\pi}{2}$ (k an odd integer). This is because tan is not defined at these values.

Just as was the case with the logarithmic functions, there are inverses of each of the trig functions. Some care must be taken to describe the domains of such inverses, but the approach to compute the derivatives of these inverses is the same as with the logarithms. We ask the reader to consult a calculus tome for a discussion and listing of these functions and their derivatives.

An Interpretation of the Derivative

Let d and v be the functions which give the distance and velocity, respectively, of a car as a function of time. Let's start by assuming that the car is traveling at a constant 30 mph. We know from simple physics that when the velocity is constant the distance function is given as $d(t) = v_0 t$, where v_0 is the velocity. This is simply the statement distance = rate × time. In our case this becomes: $d(t) = 30t$. What do the graphs of d and v look like?

Figure 3.12: Distance graph when car is traveling 30 mph.

Figure 3.13: Velocity graph when car is traveling 30 mph.

Another way to say this is that the velocity = distance ÷ time. This last statement translates geometrically as follows: The slope of the distance graph is 30. More generally, for a car traveling at a constant velocity v_0, the slope of the distance graph is v_0. That is, at any point on the distance graph, $(t, d(t))$, its slope is the velocity at time t, $v(t)$.

What about more general velocity and distance functions. Can we related the distance function, $d(t)$, to the velocity function, $v(t)$ as we did above? The answer is yes. Take a look at the distance and velocity graphs below.

Figure 3.14: Distance function.

Figure 3.15: Velocity function.

The tangent line approximations to the slope at a point in the distance graph are changes in the distance function divided by the time interval. This gives the average velocity over that time interval. As the approximations get better, the intervals of time shrink. In the limit we get the velocity at that instant of time; that is

$$d'(t) = v(t)$$

Using the same arguments one can show the relation between velocity, v, and acceleration, a:

$$v'(t) = a(t)$$

An Application To Physics

If you throw a ball up in the air with an initial velocity v_0 (we take positive values to me up), how high up will the ball travel? This is a maximization problem again; we're maximizing the height a ball travels. From physics, discounting air resistance, the distance as a function of time is: $d(t) = v_0 t - \frac{1}{2}gt^2$, where $h(t)$ is the distance in feet at time t in seconds and g is the constant 32 ft/sec^2. To solve this problem we need to produce candidate inputs for the maximum. This is done by finding the solution set, $\{\, t \mid d'(t) = 0 \,\}$, and adding to this set the end points. There is only one end point, 0; we are not constrained by any end time condition. The solutions set $\{\, t \mid d'(t) = 0 \,\}$ is the set of solutions to $d'(t) = (v_0 t)' - (\frac{1}{2}gt^2)' = v_0 - gt = 0$. There is only one solution which is, $t = v_0/g$. So, the set of input candidates is $\{\, 0, v_0/g \,\}$. Plugging 0 in for t gives $d(0) = 0$ as it should. Plugging in v_0/g in for t gives $d(v_0/g) = v_0^2/g - gv_0^2/2g^2 = v_0^2/g - v_0^2/2g = v_0^2/2g$. If v_0 is positive, then this time corresponds to the maximum height. Positive means throwing the ball up with a nonzero speed. So, if we throw the ball up with speed 80ft/sec, (about 55 mph) then the time which maximizes the height is $80/g$; this is $80/32 = 2.5$ seconds. Plugging in this time into the distance function gives a maximum height of $80^2/(2 * 32) = 6400/64 = 100$ feet.

Where did this distance function come from? Can we derive this formula ourselves with what we know? The answer is yes. The earth's gravitation pull is constant. This is what Galileo observed: acceleration is constant for any object. For our ball this means

$$a(t) = -g$$

To find the corresponding velocity we need to find a function, v, such that $v' = a$. Notice this is the opposite of differentiation. We need to find an *anti-derivative* of a.† A function which when differentiated is a constant, c, is ct. Therefore, $v(t) = -gt$ works. We're using the power rule in reverse. One can easily check its derivative is $-g$. But is this the only function which when differentiated gives $-g$? No, anti-derivatives are not unique. If you remember in an earlier section we argued that if h is a function with derivative, h', then h_1, defined by $h_1(x) = h(x) + c$ also has derivative h'. The idea was that the constant just shifted the graph of h up or down and didn't change h's shape and therefore the slopes of the tangent lines were unchanged. So, we claim that $v(t) = v_0 - gt$ is also a solution to $v' = a$. In this case, $v(t)$ not only has the correct acceleration but also the correct initial velocity. Now, we also know that $d' = v$. So by the same arguments we need to find a function whose derivative is $v_0 - gt$. Using the power rule in reverse again, we see that d could be any one of the functions in the family $c + v_0 t - \frac{1}{2}gt^2$. The constant c is chosen so that the initial distance is right. Since our distance function is measured from the ground, initially our distance must be 0. That is, $d(0) = c + v_0 * 0 - \frac{1}{2}g * 0^2 = c = 0$. So, $d(t) = v_0 t - \frac{1}{2}gt^2$. We've just derived the distance function used in physics to describe an object under the influence of gravity – not a small feat.

Summary

We motivated the definition of the derivative through a maximization problem over an interval.

1. Finding the maximum potentially means sifting though infinitely many points – **this is hard**.
2. Found a *geometric* way to reduce this search to just a handful of points for smooth functions. The candidate solutions come from the end points of the interval and the points where the graph of the function is flat.
3. Defined the derivative as a measure of steepness of the graph - this measure is 0 at the flat points.
4. Found an *algebraic* way to compute the derivative (at *all* points) for a certain class of functions.
5. *Extended* this class to a much larger one by finding formulas for the derivative of linear combinations, products, division, or composition of functions we already know how to differentiate.
6. Now finding the maximum of a *nice* function can be found from a handful of candidate values: the end points and the points where $f'(x) = 0$ – **this is often easy**.

† This is analogous to finding an *anti-difference* of a sequence.

Chapter 4: Integral Calculus

Fundamental Theorem of Calculus

Keeping with the car theme, let $x(t)$ and $v(t)$ represent the distance and velocity of a car at time t. We found a connection between these two functions in the last chapter: the "instantaneous" difference of $x(t)$ over the "instantaneous" difference in time is $x'(t) = v(t)$. We now explore the relationship between these two functions over an *interval* of time. Recall that a similar discrete difference lead to the Fundamental Theorem of Difference Calculus. If $\Delta x_n = v_n$, then the sum over the "time" values, $0, 1, \ldots, N-1$, of the difference, Δx_n, is $\sum_{n=0}^{N-1} \Delta x_n = \sum_{n=0}^{N-1} v_n = x_N - x_0$. We seek an analog of this discrete result.

The sense of the result we would expect is: The "continuous" sum of x' over the time interval $[t_\mathrm{b}, t_\mathrm{e}]$ is equal to the $x(t_\mathrm{e}) - x(t_\mathrm{b})$. For a start, set the discrete times $t_\mathrm{b} = t_0 < t_1 < \ldots < t_N = t_\mathrm{e}$ and create a *sequence*, x, whose n^{th} value, x_n, is defined in terms of the *function*, x, by $x_n = x(t_n)$. Consider the following discrete sum:†

$$\begin{aligned} \sum_{n=1}^{N-1} x'(t_n)\Delta t_n &\approx \sum_{n=1}^{N-1} \frac{\Delta x_n}{\Delta t_n} \Delta t_n \\ &= \sum_{n-1}^{N-1} \Delta x_n = x_N - x_0 = x(t_\mathrm{e}) - x(t_\mathrm{b}) \end{aligned}$$

Note the approximation $x'(t_n) \approx \frac{\Delta x_n}{\Delta t_n}$ just says that the derivative of x at t_n is approximately the "rise" between x_n and x_{n+1}, $x_{n+1} - x_n$, divided by the "run", $t_{n+1} - t_n$.

We now associate a physical meaning with the above sum and extend this discrete sum to a "continuous" one. We start with the simplest case: Let v be the constant function having value v_0 over the time interval $[t_0, t_1]$. Then the change in distance from time t_0 to t_1 is (from high school physics) $x(t_1) - x(t_0) = v_0(t_1 - t_0)$. Another way to interpret this is that the average velocity of the time interval $[t_0, t_1]$ is $v_0 = \frac{x(t_1) - x(t_0)}{t_1 - t_0}$. Suppose instead the car velocity is a piecewise constant function with two values over the interval. The velocity has the constant value v_0 over the interval $[t_0, t_1)$‡, and has the constant value v_1 over the interval $[t_1, t_2]$. Forgetting for the moment how the car manages to jump instantly from one velocity to the next, what is the relationship between the distance function and the velocity function? We know how to compute the change in the distance function over each sub-interval, so the total change is the sum of the changes.

$$\begin{aligned} x(t_2) - x(t_0) &= x(t_1) - x(t_0) + x(t_2) - x(t_1) \\ &= \Delta x_0 + \Delta x_1 \\ &= \frac{\Delta x_0}{\Delta t_0}\Delta t_0 + \frac{\Delta x_1}{\Delta t_1}\Delta t_1 \\ &= v_0 \Delta t_0 + v_1 \Delta t_1 \end{aligned}$$

Suppose instead the car velocity is a piecewise constant function, with velocities $v_0, v_2, \ldots, v_{N-1}$ over intervals $[t_0, t_1), [t_1, t_2), \ldots, [t_{N-1}, t_N]$. Again, let x_n be the sequence

† The symbol $\approx$ means approximately equal.

‡ The half open interval $[a, b)$ is the same set as $[a, b]$ excluding the point b.

defined by $x_n = x(t_n)$. What is the change in the distance over the interval now? The answer is just the sum of all the changes produced by each interval of constant velocity.

$$\begin{aligned}
x(t_N) - x(t_0) &= \Delta x_0 + \Delta x_2 + \ldots \Delta x_{N-1} \\
&= \frac{\Delta x_0}{\Delta t_0}\Delta t_0 + \frac{\Delta x_1}{\Delta t_1}\Delta t_1 + \frac{\Delta x_2}{\Delta t_2}\Delta t_2 + \cdots + \frac{\Delta x_{N-1}}{\Delta t_{N-1}}\Delta t_{N-1} \\
&= v_0\Delta t_0 + v_1\Delta t_1 + v_2\Delta t_2 + \cdots + v_{N-1}\Delta t_{N-1}
\end{aligned}$$

This can be written in summation notation as:

$$x(t_N) - x(t_0) = \sum_{n=0}^{N-1} v_n \Delta t_n$$

Since the velocity has value v_n over the interval $[t_n, t_{n+1})$, the area under that portion of the velocity graph is $v_n\Delta t_n$ (height $\times$ width). The area under the velocity graph from t_0 to t_N is just the sum of these, which is the right hand side of the above formula.

THAT IS, THE AREA UNDER THE VELOCITY GRAPH (over an interval) IS THE CHANGE IN DISTANCE (over the same interval).

In the figure below, the dotted lines are not part of the graph of the velocity function but have been added so that the distance traveled can be seen geometrically as the *area under the velocity graph*. The distance over the interval $[0, 8]$ is the sum of the areas of each of the boxes in the figure below. That is, $x(t_8) - x(t_0) = 30 * 2 + 60 * 2 + 40 * 2 + 80 * 2 = 420$.

Figure 4.1: Piecewise constant velocity.

The corresponding graph of the distance function, x, appears below.

Figure 4.2: Piecewise linear distance function.

We repeat what we have discovered: No matter how complicated a piecewise constant velocity function is, the area under a portion of its graph represents the change in distance over that segment.

But what about velocity functions which aren't piecewise constant? How do we find the distance traveled? Well, we could approximate the velocity function by a piecewise constant function, v_{approx}. We will call the corresponding distance function x_{approx}. The smaller the intervals of constancy, the better the approximation would be to the actual velocity function. Specifically, we divide the interval $[t_{\text{b}}, t_{\text{e}}]$ into N intervals, each with width $h_N = (t_{\text{e}} - t_{\text{b}})/N$. Letting $t_n = t_{\text{b}} + n\,h_N$, for $n \in [0, N-1]$, we define the piecewise constant approximation function to take the value $v(t_n)$ on the interval $[t_n, t_{n+1})$. That is, the approximation takes the value of the function v at the left end point of a given interval and stays at this value for the duration of the interval. In other words, the approximation is a piecewise constant function. For an example, see the figure below where N is 10.

Figure 4.3: Piecewise constant approximation to velocity function, v.

From the figure, the distance traversed by an object with the piecewise constant function, v_{approx} is as before, the sum of the velocities $\times$ the corresponding width of the time interval, which is 10. That is, $x_{\text{approx}}(100) - x_{\text{approx}}(0) = v(0) * 10 + v(1) * 10 + v(2) * 10 + \cdots + v(80) * 10 + v(90) * 10$. Note that each term in the sum is the area of one of the columns in the figure. So, the distance traversed when the velocity profile is v_{approx} is the sum of all the boxes under the curve. In summation notation, we may rewrite the expression for x_{approx} as

$$x_{\text{approx}}(100) - x_{\text{approx}}(0) = \sum_{n=0}^{9} v(10\,n)10$$

Here is a piecewise constant function which does a better job of approximating v.

Figure 4.4: A better approximation to v.

The total distance traversed by an object with velocity v_{approx2} is again the area of its graph, which is the sum of all the rectangular dashed boxes. That is,

$$x_{\text{approx2}}(100) - x_{\text{approx2}}(0) = \sum_{n=0}^{19} v(5\,n)5$$

You can see that if we make more refined approximations to v, the value of the corresponding distance function gets closer and closer to the area under the graph of v and I hope you would agree that this distance function will get closer and closer to x. Therefore, the area under the curve of v is the limit of these refined sums. If we divide the interval $[0,100]$ into n intervals then the approximation would be $\sum_{n=0}^{N-1} v(h_N\,n)h_N$ where $h_N = (100-0)/N$. Just as when we defining derivatives, it would be useful for us to have a notation for the limit of this process.

Definition(Integral). *Given a function f on an interval $[a,b]$. For any integer N we chop up the interval $[a,b]$ into N uniform pieces of length $h_N = (b-a)/N$. By the notation $\int_a^b f(t)\,dt$ we mean†:*

$$\int_a^b f(t)\,dt \stackrel{\text{def}}{\equiv} \lim_{N\to\infty} \sum_{n=0}^{N-1} f(t_n)\Delta t_n$$

where $t_n = a + n\,h_N$ is the left end point of the n^{th} interval.

The sum on the right represents an increasingly refined approximation to the area under the curve using stepwise approximations.‡

We say that $\int_a^b f(t)\,dt$ is the *integral* of f over the interval $[a,b]$. We call the process of computing the integral *integration*.

Note: The symbol, $\int$, is a stretched out 'S', the first letter of "Sum", while dt is interpreted as a single symbol – not multiplication of d times t – and is an echo of the symbols Δt_n used in the approximating discrete sums.

Note: The variable t(and its cousin dt in the expression $\int_a^b f(t)\,dt$ are "dummy" variables. Their notational purpose is to represent the "continuous" looping variable used in this "sum". Just as in programming languages, the name is not important. It is important, however, that it not be a name which causes confusion. For instance, if the interval over which we are we are integrating is $[a,t]$ then the integral, $\int_a^t f(t)\,dt$, doesn't make sense. This is directly analogous to writing a "for loop" that confuses the looping variable with a variable used to determine the end of the loop. The confused code might look like this in the language 'C': `for(i = 0; i < i; i++) {...}`. This loop can be written correctly by using a different looping variable:
`for(j = 0; j < i; j++) {...}`. The incorrectly written integral, $\int_a^t f(t)\,dt$, can be fixed in the same way: $\int_a^t f(s)\,ds$. To make the point clear, let us note that all of the integrals below mean the same thing:

$$\int_a^b f(t)\,dt = \int_a^b f(s)\,ds = \int_a^b f(u)\,du = \int_a^b f(x)\,dx$$

† This definition only works for well behaved functions. For example, it works for continuous functions or a patch work of continuous functions. See the "Computational Strategy for Limits" section of the appendix "Infinite Sequences and Limits" for a definition of a continuous function.

‡ The right hand side is really $\lim_{N\to\infty} S_N$, where $S_N \stackrel{\text{def}}{=} \sum_{n=0}^{N-1} f(t_n)\Delta t_n$. That is, this is really the limit of a sequence, something we have already discussed.

Our new shiny notation says that $\int_{t_\mathrm{b}}^{t_\mathrm{e}} v(t)\,dt =$ *the area under the graph of* v†. This is also the limit of $x_{\mathrm{approx}}(t_\mathrm{e}) - x_{\mathrm{approx}}(t_\mathrm{b})$ as the approximations get closer to x. Therefore, if we believe that the x_{approx} values at t_b and t_e approach the values of x at t_b and t_e, then we must have

$$\int_{t_\mathrm{b}}^{t_\mathrm{e}} v(t)\,dt = x(t_\mathrm{e}) - x(t_\mathrm{b})$$

So, just as with the piecewise constant functions it is the case that the area under v is just the difference in the value of the corresponding distance function, x, at the endpoints.

We can also relate the point-wise relationship of v to x with the above interval relationship. Since $x'(t) = v(t)$, we may write the above as

$$\int_{t_\mathrm{b}}^{t_\mathrm{e}} x'(t)\,dt = x(t_\mathrm{e}) - x(t_\mathrm{b})$$

It turns out that this formula is true even if x is not a distance function. That is, given any function F, with corresponding derivative, F', we may write what is called *the Fundamental Theorem of Calculus*:

Theorem(Fundamental Theorem of Calculus). *If F is a function such that F' exists and is continuous over the interval $[a, b]$, then*

$$\int_a^b F'(t)\,dt = F(b) - F(a)$$

This is really the continuous analog of the Fundamental Theorem of Difference Calculus:

$$\overbrace{\sum_{n=0}^{N-1}}^{\text{Discrete Sum}} \quad \overbrace{\Delta F_n}^{\text{Telescoping Difference}} \quad = \quad \overbrace{F_N - F_0}^{\text{Difference at End Points}}$$

We write the Fundamental Theorem of Calculus (FTC) again with suggestive markup to show the analogy with the above discrete equation:

$$\overbrace{\int_a^b}^{\text{Continuous Sum}} \quad \overbrace{F'(t)\,dt}^{\text{Infinitesimal Telescoping Difference}} \quad = \quad \overbrace{F(b) - F(a)}^{\text{Difference at End Points}}$$

We will show how this result follows by breaking up the interval $[a, b]$ into N pieces (where N is large) and approximating the integral with a sum of telescoping differences of a related sequence. Then we can use the Fundamental Theorem of *Difference* Calculus (FTDC) to show the FTC.

To implement this program, note that each interval will have length $h_N = (b - a)/N$. Let F be a sequence (we use the same name as the function) whose n^{th} element, F_n, is the value of left end point of the n^{th} interval (intervals are labeled starting at zero). That is,

† For this statement to be true we are implicitly assuming that the velocity function is non-negative.

$F_n = F(t_n)$ where t_n represents the left end point of the n^{th} interval; i.e., $t_n = a + n\,h_N$. Notice that $F_0 = F(a)$ and $F_N = F(b)$. In addition, note that $t_{n+1} = t_n + h_N$ $(\Delta t_n = h_N)$ so that $F_{n+1} = F(t_{n+1}) = F(t_n + h_N)$. We also know that $F'(t_n) \approx \frac{\Delta F_n}{\Delta t_n}$.† We now perform a crude calculation to show why the Fundamental Theorem is true:

$$\int_a^b F'(t)\,dt \approx \sum_{n=0}^{N-1} F'(t_n)\Delta t_n \approx \sum_{n=0}^{N-1} \frac{\Delta F_n}{\Delta t_n}\Delta t_n = \overbrace{\sum_{n=0}^{N-1} \Delta F_n = F_N - F_0}^{\text{FTDC}} = F(b) - F(a)$$

This approximation can be seen by performing a slightly more careful analysis:

$$\begin{aligned}
\int_a^b F'(t)\,dt &\approx \sum_{n=0}^{N-1} F'(t_n)\Delta t_n \\
&= \sum_{n=0}^{N-1} \Big(\frac{\Delta F_n}{\Delta t_n} + \text{error}(t_n)\Big)\Delta t_n = \sum_{n=0}^{N-1} \Delta F_n + \overbrace{\sum_{n=0}^{N-1} \text{error}(t_n)\Delta t_n}^{\text{very small term}} \\
&\approx \sum_{n=0}^{N-1} \Delta F_n = F_N - F_0 = F(b) - F(a)
\end{aligned}$$

Here, $\text{error}(t_n)$ is the correction term to make $\Delta F_n / \Delta t_n$ equal to $F'(t_n)$. The only hard part of our crude calculation is showing that $\sum_{n=0}^{N-1} \text{error}(t_n)\Delta t_n$ goes to zero as the N gets large. If we call $\text{error}_{\max}$ the largest of the errors in all the intervals, then we may bound the "very small term" expression above by $\text{error}_{\max} \sum_{n=0}^{N-1} \Delta t_n$. This, in turn, is bounded by $\text{error}_{\max}(b-a)$‡ We need now only believe that the maximum error term will go to zero as N goes to infinity.

How do we use this theorem? For instance, suppose I want to compute the area under the curve of a function f over the interval $[a, b]$, how does this theorem help? Well, we need to find an F such that $F' = f$. Then

$$\int_a^b f(x)\,dx = \int_a^b F'(x)\,dx = F(b) - F(a)$$

We call F an *anti-derivative* of f. It turns out that F is not unique; but all anti-derivatives differ by a constant. It doesn't matter which one you use, you will get the same answer*. Now we need to be able to find an anti-derivative of a function and we'll be done.

Let's see how to apply these ideas to a problem. Keeping with the car theme, suppose that you didn't have sensors in your car, but you were told that the velocity of the car over the interval $[0, 10]$ (in hours) would be $v(t) = t^2$ mph. How far will the car have traveled over this time period. With our new notation we can write down the answer $\int_0^{10} t^2\,dt$. Although this answer is concise, the notation hides the real work involved in its computation. That is, to

† The approximate derivative is obtained by dividing the "rise" in F, ΔF_n, by the "run", Δt_n.

‡ $\sum_{n=0}^{N-1} \Delta t_n = t_N - t_0 = b - a$ by the FTDC.

* If G is another anti-derivative then $G(x) = F(x) + c$ for some constant c. So $G(b) - G(a) = (F(b) + c) - (F(a) + c) = F(b) - F(a)$.

compute this value from its definition would be great deal of work. But, the Fundamental Theorem of Calculus says we may avoid that work by evaluating a certain function x at 10 and 0 and taking the difference. It is a function that has the property that $x'(t) = v(t) = t^2$. Can we guess what functions have this property? We are looking for an *anti-derivative* of v. We know that from the chapter on differentiation that differentiation of power functions reduces their degree by 1. To find an anti-derivative we should do the opposite; that is, look for a function of degree one more than 2. Let's try $x(t) = t^3$. In this case, $x'(t) = 3t^2$. This isn't quite right, we're off by a factor of 3. If we let $x(t) = t^3/3$ then $x'(t) = t^2$ and we can find the distance traveled, it is $x(10) - x(0) = 10^3/3 - 0^3/3 = 1000/3 = 333\,\frac{1}{3}$ miles. That is,†

$$\int_0^{10} t^2\,dt = t^3/3\,|_{t=10} - t^3/3\,|_{t=0} = 333\,\frac{1}{3}$$

We make a formal statement of the anti-derivative style FTC.

Theorem(FTC Anti-derivative Style). *If f is a function and F is an anti-derivative of f, then the following is true:*

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

proof: This just says $\int_a^b f(x)\,dx = \int_a^b F'(x)\,dx = F(b) - F(a)$ (since $F' = f$). And the last equality is just the statement of the fundamental theorem of calculus.

So, in order to compute an integral of a function we need to find an anti-derivative of that function.

Exercise 4.1: Think about why all anti-derivatives of a function f differ by a constant. Hint: Let F and G be two anti-derivatives of f. Show that the derivative of $F - G$ is zero. What kind of functions have zero slope at every point?

Integral Calculus: Computational Strategy

We would like to be able to compute an *anti-derivative* of a function in order to be able to integrate it. We use the notation $\int f(x)\,dx$ to indicate an anti-derivative. That is, when we omit the boundaries of the integral we mean an anti-derivative; with the boundaries we mean the integral over an interval.

We proceed in the same manner as with the differential calculus. First, we find anti-derivatives for a base class of functions. Next, we find formulas for anti-derivatives of functions composed in some way from simpler functions. Just as with the differential calculus, we use the power functions as a base set.

A Base – a Family of Functions and their Anti-derivatives

Theorem(Anti-derivative of Power Functions). *An anti-derivative of the power function $f(x) = x^n$ is $F(x) = x^{n+1}/(n+1)$; valid for integer n, $n \neq -1$; valid for all x when $n \geq 0$ and for*

† One last bit of notation, we use $f(t)\,|_{t=a}$ to mean $f(a)$. This is similar to notation you would see in other texts.

all $x \neq 0$ when $n < 0$. This may be written:

$$\int x^n \, dx = x^{n+1}/(n+1)$$

Note that when $n = -1$ we already know an anti-derivative (only valid for $x > 0$): $\int 1/x \, dx = \ln(x)$

Enlarging the Class of Functions to Integrate

Here is a result for the second stage of our program. It is a linearity theorem similar to the one for summation calculus.

Theorem(Linearity of the Anti-derivative). *An anti-derivative of the function f, where $f(x) = c_1 f_1(x) + c_2 f_2(x)$ is c_1 times an anti-derivative of f_1 plus c_2 times an anti-derivative of f_2. This may be written:*

$$\int f(x) \, dx = c_1 \int f_1(x) \, dx + c_2 \int f_2(x) \, dx$$

Note that the when $f_2(x) = 0$, the above implies $\int cf(x) \, dx = c \int f(x) \, dx$.

Corollary(Linearity of the Anti-derivative). *An anti-derivative of the function $f(x) = c_1 f_1(x) + c_2 f_2(x) + \ldots + c_n f_n(x)$ is*

$$\int f(x) \, dx = c_1 \int f_1(x) \, dx + c_2 \int f_2(x) \, dx + \ldots + c_n \int f_n(x) \, dx$$

This follows from the theorem on linearity and mathematical induction. The corollary and the results for the base family can be combined to find an anti-derivative of *any* polynomial.

Theorem(Anti-derivative of a Polynomial). *An anti-derivative of the polynomial $p(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n$, ($n$ a non-negative integer) is*

$$\int p(x) \, dx = c_0 x + c_1 x^2/2 + c_2 x^3/3 + \cdots + c_n x^{n+1}/(n+1)$$

Example 4.1: Compute the integral: $\int_1^2 1+2x+x^2 \, dx$. We can use the last theorem to find an anti-derivative of $1+2x+x^2$; it gives $x+2x^2/2+x^3/3 = x+x^2+x^3/3$. Therefore, by the fundamental theorem of calculus: $\int_1^2 1+2x+x^2 \, dx = (x+x^2+x^3/3) \mid_{x=2} -(x+x^2+x^3/3) \mid_{x=1} = (2+4+8/3) - (1+1+1/3) = 6$.

Here is another result from stage 2 of our program.

Integration by Parts

This is a technique which is essentially a way of computing an integral by using the product rule from differential calculus. The product rule states that $(f\,g)' = f'\,g + f\,g'$. We rewrite this as: $f\,g' = (f\,g)' - f'\,g$. Integrating both sides over the interval $[a, b]$ gives

$$\int_a^b f(x)g'(x)\,dx = \int_a^b (f(x)\,g(x))'\,dx - \int_a^b f'(x)g(x)\,dx$$

The first integral on the right hand side is easy to compute by the fundamental theorem. Since this integral s easy to evaluate by the FTC, we have direct relationship between the other two integrals.

$$\int_a^b f(x)g'(x)\,dx = \big(f(b)g(b) - f(a)g(a)\big) - \int_a^b f'(x)g(x)\,dx$$

We call this result *Integration by Parts* which is the continuous analog of *Summation by Parts* introduced in chapter 2. It relates one integral to another. This is useful (just as it was for summing sequences) because if one of the integrals can be easily evaluated we get the value of the other one for free.

Example 4.2: Compute the integral $\int_0^{\pi/2} x\sin(x)\,dx$.

We can use the integration by parts formula here. Notice it would be convenient to treat sin as the derivative of a function and then pass the derivative onto x. Let $g(x) = -\cos(x)$ and $f(x) = x$. Then $g'(x) = \sin(x)$. Integration by parts gives us:

$$\begin{aligned}
\int_0^{\pi/2} \overbrace{x}^{f(x)}\,\overbrace{\sin(x)}^{g'(x)}\,dx &= \overbrace{\big(\pi/2 * -\cos(\pi/2)\ -\ 0 * -\cos(0)\big)}^{f(\pi/2)g(\pi/2)\ -\ f(0)g(0)} - \int_0^{\pi/2} \overbrace{x'}^{f'(x)}\ \overbrace{-\cos(x)}^{g(x)}\,dx \\
&= -\int_0^{\pi/2} -\cos(x)\,dx = \int_0^{\pi/2} \cos(x)\,dx = \sin(\pi/2)\ -\ \sin(0) \\
&= 1
\end{aligned}$$

Note that the values 0 and $\pi/2$ are in radians. The value $\pi/2$ radians is 90 degrees. This means that $\sin(\pi/2) = 1$ and $\cos(\pi/2) = 0$.

Substitution

This technique uses the chain rule (from differential calculus) to compute an integral. The chain rule states that $f(u(x))' = f'(u(x))u'(x)$. integrating both sides gives

$$\int_a^b f'(u(x))u'(x)\,dx = \int_a^b f(u(x))'\,dx = f(u(b)) - f(u(a))$$

Successful use of the above techniques is based on how well one can pattern match. It takes practice, but it is not our focus here. We demonstrate the idea through a few examples.

Example 4.3: Compute $\int_a^b (x-a)^n\,dx$ when n is a non-negative integer.

Letting $u(x) = (x-a)$, we see that the integral can be written as $\int_a^b u(x)^n u'(x)\,dx$. In this case letting $f'(z) = z^n$ (which has $z^{n+1}/(n+1)$ as an anti-derivative), the substitution formula gives $\int_a^b f'(u(x))u'(x)\,dx = f(u(b)) - f(u(a)) = \frac{1}{n+1}(b-a)^{n+1} - \frac{1}{n+1}(a-a)^{n+1} = \frac{1}{n+1}(b-a)^{n+1}$.

Example 4.4: Compute $\int_a^b (a-x)^n\,dx$ when n is a non-negative integer.

Letting $u(x) = (a-x)$, we see that the integral can be written as $-\int_a^b u(x)^n u'(x)\,dx$. Setting $f'(z) = z^n$ as in the previous example the substitution formula gives $-\int_a^b f'(u(x))u'(x)\,dx = -(f(u(b)) - f(u(a))) = -\big(\frac{1}{n+1}(a-b)^{n+1} - \frac{1}{n+1}(a-a)^{n+1}\big) = -\frac{1}{n+1}(a-b)^{n+1}$.

Example 4.5: Compute $\int_0^1 \sin^2(x)\cos(x)\,dx$.

Letting $u(x) = \sin(x)$, and therefore $u'(x) = \cos(x)$, we see that the integral can be written as $\int_0^1 u(x)^2 u'(x)\,dx$. In this case $f'(z) = z^2$. Using the substitution formula gives $\int_0^1 f'(u(x))u'(x)\,dx = f(u(1)) - f(u(0)) = u(1)^3/3 - u(0)^3/3 = \sin(x))^3/3\ |_{x=1} - \sin(x)^3/3\ |_{x=0} = \sin^3(1)/3 - 0 = \sin^3(1)/3$. Note, as always the angle measure is in radians. So, the 1 in our answer, $\sin^3(1)/3$, means 1 radian, which is approximately 57 degrees.

Can't always find Formulas for Anti-derivatives

As noted in the last chapter, not every function has a derivative. And even when it does, it may not be easy to find a formula for the derivative. However, if the function has a nice formula it is often the case that we can find a formula for the derivative. This is not the case with anti-derivatives. For nice continuous functions we always have an anti-derivative, but finding a formula is not always possible – even for relatively simple functions. That is, there are functions with nice formulas for which there is no formula for an anti-derivative. One example is the bell shaped functions associated with the subjects of probability and statistics. The graph of the function $f(x) = e^{-x^2}$ is bell shaped and its formula is simple. The area under the graph of this function exists and we can *create* a function which is the area under this curve from 0 to x. We can write the value down in our notation as: $A(x) = \int_0^x e^{-s^2}\,ds$. However, there is no formula for this function in terms of powers of x, logarithmic, exponential, or trig functions because there is no formula for an anti-derivative of e^{-x^2}. The integral (the area under the graph) exists; that is, the function $A(x)$ is a valid function – there is just no formula for it. See the appendix "Functions, Formulas, Graphs, and Notation" for other examples of functions without formulas.

Summary

The fundamental theorem of calculus can be stated succinctly: the integration operation is the inverse of the differentiation operation. This can be seen from the FTC: $\int_a^b f'(t)\,dt = f(b) - f(a)$, and writing it more crudely as $\int f' = f$. What makes this result useful is if for a given integral, $\int_a^b v(t)\,dt$, one can recognize v as the derivative of another function f, then this integral is $\int_a^b f'(t)\,dt$ and by the FTC this is just $f(b) - f(a)$.

The process of recognition is helped greatly if there is a well developed stock of differentiation formulas. Then when confronted with integrating a function, we may be able to recognize, or more easily deduce, a function which is an *anti-derivative* and then easily compute the integral. As with differencing and summing, finding formulas for derivatives and anti-derivatives of functions is much like multiplying and dividing numbers. Computing derivatives is relatively easier – like multiplying, while finding anti-derivatives is harder – like division. To do long division it helps to be fluent with multiplication, because trial and error multiplications are necessary. The same is true when computing anti-derivatives, it helps first to be fluent with the formulas for the derivatives of a number of functions. To this end, one would

1. Develop formulas for the derivatives of a base collection of functions.
2. Develop formulas for the derivatives of functions formed in various ways from the base.

This is exactly what we did in the last chapter.

Sometimes, in fact many times, it may not be clear or even possible to evaluate an integral. It is often the case that one can at least approximate the integral by bounding the value from above and below. For instance, if f is a continuous function† and non-negative on an interval $[a, b]$, then f attains a maximum, say M, and a minimum, say m on this interval. Therefore, the area under the graph of f (the integral $\int_a^b f(x)\, dx$) is bounded above by the area of the rectangle with height M and width $b - a$ and bounded below by the area of the rectangle with height m and width $b - a$. That is,

$$m\,(b-a) \leq \int_a^b f(x)\, dx \leq M\,(b-a)$$

This result is true even if f takes on negative values. This is true since the integral is a limit of sums of samples of f along the interval, $[a, b]$, times small differences – which sum to $b-a$. But each of these sample values of f is bounded above and below by M and m respectively. Therefore, the sum is bounded above and below by $M\,(b-a)$ and $m\,(b-a)$. We can show something just a little more interesting with almost no extra work.

Mean Value Theorem

Again, consider the area under the graph of f from a to b $(a < b)$‡ : $\int_a^b f(s)\, ds$. You can think of the graph as the outline of fine dust in a very narrow rectangular glass container. If we shift the dust around so that the dust is level with height, say h, we must have (since the area of the dust is conserved)

$$\int_a^b f(s)\, ds = h\,(b-a)$$

If m and M are the minimum and maximum values that f takes on in the interval $[a, b]$, then $m \leq h \leq M$. It turns out that continuous functions have the property that they attain all intermediate values. That is, if f is a continuous function which hits the values m and M over the interval $[a, b]$, then for any $m \leq h \leq M$ there exists a value $\mu \in [a, b]$ such that $f(\mu) = h$. Collecting these ideas we have:

Theorem(Intermediate Value Theorem). *If f is a continuous function on the interval $[a, b]$ attaining values m and M with $m \leq M$, then for any intermediate value h with $m \leq h \leq M$ there exists a value $\mu \in [a, b]$ such that $h = f(\mu)$.*

Theorem(Mean Value Theorem). *If f is a continuous function on the interval $[a, b]$, then there exists a value, $\mu \in [a, b]$, such that*

$$\int_a^b f(s)\, ds = f(\mu)(b-a)$$

† Continuous functions are discussed in the section "Computational Strategy for Limits" in the appendix "Infinite Sequences and limits". They have the property that when drawing their graph, one never need take the pen off the page.

‡ The following argument works for non-negative f, but can be easily extended to all integrable f.

If f' is also continuous then the theorem applies to f':

$$\int_a^b f'(s)\,ds = f'(\mu)(b-a)$$

By the Fundamental Theorem of Calculus this can be written

$$f(b) - f(a) = f'(\mu)(b-a)$$

which is

$$f(b) = f(a) + f'(\mu)(b-a) \tag{1}$$

This can be written in a way that has an immediate geometric interpretation:

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

This says that when f' is continuous there is a point $\mu \in [a,b]$ where the slope of the tangent line is the same as the slope of the line passing through the points $(a, f(a))$ and $(b, f(b))$.

Here is another way to view this. Consider the interval $[x-h, x+h]$ for some $h > 0$. Substituting $x+h$ for b and x for a in (1) gives $f(x+h) = f(x) + f'(\mu_1)h$ – for some $\mu_1 \in [x, x+h]$. Next substituting x for b and $x-h$ for a in (1) gives $f(x) = f(x-h) + f(\mu_2)h$, or $f(x-h) = f(x) - f(\mu_2)h$ – for some $\mu_2 \in [x-h, x]$. That is, we may write $f(x+h)$ as $f(x+h) = f(x) + f(\mu)h$ for h independent of the sign of h and $\mu \in [x - |h|, x + |h|]$. If h is small, then $f(\mu)$ is close to $f(x)$; therefore, an approximation to $f(x+h)$ is

$$f(x+h) \approx f(x) + f'(x)h \tag{1a}$$

Approximation Application

Suppose that you needed to build a concrete wall 100 ft long and 5 ft high. Builders tell you to set the depth of the wall based on the height of the wall by the following formula: depth (in inches) = height (in inches) / 10. So, for a given height x in inches the volume of concrete needed is $V(x) = 100ft \times x\,in \times (x/10)in$. Using the units of inches, the formula for V (in cubic inches) in terms of height, x, in inches is $V(x) = 12 \times 100x^2/10 = 120x^2$. To pour a 5ft wall we need (x = 60 inches) $V(60) = 864000$ cubic inches of concrete, which is 18 1/2 yards. But there is always error in measurement. What if, due to the ground conditions, we are off by 1 inch when measuring the height? We need to order extra concrete to be sure we have enough. How much more should we order to take care of this 1 inch error? We don't want to order too much extra because we want to control our costs. We could simply compute $V(60+1)$ easily enough – this gives an upper bound on the amount of concrete to use. However, we would like to estimate the error for *any* small error in height.‡

Equation (1a) tells us that if there is a small error in the height then the approximate change in the volume is given by $V(x+h) \approx V(x) + V'(x)h$. Let's say that we are concerned with an error of 1in when the height is 5 ft (60 inches). The magnitude of the change is $\mid V'(x)h \mid$. It is easy to see that $V'(x) = 240x$. When $x = 60$ we have $V'(60) = 14400$; if, in addition, $h = 1$, $\mid V'(x)h \mid$ becomes 14400 cubic inches. This is 0.309 cubic yards. Therefore, we should order an extra 1/3 of a yard of concrete.

‡ In our case, $V(60+h) - V(60)$ is the extra amount of concrete to use. However, in many similar problems the approximation formula (1a) is often simpler and provides more insight into how sensitive the original function is to *all* small errors at a given input.

Taylor's Theorem

We now describe a result that is more difficult but which is quite impressive†. Let us first describe a notation for differentiation. We have used f' to denote the derivative of f. We will use the notation f'' to denote the derivative of the function f'. Similarly, we use the notation f''' to denote the derivative of f''. In general, we use the notation $f^{(n)}(x)$ to denote the n^{th} derivative of f evaluated at x.

Now, if f' is a continuous function we know by the mean value theorem that

$$f(b) = f(a) + f'(\mu)(b - a)$$

But we could invoke the theorem for any $x \in [a, b]$ because $[a, x]$ is an interval where the theorem applies. In this case the conclusion of the theorem is

$$f(x) = f(a) + f'(\mu_1)(x - a) \tag{2}$$

Note that we changed μ to μ_1 because the interval $[a, x]$ is different than the interval $[a, b]$. All that the Mean Value Theorem says is that there is *some* value in $[a, x]$; it may not be the same one that applied on the interval $[a, b]$. However, for the time being, let us assume that it is the same; that is, we assume that $\mu_1 = \mu$.

If f'' is also continuous on $[a, b]$, then we may apply this formula to f' (its a function which satisfies the requirements of the theorem) giving

$$f'(x) = f'(a) + f''(\mu_2)(x - a)$$

for some value μ_2 which depends on the interval $[a, x]$. Why is this different that μ_1? Because f' is a different function from f so we should not expect μ_2 to be the same as μ_1. Let us assume, however, that $\mu_2 = \mu_1$, which we are assuming is μ.

Integrating this function over the interval $[a, x]$ we have

$$\int_a^x f'(s)\,ds = \int_a^x f'(a)\,ds + \int_a^x f''(\mu)(s - a)\,ds$$

Since the integration is linear, we may place the constants outside of the integral‡.

$$\int_a^x f'(s)\,ds = f'(a)\int_a^x 1\,ds + f''(\mu)\int_a^x (s - a)\,ds$$

Using the FTC on the left hand and first term of the right hand side gives

$$f(x) - f(a) = f'(a)(x - a) + f''(\mu)\int_a^x (s - a)\,ds$$

As an anti-derivative of $(s - a)$ is $\frac{(s-a)^2}{2}$, we have

$$f(x) = f(a) + f'(a)(x - a) + f''(\mu)\int_a^x \left(\frac{(s - a)^2}{2}\right)' ds$$

† This section stands by itself and may be skipped if you wish.

‡ Anything that does not depend on the integration variable, s, is a constant with respect to the integration. Therefore, it may be pulled outside of the integral.

Using the FTC on the last term, we may write

$$f(x) = f(a) + f'(a)(x-a) + f''(\mu)\frac{(x-a)^2}{2} \tag{3}$$

Now, just as we plugged f' into equation (2), we can do the same with equation (3). If the function f''' is continuous, then plugging f' into (3) gives

$$f'(x) = f'(a) + f''(a)(x-a) + f'''(\mu_4)\frac{(x-a)^2}{2}$$

Again, we need a new "μ" because we are using a new function, f'. As before, we integrate this equation over the interval $[a,x]$ giving

$$f(x) - f(a) = f'(a)(x-a) + f''(a)\frac{(x-a)^2}{2} + \int_a^x f'''(\mu_4)\frac{(s-a)^2}{2}\,ds$$

As an anti-derivative of $(s-a)^2/2$ is $\frac{(s-a)^3}{3*2}$, the above becomes

$$f(x) - f(a) = f'(a)(x-a) + f''(a)\frac{(x-a)^2}{2} + f^{(3)}(\mu_4)\int_a^x \frac{(s-a)^2}{2}\,ds$$

Which is

$$f(x) = f(a) + f'(a)(x-a) + f''(a)\frac{(x-a)^2}{2} + f^{(3)}(\mu_4)\frac{(x-a)^3}{3*2} \tag{4}$$

We can keep plugging f' into these successively more complicated equations (equations (2), (3), and (4)) provided f has enough derivatives and they are continuous. The pattern that emerges is called Taylor's Theorem:

Theorem(Taylor's Theorem). *If f is a function such that $f^{(n)}$ exists and is a continuous function on the interval $[a,b]$, then for any $x \in [a,b]$*†

$$\begin{aligned} f(x) = f(a) + f'(a)(x-a) + f^{(2)}(a)\frac{(x-a)^2}{2!} + f^{(3)}(a)\frac{(x-a)^3}{3!} \\ + \ldots + f^{(n-1)}(a)\frac{(x-a)^{n-1}}{(n-1)!} + f^{(n)}(\mu(x))\frac{(x-a)^n}{n!} \end{aligned}$$

where μ is a function such that $\mu(x) \in [a,x]$.

Our derivations for Taylor's Theorem in the cases $n = 2$ and $n = 3$ are not correct since we assumed at each stage that the various "μ"'s were the same. A mathematically correct derivation can be found in the appendix on Taylor's Theorem. The derivation has the same structure as our derivation above but uses a generalization of the Mean Value Theorem to prove the result. One can also show without much work that the formula is true when x is less than a:

† *The notation $n!$ means the product of all the integers from 1 to n.*

Corollary. *If f is a function such that $f^{(n)}$ exists and is a continuous function on the interval $[a,b]$, with $c \in [a,b]$, then for any $x \in [a,b]$*

$$f(x) = f(c) + f'(c)(x-c) + f^{(2)}(c)\frac{(x-c)^2}{2!} + f^{(3)}(c)\frac{(x-c)^3}{3!} + \ldots + f^{(n-1)}(c)\frac{(x-c)^{n-1}}{(n-1)!} + f^{(n)}(\mu(x))\frac{(x-c)^n}{n!}$$

where μ is a function such that $\mu(x)$ is a value between x and c.

Taylor's Theorem says that we may approximate a function f with n continuous derivatives in a region containing a point near c with a *polynomial* of degree $n-1$. The error in the approximation is the term $f^{(n)}(\mu(x))\frac{(x-c)^n}{n!}$.

Power Series Representation

In summation notation Taylor's formula may be written:

$$f(x) = f(a) + \sum_{k=1}^{n-1} f^{(k)}(a)\frac{(x-a)^k}{k!} + f^{(n)}(\mu(x))\frac{(x-a)^n}{n!}$$

If a function has an infinite number of derivatives and the last term in the above formula goes to zero in a sufficiently "well behaved" manner (which is beyond the scope of this book to describe) as n goes to ∞, then we may write

$$f(x) = f(a) + \sum_{n=1}^{\infty} f^{(n)}(a)\frac{(x-a)^n}{n!}$$

One function that has an infinite number of derivatives is the exponential function, $f(x) = e^x$. We know that $f'(x) = e^x$, and consequently $f^{(n)}(x) = e^x$. It is also a function that is sufficiently "well behaved". Letting $a = 0$ we may write‡

$$e^x = e^0 + \sum_{n=1}^{\infty} e^0 \frac{x^n}{n!} = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

This says that e^x can be thought of as an "infinite polynomial". Many other such functions can be represented in this way, for instance the trig functions, *sin* and *cos*. Such representations can be used to find (among other things) an approximation to a function, its derivative, or its integral.

Finally, notice that the value e itself can be approximated for a given N by

$$e = e^1 = \sum_{n=0}^{\infty} \frac{1^n}{n!} = \sum_{n=0}^{\infty} \frac{1}{n!} \approx 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \ldots + \frac{1}{N!}$$

‡ The notation $0!$ is defined to be 1 as is x^0 for all $x \neq 0$.

Chapter 6: Introduction to Difference Equations

Motivational Problem

If a person buys a house they will almost surely need a loan from a bank to pay for it. Once you borrow the money, the bank will have you spread the payments out over 15 to 30 years. You're not simply paying back the money to the bank, they charge interest on the use of their money. So, each payment can be thought of as consisting of two pieces: a payment for the principal borrowed and a second piece which is interest on the outstanding debt. We describe what happens at any given month. At month n let x_n be the outstanding debt we owe to the bank at month n. The bank has decided that at then end of each month we pay p dollars for the loan. Also, during the month, we are charged interest (for that month) on the outstanding debt as of the beginning of that month. The rate of interest is given as an annual interest rate (AIR) and so must be divided by 12 to get the monthly interest rate, $r = \text{AIR}/12$. At the beginning (month zero) the outstanding debt is known, it is just the loan amount; we'll call it L.

We would like to know what the outstanding debt is at any given month. Our strategy is similar to what one would do in algebra class: rather than writing down $x = \textit{answer}$, one first found a relation that x satisfied and then solved for x. We will do the same thing here. Rather than write down a formula for the outstanding debt at month n, x_n, we first find a relation that x_n satisfies and then solve for x_n. Such a relation can often be stated as a change or a *difference* from one time step to the next. This change is often easy to describe. For the problem at hand, the change in debt from one month to the next is the amount of new debt accumulated by interest less the monthly payment. Mathematically, this can be written as:

$$\begin{aligned}
\overbrace{x_{n+1}}^{\text{Debt at month n+1}} &= \overbrace{x_n}^{\text{Debt at month n}} + \overbrace{rx_n}^{\text{Interest on Debt over the month}} - \overbrace{p}^{\text{monthly Payment}} \\
x_0 &= L
\end{aligned}$$

This can be rewritten as:

$$\begin{aligned}
x_{n+1} &= (1+r)x_n - p \\
x_0 &= L
\end{aligned}$$

Recall that in algebra an unknown numeric value was denoted x and the game was to solve for x. Here, x is a *sequence*; so, what does it mean to solve for x? Ideally, we would like a *formula* for the n^{th} element of the sequence x in terms of this n. Then, when someone asked for the debt at month 17, we could just plug 17 into this formula and out would pop the debt.

Basic Solution Strategy

The equation we wrote down for the debt as a function of the month was an example of a recurrence relation: each term in the sequence is a function of previous terms. This is also know as a difference equation; the difference and summation calculus of the first two chapters will play an important role in solving such equations. In the equation above, a given element of the sequence is a function of only the previous element. Such an equation is called a first order difference equation. The form of this relation between previous elements determines how easy it is to solve, depending on what the meaning of solve is. Before we start, let us be clear: The solution to any difference equation is not a number; that is, the unknown we are looking to find is not a number but an *entire sequence* – which we have said is a *function* on the integers.†

Let's start with a simpler difference equation. In fact, let's start with the simplest first order difference equation relating a given value of a sequence with its previous value:

$$\begin{aligned} x_{n+1} &= x_n \\ x_0 &= L \end{aligned}$$

We can solve this difference equation by using the equation to write down the solution term by term. That is, we know $x_0 = L$. The first equation tells us that $x_1 = x_0$; and since we know x_0, we know x_1. Continuing, we have $x_2 = x_1$; and since we know x_1, we know x_2, it is just $x_1 = x_0 = L$. We can continue and produce each value of the sequence. It is tedious, but effective. Every difference equation can be solved by doing this sort of computation. In theory, any element of the sequence could be produced this way. In practice, one would compute as many elements of the sequence as one needed.

A better solution though, would be to write down a *formula* for any given element of the sequence. In our present example, we find that the formula for the n^{th} element of the sequence is $x_n = L$. Having a formula means one can, on request, produce any element of the sequence without having to compute *any* of the previous elements. The question is

"When presented with a difference equation, can we write down a formula for the n^{th} term of the sequence satisfying such an equation?" The answer is no in general. We will, however, be able to find a solution (in this sense) to our motivational problem.

The problem with finding a formula, in general, is that the n^{th} value of the sequence which satisfies such an equation is tangled up with the previous values. Because of this, it is not a simple matter to write down a formula for the n^{th} value. We introduce a technique which will allow us to solve a larger array of equations. Consider the previous simple difference equation; we know by the Fundamental Theorem of Difference Calculus that it is easy to compute the sum of telescoping differences. Therefore, rewrite the last equation as:

$$\begin{aligned} x_{n+1} - x_n &= 0 \\ x_0 &= L \end{aligned}$$

This is

$$\begin{aligned} \Delta x_n &= 0 \\ x_0 &= L \end{aligned}$$

Summing both sides of the first equation from 0 to $N-1$ we have:

$$\sum_{n=0}^{N-1} \Delta x_n = \sum_{n=0}^{N-1} 0$$

After summing (using the FTDC):

$$X_N - x_0 = 0$$

† Described in the sub-section "Sequences as Functions" in the first chapter.

Or, $x_N = L$. We were able to untangle the connections of the sequence by summing them out. This left an element of our choosing, N, and the 0^{th} term plus whatever the sum was on the right hand side of the equation.

If we modify the difference equation so that it becomes:

$$x_{n+1} = x_n + b_n$$
$$x_0 = L$$

This is

$$\Delta x_n = b_n$$
$$x_0 = L$$

Summing both sides of the first equation from 0 to $N-1$ gives:

$$\sum_{n=0}^{N-1} \Delta x_n = \sum_{n=0}^{N-1} b_n$$

After summing (using the FTDC):

$$X_N - x_0 = \sum_{n=0}^{N-1} b_n$$

Or, $x_N = L + \sum_{n=0}^{N-1} b_n$. If $b_n = b$, that is, b_n is constant, then $x_N = L + bN$.

Solution of Motivational Problem

We are moving toward the solution of our motivational problem. To get a step closer, consider the following difference equation:

$$x_{n+1} = a\, x_n$$
$$x_0 = L$$

A solution to this difference equation is a sequence which has L for its initial value and then multiplies the current value by a to get the next. It is not too hard to guess that the solution† is $x_n = a^n L$. Let us check this is indeed a solution‡. We need to check two things. First, that $x_0 = L$. This is so, since $x_0 = a^0 L = 1 * L = L$. We also need to check $x_{n+1} = a\, x_n$. But $x_{n+1} = a^{n+1} L = a(a^n L) = a\, x_n$. So this is a solution. Remember, a solution to a difference equation is a *sequence*, not a number.

Now let's look at an equation that has the same form as our original problem:

$$x_{n+1} = a\, x_n + b_n$$
$$x_0 = L$$

It is not as easy to guess a solution to this. Can we also use the differencing technique we used earlier? Well, if we rewrite the equation so that the left hand side is a difference as before, then:

$$x_{n+1} - x_n = (a-1)x_n + b_n$$
$$x_0 = L$$

† The pattern emerges quickly. $x_1 = a\, x_0 = a\, L$; $x_2 = a\, x_1 = a\, a\, L = a^2\, L$; $x_3 = a\, x_2 = a\, a^2 L = a^3\, L$..

‡ In fact, it is *the* solution.

This does not help us untangle the sequence of unknowns, x_n. If we try to sum this equation, the left hand side reduces to x_N and x_0. However, the right hand side contains the unknowns for which we are trying to solve. For this technique to work we need to get all of the x_n's on the left as part of *some* telescoping difference (not necessarily a difference in x_n alone), leaving the right hand side containing the rest of the *known* quantities. Summing the left hand side will reduce to two terms; from there we can solve for x_N. To start, let's rewrite the equation as a difference of the unknowns (not yet a telescoping difference):

$$\begin{aligned} x_{n+1} - a\,x_n &= b_n \\ x_0 &= L \end{aligned}$$

The key idea is to multiply both sides of the first equation by a known sequence so that the left hand side is a quantity which can be represented as a *telescoping* difference. Given a sequence $\{y_n\}_{n=0}^{\infty}$, multiply the first equation by y_{n+1}. This gives:

$$\begin{aligned} y_{n+1}x_{n+1} - (a\,y_{n+1})x_n &= b_n y_{n+1} \\ x_0 &= L \end{aligned}$$

If we chose the sequence y so that $a\,y_{n+1} = y_n$, then we would have a telescoping difference on the left hand side

$$\begin{aligned} y_{n+1}x_{n+1} - y_n x_n &= b_n y_{n+1} \\ x_0 &= L \end{aligned}$$

Summing from 0 to $N-1$ gives us $y_N x_N - y_0 x_0 = \sum_{n=0}^{N-1} b_n y_{n+1}$. Solving for x_N gives

$$x_N = \left(\frac{y_0 L + \sum_{n=0}^{N-1} b_n y_{n+1}}{y_N} \right)$$

So, we're done provided we know how to find the sequence y. But the sequence y satisfies its own difference equation: $y_{n+1} = a^{-1} y_n$. From what we discovered above, the solution is $y_n = a^{-n} y_0$. There are no further restrictions so we may choose y_0 to be anything; let's set it to 1. Therefore, we may write down a solution for the N^{th} element of the sequence x to be

$$x_N = a^N L + a^N \sum_{n=0}^{N-1} a^{-(n+1)} b_n$$

When b_n has the same value, say b, for all n this formula simplifies to

$$x_N = a^N L + b\,a^{N-1} \sum_{n=0}^{N-1} a^{-n}$$

Which simplifies to

$$x_N = a^N L + b\,a^{N-1} \left(\frac{a^{-N} - 1}{a^{-1} - 1} \right)$$

provided $a \neq 1$ (we used the following fact about geometric series: $\sum_{n=0}^{N-1} z^n = \frac{z^N - 1}{z - 1}$ $z \neq 1$; used above with $z = a^{-1}$). This can also be written as:

$$x_N = a^N L + b \left(\frac{1 - a^N}{1 - a} \right)$$

Applying this formula to the motivational problem with $a = (1 + r)$, and $b = -p$ yields a formula for the debt owed at month N:‡

$$x_N = (1+r)^N L - p\left(\frac{1-(1+r)^N}{-r}\right)$$

which is

$$x_N = (1+r)^N L - p\left(\frac{(1+r)^N - 1}{r}\right)$$

Example 6.1: If a 30 year loan is taken out on \$100,000 at an annual interest rate of 6% with monthly payment \$600, what is the outstanding debt after 5 years?

The monthly rate is $r = 0.06/12 = 0.005$, $L = 100,000$, $p = 600$, and $N = 5 * 12 = 60$. Plugging these values into the solution to our difference equation gives

$$x_{60} = (1.005)^{60} * 100000 - 600 * \left(\frac{1.005^{60} - 1}{0.005}\right)$$

After evaluating the expression on the right hand side we obtain $x_{60} = 93,023$. This is the outstanding debt after 5 years of monthly payments. During this time you will have paid $5 * 12 * 600 = 36,000$ dollars to the bank.

Summary

We are able to solve a simple difference equation by first rewriting the tangling of x's as a telescoping difference and then use the Fundamental Theorem of Difference Calculus to "sum away" the tangles. This leaves one element of the sequence, x_N, on one side of the equation and the sum of the known values on the other side of the equation.

What's Left

A difference equation is effectively any equation which relates a given element of a sequence with previous values of the sequence. Obviously, they can be much more complicated than the ones we've examined. A careful study of the subject classifies difference equations into various categories and then provides recipes for solutions in some cases. The main solution technique is to "sum away" the differences; that is, use the Fundamental Theorem of Difference Calculus.

‡ Note $a \neq 1$ so we may use the just derived formula.

Chapter 7: Introduction to Differential Equations

Motivational Problem

The population of a species of bacteria, represented by $p(t)$, is the number of a bacteria on a given surface. The evolution of the population is modeled by the assumption that its "instantaneous" growth rate is proportional to the amount of bacteria currently present. The growth is impeded by a chemical on the surface that kills bacteria at a rate of b bacteria per second. These facts may be collected into an equation:

$$\begin{aligned} p'(t) &= r\,p(t) - b \\ p(0) &= P \end{aligned}$$

Here, r is the proportionality coefficient and has units 1/second. With this choice the units of the two sides of the equation match. This is the case since the left hand side is the limit, $\lim_{n\to\infty} \frac{p(t+h_n)-p(t)}{h_n}$, (where $h_n \to 0$) which has units 1/second.

This equation is called a differential equation which you can think of as an infinitesimal difference equation. How do we solve this and what does a solution mean? The solution is a *function* p which satisfies the above equations. We will find an explicit formula for the solution.

Basic Solution Strategy

Let's start with a simpler differential equation. In fact, let's start with the simplest differential equation that we can think of.

$$\begin{aligned} x'(t) &= 0 \qquad (1) \\ x(0) &= P \end{aligned}$$

To solve this equation we must find a *function* such that its derivative is 0. That is, the tangent at each point is flat. We know x must be a constant function. Since the initial value is P, the function must be $x(t) = P$. Another way to solve this equation is to use the Fundamental Theorem of Calculus. We do this by integrating out the derivative. Integrating (1) from 0 to t gives

$$\begin{aligned} \int_0^t x'(s)\,ds &= \int_0^t 0\,ds \\ x(t) - x(0) &= 0 \end{aligned}$$

Therefore, $x(t) = x(0)$, and since $x(0) = P$, we have $x(t) = P$.

As with difference equations, there is no technique for writing a formula for the solution of a general differential equation. Again, as with difference equations, there are limited classes of equations for which we can write down a solution. However, unlike difference equations there is a question of existence and uniqueness of a solution. It turns out there are differential equations which don't have solutions, and others for which the solution is not unique. When we say that a solution doesn't exist, we don't just mean that we can't find a formula to express the solution; we mean that there are some equations for which there is no function which satisfies the differential equation. And, there are differential equations which have more than one solution.†

The way to find formulas for the solutions to these problems is to use the Fundamental Theorem of Calculus to integrate away the derivative. This is analogous to the way we solved difference equations: we used a summation operator to get rid of the telescoping differences. That is, we used the Fundamental Theorem of Difference Calculus to find solutions.

Solution of Motivational Problem

Let's look at an equation that is closer to our motivational problem. Consider the simpler problem:

$$\begin{aligned} p'(t) &= r\,p(t) \\ p(0) &= P \end{aligned} \tag{2}$$

If we integrate both sides we can use the Fundamental Theorem of Calculus to integrate away the derivative on the left hand side of the equation. However, we don't know what the integral of the right hand side is since it involves the unknown function, p. We need a way to combine all the p's together and view them as the derivative of some function of p. Then, the Fundamental Theorem of Calculus can be used to eliminate the derivative. We then will have an equation which does not have derivatives, and therefore, we can use algebra to find p. To do this divide both sides by p; we now have all the p's (the unknowns) on one side of the equation.

$$\begin{aligned} \frac{p'(t)}{p(t)} &= r \\ p(0) &= P \end{aligned}$$

The left hand side is the derivative of $\ln(p)$.‡ Integrating both sides from 0 to t we find:

$$\ln(p(t)) - \ln(p(0)) = r(t-0)$$

Or,

$$\ln(p(t)) = \ln(p(0)) + r\,t$$

The inverse function of $\ln(x)$ is e^x; therefore, exponentiating both sides of the equation gives us

$$p(t) = e^{\ln(p(0))+r\,t} = e^{\ln(p(0))}e^{r\,t} = p(0)\,e^{r\,t} = P\,e^{r\,t}$$

To solve our motivational problem, we try to do the same thing; that is, arrange the p's on one side in such a way that they are the derivative of some expression involving p. Then, use the Fundamental Theorem of Calculus to integrate away the derivative. We then have an ordinary equation for which we solve for p.

To do this we employ a technique analogous to what was done for the motivational problem in the chapter on difference equations. First move the p on the right hand side to the left hand side – grouping all the p's together. Next, multiply both sides by some function, y at t. Our motivational problem becomes:

$$y(t)p'(t) - ry(t)\,p(t) = -b\,y(t) \tag{3}$$

† A discussion of these issues is beyond the scope of this book.

‡ This follows from the chain rule.

The goal is to choose y so that the left hand side is the derivative of the product, $(y(t)p(t))$. The derivative of this product is

$$(y(t)p(t))' = (p(t)y(t))' = p'(t)y(t) + p(t)y'(t) = y(t)p'(t) + y'(t)p(t)$$

To match the left hand side of (3) we need y to satisfy $y'(t) = -r\,y(t)$. But this is an equation we know how to solve, the solution is $y(t) = C\,e^{-r\,t}$, for some constant C. Since there are no further restrictions on y, we take C to be 1. So, equation (3) becomes

$$\left(e^{-r\,t}p(t)\right)' = -b\,e^{-r\,t}$$

Integrating both sides of this equation from 0 to t gives:‡

$$\int_0^t \left(e^{-r\,s}p(s)\right)' ds = \int_0^t -b\,e^{-r\,s}\,ds$$

It is not clear what the anti-derivative of $e^{-r\,s}$ is however. We do know that the derivative of $e^{a\,x}$ is $a\,e^{a\,x}$ from the subsection of chapter 2 dealing with derivatives of logarithms – and their inverses. From this we can guess that the anti-derivative of $e^{a\,x}$ is $\frac{e^{a\,x}}{a}$; which can be verified by taking its derivative.

Now, the integral above can be evaluated to give

$$e^{-r\,t}p(t) - e^{-r\,0}p(0) = \frac{b}{r}e^{-r\,t} - \frac{b}{r}$$

Solving for p we obtain (since $e^0 = 1$ and $p(0) = P$)

$$p(t) = P\,e^{r\,t} + \frac{b}{r}(1 - e^{r\,t})$$

This may be written:

$$p(t) = \frac{b}{r} + (P - \frac{b}{r})e^{r\,t}$$

Notice that if $P - \frac{b}{r} < 0$ then $p(t)$ will eventually become zero. Let us assume that this is the case and find the time to extinction. That is, find t^* such that $p(t^*) = 0$. The equation to solve for t^* (put t^* on one side and everything else on the other) is

$$p(t^*) = 0 = \frac{b}{r} + (P - \frac{b}{r})e^{r\,t^*}$$

After some algebra this becomes

$$e^{r\,t^*} = -\frac{\frac{b}{r}}{P - \frac{b}{r}} = \frac{\frac{b}{r}}{\frac{b}{r} - P}$$

Taking the $\log_e$(which is named ln) of both sides and solving for t^* gives

$$t^* = \frac{1}{r}\ln\left(\frac{\frac{b}{r}}{\frac{b}{r} - P}\right) = \frac{1}{r}\ln\left(\frac{b}{b - r\,P}\right)$$

‡ We're using integration to collapse the infinitesimal telescoping differences; that is, we use the Fundamental Theorem of Calculus to get a simple difference. Notice we used a different integration variable, s, to avoid confusion with t.

So, in the case that $P < \frac{b}{r}$ (which is the same as $r\,P < b$) the population of the bacteria die out at this t^*. That makes sense since initially, the rate of change of the population is $p'(0) = r\,p(0) - b = r\,P - b < 0$. That is, at the start the rate of growth is decreasing which implies $p(t)$ gets smaller which means that $p'(t) = r\,p(t) - b < r\,P - b < 0$; that is, the growth rate decreases further. So, once the population growth is negative it becomes a death spiral for the bacteria.

It is interesting to note that the formula for $p(t)$ becomes negative for $t > t^*$. Obviously, this doesn't make physical sense – the model no longer holds after the critical time t^*. Notice also that our model predicts that there are a non integral number of bacteria at almost any given time. The model could be corrected by removing the fractional part of the predicted population size by either truncating or rounding the result. One could think of this extra step as part of an improved model.

Keep in mind that models are always an approximations to the truth. For instance, although the model that we have may work over short intervals of time, there is no population that satisfies this equation for too long because as the population rises other factors in the environment inhibit growth.

Radioactive decay

We apply this technology to a problem in physics.

A radioactive substance decays at a rate proportional to the amount present. If one starts with Q pounds of the substance, in a 1,000 years there will be $Q/4$ pounds. When will there be $\frac{1}{2}Q$ pounds?

The equation is similar to the previous problem:

$$\begin{aligned} a'(t) &= -r\,a(t) \\ a(0) &= Q \end{aligned}$$

Here, $a(t)$ is the amount in pounds at time t in years; and, r is taken to be a positive constant, whose units must be 1 / years for the equation to make sense. We don't know what r is but we know that in 1,000 years there will be a quarter of Q left. Solving the equation gives $a(t) = Q\,e^{-rt}$. When t is 1,000 the equation reads $a(1000) = Q\,e^{-1000r} = \frac{1}{4}\,Q$. We may solve this equation for r by taking the $\log_e$ (which is the ln function) of the last equation. This gives us $\ln(Q) - 1000\,r = \ln(\frac{1}{4}Q) = \ln(1) - \ln(4) + \ln(Q) = \ln(Q) - \ln(4)$ (since $\ln(1) = 0$). Or, $-1000\,r = -\ln(4)$. So, $r = \frac{\ln(4)}{1000}$. To find the time t when $a(t) = \frac{1}{2}$, we need to solve: $a(t) = Q\,e^{-rt} = \frac{1}{2}\,Q$ for t (the only unknown in the equation is t as we now know r). Taking the ln of both sides gives us $\ln(Q) - r\,t = \ln(\frac{1}{2}Q) = \ln(1) - \ln(2) + \ln(Q) = \ln(Q) - \ln(2)$. So $t = \ln(2)/r = 1000\ln(2)/\ln(4) = 500$.

That is, the time it takes the initial amount to decay to $\frac{1}{2}Q$ pounds is 500 years. Notice that the answer is independent of the initial amount Q. In other words, the "half live" of this radioactive material is an intrinsic attribute of this substance.

What's Left

A differential equation is effectively any equation which involves a derivative. Obviously, they can be much more complicated than the ones we looked at. A careful study of the subject classifies differential equations into various categories and then provides recipes for solutions in some cases. The main solution technique is to "integrate away" the derivatives; that is, use the Fundamental Theorem of Calculus.

Appendix A: Functions, Formulas, Graphs, and Notation

In this appendix we discuss a few odds and ends about functions. As mentioned earlier in this book functions are rather abstract things. Functions are not numbers or a collection of numbers; they are rules which assign one number to another. As such, we are tempted to identify anything tangibly associated with a function *as the function*. The two things that people often confuse with a function are formulas – usually used to define a function – and the graph of a function. There is a third problem which comes from the way mathematicians and others notate functions.

Before we clarify these concepts let us emphasize that although functions are not formulas, the kinds of functions which make calculus computationally tractable are the ones which are defined by formulas or are made from a patch work of formulas.

Functions are NOT Formulas

You will often hear the terms *function* and *formula* used interchangeably. One may hear "This is a function that describes such and such." Or, "Here is a formula for computing such and such." Many people think they are the same thing. In fact, mathematicians a few centuries back *did* use the two terms interchangeably. They were driven to separate these notions through the problems they faced over the last few centuries – just as we are being driven to introduce the concepts of Calculus by a particular set of problems. A few definitions will help draw this distinction.

Definition(Function). *A function, f, takes an input from a number in a set called the domain of f and produces one number as output. The set of outputs of f (using input values from the domain) is called the range of f.*

A simple example is the function which squares it's input. If we call this function f, then f is defined by $f(x) = x^2$. We can take the domain of this function to be any set where the squaring operation makes sense. One usually takes the domain to be the largest possible set one can. In this case, the domain is usually taken to be all real numbers. As a second example, consider the function which takes the square root of it's input. If we call this function g, then g is defined by $g(x) = \sqrt{x}$. We can take the domain of this function to be any set where the square-root operation makes sense. Since taking square roots of a number is only valid for non-negative numbers, we can take the domain of this function to be any subset of the non-negative numbers. One usually takes the domain to be the non-negative numbers. Lastly, consider the family of functions, $f(x) = x^n$ with integer n, $n < 0$. The domain of these functions is taken to be the set of all numbers except 0; that is, all places where the function makes sense.

A formula doesn't have much of a formal definition, but we can define it roughly as:

Definition(Formula). *A formula is an "algebraic" implementation of a function. That is, a formula is an "algebraic rule" which takes a numeric input and produces a numeric output.*

Examples of formulas are:

1. $1 + 2x$
2. $x^2 + 3x + 2$
3. $\sin^2(x) + x + \tan(x)$
4. $\frac{1+x^2}{2x-7}e^x + 2\tan(x)$

These formulas can be used to define functions. Listed below are the associated functions from the formulas above.

1. $f_1(x) = 1 + 2x$
2. $f_2(x) = x^2 + 3x + 2$
3. $f_3(x) = \sin^2(x) + x + \tan(x)$
4. $f_4(x) = \frac{1+x^2}{2x-7}e^x + 2\tan(x)$

It may seem like we're quibbling over trivial matters. To drive home the point that they are different we present below functions which aren't described by a single formula. Some can be viewed as a patch-work of formulas. In some examples, a graph picturing the function is given.

Example 1. f is the absolute value function defined on the real line. There is a simple formula for this function. $f(x) = |x|$. But it may be defined as a compilation of even simpler formulas as:

$$f(x) = \begin{cases} x & \text{if } x \geq 0; \\ -x, & \text{otherwise.} \end{cases}$$

Figure 1: The absolute value function $f(x) = |x|$.

Example 2. g is referred to as a piecewise linear function. It is defined on the real line as:

$$g(x) = \begin{cases} 1, & \text{if } x < 0; \\ 1 + x, & \text{if } 0 \leq x < 3; \\ 4 + (3 - x), & \text{if } 3 \leq x < 5; \\ 2, & \text{if } x \geq 5. \end{cases}$$

Figure 2: A sample piecewise linear function.

Example 3. h is defined only on the interval $[0, 10]$. That is, its domain is the set $[0, 10]$. This is an example of a discontinuous function; it makes a sudden jump when $x = 5$.

$$h(x) = \begin{cases} x & \text{if } 0 \leq x < 5; \\ 7, & 5 \leq x \leq 10. \end{cases}$$

Figure 3: A discontinuous function.

Example 4. d is defined only on the input set $\{1, 2, 3\}$ (its domain) producing values in the output set $\{4, 10, 21\}$ (its range).

$$d(x) = \begin{cases} 10 & \text{if } x = 1; \\ 21, & \text{if } x = 2; \\ 4, & \text{if } x = 3. \end{cases}$$

Example 5. f is defined only on the set of positive integers producing the even numbers.

$$f(n) = 2\,n$$

Example 6. f is defined only on the set of positive integers producing the prime numbers.

$$f(n) = p_n \quad \text{where } p_n \text{ is the } n^{\text{th}} \text{ prime}$$

Several of the formulas above, while not represented as formulas, are patch-works of formulas, like examples 3 and 4. Example 5 can be expressed as a simple formula: $f(n) = 2n$. While Example 6 is a function for which one can't write down a formula; nor a (finite) patch-work of formulas as there is no known formula for expressing the sequence of prime numbers.†

Let us also try to remove the confusion about functions and their descriptions. The function f defined by $f(x) = x^2$ can also be defined by $f(z) = z^2$. Why? Because both descriptions tell us to do the same thing to their inputs. That is, the input variable is a dummy variable, not unlike a dummy variable used in a "for loop" of a programming language. For instance, consider the two "for loops" below:

```
sum = 0; for(i = 0; i < 10; i++) { sum += i; }

sum = 0; for(j = 0; j < 10; j++) { sum += j; }
```

The results don't depend on i or j. In the same way, be careful not to get too attached to the dummy variables used in function definitions. We state this one more time with emphasis:

NOTE: The following all define a function named f that squares its input:

$$f(x) = x^2 , \quad f(z) = z^2 , \quad f(a) = a^2 , \quad f(c) = c^2$$

Functions are NOT Graphs

Functions are abstract things, they are not numbers, but a thing that takes a number in to produce another; it is difficult to get a picture of this. Graphs of functions give us a picture. However, it is worth mentioning again that graphs are NOT functions. A function has a graph, but a graph may not correspond to a function. To make the distinction clearer we need to define what we mean by a graph.

Definition(Graph). *A graph is a set of pairs of numbers. That is, a graph is a set of the form:* $\{(x, y) \mid x, y \in R\}$.‡

Definition(Graph of a Function). *The graph of a function, f, is the set* $\{(x, f(x)) \mid x \in \text{Domain}(f)\}$.

This leads to a question: Given a graph, G, is there a corresponding function, f, for which the graph of f is G?

The answer, in general, is no. Consider, for example, a graph that is a circle. The graph of a function has the property that for any given input, x, (provided it's in the function's

† Actually, a relatively concise formula was produced in the 1960s but it is not a practical means of computing primes.

‡ *The symbol R stands for the set of all real numbers.*

domain) only one output is produced. What does this mean for the graph of a function? It means that if we take a vertical line through x, it should only hit the graph of the function at one point. Because otherwise, if it hit the graph at more than one point, there would be more than one value produced by f – and this can't be. So, in the case of the circle graph, we know that this is not the graph of a function because a vertical line hits more than one point for at least one value of x. In fact, a vertical line hits two points of the circle at all but the bounding x values.

Function Notation

In many respects mathematics is a triumph of notation. We use symbols to name and manipulate abstract ideas. It is often not as precise as mathematicians would like, however. There are cases where the short hand of notation is confusing or misleading. One case in particular is the way people refer to functions. In this book you will see functions notated as variables are notated; for example, we might call a function f. In some texts there is a concern that this simple name may confuse a function name with the name of a variable that has a numeric value. In order to add information (in this case to indicate that f is a function) some people use the notation $f(x)$ to try to make things clear. The sense which is to be conveyed by this notation is that f is a function because it takes inputs. The symbol x is often used because it conveys the idea of "variable input"; that is, f takes on "any number" of values. This leads to confusion because the usual interpretation of $f(x)$ is f evaluated at x.† Other people will use the notation $f(\cdot)$ to more clearly indicate that f is a function. The "dot" serves as a place holder indicating that f feeds on something; that is, it is a function.

Reiterating, we will use the notation $f(x)$ to mean: evaluate the function f at x.

† The only place where $f(x)$ has the sense of f taking on any number of values is when we *define* the function f as in $f(x) =$ value, or expression involving x.

Appendix B: Straight Line Functions

Given a function, we know how to find the graph. What about the reverse? Namely, for a given graph can we find the function which has that graph? In appendix A it is shown that a graph doesn't necessarily correspond to a function. Since a function is just a mapping from one value to another and the graph of a function is just a visual representation of this fact in some sense we are already done. What I mean when I say find the function which corresponds to a graph is really this: Can you find a *formula* for the function represented by a given graph? This is not possible in general even when the graph comes from a function, but it is for certain graphs. Perhaps the simplest graphs are straight lines, and in this case there are corresponding functions *and* we can find simple formulas for them. We make one restriction on the straight lines we consider: we assume that they are not vertical.

First, straight lines are the graphs of functions. Why, because they pass the vertical straight line test from appendix A: a vertical line intersects a straight line in at most one point.

Next, to find a formula for the corresponding function, let's start with the graph below. Pick an arbitrary point on the graph, (x, y).

Figure 1: Straight line through the origin.

The strategy will be to examine the constraint of being on the line and see if this translates into a formula for y in terms of the x. Notice, in this case, y is a "rise" of the line and x is the corresponding "run". Because the line is straight the ratio of the "rise" to the "run" is a constant we'll call m. Therefore, $y/x = m$, or

$$y = m\,x$$

Thus the function whose graph is the straight line passing through the origin is $f(x) = m\,x$. Such a function is called a *linear function.*

Figure 2: A linear function.

What about the graph in the figure below?

Figure 3: A straight line.

Can we find an equation for this more general line? Let's assume the slope of this line is m and it hits the y axis at the point $(0, b)$. You should convince yourself this defines the line uniquely. To find the formula whose graph is the line, we take an arbitrary point on the line (x, y) as before and translate the geometric relationship of being on the line into an algebraic one. Again, we will end up with y expressed as a formula in terms of x.

Notice the "rise" from the point $(0, b)$ to (x, y) is $y - b$ while the "run" is x. Since the slope of the line is m and is the ratio of any such "rise" to "run" we have $(y - b)/(x - 0) = m$. Or,

$$y = m\,x + b$$

So the function whose graph is a straight line is given by the function $f(x) = m\,x + b$. Such a function is called a *rectilinear function*. In this formula, b is referred to as the intercept and m is called the slope. What we derived is a representation for non-vertical straight lines.

Theorem(Straight Line Representation). *A non-vertical straight line is the graph of a function of the form*

$$f(x) = m\,x + b$$

for some choice of m *and* b. *The values* m *and* b *are unique.*

Appendix C: Infinite Sequences and Limits

In this section we will invent a definition for what we've been calling the limit of a sequence. The definition will come from looking at examples. I hope these examples will convince you that the definition given below is reasonable.

First to make life easier when discussing sequences let us define the *tail*, T_N, of a sequence $\{x_n\}_{n=1}^{\infty}$ at N to be the set $\{\, x_n \mid n > N \,\}$. You can think about forming the tail by visualizing the sequence as a list of numbers written out sequentially from left to right and then chopping the sequence at a location N; this leaves the list $[x_1 \ldots x_N]$ and the rest which we are calling the tail at N, T_N.

Example 1: Let $x_n = 1/n$. The tail of this sequence at 10, T_{10}, is $\{\, x_n \mid n > 10 \,\} = \{1/11,\ 1/12,\ 1/13,\ \ldots\}$.

You should draw a picture of the tail.

Notice that the sequence in the example seems to be "headed toward" 0. Question: Do you believe this, if so, how do we say this precisely?

We need to gain some experience with sequences in order to extract the essence of the notion of "heading towards". To this end, let's look at some examples to help us sharpen our idea of what this "heading toward" or "goes to" might mean. We will use the term *convergence* to mean that a sequence "goes to" a particular value. In this case we say the *limit* of the sequence exists. Our goal will be to give a precise definition of convergence.

Example 2: This example gets close and then "bounces away", but eventually it "goes to" zero. We would say that this sequence converges to 0.

$$1/1,\ 100,\ 1/2,\ 100/2,\ 1/3,\ 100/3,\ 1/4,\ 100/4,\ \ldots$$

Example 3: This example oscillates about 0, but gets smaller and smaller in absolute value, moving closer and closer to zero. In this case we would say the sequence converges to 0.

$$-1,\ 1/2,\ -1/3,\ 1/4,\ -1/5,\ 1/6,\ 1/7,\ -1/8,\ \ldots$$

Example 4: This sequence has elements that get close but then "bounce away". Therefore, this sequence does not converge as there is always an element in the tail of the sequence which is "far" away from 0.

$$1,\ 1,\ 1/2,\ 1,\ 1/3,\ 1,\ 1/4,\ 1,\ 1/5,\ 1,\ 1/6,\ 1,\ \ldots$$

Example 5: This sequence has elements that get close but then "bounce away". Therefore, this sequence does not converge to zero as there is always an element in the tail of the sequence which is "far" away from 0.

$$1,\ 0.0001,\ 1/2,\ 0.0001,\ 1/3,\ 0.0001,\ 1/4,\ 0.0001,\ 1/5,\ 0.0001,\ \ldots$$

Example 6: This sequence has elements that tend to 1. The farther out one goes in the sequence, the tail gets close and closer to 1. In this case we would say this sequence converges to 1.

$$1+1,\ 1+1/2, 1+1/3,\ 1+1/4, 1+1/5,\ 1+1/6\ \ldots$$

In the examples where the sequence "goes to" a point we see that the tails get closer and closer to this point. In the examples were there wasn't convergence we saw that getting close is a relative statement. Even though the sequence in example 5 seems to eventually get within 0.0001 of 0, it doesn't get any closer than this. So, it is not enough for *elements* of a sequence to occasionally get closer to a limit point. For convergence we need the *tails* to get closer to the limit point; and, not just closer, I think you will agree that we need the tails to get *arbitrarily close* to the limit point. How do we precisely define the phrase *arbitrarily close*?

One way to say this is that a sequence converges to a number a if the sequence eventually gets within *every* prescribed "closeness" tolerance of a.

This leads to a definition of convergence for infinite sequences.

Definition. *We say the sequence $\{x_n\}_{n=1}^{\infty}$ has the limit a, written $\lim_{n\to\infty} x_n = a$, if for every $\varepsilon > 0$ (closeness tolerance), there is a point in the sequence, N, (which may depend of ε) at which all the elements of its tail, T_N, are within ε of a.*

We now have a precise definition of what it means for a sequence to have a limit, but how do we actually show a given sequence converges to a given limit point? Notice work has to be done for *every* $\varepsilon > 0$. The problem is that there are an infinite number of such ε's!

The only way to be able to handle an infinite collection of such ε's is to automate the process of finding an N for a given ε. That is, we need to be able to find a function for N in terms of ε. Then, given an ε, we use this function to produce an N and then argue that any element in the corresponding tail is within this ε of the limit value. In effect, we come up with one (or maybe a handful) of generic proofs which cover all possible tolerances.

Note: This is, in general, a quantum leap in difficulty from what you are probably used to seeing. Although I hope you have a good idea of the what and why of a limit, don't be discouraged if you find the next example and the theorem after that hard. It took mankind a long time to figure out this definition and use it. You should also realize this definition is based on the examples we considered. The definition is a formalization of an agreement that was felt to embody convergence; that is, it is a human activity. Definitions are man made constructions based on concrete problems.

Example 5. *Prove that $\lim_{n\to\infty} x_n = 0$ when $x_n = 1/n$.*

proof: We need to write a program to handle *any* ε as input. That is, if someone were to hand us a closeness tolerance, ε, we would need to produce an N such that all elements in the tail are within this ε of 0. This means all elements of the tail satisfy $| x_n - 0 | < \varepsilon$, which is $| x_n | < \varepsilon$; or, $x_n < \varepsilon$, since all of the x_n are positive. What do we do?

Let's first start with a particular ε and see if we can handle that. Then maybe we can see how to *automate* our procedure to handle *any* ε. If ε is $1/10$, for instance, what would our choice be for N? Well, there are many choices for N. We are asking for a place to cut the sequence so that all elements of its tail are within $1/10$ of the limit point 0. We claim $N = 10$ works. Why, because the tail in this case is: $T_{10} = \{1/11, 1/12, 1/13, \ldots\}$. It's clear that choosing N to be 100 or 137 or 1000 works as well. This is because each of their tails: $T_{100} = \{1/101, 1/102, 1/103, \ldots\}$ or $T_{137} = \{1/138, 1/139, 1/40, \ldots\}$ or $T_{1000} = \{1/1001, 1/1002, 1/1003, \ldots\}$ all have the property that their elements are within $1/10$ of the limit point 0. We only need to find *an* N that works, let's pick N to be 10. So, we have found an N for a *particular* ε. Can we automate this process so that we can find an N for *any* ε? Notice that we found our N by finding the first N such that $1/(N+1)$ was less than $1/10$. If we do the same thing for an arbitrary ε we will have handled the automation step. The choice for N is then† $N = \text{floor}(1/\varepsilon) + 1$. Try a few choices of ε and see if this makes sense. For this choice of N every element in the tail T_N is within $1/N$ of 0 which is within ε of the 0.

We have done what was asked:

"Given any closeness tolerance $\varepsilon > 0$ find an N (which usually depends on ε) such that all elements in the tail of the sequence are within ε of the limit value."

Proofs of this kind are probably different from what you are used to. You are asked to automate a proof of a generic example; i.e., that of a given ε. Further, the choice of N is not unique. In the example above letting $N = 100 * (\text{floor}(1/\varepsilon) + 1)$ is also a valid choice for N. In addition, the way you find this N (based on a given ε) will change from proof to proof.

One thing I think we all assume from our definition is that if a sequence converges to a number a, then a is unique. That is, a sequence can't converge to two different numbers. This turns out to be true under our definition, but it requires a proof.

Theorem(Limits are Unique). *If the limit of a sequence exists, it is unique.*

proof: We will show that if the sequence x_n converges to a, $\lim_{n\to\infty} x_n = a$, and x_n converges to b, $\lim_{n\to\infty} x_n = b$, then a and b are equal.

We first sketch the outline of our argument. The key idea is that as we go farther and farther out in the sequence, the tail should be very close to both a *and* b; in fact *arbitrarily close*. This suggests that, in turn, a and b must be very close; in fact *arbitrarily close*. But to be arbitrarily close means that they are the same number. Therefore, our strategy is to show that $| a - b |$ is less than any prescribed tolerance. This will then mean that $| a - b | = 0$; that is, they are the same number. We now show that when $| a - b | < \varepsilon$ for *every* $\varepsilon > 0$, then $a = b$. This is really the same as showing that when a number c satisfies $| c | < \varepsilon$ for *every* $\varepsilon > 0$, then $c = 0$.

claim: Suppose we have a non-negative number, c, which has the property that $c < \varepsilon$ for all $\varepsilon > 0$, then $c = 0$. One might say that c is arbitrarily close to 0. If c isn't 0, then let $\varepsilon = c/2$. In this case c must be less than this ε, so $c < c/2$. Since c is positive the inequality

† The floor functions takes a number, x, and returns the greatest integer which is less than or equal to x. In effect, it "rounds down" to the nearest integer.

is preserved after dividing by c yielding $1 < 1/2$. This is a contradiction; therefore c must be zero.

We now continue with our proof and show that $\mid a-b \mid < \varepsilon$ for all $\varepsilon > 0$. Thus showing a and b are equal.

Let $\varepsilon > 0$ be a given number, then since we know x_n converges to a, given an ε_1 we know that there exists an N_1 such that $\mid x_n - a \mid < \varepsilon_1$ for all x_n in the tail T_{N_1}. Similarly, since x_n converges to b, given an ε_2 we know there exists an N_2 such that $\mid x_n - b \mid < \varepsilon_2$ for all x_n in the tail T_{N_2}. If we let $N = \max(N_1, N_2)$, then

$$\mid a-b \mid = \mid a - x_N + x_N - b \mid \leq \mid a - x_N \mid + \mid x_N - b \mid < \varepsilon_1 + \varepsilon_2$$

Now we need to choose ε_1 and ε_2 so that their sum is less than or equal ε. This will certainly be the case if both are chosen to be equal to $\varepsilon/2$. So, for an arbitrary $\varepsilon > 0$ the relation, $\mid a-b \mid < \varepsilon$ holds. Therefore, a is equal to b and we have shown that the limit of a sequence is unique.

Computational Strategy for Limits

One of the things we would like to be able to do, just as was done when computing derivatives, is to come up with ways to find the limits of sequences formed from simpler sequences. When computing derivatives we had a basic set of functions and computed their derivatives from the definition. We then had rules for how to find the derivatives of functions that were composed of these basic functions. We would like to invoke a similar strategy when computing limits. To this end, we would like to compute the limit of a sequence by recognizing it as a “base” sequence or by recognizing it as a composition of base sequences and use some set of rules to determine the limit in terms of limits of its base component sequences. Hopefully, we could avoid using the definition directly to work out a given limit just as we were able to avoid using the direct definition of the derivative to compute derivatives of certain classes of functions.

In what follows we assume that $\{x_n\}_{n=1}^{\infty}$ and $\{y_n\}_{n=1}^{\infty}$ are sequences with limits a and b respectively.

Here are some rules for combining sequences stated without proof:

Theorem(Linearity of Limits). *If* $z_n = c_1 x_n + c_2 y_n$ *then*

$$\lim_{n\to\infty} z_n = c_1 a + c_2 b$$

Theorem(Product of Limits). *If* $z_n = x_n * y_n$ *then*

$$\lim_{n\to\infty} z_n = a * b$$

Theorem(Quotient of Limits). *If* $z_n = x_n / y_n$ *and* $b \neq 0$ *then*

$$\lim_{n\to\infty} z_n = a/b$$

Theorem(Squeeze Limit). *If $x_n \le z_n \le y_n$ and $a = b$ then*

$$\lim_{n\to\infty} z_n = a = b$$

To this we add a specific set of sequences for which we can compute the limit. In what follows, $\{x_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ are sequences. The sequence $\{b_n\}_{n=1}^{\infty}$ is assumed to be bounded. By bounded we mean there is a number M such that the absolute value of every element in the sequence is less than or equal to M.

Theorem(Constant Limit). *If $x_n = k$ for some number k, then*

$$\lim_{n\to\infty} x_n = k$$

Theorem(Generalized Decreasing Sequence). *If $x_n = b_n/n$ then*

$$\lim_{n\to\infty} x_n = 0$$

Theorem(Power Limit). *If $x_n = k^n$ with $\mid k \mid < 1$, then*

$$\lim_{n\to\infty} x_n = 0$$

As an example of the limit of the generalized decreasing sequence pattern consider the sequence $\{x_n\}_{n=1}^{\infty}$, where $x_n = \sin(n)/n$. In this case $b_n = \sin(n)$. The only thing we need to do is check that $\mid b_n \mid$ is bounded. This is easy as the absolute value of the sin of anything is less than or equal to 1, for instance. Therefore,

$$\lim_{n\to\infty} \sin(n)/n = 0$$

As another example, consider the sequence $z_n = 1/n + \sin(n)/n$. We can view z_n as a sum of two sequences $1/n$ and $\sin(n)/n$. Since each of these sequences have limits their sum does also and the value of the limit is the sum of the limits of $1/n$ and $\sin(n)/n$. Therefore,

$$\lim_{n\to\infty} z_n = \lim_{n\to\infty} 1/n + \lim_{n\to\infty} \sin(n)/n = 0 + 0 = 0$$

Finally, if a given sequence, $\{x_n\}_{n=1}^{\infty}$, converges to a, is it the case that $\{f(x_n)\}_{n=1}^{\infty}$ converges to $f(a)$? This is a nice computational property to have as its says: $\lim_{n\to\infty} f(x_n) = f(\lim_{n\to\infty} x_n) = f(a)$. In other words, if a function has this property, limits "pass through" it. This property holds for "nice" functions. In fact, we will call a function *continuous* at a if this property holds. Functions that are continuous at every input value in their domain are called continuous functions.

So, what do continuous functions look like and how can one spot them. They have the property that while drawing their graph, one would not have to lift the pen off the page. All polynomials are continuous for all inputs; the logarithm functions are continuous for positive inputs; the exponential functions, a^x, for $a > 0$ are all continuous for all inputs. Sums, products, and differences of continuous functions are continuous. Ratios of continuous functions are continuous provided the denominator stays away from 0. For an example of a discontinuous function see example 3 in appendix A.

Convergence Based on Properties of a Sequence

We have seen how to show a sequence converges to a given value. What if we don't have a candidate for the limit? Can we tell if a sequence will converge just by examining properties of the sequence itself? The answer is yes, in certain situations we know that a limit of the sequence exists.

Theorem (Monotone Sequence). *Suppose $\{x_n\}_{n=0}^{\infty}$ is a sequence which is bounded above by M; meaning, $x_n \leq M$ for all n. If the sequence is also non-decreasing ($x_{n+1} \geq x_n$ for all n) then the limit of the sequence exists. Similarly, if a sequence is bounded below and is non-increasing it also has a limit.*

There are other criteria, but they are outside the scope of this book.

Sequences without Limits

Not all sequences have limits, in fact most do not. In this section we give examples of sequences which don't have limits.

Example 7: Let $x_n = (-1)^n$. This sequence bounces between 1 and -1 and so doesn't converge to anything.

Example 8: Let $x_n = n$. This sequence gets larger and larger and so it cannot converge to anything.

Example 9: Let $x_n = \sin(n)$. This sequence bounces around the interval $[-1, 1]$ but doesn't converge to anything.

Appendix D: Mathematical Induction

Domino Theory

Mathematical Induction is a technique used to determine the truth of a sequence of statements. The sequence is usually large or infinite. Normally, determining the truth of such a collection would be hard – if not time consuming. However, if there is a strong connection between successive elements of the sequence, then it can be much easier. By connection I mean that when a given statement in the sequence is true it is relatively easy to show that the next statement is true.

In this sense, Induction is a generalization of what might be called "The Domino Theory". If you have ever played with dominoes then, in a sense, you already understand Mathematical Induction. To see why, consider placing a number of (upright) dominoes in a room. Knocking them down may be easy or hard depending on how they are laid out. If they are put next to each other in the "standard domino way", then knocking one domino down will knock all the rest down. How would you describe this to a person who has never seen dominoes fall in this way? You might start by labeling the dominoes; that is, after setting up the dominoes they have a natural ordering – they form a sequence. Let's label them $D_1, D_2, \ldots, D_n$. You then tell the domino newcomer the dominoes are arranged so that whenever any one domino falls the domino after it falls. By this we mean domino D_2 is placed after domino D_1 so that when D_1 falls D_2 falls. In general, if domino D_k falls then D_{k+1} falls as well, if $1 \leq k < n$. Now you ask the newcomer: "Without looking into the room with the dominoes, if I told you that the first domino, D_1, has fallen, what can you say about the rest of dominoes?" So, you've given the newcomer two facts:

1. Domino D_1 has fallen.
2. Whenever domino D_k falls, domino D_{k+1} falls provided $1 \leq k < n$. That is, when one domino falls the next in line falls.

This is all he needs to conclude that all of the dominoes in the room have fallen.

This is the essence of mathematical Induction. Mathematical Induction concerns the *truth* of propositions labeled $P_1, P_2, \ldots, P_n$. Like Dominoes, the propositions are either easy or hard to prove depending on how they are "arranged". If the propositions are arranged like dominoes; that is, if

1. Proposition P_1 is true;
2. Whenever proposition P_k is true, proposition P_{k+1} is true provided $1 \leq k < n$;

then the conclusion is that all of the propositions are true.

In moving from dominoes to propositions all we've done is to replace statements like "domino D_k falls" with "proposition P_k is true".

This result extends to the case where there are an infinite number of propositions (dominoes). In this case the induction conditions (domino conditions) are:

1. Proposition P_1 is true.
2. Whenever proposition P_k is true, proposition P_{k+1} is true provided $1 \leq k$.

The conclusion under these conditions is that *all* of the P statements are true.

One may ask how easy is it to show that any given proposition implies the next? Just as one may ask how one verifies that a given domino falling knocks the next one down? For many collections of statements, P_n, it is the case that the proof of one proposition implying the next is generic, and essentially one only has to do one proof!

Applications

We apply Mathematical Induction to a few problems below. It is worth keeping in mind what must be shown in the second step of induction. You must show that P_{k+1} is true *assuming* that P_k is true *for all* $k \geq 1$. This does not mean P_{k+1} is true or P_k is true. If you get confused think of dominoes. To show domino D_{k+1} falls when D_k falls you would need to *assume* domino D_k has fallen and show that D_{k+1} falls as a consequence.

A Formula for Adding the First n Integers

By adding up the numbers from 1 to 2, then 1 to 3, and so forth, someone guesses that the formula for the sum of the first n integers is:

$$\sum_{i=1}^{n} i = n(n+1)/2$$

How can one prove this? We do it by Mathematical Induction. Let P_n be the statement* "$\sum_{i=1}^{n} i = n(n+1)/2$". We want to show that every statement, $P_1, P_2, \ldots$, is true. Mathematical Induction tells us that if P_1 is true and it is the case that whenever P_k is true it follows that P_{k+1} is true as well, then all of the statements in the sequence $\{P_n\}_{n=1}^{\infty}$ are true.

First we show P_1 is true. But P_1 is the statement "$\sum_{i=1}^{1} i = 1(1+1)/2$". The left hand side is 1 and the right hand side is 1; therefore, P_1 is true, since "$1 = 1$" is a true statement.

The next thing to do is to *assume* for any given k ($k \geq 1$) P_k is true. This means we assume "$\sum_{i=1}^{k} i = k(k+1)/2$" is a true statement; that is we assume: $\sum_{i=1}^{k} i = k(k+1)/2$. We then use this to show "$\sum_{i=1}^{k+1} i = (k+1)(k+2)/2$" is a true statement; that is, $\sum_{i=1}^{k+1} i = (k+1)(k+2)/2$. We proceed as follows: Take the sum on the left and relate it the P_k.

$$\sum_{i=1}^{k+1} i = (k+1) + \sum_{i=1}^{k} i$$

We use the assumed truth of P_k to rewrite the sum on the right hand side so that

$$\sum_{i=1}^{k+1} i = (k+1) + \sum_{i=1}^{k} i = (k+1) + k(k+1)/2 = \big(2(k+1) + k(k+1)\big)/2 = (k+1)(k+2)/2$$

That is,

$$\sum_{i=1}^{k+1} i = (k+1)(k+2)/2$$

---

* We surround statements in double quotes to indicate that they are statements with no assigned truth value. You can think of our usual mathematical statements such as, $3 = 2 - 1$, as consisting of two things: The first is the statement "3 is equal to 2 minus 1". The second thing that is conveyed is that the statement is true. In this section we wish to make the separation of a statement from its truth value. So we will refer to statements, P_n, in quotes to indicate that they are statements without (yet) an assigned truth value.

We have just shown the truth of the statement P_{k+1}, "$\sum_{i=1}^{k+1} i = (k+1)(k+2)/2$", from the assumed truth of P_k. We see that we have done so for every $k \geq 1$; the same argument works for *every* k.† Mathematical Induction tells us that P_n is true for all n. That is, we have shown that

$$\sum_{i=1}^{n} i = n(n+1)/2 \quad \text{for } 1 \leq n$$

Notice what helped here was that we had a natural ordering of the propositions (like our room of dominoes) and showing the truth of a given P_k implying P_{k+1} was the same for all k. Most successful applications of induction behave similarly. If the proofs were different for each k, then induction wouldn't be as useful; we would have a lot of work to do.

The Derivative of the Power Functions

We know the derivative of the function x; it is 1. What is the derivative of x^n when n is an integer? We stated earlier in the book that the answer is $(x^n)' = n\,x^{n-1}$. We can use Mathematical Induction to show this for $n \geq 1$. Let P_n be the statement: "$(x^n)' = n\,x^{n-1}$" First, $(x^1)' = x' = 1$; while $1x^{1-1} = 1 * x^0 = 1 * 1 = 1$. So, P_1 is true. We *assume* now for any k ($k \geq 1$), P_k is true and try to show P_{k+1} is true. Statement P_k true means $(x^k)' = k\,x^{k-1}$. Now, the derivative of x^{k+1} can be related to x^k since $x^{k+1} = x\,x^k$. Using the product rule and the assumed truth of the formula for the derivative of x^k gives: $(xx^k)' = x'x^k + x(x^k)' = 1x^k + x(k\,x^{k-1}) = x^k + k\,x^k = (k+1)x^k$. Or, $(x^{k+1})' = (k+1)x^{(k+1)-1}$. That is, the statement P_{k+1} is true when P_k is true (for *any* $k \geq 1$). By Mathematical Induction, P_n is true for all $n \geq 1$. That is,

$$(x^n)' = n\,x^{n-1} \quad \text{for } n \geq 1$$

What about when n is 0 or negative, we haven't proved the formulas in these cases? One can easily check the formula for $n = 0$. For negative n we can use induction again. In this case let P_n be the statement "$(x^{-n})' = -nx^{-n-1}$" for n 1 and $x \neq 0$. Induction requires statements labeled from 1, so we map our labels to the sequence of statements by negating n. To check the truth of P_1: $(x^{-1})' = -1x^{-1-1} = -x^{-2}$ one can use the definition of the derivative directly. This requires work, but once done, one can mimic the induction proof above. Given that P_1 has been proved true we next show that P_k implies P_{k+1}. Consider an $x \neq 0$, then

$$(x^{-(k+1)})' = (x^{-1}x^{-k})' = (x^{-1})'x^{-k} + x^{-1}(x^{-k})'$$

The last statement is true by the product rule. Using the assumed truth of P_1 and P_k we may write

$$(x^{-(k+1)})' = -1x^{-1-1}x^{-k} + x^{-1}(-kx^{-k-1}) = -x^{k-2} - kx^{-k-2} = -(k+1)x^{-(k+1)-1}$$

We have just shown that P_{k+1} is true given that P_k is true. By mathematical induction P_n must be true for $n \geq 1$. The only thing we left out was the proof that P_1 is true. We leave this as an exercise for the reader.

† Notice that this is the same *magic* we have when finding formulas for the derivative of certain functions; namely, we find a way to compute the formula for the derivative at a given "x" and that same formula works for all other "x"s. If this wasn't the case for a lot of elementary functions, calculus wouldn't be useful. This is the same for induction: If the way to show the truth of statement P_k implies that P_{k+1} is true *changes* for each k, then induction isn't useful. This is analogous to writing a function in a programming language that is parameterized. Sometimes rather than writing many many such functions, one can write *one function* that takes one or more parameters while the body of the function is *generic*. This generic logic doesn't always happen in programming nor does it always happen in mathematics, but when it does and we can identify the generic argument, Induction gives us tremendous power.

Appendix E: Taylor's Theorem

We start by stating and proving a generalization of the Mean Value Theorem:

Theorem(Generalized Mean Value Theorem). *Let f and w be continuous functions on the interval $[a,b]$. If f has minimum and maximum values m and M while $w(s) \geq 0$ for $s \in [a,b]$ and $\int_a^b w(s)\,ds > 0$, then there exists a value $\mu \in [a,b]$ such that*

$$\int_a^b f(s)w(s)\,ds = f(\mu)\int_a^b w(s)\,ds$$

proof: Since f is greater or equal to m and less than or equal M we must have

$$\int_a^b m\,w(s)\,ds \leq \int_a^b f(s)w(s)\,ds \leq \int_a^b Mw(s)\,ds$$

Since m and M are constants they may be taken out of the integrals (linearity property of integrals), so that

$$m\int_a^b w(s)\,ds \leq \int_a^b f(s)w(s)\,ds \leq M\int_a^b w(s)\,ds$$

Dividing by the integral in w gives

$$m \leq \frac{\int_a^b f(s)w(s)\,ds}{\int_a^b w(s)\,ds} \leq M$$

Note that the middle term is an intermediate value of the function f. Since f is a continuous function on $[a,b]$, the intermediate value theorem tells us that there is a $\mu \in [a,b]$ such that $f(\mu) = \frac{\int_a^b f(s)w(s)\,ds}{\int_a^b w(s)\,ds}$. This can be rewritten as

$$\int_a^b f(s)w(s)\,ds = f(\mu)\int_a^b w(s)\,ds$$

Corollary(Mean Value Theorem). *If f is a function such that f' exists and is a continuous function on the interval $[a,b]$, then there exists a $\mu \in [a,b]$ such that*

$$f(b) = f(a) + f'(\mu)(b-a)$$

This follows from the generalized Mean Value result substituting f' for f and setting $w(x) \equiv 1$ for all $x \in [a,b]$. That is, we have for some $\mu \in [a,b]$:

$$\int_a^b f'(s)1\,ds = f'(\mu)\int_a^b 1\,ds$$

Or,

$$f(b) - f(a) = f'(\mu)(b-a)$$

Taylor's Theorem

If f' is a continuous function we know by the Mean Value Theorem that

$$f(b) = f(a) + f'(\mu)(b - a)$$

But we could invoke the theorem for any $x \in [a, b]$ because $[a, x]$ is an interval where the theorem applies. In this case the conclusion of the theorem is

$$f(x) = f(a) + f'(\mu_1(x))(x - a) \tag{1}$$

Note that since the value μ_1 depends on the interval $[a, x]$, and as we are treating a as a constant and varying x, μ will potentially change as x changes; therefore, it is a function of x.

If f'' is also continuous on $[a, b]$, then we may apply this formula to f' giving

$$f'(x) = f'(a) + f''(\mu_2(x))(x - a)$$

for some function μ_2(most likely different than μ_1). If we now integrate this function over the interval $[a, x]$ we obtain

$$\int_a^x f'(s)\,ds = \int_a^x f'(a)\,ds + \int_a^x f''(\mu_2(s))(s - a)\,ds$$

Or,

$$f(x) - f(a) = f'(a)\int_a^x 1\,ds + \int_a^x f''(\mu_2(s))(s - a)\,ds = f'(a)(x - a) + \int_a^x f''(\mu_2(s))(s - a)\,ds$$

Which is

$$f(x) = f(a) + f'(a)(x - a) + \int_a^x f''(\mu_2(s))(s - a)\,ds$$

But by the Generalized Mean Value Theorem the last term is just $f''(\mu_3(x))\int_a^x (s - a)\,ds$ (for some functoin μ_3). Since $(s - a)^2/2$ is an anti-derivative of $(s - a)$ we have

$$f(x) = f(a) + f'(a)(x - a) + f''(\mu_3(x))\int_a^x \left(\frac{(s - a)^2}{2}\right)'\,ds$$

The last integral is evaluated using the FTC giving

$$f(x) = f(a) + f'(a)(x - a) + f''(\mu_3(x))\frac{(x - a)^2}{2} \tag{2}$$

Again, if the function f''' is continuous, then we may plug f' into (2) giving

$$f'(x) = f'(a) + f''(a)(x - a) + f'''(\mu_4(x))\frac{(x - a)^2}{2}$$

Again, for some function μ_4. As before, we integrate this equation over the interval $[a, x]$ giving

$$f(x) - f(a) = f'(a)(x - a) + f''(a)\frac{(x - a)^2}{2} + \int_a^x f'''(\mu_4(s))\frac{(s - a)^2}{2}\,ds$$

The last term can be replaced using the Generalized Mean Value Theorem giving (which means we get a new function, μ_5).

$$f(x) - f(a) = f'(a)(x-a) + f''(a)\frac{(x-a)^2}{2} + f^{(3)}(\mu_5(x))\int_a^x \frac{(s-a)^2}{2}\,ds$$

An anti-derivative of $(s-a)^2/2$ is $(s-a)^3/(3*2)$ so the integral may be replaced using the FTC

$$f(x) = f(a) + f'(a)(x-a) + f''(a)\frac{(x-a)^2}{2} + f^{(3)}(\mu_5(x))\frac{(x-a)^3}{3*2}$$

Plugging in f' into this equation (provided $f^{(4)}$ exists and is continuous) and repeating the previous calculations yields:

$$f(x) = f(a) + f'(a)(x-a) + f''(a)\frac{(x-a)^2}{2} + f'''(a)\frac{(x-a)^3}{3*2} + f^{(4)}(\mu_6(x))\frac{(x-a)^4}{4*3*2}$$

One can continue plugging in f' into these increasingly complicated expressions. The pattern that emerges is called Taylor's Theorem:

Theorem(Taylor's Theorem). *If f is a function such that $f^{(n)}$ exists and is a continuous function on the interval $[a,b]$, then for any $x \in [a,b]$*†

$$f(x) = f(a) + \sum_{i=1}^{n-1} f^{(i)}(a)\frac{(x-a)^i}{i!} + f^{(n)}(\mu(x))\frac{(x-a)^n}{n!}$$

where μ is a function such that $\mu(x) \in [a,x]$.

How does one prove this? We can use induction (see the appendix on induction). If P_n is the statement found in the theorem, then the main work that has to be done is to take our calculations that showed that P_1 implies P_2 and generalize it to show that P_k implies P_{k+1}.

Proof by induction means that to prove that P_n is true for any n we must show P_1 to be true and to show that P_k implies P_{k+1} for any $k \geq 1$.

P_1 is true since this is just the equation (1) which we know to be true. Next, if $f^{(k+1)}$ is continuous, then assuming P_k is true we need to show that P_{k+1} is true. If P_k is true we have

$$f(x) = f(a) + \sum_{i=1}^{k-1} f^{(i)}(a)\frac{(x-a)^i}{i!} + f^{(k)}(\mu(x))\frac{(x-a)^k}{k}$$

To show P_{k+1} we are given that $f^{(n+1)}$ exists and is continuous. Using the assumed truth of P_k with the function f' gives

$$f'(x) = f'(a) + \sum_{i=1}^{k-1} f^{(i+1)}(a)\frac{(x-a)^i}{i!} + f^{(k+1)}(\mu_1(x))\frac{(x-a)^k}{k}$$

As before, integrate this equation from $[a,x]$

$$\int_a^x f'(s)\,ds = \int_a^x f'(a)\,ds + \int_a^x \sum_{i=1}^{k-1} f^{(i+1)}(a)\frac{(s-a)^i}{i!}\,ds + \int_a^x f^{(k+1)}(\mu_1(s))\frac{(s-a)^k}{k!}\,ds$$

† Note, when $n=1$, $\sum_{i=1}^{n-1} g_i = \sum_{i=1}^{0} g_i = 0$, regardless of the values of g_i.

We use the FTC on the left hand side and the first term of the right hand side. We rely on the linearity of the integration to pass the integral through the sum and past the constants $f^{i+1}(a)$ giving

$$f(x) - f(a) = f'(a)(x-a) + \sum_{i=1}^{k-1} f^{(i+1}(a) \int_a^x \frac{(s-a)^i}{i!}\, ds + \int_a^x f^{(k+1)}(\mu_1(s)) \frac{(s-a)^k}{k!}\, ds$$

We note that an anti-derivative of $\frac{(s-a)^i}{i!}$ is $\frac{(s-a)^{i+1}}{(i+1)i!} = \frac{(s-a)^{i+1}}{(i+1)!}$ ‡. Therefore, the last equation may be written

$$f(x) - f(a) = f'(a)(x-a) + \sum_{i=1}^{k-1} f^{(i+1)}(a) \int_a^x \left(\frac{(s-a)^{i+1}}{(i+1)!} \right)' ds + \int_a^x f^{(k+1)}(\mu_1(s)) \frac{(s-a)^k}{k!}\, ds$$

Using the FTC on the middle integrals gives

$$f(x) - f(a) = f'(a)(x-a) + \sum_{i=1}^{k-1} f^{(i+1)}(a) \frac{(x-a)^{i+1}}{(i+1)!} + \int_a^x f^{(k+1)}(\mu_1(s)) \frac{(s-a)^k}{k!}\, ds$$

Using the generalized mean value theorem on the last term yields

$$f(x) = f(a) + f'(a)(x-a) + \sum_{i=1}^{k-1} f^{(i+1)}(a) \frac{(x-a)^{i+1}}{(i+1)!} + f^{(k+1)}(\mu_2(x)) \int_a^x \frac{(s-a)^k}{k!}\, ds$$

An anti-derivative of $\frac{(s-a)^k}{k!}$ is $\frac{(s-a)^{k+1}}{(k+1)!}$ so that integral above can be replaced using the FTC

$$f(x) = f(a) + f'(a)(x-a) + \sum_{i=1}^{k-1} f^{(i+1)}(a) \frac{(x-a)^{i+1}}{(i+1)!} + f^{(k+1)}(\mu_3(x)) \frac{(x-a)^{k+1}}{(k+1)!}$$

Making the substitution $j = i + 1$ in the middle sum gives

$$f(x) = f(a) + f'(a)(x-a) + \sum_{j=2}^{k} f^{(j)}(a) \frac{(x-a)^{j}}{j!} + f^{(k+1)}(\mu_3(x)) \frac{(x-a)^{k+1}}{(k+1)!}$$

Combining the second term with the middle sum yields

$$f(x) = f(a) + \sum_{j=1}^{k} f^{(j)}(a) \frac{(x-a)^{j}}{j!} + f^{(k+1)}(\mu_3(x)) \frac{(x-a)^{k+1}}{(k+1)!}$$

We have established the truth of statement P_{k+1}. Note that the summation variable, j, is a dummy variable and could be relabeled i. And, μ_3 is a function with the requisite property: $\mu_3(x) \in [a, x]$. The two requirements for induction have been fulfilled; therefore the theorem is proved.

‡ The anti-derivative is with respect to the variable s.