An Introduction to Proofs and the Mathematical Vernacular [1]

Martin V. Day
Department of Mathematics
Virginia Tech
Blacksburg, Virginia 24061
http://www.math.vt.edu/people/day/ProofsBook

Dedicated to the memory of my mother:
Coralyn S. Day, November 6, 1922 – May 13, 2008.

December 30, 2015

Contents

Preface: To the Student

In the standard year (or two) of university calculus and differential equations courses you have learned a lot of mathematical techniques for solving various types of problems. Along the way you were offered "proofs" of many of the fundamental relationships and formulas (stated as "theorems"). Perhaps occasionally you were asked to "show" or "prove" something yourself as a homework problem. For the most part, however, you probably viewed the proofs as something to be endured in the lectures and skimmed over in the book. The main emphasis of those courses was on learning *how to use* the techniques of calculus, and the proofs may not have seemed very helpful for that.

Historically, techniques of calculation were the principal concern of mathematics. But as those techniques became more complex, the concepts behind them became increasingly important. You are now at the stage of your mathematical education where the focus of your studies shifts from techniques to ideas. The goal of this book is to help you make the transition from being a mere user of mathematics to becoming conversant in the language of mathematical discussion. This means learning to critically read and evaluate mathematical statements and being able to write mathematical explanations in clear, logically precise language. We will focus especially on mathematical proofs, which are nothing but carefully prepared expressions of mathematical reasoning.

By focusing on how proofs work and how they are expressed we will be learning to think about mathematics as mathematicians do. This means learning the language and notation (symbols) which we use to express our reasoning precisely. But writing a proof is always preceded by *finding* the logical argument that the proof expresses, and that may involve some exploration and experimentation, trying various ideas, and being creative. We will do some practicing with mathematics that is familiar to you, but it is important to practice with material that you don't already know as well, so that you can really have a try at the creative exploration part of writing a proof. For that purpose I have tried to include some topics that you haven't seen before (and may not see elsewhere in the usual undergraduate curriculum). On the other hand I don't want this course to be divorced from what you have learned in your previous courses, so I also have tried to include some problems that ask you to use your calculus background. Of course the book includes many proofs which are meant to serve as examples as you learn to write your own proofs. But there are also some which are more involved than anything you will be asked to do. Don't tune these out — learning to read a more complicated argument written by someone else is also a goal of this course.

Some consider mathematics to be a dry, cold (even painful) subject. It certainly is very difficult in places. But it can also be exciting when we see ideas come together in unexpected ways, and see the creative ways that great minds have exploited unseen connections between topics. To that end I have included a few examples of really clever proofs of famous theorems. It is somewhat remarkable that a subject with such high and objective standards of logical correctness should at the same time provide such opportunity for the expression of playful genius. This is what attracts many people to the study of mathematics, particularly those of us who have made it our life's work. I hope this book communicates at least a little of that excitement to you.

Observe that theorems, lemmas, and propositions are numbered consecutively within chapters. For instance in Chapter 4 we find Theorem 4.6 followed by Lemma 4.9. Many theorems also have names (e.g. "The Division Theorem"). You can locate them using the index. Definitions are not numbered, but are listed in the index by topic. Problems are numbered within chapters similarly; for instance you will find Problem 2.8 in Chapter 2. The end of each problem statement is marked by a dotted line followed by a cryptic word in a box, like this:

..... `whatever`

The word in the box is of no significance to you; it is just a reminder for me of the computer file which contains the text of the problem.

References are marked by a number in brackets like this: [9]. You can find the full reference for [9] in the Bibliography at the back of the book. (This is the customary format for citing references in mathematics.) Sometimes a page number or other information is included after the number to help you find the relevant place in the reference, like [9, page 99].

It is unlikely that every problem in the book will be assigned. If a problem you are assigned refers to another which was not assigned, you are *not* expected to work the unassigned problem, but can just take its statement for granted and use it in your solution of the assigned problem. For instance, if you were assigned Problem 4.17 (which asks you to use Problem 4.16), you are free to use the fact stated in Problem 4.16 whether or not it was assigned.

I want to thank the students of Math 3034, Spring 2008, who had to deal with the rather rough first version of this book. A special thanks to those who took the time to give me constructive criticism and suggestions: Aron Chun, Chelsey Cooper, Justin Cruz, Anna Darby, Victoria Dean, Mark McKinley, Drew Nagy, David Richardson, Ryan Ritch, Diana Schoof, Samantha Simcik, Jennifer Soldan, John Stevens, Dean Stevenson, and Kendra Valencia. Thanks also to Dawson Allen, Michele Block, Todd Gasparello, Micah Hafich, Michael Overson, Kelsey Pope, Kehong Shi, Melissa Tilashalski, and Michael Wolf from Fall 2008 for their comments and suggestions. Thanks also to William Gunther for correcting some misstatements about the continuum hypothesis.

Martin Day, Blacksburg, VA

Chapter 1

Some Specimen Proofs

This chapter begins our study of proofs by looking at numerous examples. In the *next* chapter we will try to summarize the logic which underlies typical proofs and the special ways the English language is used in precise mathematical discussion. This is the way most people learn a new language — learn to say a few simple things first and *after* that start to learn the rules of grammar. But not everyone is comfortable with this jumping–right–in approach. So if you want more up–front explanation, feel free to skip ahead to Chapter 2 and read it now. In particular you might look at the chart on page 37 which catalogues some basic types of proofs, and the advice for writing proofs on page 50. Consulting those as we work through this chapter may be helpful.

Along with the proof specimens in this chapter we include a couple *spoofs*, by which we mean arguments that seem like proofs on their surface, but which in fact come to false conclusions. The point of these is that the style or language of an argument does not make it a proof; what matters is that its logic stands up to close scrutiny. Learning to look past the language and critically examine the content of an argument is important for both reading and writing proofs.

The mathematical topics in this chapter don't fit together in any particular way, so don't look for some mathematical theme which connects them. Instead, you should view this chapter as a sampler of different types of proofs. In fact, much of the material of this chapter will be familiar to you. As we discuss each topic, ask yourself not whether you already know it, but whether you know *why* it is true. Could you write a convincing justification of it to someone who is skeptical?

A Inequalities and Square Roots

A.1 Absolute Value and the Triangle Inequality

Our first examples are inequalities involving the absolute value. Before anything else, we need to be sure everyone understands what the absolute value refers to. Here is its definition.

Definition. For a real number x, the *absolute value* of x is defined to be

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0. \end{cases}$$

This definition is doing two things. It is stating precisely what we mean by the words "the absolute value of x" and it is announcing the notation "$|x|$" which is used to refer to it. (Don't let the minus sign in second line of the definition confuse you. In the case of $x < 0$ observe that $|x| = -x$ is a *positive* value. For instance, if $x = -2$ then $-x = +2$.)

Here is an elementary property of the absolute value.

Lemma 1.1. *For every real number x, $x \leq |x|$.*

We have called this a "lemma." The words "lemma," "theorem," "proposition," and "corollary" all refer to a statement or announcement of a mathematical fact. You have seen theorems before. Later on[1] we will talk about how "lemma" is different from "theorem." For now, just think of it as a sort of mini-theorem. This particular lemma is saying that no matter what value of x you pick, if you compare the values of x and $|x|$ you will always find that $|x|$ is at least as big as x. For instance, considering $x = -3.227$ the lemma is telling us that $-3.227 \leq |-3.227|$. Or considering $x = \frac{1}{\sqrt{2}}$, the lemma is telling us that $\frac{1}{\sqrt{2}} \leq \left|\frac{1}{\sqrt{2}}\right|$. You probably have no doubt that this lemma is true. The question for us is how we can write a proof of it. We can't check all possible values of x one at a time. We need to give reasons that apply simultaneously to all x. Since the definition of $|x|$ falls into two cases, it is natural to write a proof that provides reasons in two cases; one line of reasoning that works if $x \geq 0$ and a second line of reasoning that works if $x < 0$. Let's talk through these two lines of reasoning.

The first case we want to consider accounts for those x for which $x \geq 0$. According to the definition above, for these x we know $|x| = x$. So what we are trying prove is just that $x \leq x$, which is certainly true.

The second case considers those x for which $x < 0$. In this case the definition says that $|x| = -x$. So what we are trying to prove is that $x < -x$. What reasons can we give for this? For this case we know that $x < 0$, and that implies (multiplying both sides by -1) that $0 < -x$. So we can string the two inequalities

$$x < 0 \quad \text{and} \quad 0 < -x$$

together to conclude that $x < -x$, just as we wanted.

Now we will write out the above reasoning as a proof of Lemma 1.1 in a typical style. Observe that the beginning of our proof is announced with the italicized word "*Proof*" and the end is marked with a small box[2], so you know exactly where the proof starts and stops.

Proof. We give the proof in two cases.

Case 1: Suppose $x \geq 0$. In this case, the definition of absolute value says that $|x| = x$. So the inequality $x \leq |x|$ is equivalent to $x \leq x$, which *is* true.

Case 2: Suppose $x < 0$. In this case, the definition of absolute value says that $|x| = -x$. Since $x < 0$, it follows that $0 < -x$, and therefore

$$x \leq -x = |x|,$$

proving the lemma in the case as well.

Since the two cases exhaust all possibilities, this completes the proof. □

Here is another property of the absolute value which everyone knows as the Triangle Inequality, stated here as a theorem.

Theorem 1.2 (Triangle Inequality)**.** *For all real numbers a and b,*

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

We can test what the theorem is saying by just picking some values for a and b.

Example 1.1.

- For $a = 2$, $b = 5$: $|a + b| = 7$, $|a| + |b| = 7$, and $7 \leq 7$ is true.
- For $a = -6$, $b = 17/3$: $|a + b| = |-1/3| = 1/3$, $|a| + |b| = 35/3$, and $1/3 \leq 35/3$ is true.
- For $a = -\pi$, $b = 1.4$: $|a + b| = 1.741595\ldots$, $|a| + |b| = 4.54159\ldots$, and $1.741595\ldots \leq 4.54159\ldots$ is true.

[1] See the footnote on page 25 and the description in Appendix: Mathematical Words.
[2] Called a halmos, after Paul Halmos who initiated that convention.

In each case the inequality works out to be true, as the theorem promised it would. But we have only checked three of the infinitely many possible choices of a and b. We want to write a proof that covers all possible choices for a and b. There are several approaches we might follow in developing a proof. One would be to consider all the possible ways the three absolute values in the triangle inequality could work out. For instance if $a \geq 0$, $b < 0$ and $a + b < 0$, then we could make the replacements

$$|a + b| = -(a + b), \quad |a| = a, \quad |b| = -b.$$

Then we would need to show that $-(a + b) \leq a - b$, using the assumptions that $a \geq 0$, $b < 0$, and $a + b < 0$. We can develop a valid proof this way, but it would have many cases.

1: $a \geq 0$, $b \geq 0$.

2: $a \geq 0$, $b < 0$, and $a + b \geq 0$.

3: $a \geq 0$, $b < 0$, and $a + b < 0$.

4: $a < 0$, $b \geq 0$, and $a + b \geq 0$.

5: $a < 0$, $b \geq 0$, and $a + b < 0$.

6: $a < 0$, $b < 0$.

This would lead to a tedious proof. There is nothing wrong with that, but a more efficient argument will be more pleasant to both read and write. We will produce a shorter proof using just two cases: Case 1 will be for $a + b \geq 0$ and Case 2 will account for $a + b < 0$.

As we think through how to write this (or any) proof, **it is important to think about what our reader will need in order to be convinced.** Picture a reader who is another student in the class and has followed everything so far, but is skeptical about the triangle inequality. They know the definition of absolute value and have read our proof of Lemma 1.1 and so will accept its use as part of the proof. They have seen our test cases above but are thinking, "Well, how do I know there aren't some other clever choices of a and b for which $|a + b| > |a| + |b|$?" We expect the reader to to look for gaps or flaws in our proof but can be confident they will agree with our reasoning when there is no logical alternative[3]. Our proof needs to be written to convince this reader.

With this reader in mind, lets think about Case 1. In that case $|a + b| = a + b$, so we need to provide reasons to justify $a + b \leq |a| + |b|$. But we can deduce that by adding the two inequalities

$$a \leq |a|, \quad b \leq |b|,$$

both of which our reader will accept once we point out that these are just restatements of Lemma 1.1. What about Case 2, in which $|a + b| = -(a + b)$? Since $-(a + b) = (-a) + (-b)$, we can do something similar to what we did above: add the two inequalities

$$-a \leq |a|, \quad -b \leq |b|.$$

Now, what should we write to convince our skeptical reader of these? Using the fact that $|-a| = |a|$ we can point out to our reader that by Lemma 1.1 $-a \leq |-a|$ and therefore $-a \leq |-a| = |a|$. Likewise for $-b \leq |b|$. We should ask ourselves whether our reader knows that $|-a| = |a|$. Well they probably do, but we can't be absolutely sure. There are always judgments like this to make about how much explanation to include. (Ask yourself how much *you* would want written out if you were the skeptical reader.) We could give them a proof that $|-x| = |x|$ for all real numbers x. Or as a compromise, we can just put in a reminder of this fact — that's what we will do. Here then is our proof, written with these considerations in mind.

Proof. Consider any pair of real numbers a and b. We present the proof in two cases.

[3] We must assume that our reader is rational and honest. If they are adamant in doubting the triangle inequality, even when faced with irrefutable logic, there is nothing we can do to convince them.

Case 1: Suppose $a + b \geq 0$. In this case $|a + b| = a + b$. Note that Lemma 1.1 implies that $a \leq |a|$ and $b \leq |b|$. Adding these two inequalities yields $a + b \leq |a| + |b|$. Therefore we have

$$|a + b| = a + b \leq |a| + |b|,$$

which proves the triangle inequality in this case.

Case 2: Suppose $a + b < 0$. In this case $|a + b| = -a - b$. Applying Lemma 1.1 with $x = -a$ we know that $-a \leq |-a|$. But because $|-a| = |a|$, this implies that $-a \leq |a|$. By the same reasoning, $-b \leq |b|$. Adding these two inequalities we deduce that $-a - b \leq |a| + |b|$ and therefore

$$|a + b| = -a - b \leq |a| + |b|.$$

This proves the triangle inequality in the second case.

Since these cases exhaust all possibilities, this completes the proof. □

The proofs above are simple examples of proofs by cases, described in the "For all x ..." row of the table on page 37. We gave different arguments that worked under different circumstances (cases) and every possibility fell under one of the cases. Our written proofs say the same things as our discussions preceding it, but are more concise and easier to read.

Problem 1.1 Write out proofs of the following properties of the absolute value, using whatever cases seem most natural to you.

a) $|x| = |-x|$ for all real numbers x.

b) $0 \leq |x|$ for all real numbers x.

c) $|x|^2 = x^2$ for all real numbers x.

d) $|xy| = |x|\,|y|$ for all real numbers x and y.

..... `absv`

Problem 1.2 At the top of page 3 a proof of the triangle inequality in six cases was contemplated. Write out the proofs for cases 4, 5, and 6.

..... `trialt`

Problem 1.3 Prove that for any two real numbers a and b,

$$||a| - |b|| \leq |a - b|.$$

(There are four absolute values here; if you split each of them into two cases you would have 16 cases to consider! You can do it with just two cases: whether $|a| - |b|$ is positive or not. Then take advantage of properties like the triangle inequality which we have already proven.)

..... `absv3`

Problem 1.4 For two real numbers a, b, the notation $\max(a, b)$ refers to the maximum of a and b, i.e. the larger of the two:

$$\max(a, b) = \begin{cases} a & \text{if } a \geq b \\ b & \text{if } a < b. \end{cases}$$

The smaller of the two is denoted $\min(a, b)$.

Prove that for any two real numbers a and b, $\max(a,b) = \frac{1}{2}(a+b+|a-b|)$. Find a similar formula for the minimum. (You don't need to prove the minimum formula. Just find it.)

..... max

Problem 1.5 Define the *sign* or *signum* function of a real number x to be

$$\operatorname{sgn}(x) = \begin{cases} +1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}.$$

Prove the following facts, for all real numbers a and b.

a) $\operatorname{sgn}(ab) = \operatorname{sgn}(a)\operatorname{sgn}(b)$.

b) $\operatorname{sgn}(a) = \operatorname{sgn}(1/a)$, provided $a \neq 0$.

c) $\operatorname{sgn}(-a) = -\operatorname{sgn}(a)$.

d) $\operatorname{sgn}(a) = a/|a|$, provided $a \neq 0$.

..... sgn

A.2 Square Roots

Next we consider some basic properties of the square root function, $\sqrt{s}$. Although these facts are well-known to you, we want to make ourselves stop long enough to ask *why* they are true.

First, what exactly do we mean by the square root $\sqrt{s}$ of a real number s? For instance is $\sqrt{4} = 2$ or is $\sqrt{4} = -2$? You might be inclined to say "both," but that is incorrect. The notation $\sqrt{s}$ refers only to the *nonnegative*[4] number r which solves $r^2 = s$. Thus $\sqrt{4} = 2$, *not* -2.

Definition. Suppose s is a nonnegative real number. We say r is the *square root* of s, and write $r = \sqrt{s}$, when r is a nonnegative real number for which $r^2 = s$.

We all know that for $s \geq 0$, $\sqrt{s}$ does exist and there is only one value of r that qualifies to be called $r = \sqrt{s}$, but for $s < 0$ there is no (real[5]) value that qualifies to be called $\sqrt{s}$. For nonnegative s these facts are stated in the following proposition.

Proposition 1.3. *If s is a nonnegative real number, then $\sqrt{s}$ exists and is unique.*

You will prove the existence as Problem 1.6; we will just write a proof of the uniqueness assertion here. When we say a mathematical object is *unique* we mean that there is only one of them (or none at all). This proof is typical of uniqueness proofs. We assume there are two values (r and $\tilde{r}$ here) which both satisfy the requirements of the definition, and then based on that show that the two values must be the same.

Proof. To prove uniqueness, suppose both r and $\tilde{r}$ satisfy the definition of $\sqrt{s}$. In other words, $r \geq 0$, $\tilde{r} \geq 0$, and $r^2 = s = \tilde{r}^2$. We need to show that $r = \tilde{r}$. Since $r^2 = \tilde{r}^2$ we know that

$$0 = r^2 - \tilde{r}^2 = (r - \tilde{r})(r + \tilde{r}).$$

Thus either $r - \tilde{r} = 0$ or $r + \tilde{r} = 0$. If $r - \tilde{r} = 0$ then $r = \tilde{r}$. If $r + \tilde{r} = 0$ then $r = -\tilde{r}$. Since $\tilde{r} \geq 0$ this means that $r \leq 0$. But we also know that $r \geq 0$, so we conclude that $r = 0$, and that $\tilde{r} = -r = 0$ as well. Thus in either case $r = \tilde{r}$. □

[4] "nonegative" means ≥ 0; "positive" means > 0.

[5] We are considering only real square roots here. If we allow complex numbers for r then $r^2 = s$ can always be solved for r (even if s is complex). But it may be that neither of them is nonnegative, so that the defintion of "square root" above is no longer appropriate. Complex square roots are important in many subjects, differential equations for instance.

This is a typical proof of uniqueness; see the last row of the table on page 37. Observe that embedded in it is another argument by cases. The proof does not start by saying anything about cases. Only when we get to $(r - \tilde{r})(r + \tilde{r}) = 0$ does the argument split into the two cases: $r - \tilde{r} = 0$ or $r + \tilde{r} = 0$. Since those are the only ways $r^2 - \tilde{r}^2 = 0$, the cases are exhaustive.

Problem 1.6 In this problem you will use some of what you know about calculus to prove the existence of square roots.

a) Look up the statement of the Intermediate Value Theorem in a calculus book. Write it out carefully and turn it in. (Use an actual book, not just a web page. Include the author, title, edition number, publisher and date of publication for the specific book you used.) The statement of a theorem can be divided into *hypotheses* and *conclusions*. The hypotheses are the part of the theorem that must be true in order for the theorem to be applicable. (For example, in Propositon 1.3 the hypotheses are that s is a real number and that s is nonnegative.) The conclusions are things that the theorem guarantees to be true when it is applicable. (In Propositon 1.3 the conclusions are that $\sqrt{s}$ exists and that it is unique.) What are the hypotheses of the Intemediate Value Theorem? What are its conclusions? Does the book you consulted provide a proof of the Intemediate Value Theorem? (You don't need to copy the proof, just say whether or not it included one.) If it doesn't provide a proof does it say anything about how or where you can find a proof?

b) Suppose that $s > 0$ and consider the function $f(x) = x^2$ on the inteval $[a, b]$ where $a = 0$ and $b = s + 1/2$. By applying the Intermediate Value Theorem, prove that $\sqrt{s}$ does exist. (To "apply" the theorem, you need to make a particular choice of function $f(x)$, interval $[a, b]$ and any other quantities involved in the theorem. We have already said what we want to use for $f(x)$ and $[a, b]$; you need to specify choices for any other quantities in the theorem's statement. Next explain why your choices satisfy all the hypotheses of the theorem. Then explain why the conclusion of the Intermediate Value Theorem means that an r exists which satisfies the defintion of $r = \sqrt{s}$. Your are *not* writing a proof of the Intermediate Value Theorem. You are taking the validity of the Intemediate Value Theorem for granted, and explaining how it, considered for the specific $f(x)$ that you selected, tells us that that an r exists which satisfies the definition of $\sqrt{s}$.)

..... `IVTsqrt`

In general it is *not* true that the inequality $a^2 \leq b^2$ implies $a \leq b$. You can easily think of an example with b negative for which this doesn't work. So under what circumstances *can* we legitimately say that $a \leq b$ *is* a valid deduction from $a^2 \leq b^2$? If we rule out negative values by insisting that both a and b be nonnegative, that will do it. We state this property as a lemma. Note that because square roots are by definition nonnegative, the lemma does not need to say something like "assume a and b are nonnegative" — that is implicit in $a = \sqrt{x}$ and $b = \sqrt{y}$.

Lemma 1.4 (Square Root Lemma)**.** *If x and y are nonnegative real numbers with $x \leq y$, then*

$$\sqrt{x} \leq \sqrt{y}. \tag{1.1}$$

The reasoning for our proof will go like this. Let $a = \sqrt{x}$ and $b = \sqrt{y}$. We know these exist, are nonnegative, and that $a^2 = x \leq y = b^2$. Working from these facts we want to give reasons leading to the conclusion that $a \leq b$. Now for any pair of real numbers either $a \leq b$ or $b < a$ must be true. What our proof will do is show that $b < a$ is impossible under our hypotheses, leaving $a \leq b$ as the only possibility. In other words, instead of a sequence of logical steps leading to $a \leq b$, we provide reasoning which eliminates the only alternative.

Proof. Suppose $0 \leq x \leq y$. Let $a = \sqrt{x}$ and $b = \sqrt{y}$. We know both a and b are nonnegative, and by the hypothesis that $x \leq y$,

$$a^2 \leq b^2.$$

Suppose it were true that $b < a$. Since $b \geq 0$ it is valid to multiply both sides of $b < a$ by b to deduce that

$$b^2 \leq ab.$$

We can also multiply both sides of $b < a$ by a. Since $a > 0$ we deduce that

$$ab < a^2.$$

Putting these inequalities together it follows that

$$b^2 < a^2.$$

But that is impossible since it is contrary to our hypothesis that $a^2 \leq b^2$. Thus $b < a$ *cannot* be true under our hypotheses. We conclude that $a \leq b$. $\square$

This may seem like a strange sort of proof. We assumed the opposite of what we wanted to prove (i.e. that $b < a$), showed that it lead to a logically impossible situation, and called that a proof! As strange as it might seem, this *is* a perfectly logical and valid proof technique. **This is called proof by contradiction**; see the second row of the table on page 37. We will see some more examples in Section E.

Problem 1.7 Prove that $|x| = \sqrt{x^2}$, where x is any real number. (You are free to use any of the properties from Problem 1.1.)

..... altfor

Problem 1.8

a) Using Problem 1.6 as as a guide, prove that for *any* real number x there exists a real number y with $y^3 = x$. (Observe that there is no assumption that x or y are positive.)

b) Show that y is unique, in other words that if $y^3 = z^3 = x$ then $y = z$. (Although there are a number of ways to do this, one of the fastest is to use calculus. Can you think of a way to write $z^3 - y^3$ using caluclus, and use it to show that $y^3 = z^3$ implies that $y = z$?)

..... CubeRt

Problem 1.9 Prove (by contradiction) that if $1 \leq x$ then $\sqrt{x} \leq x$.

..... ri

A.3 Another Inequality

Here is another inequality, whose proof will be instructive for us.

Proposition 1.5. *For any real numbers x and y, the following inequality holds.*

$$\sqrt{x^2 + y^2} \leq |x| + |y|.$$

Any proof must exhibit a connection between things we already know and the thing we are trying to prove. Sometimes we can discover such a connection by starting with what we want to prove and then manipulate it into something we know. Let's try that with the inequality of the lemma.

$$\sqrt{x^2 + y^2} \overset{?}{\leq} |x| + |y| \tag{1.2}$$

$$x^2 + y^2 \overset{?}{\leq} (|x| + |y|)^2 \tag{1.3}$$

$$x^2 + y^2 \overset{?}{\leq} |x|^2 + 2|x|\,|y| + |y|^2$$

$$x^2 + y^2 \overset{?}{\leq} x^2 + 2|x|\,|y| + y^2$$

$$0 \overset{\text{Yes!}}{\leq} 2|x|\,|y|.$$

The question marks above the inequalities are to remind us that at each stage we are really asking ourselves, "Is this true?" The last line is certainly true, so we seem to have succeeded! So is this a proof? No — the logic is backwards. As presented above, we started from what we *don't* know (but want to prove) and deduced from it something we *do* know. To make this into a proof we need to reverse the logical direction, so that we start from what we do know (the last line) and argue that each line follows from the one below it. In other words we need to check that our reasoning still works when the order of steps is reversed. For most of the steps that just involves simple algebraic manipulations. But there is one place where where we should be careful: in deducing (1.2) from (1.3) we need to appeal the Square Root Lemma. Here then is what the proof looks like when written out in the correct sequence.

Proof. Suppose x and y are real numbers. We know from Problem 1.1 that $0 \leq |x|$ and $0 \leq |y|$, so that

$$0 \leq 2|x|\,|y|.$$

Adding $x^2 + y^2$ to both sides, and using the fact that $x^2 + y^2 = |x|^2 + |y|^2$ we find that

$$x^2 + y^2 \leq x^2 + 2|x|\,|y| + y^2 = |x|^2 + 2|x|\,|y| + |y|^2 = (|x| + |y|)^2.$$

By the Square Root Lemma, it follows from this that

$$\sqrt{x^2 + y^2} \leq \sqrt{(|x| + |y|)^2}.$$

Since $0 \leq |x| + |y|$ we know that $\sqrt{(|x| + |y|)^2} = |x| + |y|$, and so we conclude that

$$\sqrt{x^2 + y^2} \leq |x| + |y|.$$

□

The lesson of this proof is that we need to be sure our reasoning starts with our assumptions and leads to the desired conclusion. **When we discover a proof by working backwards from the conclusion, we will need to be sure the logic still works when presented in the other order.** (See the spoof of the next section for an example of an argument that works in one direction but not the other.) This is an example of a direct proof; see the top row of the table on page 37.

Problem 1.10 Suppose that a and b are both positive real numbers. Prove the following inequalities:

$$\frac{2ab}{a+b} \leq \sqrt{ab} \leq \frac{a+b}{2} \leq \sqrt{(a^2+b^2)/2}.$$

The first of these quantities is called the *harmonic mean* of a and b. The second is the *geometric mean*. The third is the *arithmetic mean*. The last is the *root mean square*. (You should prove this as three separate inequalities, $\frac{2ab}{a+b} \leq \sqrt{ab}$, $\sqrt{ab} \leq \frac{a+b}{2}$, and $\frac{a+b}{2} \leq \sqrt{(a^2+b^2)/2}$. Write a separate proof for each of them. Follow the way we found a proof for Proposition 1.5 above: work backwards to discover a connection and then reverse the order and carefully justify the steps to form a proof.)

..... `means`

B Some Spoofs

To read a proof you need to think about the steps, examining each critically to be sure it is logically sound, and convince yourself that the overall argument leaves nothing out. To help you practice that critical reading skill, we offer a couple "spoofs," arguments which seem valid on careless reading, but which come to a false conclusion (and so must have at least one logically flawed step).

Here is an old chestnut, probably the best known spoof, which "proves" that $1 = 2$.

Spoof! Let $a = b$. It follows that

$$\begin{aligned} ab &= a^2, \\ a^2 + ab &= a^2 + a^2, \\ a^2 + ab &= 2a^2, \\ a^2 + ab - 2ab &= 2a^2 - 2ab, \text{ and} \\ a^2 - ab &= 2a^2 - 2ab. \end{aligned}$$

This can be written as

$$1(a^2 - ab) = 2(a^2 - ab),$$

and canceling $a^2 - ab$ from both sides gives $1 = 2$. $\boxed{\textit{Ha Ha!}}$

See if you can spot the flaw in the reasoning before we talk about it in class. **This shows that is it important to examine each step of a sequence of deductions carefully.** All it takes is one false step to reach an erroneous conclusion.

This spoof also reinforces the point of page 8: the order of the logic in an argument matters. Here is the argument of the spoof but in the opposite order. Let $a = b \neq 0$. Then we check the truth of $1 = 2$ by the following argument.

$$\begin{aligned} 1 &= 2 \\ 1(a^2 - ab) &= 2(a^2 - ab) \\ a^2 - ab &= 2a^2 - 2ab \\ a^2 + ab - 2ab &= 2a^2 - 2ab \\ a^2 + ab &= 2a^2, \\ ab &= a^2, \\ a &= b. \end{aligned}$$

This time every line *really does* follow from the line before it! (The division by a in the last line is valid because $a \neq 0$.) So does this prove that $1 = 2$? Of course not. We can derive true statements from false ones, as we just did. That doesn't make the false statement we started from true. A valid proof must start from what we *do* know is true and lead to what we want to prove. To start from what we want to prove and deduce something true from it is not a proof. To make it into a proof you must reverse the sequence of the logic so that it leads from what we know to what we are trying to prove. In the case of the above, that reversal of logic is not possible.

Here are a couple more spoofs for you to diagnose.

Problem 1.11 Find the flawed steps in each of the following spoofs, and explain why the flawed steps are invalid.

a)

$$\begin{aligned} -2 &= -2 \\ 4 - 6 &= 1 - 3 \\ 4 - 6 + 9/4 &= 1 - 3 + 9/4 \\ (2 - 3/2)^2 &= (1 - 3/2)^2 \\ 2 - 3/2 &= 1 - 3/2 \\ 2 &= 1 \end{aligned}$$

Notice once again that if we started with the last line and worked upwards, each line *does* follow logically from the one below it! But this would certainly not be a valid proof that $2 = 1$. This again

makes our point that starting from what you want to prove and logically deriving something true from it *does not constitute a proof*!

b) Assume $a < b$. Then it follows that

$$\begin{aligned} a^2 &< b^2 \\ 0 &< b^2 - a^2 \\ 0 &< (b-a)(b+a). \end{aligned}$$

Since $b - a > 0$ it must be that $b + a > 0$ as well. (Otherwise $(b-a)(b+a) \leq 0$.) Thus

$$-a < b.$$

But now consider $a = -2$, $b = 1$. Since $a < b$ is true, it must be that $-a < b$, by the above reasoning. Therefore

$$2 < 1.$$

..... more

Problem 1.12 We claim that 1 is the largest positive integer. You laugh? Well here's the spoof (taken from [5]). Let n be the largest positive integer. Since $n \geq 1$, multiplying both sides by n implies that $n^2 \geq n$. But since n is the biggest positive integer, we must also have $n^2 \leq n$. It follows that $n^2 = n$. Dividing both sides by n implies that $n = 1$. Explain what's wrong!

..... largest1

C Proofs from Geometry

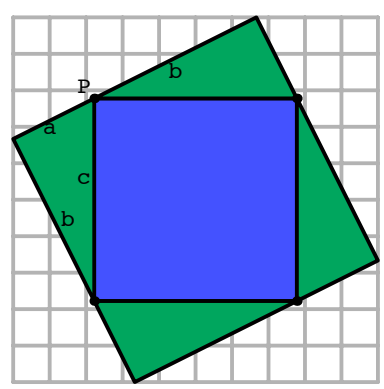
C.1 Pythagoras

The Pythagorean Theorem is perhaps the best known theorem in mathematics. You can find some of its history in Velijan [26] — there are roughly 400 known proofs! The one we consider below is attributed to Chou-pei Saun-ching around 250 B.C.

Theorem 1.6 (Pythagorean Theorem)**.** *Suppose a right triangle has hypotenuse of length c and sides of length a and b adjacent to the right angle. Then*

$$a^2 + b^2 = c^2.$$

Proof. Start with a square with sides of length c (blue in the figure) and draw four copies of the right triangle (green in the figure), each with its hypotenuse along one of the sides of the square, as illustrated. This produces a larger square with sides of length $a+b$, with the corners P of the original square touching each of the sides. We can calculate the area of the larger square in two ways. On one hand the area must be $(a+b)^2$ because it is a square with sides of length $a+b$. On the other hand it must be the sum of the areas of the smaller square and the four right triangles. It follows that



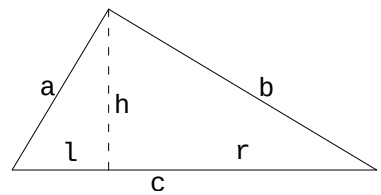
$$\begin{aligned} (a+b)^2 &= c^2 + 4(\frac{1}{2}ab) \\ a^2 + 2ab + b^2 &= c^2 + 2ab \\ a^2 + b^2 &= c^2. \end{aligned}$$

The last of these proves the theorem. □

This is rather different than the other proofs we have looked at so far. The proof starts by instructing the reader to draw a particular figure. Properties of that figure are then taken to be facts that can be used in the proof without further justification. The proof depends on the formulas for areas of squares and triangles. Especially, it depends on the fact that if a plane figure is partitioned into the (disjoint) union of several other figures, then the areas of the figures in the partition must add up to the area of the original figure. That is the basis of the equation on which the whole proof rests. The fundamental idea behind the proof is the idea of finding the Pythagorean relationship as a consequence of this equality of area calculations. This took a spark of creative insight on the part of Chou-pei Saun-ching. That is the hardest part of writing proofs for many students, finding an idea around which the proof can be built. There is no way someone can teach you that. **You have to be willing to try things and explore until you find a connection.** In time you will develop more insight and it will get easier.

Problem 1.13

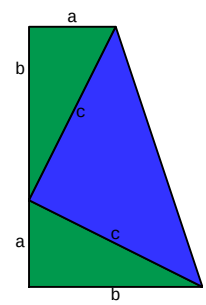
Consider a right triangle with sides a, b and hypoteneuse c. Rotate the triangle so the hypotenuse is its base, as illustrated. Draw a line perpendicular to the base up through the corner at the right angle. This forms two right triangles similar to the original. Calculate the lengths of the sides of these two trianlges (h, l, and r in the figure). When you subtitute these values in the equation $l + r = c$, you will have the outline of another proof of the Pythagorean Theorem. Write out this proof.



..... pyth2

Problem 1.14

There is a proof of the Pythagorean Theorem attributed to James A. Garfield (the 20th president of the United States); see [22]. It is based on calculating the area of the figure at right in two different ways. Write out a proof based on this figure.

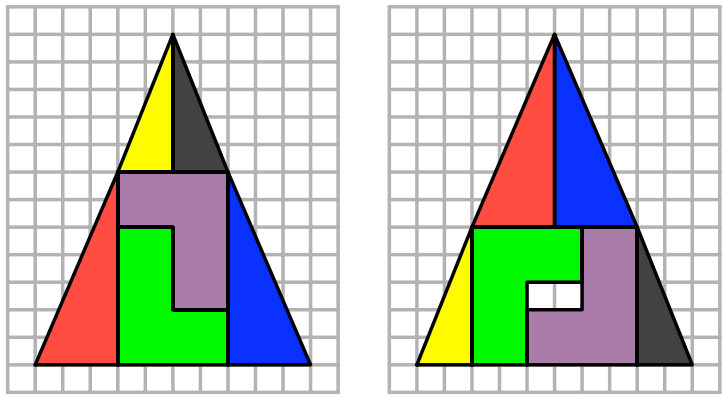


..... Garfield

C.2 The Curry Triangle

Our proof of the Pythagorean Theorem takes for granted properties of the figure which it describes. Look at the picture below, a famous example called the Curry Triangle[6]. The triangle on the left is decomposed into 6 smaller regions, which are then rearranged on the right to cover *only part* of the original triangle! This would seem to cast doubt on the principle of adding areas which our proof of Pythagorean Theorem is based. It is in fact a fraud; see Problem 1.15 below. But it serves to make the point that **your eyes can deceive you; what seems apparent to you visually need not be true.** There *are* such things as optical illusions. For that reason, proofs that are based on pictures are always a little suspect. In the case of the Pythagorean Theorem above the figure is honest. But the Curry Triangle tricks you into thinking you see something that isn't really there.

[6]See [3] for more on this dissection fallacy.



Problem 1.15 Explain the apparent inconsistency between the sums of the areas in the Curry Triangle. Based on your understanding of it, what features of figure for the Pythagorean Theorem would you want to check to be sure our proof of the Pythagorean Theorem is valid?

..... `Curry`

D From Calculus

This section considers some proofs which involve ideas from calculus. In particular we will consider some aspects of ordinary differential equations. As our primary example we consider the following equation:

$$y''(x) + y(x) = 0. \tag{1.4}$$

Some other equations will be considered in the problems.

For readers not familiar with differential equations, we provide a brief introduction. The equation (1.4) specifies that the function $y(x)$ and its second derivative $y''(x)$ be connected to each other in a specific way, namely that $y''(x) = -y(x)$ for all x. Some functions $y(x)$ will satisfy this, while most will not. To *solve* the differential equation means to find the function (or functions) $y(x)$ which have this property.

We can try guessing a solution to (1.4). For instance if we try $y(x) = x^2$, we have $y''(x) = 2$ so $y''(x) + y(x) = 2 + x^2$. That is *not* $= 0$ for all x, so $y(x) = x^2$ is *not* a solution. If keep trying things, we will eventually discover that $y(x) = \sin(x)$ works. To see this observe that $y'(x) = \cos(x)$ and $y''(x) = -\sin(x)$, so we have

$$y''(x) + y(x) = -\sin(x) + \sin(x) = 0 \text{ for all } x.$$

We have now guessed one solution: $y(x) = \sin(x)$. The next obvious guess is $y(x) = \cos(x)$, which we find also works. In fact we can put $\sin(x)$ and $\cos(x)$ together in various ways to get even more solutions, like

$$y(x) = \sin(x) + \cos(x) \text{ and } y(x) = 3\sin(x) - \sqrt{2}\cos(x), \ldots$$

In fact if you choose any two constants a and b the function

$$y(x) = a\cos(x) + b\sin(x) \tag{1.5}$$

is a solution. We now know that there are many solutions of (1.4). Could there be even more, perhaps involving something other than $\sin(x)$ or $\cos(x)$? What we are going to prove is that there are no others; every solution of (1.4) is of the form (1.5) for some choice of values for a and b.

Proposition 1.7. *The functions*

$$y(x) = a\cos(x) + b\sin(x),$$

where a and b are constants, are the only *twice differentiable functions which solve the differential equation*

$$y''(x) + y(x) = 0 \textit{ for all } x.$$

The next example illustrates how this proposition is used.

Example 1.2. Find a solution of (1.4) for which $y(\pi/4) = 3$ and $y'(\pi/4) = -1$. According to the proposition $y(x) = a\cos(x) + b\sin(x)$ are the only solutions; we just need to find values for a and b so that

$$\begin{aligned} a\cos(\pi/4) + b\sin(\pi/4) &= 3 \\ -a\sin(\pi/4) + b\cos(\pi/4) &= -1 \end{aligned}$$

A bit of algebra leads to $a = 2\sqrt{2}$ and $b = \sqrt{2}$, so that

$$y(x) = 2\sqrt{2}\cos(x) + \sqrt{2}\sin(x)$$

solves the problem.

We turn our attention now to proving the proposition. Here is the plan. Suppose $y(x)$ is a solution. The proposition implies that if we use $y(x)$ to determine the values $a = y(0)$ and $b = y'(0)$ and use these values define the function

$$\tilde{y}(x) = a\cos(x) + b\sin(x),$$

then $y(x) = \tilde{y}(x)$ for all x. Our approach to proving this consists of two parts.

1) First verify that $\tilde{y}(x)$ is a solution of (1.4), and in fact so is $\phi(x) = y(x) - \tilde{y}(x)$. Moreover, $\phi(0) = 0$ and $\phi'(0) = 0$.

2) Apply the following lemma to conclude that $\phi(x) = 0$ and therefore $y(x) = \tilde{y}(x)$ for all x.

Lemma 1.8. *Suppose $\phi(x)$ is a twice differentiable function of a real variable x satisfying $\phi''(x) + \phi(x) = 0$ for all x and $\phi(0) = \phi'(0) = 0$. Then $\phi(x) = 0$ for all x.*

Let's pause and put this in perspective. A fair bit of what we have said above is specific to the equation (1.4), especially the form (1.5) of solutions. We just arrived at (1.5) by guessing. Don't worry about whether you might be able to guess the solutions to differential equations other than (1.4); you are not going to be asked to do that for this course. (If you take a course in differential equations you will learn techniques for finding formulas like (1.5) for differential equations other than (1.4).) What is more important here is how we are approaching the task of proving that there are no solutions other than those of (1.5): we find the solution $\tilde{y}(x)$ from (1.5) so that $y(0) = \tilde{y}(0)$ and $y'(0) = \tilde{y}'(0)$, and then use Lemma 1.8 to show that $y(x) = \tilde{y}(x)$. What we have done is reduce all the situations covered by Proposition 1.7 to the special case considered in the lemma. Everything depends on our ability to prove the lemma.

Proof of the Lemma. Define the function $f(x)$ by $f(x) = (\phi(x))^2 + (\phi'(x))^2$. By hypothesis this is differentiable. Using the the chain rule and the differential equation we find

$$f'(x) = 2\phi(x)\phi'(x) + 2\phi'(x)\phi''(x) = 2\phi'(x)[\phi(x) + \phi''(x)] = 0 \text{ for all } x.$$

Thus $f(x)$ is a constant function. Therefore its value for any x is the same as its value at $x = 0$: $f(x) = f(0) = 0^2 + 0^2 = 0$. Thus $\phi(x)^2 + \phi'(x)^2 = 0$ for all x. Now since $\phi(x)^2 \geq 0$ and $\phi'(x)^2 \geq 0$, for their to sum to be 0 both terms must be 0 individually: $\phi(x)^2 = 0$ and $\phi'(x)^2 = 0$, both for all x. But $\phi(x)^2 = 0$ implies that $\phi(x) = 0$. Since this holds for all x, we have proven the lemma. □

This proof is based on introducing a new object, the function $f(x)$. To read the proof you don't need to know where the idea came from, you just have to be able to follow the reasoning leading to $f(x) = 0$. However to write such a proof in the first place you would need to come up with the idea, which requires some creativity. The proof we have given is specific to the equation (1.4); it won't work for different differential equations. **But sometimes you can modify the idea of a proof you have seen before to get it to work in new circumstances. That can be a way you come up with a new proof.** The problems will give you examples of different differential equations for which different sorts of modifications of the proof of the lemma above will work. We should also say that there are ways to prove a general version of the lemma that is not tied to one specific differential equation. Such a proof would be too much of a diversion for us; you will learn about that if you take a more advanced course on differential equations.

With the lemma established we can now write out the proof of our original proposition as we outlined it above.

Proof of Proposition 1.7. Suppose $y(x)$ solves $y''(x) + y(x) = 0$ for all x. Let

$$\tilde{y}(x) = a\cos(x) + b\sin(x), \text{ where } a = y(0), b = y'(0).$$

We first verify that $\tilde{y}(x)$ is a solution of (1.4).

$$\tilde{y}''(x) + \tilde{y}(x) = [-a\cos(x) - b\sin(x)] + [a\cos(x) + b\sin(x)] = 0, \text{ for all } x.$$

Next, define

$$\phi(x) = y(x) - \tilde{y}(x).$$

Observe that

$$\begin{aligned}\phi''(x) + \phi(x) &= y''(x) - \tilde{y}''(x) + y(x) - \tilde{y}(x)\\ &= [y''(x) + y(x)] - [\tilde{y}''(x) + \tilde{y}(x)]\\ &= 0.\end{aligned}$$

In addition,

$$\begin{aligned}\phi(0) &= y(0) - \tilde{y}(0) = a - [a\cos(0) + b\sin(0)] = 0\\ \phi'(0) &= y'(0) - \tilde{y}'(0) = b - [-a\sin(0) + b\cos(0)] = 0.\end{aligned}$$

Thus ϕ satisfies the hypotheses of Lemma 1.8. We conclude that $\phi(x) = 0$ for all x, which means that $y(x) = \tilde{y}(x)$. □

Problem 1.16 Here is a different differential equation:

$$y'(x) + y(x) = 0 \text{ for all } x. \tag{1.6}$$

a) Check that $y(x) = ce^{-x}$ is a solution, for any constant c.

b) Suppose $\phi'(x) + \phi(x) = 0$ with $\phi(0) = 0$. Write a proof that $\phi(x) = 0$ for all x by considering the function $f(x) = e^x\phi(x)$.

c) State and prove a version of Proposition 1.7 for the equation (1.6), using part b) in place of the lemma.

..... exp

Problem 1.17 If we change the equation in Lemma 1.8 to $\phi''(x) - \phi(x) = 0$, the conclusion is still true, but the same proof no longer works.

a) Write the statement of a revision of the lemma based this new differential equation.

b) Write a proof of your new lemma based on the following idea. Show that both of the following must equal to 0 for all x:

$$f(x) = e^{-x}(\phi(x) + \phi'(x)), \quad g(x) = e^x(\phi(x) - \phi'(x)).$$

Now view $f(x) = 0$ and $g(x) = 0$ as two equations for $\phi(x)$ and $\phi'(x)$. By solving those equations deduce that $\phi(x) = 0$ for all x.

..... htrig

Problem 1.18 Consider this differential equation:

$$\phi''(x) + \phi(x)^3 = 0; \quad \text{with } \phi(0) = \phi'(0) = 0.$$

Prove that $\phi(x) = 0$ for all x in a manner similar to Lemma 1.8, but using $f(x) = c\phi(x)^n + (\phi'(x))^2$ for appropriate choices of c and n.

..... nonlin

Problem 1.19 An interesting application of Proposition 1.7 is to prove trigonometric identities. For example, suppose α is a constant and consider the following two functions.

$$y_1(x) = \sin(\alpha)\cos(x) + \cos(\alpha)\sin(x), \text{ and } y_2(x) = \sin(\alpha + x).$$

Explain how Proposition 1.7 and Lemma 1.8 can be used to prove that $y_1(x) = y_2(x)$ for all x. What trigonometric identity does this prove? Write a proof of

$$\cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) - \sin(\beta)\sin(\alpha)$$

in a similar way.

..... TrigId

Problem 1.20 Prove that if $\phi'(x) = \phi(x)$ and $\phi(0) = 0$ then $\phi(x) \equiv 0$. [Hint: Consider $f(x) = e^{-x}\phi(x)$.] Use this to prove the identity

$$e^{a+b} = e^a e^b.$$

[Hint: Do something like Problem 1.19.]

..... expsol

E Irrational Numbers

The study of properties of the integers, especially related to prime numbers, is called *number theory*. Because we are very familiar with integer arithmetic, elementary number theory is a nice source of topics to illustrate and practice our proofs skills on. In this section we will look at two proofs which are often cited as favorite examples[7].

We begin with the definition of divisibility for integers. All the interesting structure of the integers is based on the fact that we can not always divide one integer by another. For instance, we cannot divide 3 by 2. Now you may say, "yes we can; we get a fraction $\frac{3}{2}$." What we mean by divisibility of integers is that the result of the division *is another integer*, i.e. we can perform the division without going outside the set of integers for the result. Here is the formal definition.

Definition. If m and n are integers, we say that m is *divisible* by n when there exists an integer k for which $m = kn$. This is denoted $n|m$.

We often use the alternate phrasings "n divides m" and "n is a divisor of m". Observe that the definition does not refer to an actual operation of division ($\div$). Instead it refers to the solvability of the equation $m = kn$ for for an integer k. We see that 1 divides all integers, and 0 divides only itself.

Definition. An integer $p > 1$ is called *prime* if p has no positive divisors other than 1 and itself. An integer $n > 1$ is called *composite* if it has a positive divisor other than 1 and itself.

As an example, 6 is composite since in addition to 1 and 6 it has positive divisors 2 and 3. On the other hand the only positive divisors of 7 are 7 and 1, so 7 is prime.

Observe that the terms "prime" and "composite" are only defined for integers larger than 1. In particular 1 is neither prime nor composite. (It's what we call a *unit*.) For integers $n > 1$ prime and composite are complementary properties; for n to be composite is the same as being not prime. If $n > 1$ is composite then can be factored (in a unique way) as a product of primes: $n = p_1 p_2 \cdots p_k$ where each p_i is a prime. This

[7] See for instance G. H. Hardy's famous little book [13].

fact is the Fundamental Theorem of Arithmetic — we will talk about it more in Chapter 4. In particular a composite number is always divisible by some prime number. We will take these things for granted as we prove the theorems of this section.

Theorem 1.9. *There are infinitely many prime numbers.*

How are we going to prove this? We can't hope to write out all infinitely many primes for someone to examine. Since "infinite" means "not finite," we will show that any *finite* list of prime numbers does *not* include all the primes. Here is the proof.

Proof. Consider a finite list of prime numbers,

$$p_1,\ p_2,\ \ldots,\ p_n. \tag{1.7}$$

We are going to show that there must be a prime number not in this list. To prove this by contradiction, assume that our list includes all the primes. Now consider the positive integer

$$q = 1 + p_1 p_2 \cdots p_n.$$

Observe that because all $p_k > 1$, it follows that q is bigger than all the primes p_k and so is not in our list (1.7) and is therefore not prime. So q is composite and must be divisible by some prime number. By our hypothesis every prime number appears in our list. So $q = mp_k$ for some integer m and one of the primes p_k from out list. But this means that we can write 1 as

$$1 = q - p_1 p_2 \cdots p_n = (m - p_1 \cdots p_{k-1} p_{k+1} \cdots p_n) p_k.$$

I.e. 1 is divisible by p_k. But this is certainly false, since 1 is not divisible by any positive integer except itself. Thus our assumption that all primes belong to our list must be false, since it leads to a contradiction with a known fact. There *is* therefore a prime number not in the list. Thus no finite collection of prime numbers contains all primes, which means that there are be infinitely many prime numbers. □

Some mathematicians try avoid proofs by contradiction. The next problem brings out one reason some people feel that way.

Problem 1.21 Let $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, $p_4 = 7$, $\ldots p_n$ be the first n prime numbers. Is it always true that $q = 1 + p_1 p_2 \cdots p_n$ is also a prime number? (For instance, for $n = 4$ we get $q = 1 + 2 \cdot 3 \cdot 5 \cdot 7 = 211$, which *is* a prime. The question is whether this is true for all n. You might want to experiment with other n. Mathematica will quickly check[8] whether a given number is prime or not for you.) Does your answer cast doubt on the proof above?

..... FP

Now we move beyond the integers to real numbers. We said above that 3 is not divisible by 2, because the division cannot be carried out within the integers. But we *can* do the division if we allow real numbers as the result, getting $\frac{3}{2}$. Such real numbers, those which are expressible as the ratio of two integers, are what we call rational numbers.

Definition. A real number r is called *rational* if it can be expressed as the ratio of two integers: $r = m/n$ where m and n are integers. A real number which is not rational is called *irrational*.

Are there any irrational numbers? Yes, and the next theorem identifies one. As we will see, the proof again uses the fact that every integer bigger than 1 is either prime or has a unique factorization into to a product of primes.

Theorem 1.10. $\sqrt{2}$ *is an irrational number.*

[8] The web site `http://www.onlineconversion.com/prime.htm` will do it also.

Proof. Suppose the contrary, namely that $\sqrt{2}$ is rational. Then $\sqrt{2} = \frac{m}{n}$, where m and n are positive integers. By canceling any common factors, we can assume that m and n share no common factors. Then $2 = \frac{m^2}{n^2}$, which implies

$$2n^2 = m^2.$$

Therefore 2 divides m^2 and so 2 must be one of the prime factors of m. But that means m^2 is divisible by 2^2, which implies that n^2 is divisible by 2. Therefore n is also divisible by 2. Thus both n and m are divisible by 2, contrary to the fact that they share no common factors. This contradiction proves the theorem. □

This, and the proof of Theorem 1.9, are examples of proof by contradiction. (See the chart on page 37.) It may seem strange that the proof is more concerned with what is not true than what is. But given our understanding of infinite as meaning "not finite," the proof has to establish that "there are a finite number of primes" is a false statement. So the proof temporarily assumes that the collection of all primes *is* finite, for the purpose of logically shooting that possibility down. That's the way a proof by contradiction works: temporarily assume that what you want to prove is *false* and explain how that leads to a logical impossibility. This shows that what you want to prove is *not* false, which means it is true!

Problem 1.22 If you think about it, the same proof works if we replace 2 by any prime number p. In fact it's even more general than that. As an example, revise the proof to show that $\sqrt{60}$ is not rational. So how far does it go? Can you find an easy way to describe those positive integers n for which $\sqrt{n}$ is irrational? (In other words you are being asked to complete this sentence, "The positive integers n for which $\sqrt{n}$ is irrational are precisely those n for which ..." You are not being asked to write a proof of your conclusion for this last part, but just to try to figure out what the correct answer is.)

..... `sqir`

The familiar constants π and e are also irrational, but the proofs are harder. We need a precise way to identify these numbers in order to prove anything about them. That was easy for $\sqrt{2}$ because it is the number $x > 0$ for which $x^2 = 2$. There are no such polynomial equations that π or e are solutions to, which is one reason the proofs are harder. There is however an infinite series[9] representation for e:

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

A short proof of the irrationality of e can be based on this. The irrationality of π is harder, but you can find proofs of both in Nagell [20], as well as Spivak [24].

F Induction

This section will consider some algebraic formulas whose proofs illustrate the technique of proof by induction. See the "For all positive integers n ..." row of the table on page 37. Like proof by contradiction, this can be hard to understand at first, so we offer several examples.

F.1 Simple Summation Formulas

You have probably encountered the following formulas for the sums of the first n integers, squares of integers, and cubes of integers.

Proposition 1.11. *For each positive integer n the following formulas hold:*

a) $\displaystyle\sum_{1}^{n} k = \frac{n(n+1)}{2}$,

[9] You may recall the power series representation of the exponential function: $\exp(x) = \sum_0^\infty \frac{x^n}{n!}$, converging for all x. The value $x = 1$ gives the formula for e.

b) $\displaystyle\sum_1^n k^2 = \frac{n(n+1)(2n+1)}{6}$,

a) $\displaystyle\sum_1^n k^3 = \frac{n^2(n+1)^2}{4}$.

These formulas are often used in calculus books for working out examples of the definite integral. Similar formulas exist for every integer power m: $\sum_{k=1}^n k^m =$ (formula).

What we want to do here is consider how such formulas can be proven. Let's focus on the first formula: $\sum_1^n k = \frac{n(n+1)}{2}$. We can check it for various values of n. For $n = 1$ it says that

$$1 = \frac{1(1+1)}{2},$$

which is certainly true. For $n = 2$ it says

$$1 + 2 = \frac{2(2+1)}{2},$$

which is correct since both sides are $= 3$. For $n = 3$,

$$1 + 2 + 3 = 6 \text{ and } \frac{3(3+1)}{2} = 6,$$

so it works in that case too. We could go on for some time this way. Maybe we could use computer software to help speed things up. But there are an infinite number of n values to check, and we will never be able to check them all individually.

What we want is a way to prove that the formula is true for all n at once, without needing to check each value of n individually. There are a couple ways to do that. We want to focus on one which illustrates the technique of mathematical induction. The idea is to link the truth of the formula for one value of n to the truth of the formula for the next value of n. Suppose, for instance, you knew that *if* the formula worked for $n = 50$, then it was *guaranteed* to work for $n = 51$ as well. Based on that, if at some point you confirmed that it *was* true for $n = 50$, then you could skip checking $n = 51$ separately. Now how could we connect the $n = 50$ case with the $n = 51$ case without actually knowing if either were true yet? Well, consider this:

$$\sum_1^{51} k = 1 + 2 + \cdots + 50 + 51 = \left(\sum_1^{50} k\right) + 51.$$

Now *if* it does turn out that $\sum_1^{50} k = \frac{50\cdot(50+1)}{2}$, then we will automatically know that

$$\sum_1^{51} k = \frac{50\cdot(50+1)}{2} + 51 = \frac{50\cdot 51 + 2\cdot 51}{2} = \frac{51\cdot 52}{2} = \frac{51\cdot(51+1)}{2},$$

which is what the formula for $n = 51$ claims! Now be careful to understand what we have said here. We have *not* yet established that the formula is true for either of $n = 50$ or $n = 51$. But what we *have* established is that *if at some point in the future* we are able to establish that the $n = 50$ formula is true, then *at the same moment we will know* that the $n = 51$ formula is also true, without needing to check it separately. We have *connected* the truth of the $n = 50$ case to the truth of the $n = 51$ case, but have not established the truth of either by itself. We have established a logical implication, specifically the statement

if the formula holds for $n = 50$ *then* the formula holds for $n = 51$.

We will talk more about logical implications in the next chapter.

What we did for $n = 50$ and $n = 51$ we can do for any pair of successive integers. We can show that the truth of the formula for any n will automatically imply the truth of the formula for $n + 1$. The reasoning is essentially the same as what we said above.

$$\sum_1^{n+1} k = \left(\sum_1^{n} k\right) + (n+1).$$

So *if* it turns out that $\sum_1^n k = \frac{n(n+1)}{2}$ is true for this particular value of n, *then it will also* be true that

$$\sum_1^{n+1} k = \frac{n(n+1)}{2} + (n+1) = \frac{n(n+1)+2(n+1)}{2} = \frac{(n+1)((n+1)+1)}{2}.$$

In other words once the formula is true for one n it will automatically be true for all the values of n that come after it. If it's true for $n = 1$ then it is true for $n = 2$, and being true for $n = 2$ automatically makes it true for $n = 3$, which in turn automatically makes it true for $n = 4$ and so on. This argument doesn't show the formula is true for any specific n, but it does show that if we can check it for $n = 1$ then all the other values of n are automatically true without needing to be checked separately. Well, we did check it for $n = 1$, so it must be true for all n. That's proof by induction: we check just the first case and link all the others cases to it logically. Here is a finished version of the proof we just described.

Proof (of first formula). We prove the formula by induction. For $n = 1$ we have

$$\sum_1^1 k = 1 \text{ and } \frac{1(1+1)}{2} = 1,$$

verifying the case of $n = 1$. Suppose the formula holds for n and consider $n + 1$. Then we have

$$\begin{aligned}\sum_1^{n+1} k &= \left(\sum_1^n k\right) + n + 1 \\ &= \frac{n(n+1)}{2} + n + 1 \\ &= \frac{n(n+1)+2(n+1)}{2} \\ &= \frac{(n+1)((n+1)+1)}{2},\end{aligned}$$

verifying the formula for $n + 1$. By induction this proves the formula for all positive integers n. □

You will prove the other two formulas from the proposition as a homework problem. Here is another example.

Proposition 1.12. *For every real number $a \neq 1$ and positive integer n,*

$$\sum_{k=0}^n a^k = \frac{a^{n+1}-1}{a-1}.$$

Proof. To prove the formula by induction, first consider the case of $n = 1$:

$$\sum_{k=0}^1 a^k = a^0 + a^1 = 1 + a = \frac{(a+1)(a-1)}{a-1} = \frac{a^2-1}{a-1}.$$

Next, suppose the formula is true for n and consider $n + 1$. We have

$$\begin{aligned}\sum_{k=0}^{n+1} a^k &= \left(\sum_{k=0}^n a^k\right) + a^{n+1} \\ &= \frac{a^{n+1}-1}{a-1} + a^{n+1} \\ &= \frac{a^{n+1}-1+a^{n+1}(a-1)}{a-1} \\ &= \frac{a^{(n+1)+1}-1}{a-1}.\end{aligned}$$

Thus the formula for $n + 1$ follows. By induction this completes the proof. □

Problem 1.23 Prove that for every positive integer n

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6} \text{ and } \sum_{k=1}^{n} k^3 = \frac{n^2(n+1)^2}{4}.$$

..... Ind1

Problem 1.24 Prove that for every positive integer n, $n^3 + 5n$ is divisible by 3. (One way to do this is by induction. For the induction step, assume $n^3 + 5n$ is divisible by 3, which means $n^3 + 5n = 3k$ for some integer k. To show $(n+1)^3 + 5(n+1)$ is divisible by 3 you need to show that it is possible to write $(n+1)^3 + 5(n+1) = 3(\cdots)$ where the $(\cdots)$ is some other integer. *Aviod* using division; i.e. don't write $\frac{n^3+5n}{3} = \cdots$. You are proving a property of the integers and so should try to give a proof that does *not* depend on operations like division that require a larger number system than the integers.)

..... div3

Problem 1.25 Prove that for every positive integer n, $n^3 + (n+1)^3 + (n+2)^3$ is divisible by 9.

..... div9

Problem 1.26 Use the fact that $\frac{d}{dx}x = 1$ and the usual product rule to prove (by induction) the usual power rule: $\frac{d}{dx}x^n = nx^{n-1}$, for all $n \geq 1$.

..... pr

F.2 Properties of Factorial

Definition. If $n \geq 0$ is an integer, we define $n!$ (called n *factorial*) by

$$0! = 1, \quad \text{and } n! = n \cdot (n-1) \cdots 3 \cdot 2 \cdot 1 \text{ for } n \geq 1.$$

You may wonder why $0!$ is defined to be 1. The reason is that many formulas involving factorials work for $n = 0$ only if we take $0! = 1$. One example is the formula $(n+1)! = (n+1) \cdot n!$ of the next problem.

Problem 1.27 The formula $(n+1)! = (n+1) \cdot n!$ for $n \geq 1$ is pretty obvious, but write a proof by induction for it. Explain why the convention $0! = 1$ is necessary if we want this formula to hold for $n = 0$. Is there a way to define $(-1)!$ so that the formula holds for $n = -1$?

..... facrecur

We want to look at some bounds on $n!$. An upper bound is pretty easy.

$$n! = n \cdot (n-1) \cdots 3 \cdot 2 \cdot 1 \leq n \cdot n \cdots n \cdot n \cdot 1 \leq n^{n-1}.$$

Here is a lower bound.

Proposition 1.13. *For every positive integer n,*

$$n^n e^{1-n} \leq n!.$$

The proof of this will be another example of an induction argument. But before we start the proof we need to recall some facts about the exponential function. First,

$$e^x \geq 1 \text{ for all } x \geq 0.$$

Secondly, $\frac{d}{dx}e^x = e^x$, so integrating both sides of the above inequality we obtain

$$e^x - 1 = \int_0^x e^t\,dt \geq \int_0^x 1\,dt = x.$$

Therefore, for all $x \geq 0$ we have

$$e^x \geq x + 1.$$

(Notice that $x+1$ is the tangent line to e^x at $x = 0$, so the above inequality simply says that the exponential function is never below that tangent line.) Now we can prove the proposition.

Proof. We prove the inequality by induction. First consider $n = 1$.

$$n! = 1 \text{ and } n^n e^{1-n} = 1e^0 = 1,$$

so the inequality holds for $n = 1$. Next, suppose the inequality holds for n and consider $n + 1$.

$$(n+1)! = (n+1) \cdot n! \geq (n+1)n^n e^{1-n}.$$

Since (using the inequalities above) $e^{1/n} \geq 1 + \frac{1}{n} = \frac{n+1}{n}$, we know that $e \geq \frac{(n+1)^n}{n^n}$. Using this we deduce that

$$(n+1)! \geq (n+1)n^n \frac{(n+1)^n}{n^n} e^{-n} = (n+1)^{n+1} e^{1-(n+1)},$$

which proves the inequality for the case of $n + 1$, and completes the induction proof. □

You may be wondering where all the inequalities in this proof came from. To read the proof you just need check that they they are all correct and lead to the desired conclusion. But to *write* the proof you have to think of them — how did we do that? The answer is that there was some scratch work that we did first, but did not record as part of the proof. To help you learn how to do this on your own, here is our scratch work. The hard part is to find a way to get from the induction hypothesis $n! \geq n^n e^{1-n}$ to the inequality for $n + 1$: $(n + 1)! \geq (n + 1)^{n+1} e^{1-(n+1)}$. Using $(n + 1)! = (n + 1)n!$ and the induction hypothesis gets us as far as $(n + 1)! \geq (n + 1)n^n e^{1-n}$, as in the fourth line of the proof. So what we hope to do is show that $(n + 1)n^n e^{1-n} \geq (n + 1)^{n+1} e^{1-(n+1)}$. If we simplify this, it reduces to

$$e^1 \geq \left(\frac{n+1}{n}\right)^n. \tag{1.8}$$

So that is the additional the fact we need to complete our proof. The inequality $e^x \geq 1 + x$ that we recalled before the proof was exactly what we would need justify $e^1 \geq \left(\frac{n+1}{n}\right)^n$. Thus we had an idea of how the proof should work (combine $(n + 1)! = (n + 1)n!$ and the induction hypothesis) and asked ourselves what other fact we would need to be able to reach the desired conclusion. Equation (1.8) was what we decided we would need, and so the proof included inequalities leading to (1.8). **This sort of strategizing is a natural part of the process of writing a proof.** Try to formulate an overall plan for your proof, how it will be organized, what the main steps or stages will be. And then focus on those parts individually and see if you can find a valid way to do each of them.

Additional Problems

Problem 1.28 Observe that the successive squares differ by the successive odd integers. This can be expressed succinctly as

$$n^2 = \sum_{i=1}^{n} (2i - 1).$$

Show that this is a consequence of the formula for $\sum_1^n k$ above. Now observe that there is a somewhat different pattern for the cubes:

$$\begin{aligned} 1^3 &= 1 \\ 2^3 &= 3+5 \\ 3^3 &= 7+9+11 \\ 4^3 &= 13+15+17+19 \\ &\vdots \end{aligned}$$

Explain why this pattern is equivalent to the formula

$$\sum_1^n k^3 = \sum_{i=1}^{\frac{n(n+1)}{2}} (2i-1),$$

and use the formulas that we gave previously for $\sum_1^n k^3$ and $\sum_1^n k$ to verify this new formula.

..... `sqodd`

Problem 1.29 Observe that

$$\begin{aligned} 1+2 &= 3 \\ 4+5+6 &= 7+8 \\ 9+10+11+12 &= 13+14+15 \\ 16+17+18+19+20 &= 21+22+23+24. \end{aligned}$$

Find a formula that expresses the pattern we see here, and prove it. You may use the summation formulas above if you would like.

..... `SumoSum`

Problem 1.30 The Gamma function, $\Gamma(x)$, is defined for $x > 0$ by

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\, dt.$$

a) Verify that $\Gamma(1) = \Gamma(2) = 1$.

b) Verify that $\Gamma(x) = (x-1)\Gamma(x-1)$ for all $x > 1$. (Hint: Integrate by parts.)

c) Explain why $n! = \Gamma(n+1)$ for all integers $n \geq 0$.

This is another example of a formula involving $n!$ for which $0! = 1$ is the only way to define $n!$ for $n = 0$ that is consistent with the formula.

..... `gamma`

Problem 1.31 You are probably familiar with *Pascal's Triangle*:

$$\begin{array}{ccccccccccccc} & & & & & & 1 & & & & & & \\ & & & & & 1 & & 1 & & & & & \\ & & & & 1 & & 2 & & 1 & & & & \\ & & & 1 & & 3 & & 3 & & 1 & & & \\ & & 1 & & 4 & & 6 & & 4 & & 1 & & \\ & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & \\ 1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 \\ & & & & & & \vdots & & & & & & \end{array}$$

Except for the 1s on the "sides" of the triangle, each number is obtained by adding the numbers to its left and right in the row above. The significance of these numbers is that each row gives the coefficients of the corresponding power of $(x+y)$.

$$\begin{aligned}
(x+y)^1 &= 1x + 1y \\
(x+y)^2 &= x^2 + 2xy + y^2 \\
(x+y)^3 &= x^3 + 3x^2y + 3xy^2 + y^3 \\
&\vdots \\
(x+y)^6 &= x^6 + 6x^5y + 15x^4y^2 + 20x^3y^3 + 15x^2y^4 + 6xy^5 + y^6 \\
&\vdots
\end{aligned} \tag{1.9}$$

The numbers in Pascal's Triangle are the *binomial coefficients* defined by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

pronounced "n choose k". The nth row in Pascal's Triangle,

$$1 \; n \; \cdots \; n \; 1,$$

are the values of $\binom{n}{k}$ for $k = 0, \ldots, n$:

$$\binom{n}{0} \; \binom{n}{1} \; \cdots \; \binom{n}{n-1} \; \binom{n}{n}.$$

For instance

$$\binom{4}{0} = 1, \; \binom{4}{1} = 4, \; \binom{4}{2} = 6, \; \binom{4}{3} = 4, \; \binom{4}{4} = 1.$$

The first (and last) entry of each row is always 1 because

$$\binom{n}{0} = \frac{n!}{0!\,n!} = 1.$$

(Notice that this would be incorrect without $0! = 1$.)

The fact that each entry in Pascal's Triangle is obtained by adding the two entries in the row above it is the following formula involving binomial coefficients: for integers k and n with $1 \le k \le n$,

$$\binom{n}{k-1} + \binom{n}{k} = \binom{n+1}{k}.$$

Prove the preceeding formula using the definition of binomial coefficients. (No induction is needed; it just boils down to manipulating formulas.)

The connection of Pascal's Triangle with $(x+y)^n$ and the binomial coefficients is expressed as this famous theorem. (You are not being asked to prove it; it's just here for your information.)

Theorem 1.14 (The Binomial Theorem)**.** *For every positive integer n*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k.$$

..... `BinomT`

Chapter 2

Mathematical Language and Some Basic Proof Structures

Now that we have looked at several examples, we want to consider the language and basic logical ideas used in proofs. A good example of the kind of complicated statement we need to work with is the definition of the limit of a function (from calculus). Suppose $f(x)$ is a function and a and L are two numbers. The statement $\lim_{x\to a} f(x) = L$ might be true or false, depending on the specifics. The precise definition of this statement is the following[1].

For every $\epsilon > 0$ there exists a $\delta > 0$ so that $|f(x) - L| < \epsilon$ holds for all $0 < |x - a| < \delta$.

In order to write proofs about limits we would need to work with the above definition in a precise way. Simply understanding what this definition says takes some care — what do all the phrases like "for every" and "there exists" mean? If you think you understand the statement above, then do you think you can write down exactly what it means for this statement to be false[2]? This is an example of the kind of thing we will talk about in the this chapter.

You will have noticed that proofs are typically written in a somewhat formal literary style. They are full of "therefore," "it follows that," and other words or phrases that you probably don't use much in everyday speech. In part this is a stylistic tradition, but it also serves an important purpose: the language is designed to tell the reader how they are supposed to logically connect the facts and deductions that make up the proof. Section A starts by summarizing different kinds of logical statements and how we work with them. The appendix on page 137 will provide a sort of brief glossary of some of the English words and phrases that are common in this kind of writing. Section B will discuss the use of variables in logical statements. Section C will describe some basic types of proofs. We have already encountered many of these in Chapter 1.

Be forewarned that you cannot learn to write proofs by learning a couple basic patterns and then following them over and over. There is no algorithm for writing a proof. Writing a proof is really distilling the result of a completed thinking process into a written form, so that it can be understood by someone else. You need *first* to think carefully to formulate your reasoning, and *after* that write it out clearly. Some general advice is collected in Section D.

A Basic Logical Propositions

Written mathematical discussion is made up of several different kinds of statements. The most basic are assertions of fact, statements that certain things are true (or false). Here are some examples.

- $\sqrt{2}$ is an irrational number.
- $\sin(x)$ is a differentiable function.

[1] You will study this more in advanced calculus.
[2] See Example 2.9 below.

- 49 is a prime number.

You might object to the third of these because it is false: 49 can be factored as $7 \cdot 7$. It is important to make a distinction between statements which are clear in their mathematical meaning (but false), and statements whose meaning itself is not clear (so that we can't even tell if they are true or false). Here are some examples of statements with no clear mathematical meaning (even though they refer to mathematical topics).

- π is an interesting number.
- Algebra is more fun than differential equations.
- The proof of the Four Color Problem[3] by computer is not a proper proof.

While you may have strong opinions about statements such as these, they are not statements that have a precise mathematical meaning. Such statements of opinion and preference are fine in talking about mathematics and editorials in scientific publications, but they don't belong in an actual proof.

A precise mathematical statement, with a definite true or false "value," is what we call a *proposition*, with a small "p". The first three bullets above are examples of propositions. The second three bullets are not. ("Proposition" with a capital "P" refers to a particular named statement[4], like Proposition 1.3 in Chapter 1. In this usage it is a proper name, and so is capitalized.) To say something is a proposition, does not mean that we *know* whether it is true or false, only that its mathematical meaning is crystal clear.

Propositions which are false *do* have legitimate roles in some proofs. For instance in our proof that $\sqrt{2}$ is irrational, we considered the consequences of the following proposition.

$$\sqrt{2} \text{ is rational.}$$

Based on that we deduced that the following proposition would also be true.

$$\text{There exist integers } n \text{ and } m \text{ for which } 2n^2 = m^2.$$

These propositions are both false, but we did state them (in a sort of hypothetical way) as part of our proof. We found that "$\sqrt{2}$ is rational" could not be true because its logical consequences were impossible. In other words the proof consisted of showing that "$\sqrt{2}$ is rational" is a false proposition. For purposes of making this argument we needed to be able to state this and other (false) propositions in order to examine them.

We also make statements which are not known to be true when we make a conjecture. A *conjecture* is a statement that we hope or expect to be true, but by calling it a conjecture we are acknowledging that we don't really know for sure (yet). See for instance the Twin Primes Conjecture on page 52 below.

A.1 Compounding Propositions: Not, And, Or

Starting with one or more propositions we can form new compound propositions which express relationships between the original component propositions. The simplest of these is the *logical negation* of a proposition. If S is a proposition its negation is "not S." The truth-value (true or false) of "not S" is exactly opposite that of S. For instance, if S is the proposition "49 is prime" then its negation can be expressed several ways:

- not (49 is prime)
- 49 is not prime
- 49 is composite

These all mean the same thing, but the last two are less awkward. The original proposition was false so the negation is true.

When we have two assertions of fact to make, we can combine them in one compound proposition joined by "and." Consider this example.

[3] The Four Color Problem is described on page 52.

[4] Whether a named result is a Theorem, Lemma, or Proposition is really a matter of the author's preference. Usually the main results of a book or paper are Theorems. Lemmas are results which are tools for proving the theorems. The significance of a Proposition is more ambiguous. For some a Proposition is a result of some interest of its own, not as important as a Theorem, but not just a stepping stone to something bigger either. In *this* book, Proposition is used for things that are interesting examples of things for us to prove, but not as important as Theorems.

$$\sqrt{2} \text{ is irrational and } 47 \text{ is prime.}$$

Such an "and"-joined compound proposition is true when both of the component propositions are true individually; otherwise the compound proposition is false. The above example is a true proposition. The following one is false, even though one piece of it would be true in isolation.

$$49 \text{ is prime and } 47 \text{ is prime.}$$

We can form a different compound proposition by joining component statements with "or." Consider for instance

$$49 \text{ is prime or } 47 \text{ is prime.}$$

Such an "or"-joined compound proposition is true when at least one of the component propositions is true. So the above example is a true proposition, as is the following.

$$\sqrt{2} \text{ is irrational or } 47 \text{ is prime.}$$

Here *both* component propositions are true[5].

The English language usually gives us several alternate ways to express the same logical statement. For instance, "*both* ... and ..." expresses the same thing as "... and ..." Including the word "both" doesn't change the meaning, it just adds emphasis. Whether to use it or not is simply a stylistic choice. Instead of "... or ..." we might say "*either* ... or ...."

Problem 2.1 Determine whether the following propositions are true or false.

a) 127 is prime and 128 is prime.

b) Either $\sqrt{27}$ is rational or $\sqrt{28}$ is rational.

c) $\int_0^5 (1-x)^5\,dx \geq -1$ or $\int_0^5 (1-x)^5\,dx \leq 1$.

d) $\int_0^5 (1+x)^5\,dx \geq 0$ and $\int_0^5 (1+x)^5\,dx \leq 10000$.

..... sprop

The true/false value of a compound proposition such as

$$\underbrace{\cdots}_{\text{S1}} \text{ and } \underbrace{\cdots}_{\text{S2}}$$

is determined by the true/false values of the two component propositions "...", which we will label as S1 and S2 to refer to them more easily. For instance a compound proposition of the form

(true proposition) and (true proposition)

is true. In other words, "S1 and S2" is true when both S1 and S2 are true individually. There are four possible combinations of true/false values for the pair S1, S2. We can summarize the meaning of "S1 and S2" by listing all the possible combinations in a *truth table*.

S1, S2	T,T	T,F	F,T	F,F
S1 and S2	T	F	F	F

The top row lists the four possible true/false combinations of S1, S2: "T,T" indicates the case in which S1 is true and S2 is true; "T,F" indicates the case in which S1=true and S2=false; and so forth. The second row indicates the true/false value of the compund proposition "S1 and S2" in each case. The table is a concise way to describe the meaning of the logical connective "and."

[5] Don't confuse this usage of "or" with what is called the *exclusive or*. That is what we would mean by a proposition like, "Either ... or ... *but not both*."

Problem 2.2 Determine whether each of the following propositions is true or false and explain why. (Assume that n refers to an unspecified positive integer, and x and α refer to unspecified real numbers. To say one of these statements is true means that it is true regardless of the specific such value assigned to n, x, or α.)

a) Either $n = 2$ or (n is prime and n is odd).

b) $\int_1^\infty x^\alpha \, dx < \infty$ if and only if $\alpha < 0$.

c) A real number x is rational if and only if both $(x+1)^2$ and $(x-1)^2$ are rational.

..... `compound`

A.2 Implications

The compound statement "S1 implies S2" is called an *implication*. Here is its truth table.

S1, S2	T,T	T,F	F,T	F,F
S1 implies S2	T	F	T	T

We call S1 the *antecedent* of the implication and S2 the *consequent*. We should emphasize that **the truth of "S1 implies S2" does *not* mean that S1 is necessarily true**. What it *does* mean is that the truth of S1 and S2 are linked in such a way that truthfulness of S1 guarantees truthfulness of S2. You might say "S1 implies S2" means that there is a logical pipeline[6] through which truthfulness will flow; truthfulness of S1 will automatically flow through the implication-pipeline and become truthfulness of S2. The implication is the pipeline itself, even if the pipeline is empty of any truthfulness. (Truthfulness is *not* guaranteed to flow in the other direction, from S2 to S1. It's only a one-way pipeline)

If it seems strange to consider "S1 implies S2" to be true when S1 is false, consider once again our proof that $\sqrt{2}$ is irrational. We can view the proof as establishing the proposition,

$$\sqrt{2} = \tfrac{n}{m} \text{ implies that both } n \text{ and } m \text{ are even.}$$

This *is* a true implication. But the truth of this implication does not mean that $\sqrt{2} = \frac{n}{m}$ is true. What we actually did in the proof was show that the above implication is true in a situation in which its consequent, "both n and m are even," is false. Inspecting the truth table we see that that can only happen when the antecedent is false. That is why we can conclude that "$\sqrt{2} = \frac{n}{m}$" must be false. The point is that a true implication with a false antecedent *was* important for our argument!

Implications are particularly important in mathematics. One equivalent wording is "If S1 then S2." In that form we have seen several examples in Chapter 1. For instance Proposition 1.3 is a statement of this form, using

$$\text{S1} = \text{"}s \text{ is a nonnegative real number" and S2} = \text{"}\sqrt{s} \text{ exists."}$$

(Lemma 1.4 is another example.) Notice in this example that both S1 and S2 are statements involving a variable s. In making the statement that S1 implies S2 we don't have just one specific value of s in mind. We are saying more than just that $\sqrt{2}$ exists, for instance — that would be the meaning of the implication if we understood s to refer only to the specific value 2. Rather we mean s to refer to an arbitrary but unspecified nonnegative value, so that the statement that S1 implies S2 is intended to be a statement about all possible $s \geq 0$ simultaneously. We might word it a little more explicitly as "whenever s is a real number with $s \geq 0$ then $\sqrt{s}$ does exist." Virtually all implications of any significance in mathematics involve variables. We will look at the use of variables more carefully in Section B. For the time being we can take the view that a statement "S1(x) implies S2(x)" means that assuming the truth of S1(x) is enough for us to deduce that S2(x) is also true, without needing to know the exact value of x.

[6] It is common to use the symbolic shorthand "S1⇒S2" for "S1 implies S2," which is very suggestive of this pipeline interpretation. There are symbolic shorthands for all the basic compound statements we are discussing. I do not encourage you to use them for the work you turn in, and so I am not even mentioning the others. If you want to use them for your own notes to yourself, that is up to you.

Lemma 1.4 is another example of an implication. The antecedent (S1) is "x and y are nonnegative real numbers with $x \leq y$" and the consequent (S2) is "$\sqrt{x} \leq \sqrt{y}$." Now look back at how we proved the implication of that lemma. You should see that it followed this pattern:

$$\text{Suppose } 0 \leq x \leq y. \cdots \text{ (argument } \textit{using } x \leq y) \cdots \text{ Therefore } \sqrt{x} \leq \sqrt{y}.$$

We *assumed* S1 and only gave an argument for S2 under the presumption that S1 is true. If we look at the truth table above we can see why; if S1 is false then the implication is true regardless of S2 — there is nothing to prove in that case! In general, **to prove "S1 implies S2" you *assume* S1 and then use that to prove S2**. Such a proof will typically take the form

Suppose S1 ... Therefore S2.

When we say "Suppose S1" in such a proof we are not making the assertion that S1 is in fact true; we are saying that for the sake of establishing "S1 implies S2" we are going to consider the case in which S1 is presumed to be true, so that we can write out the reasons which then lead to S2. The other case is that S1 is false, but there is nothing to prove in that case since "S1 implies S2" is always true when S1 is false. Thus we are focusing our attention on the only set of circumstances in which the implication might be false, namely the case in which S1 is true, and explaining why S2 is then necessarily true as a consequence. You might think of it as testing the logical pipeline by artificially applying truth to the input (antecedent) end to see if truth comes out the output (consequent) end.

There are many alternate English phrases which can be used to express "S1 implies S2":

- If S1 then S2.
- S2 if S1.
- S1 only if S2.
- S2 is necessary for S1.
- For S1 it is necessary that S2.
- S1 is sufficient for S2.
- S2 whenever S1.

Remember that the "if" or "is sufficient" go with the anticedent S1. The "only if" or "is necessary" go with the consequent S2.

Converse and Contrapositive

The *converse* of "S1 implies S2" is the implication "S2 implies S1." *An implication and its converse are not interchangeable.* The truth table below lists the implication and its converse in the second and third lines; we see that they have different truth values in some cases. The *contrapositive* of "S1 implies S2" is the implication "(not S2) implies (not S1)." The last line of the truth table below records its true/false values. (The next-to-last line is included just as an intermediate step in working out the contrapositive.) We see that the contrapositive has the same true/false value as the original implication in every case. This shows that an implication and its contrapositive are logically equivalent. To prove one is the same as proving the other.

S1, S2	T,T	T,F	F,T	F,F
original: S1 implies S2	T	F	T	T
converse: S2 implies S1	T	T	F	T
not S2, not S1	F,F	T,F	F,T	T,T
contrapositive: (not S2) implies (not S1)	T	F	T	T

You might also observe that when the original implication is false then its converse is true. But this connection does *not* hold when the implication involves variables; see Section B below. Since virtually all useful implications do involve variables, this is an *un*helpful observation!

Equivalence

A two-way logical pipeline is what we mean by saying "S1 is *equivalent* to S2;" the truth of either guarantees the truth of the other. In other words the equivalence means S1 and S2 have the same truth value; they are either both true or both false. Some other ways to express equivalence are

- S1 if and only if S2 (sometimes written "S1 iff S2");
- (S1 implies S2) and (S2 implies S1).
- S1 implies S2, and conversely.
- S1 is necessary and sufficient for S2.

The second of these is the most common way to approach the proof of an equivalence. Here is an example. Observe that the proof consists of proving two implications, each the converse of the other.

Proposition 2.1. *Suppose x and c are real numbers.*

$$|x| < c \text{ if and only if } (-c < x \text{ and } x < c).$$

Proof. To prove the "if" implication, suppose that $-c < x$ and $x < c$. We consider two cases.

Case 1: $x \geq 0$. In this case $|x| = x$ and, since $x < c$, it follows that $|x| < c$.

Case 2: $x < 0$. In this case $|x| = -x$. Since $-c < x$ it follows that $-x < c$ and therefore $|x| < c$.

To prove the "only if" implication, suppose $|x| < c$. By Lemma 1.1, $x \leq |x|$. Putting these inequlaities together we conclude that $x < c$. By Problem 1.1 part a), we know $|-x| = |x|$. Using Lemma 1.1 again, we find that $-x \leq |-x| = |x| < c$, so that $-x < c$. This implies that $-c < x$. Thus we have shown that both $-c < x$ and $x < c$ are true. □

The definition of a term establishes an equivalence between a word or phrase and its mathematical meaning. Look back at the definition we gave of rational numbers on page 16. The wording of our definition was that *if* $r = m/n$ for some integers m, n *then* we say r is rational. But what we really meant was

$$r \text{ is rational if } \textit{and only if } r = m/n \text{ for some integers } m, n.$$

A literal reading of the defintion as we stated in on page 16, using "if," would allow $\sqrt{2}$ to also be called a rational number. It does *not* say that a rational number has to be expressible as m/n; it says nothing in that case. We actually meant the definition to establish an equivalence, not just an implication. Unfortunately this inconsistent usage is quite common. You simply need to remember that **when reading a definition the author's "if" might (and probably does) really mean "if and only if."**

Problem 2.3 We said that "(P implies Q) and (Q implies P)" has the same meaning as "P iff Q." Verify that with a truth table, like we did on page 34. What about "(P or Q) implies (P and Q);" does that also have the same meaning? What about "(P or not Q) implies (Q or not P)"?

..... L1

A.3 Negations of Or and And

In the introduction to this chapter we talked about the importance of false statements in proofs. If we are writing a proof by contradiction for instance, we need to be able to state the negation of what we want to prove, in order to see that it leads to an impossible situation.

Sometimes our vocabulary allows us to conveniently state the negation of a statement. For instance, consider the proposition "459811 is prime." The negation of this could be expressed as "459811 is not prime," but we have a word meaning not prime: "composite." So the negation of our statement can be stated concisely as "459811 is composite."

For compound statements built up statements with multiple uses of the above connectives we need to know how to handle the connectives when we form the negation. One simple observation is that **"not" interchanges "and" with "or" when it is distributed over them**[7].

"Not (S1 and S2)" is equivalent to "(not S1) or (not S2)".

For example the negation of our (false) proposition "49 is prime and 47 is prime" is the (true) proposition

either 49 is not prime or 47 is not prime.

You will work out the negation of an implication in Problem 2.5.

Problem 2.4 Verify using truth tables that each of the following pairs of propositions are equivalent.

a) "Not (S1 and S2)" and "(not S1) or (not S2)."

b) "Not (S1 or S2)" and "(not S1) and (not S2)."

c) "S1 or S2" and "(not S1) implies S2."

..... EqCon

Problem 2.5 Find a proposition which is equivalent to "not (S1 implies S2)" but which does not use an implication in its statement.

..... negimp

Problem 2.6 Suppose ℓ is a positive integer, and γ is a real number. For each of the following, formulate a wording of its negation.

a) Both ℓ and $\ell + 2$ are prime numbers.

b) Either γ is positive or $-\gamma$ is positive.

c) Either $\gamma \leq 0$ or both γ and γ^3 are positive.

..... negation

B Variables and Quantifiers

Theorem 1.10 is unusual in that it is a statement about a single specific quantity. All the other statements we proved in Chapter 1 refered to many different situations at once: Lemma 1.1 makes a statement about *every real number*, the Triangle Inequality makes a statement about *all pairs of real numbers*, the Pythagorean Theorem refers to *all right triangles*, Proposition 1.13 asserts certain inequalities or equations *for all positive integers*, and Proposition 1.3 says that *for every* $s \geq 0$ there exists an $r \geq 0$ for which $r^2 = s$ is true. All of these are statements that involve variables which can take a range of different specific values. Statements involving variables are sometimes called *open* statements. In general whether they are true or false will depend on the actual values of the variables. For example consider the statement "n is a prime number." We can refer to this statement as $\mathrm{P}(n)$ to emphasize that it is a different statement for each value of the variable n: $\mathrm{P}(1)$ is false; $\mathrm{P}(2)$ is true, $\mathrm{P}(3)$ is true, $\mathrm{P}(4)$ is false .... The true/false value of $\mathrm{P}(n)$ depends on the value of the variable n.

We have already used variables in some of this chapter's examples. Now we want to look more carefully at the handling of variables.

[7] We have used parentheses to clarify. English can sometimes be ambiguous. For instance "Not [(I am a thief) or (I am a liar)]" is different than "(Not I am a thief) or (I am a liar)" although the word order is the same. The first means "I am neither a thief nor a liar." The second means "Either I am *not* a thief or I *am* a liar."

B.1 The Scope of Variables

When faced with a proposition involving variables we need to understand what values the variables are allowed to take. For instance in our statement $\mathrm{P}(n)$ above what values of n are we considering? Should $n = 0$ or $n = -13$ be considered? What about $n = 1/\sqrt{2}$? We will call this the *scope* of the variable n.

The intended scope of a variable can be determined in several ways. It might be implied by the overall context of a discussion. The "context" of a proposition refers to the assumptions and hypotheses which are inherent in the overall discussion of which the particular proposition is a part. Consider for instance the statement

the polynomial $x^2 + 2$ cannot be factored as a product of two first-degree polynomials.

This is true if we are only thinking of polynomials with real numbers as coefficients, but false if we are including polynomials with complex coefficients: $x^2 + 2 = (x + i\sqrt{2})(x - i\sqrt{2})$. Suppose we were reading a book about polynomials and at the beginning of the book the author had written, "Throughout this book we consider only polynomials with real coefficients." That sets the context for the rest of the book, so that if somewhere later in the book we find the above statement about $x^2 + 2$ then we know to interpret "polynomial" to mean only polynomials with real coefficients. The context might be set at the beginning of a book or paper, or it might be set just for an individual chapter. In Section A of Chapter 1 "number" referred to *real* number, and we said so consistently. However in Chapter 4 below we will be discussing the integers exclusively, so whenever we say "number" there it refers to numbers *which are integers*, even though we may not say so for every instance.

Sometimes the propositions being considered don't even make sense outside of a certain context so that we deduce the scope of variables based on that. For instance in Example 2.9, since we were talking about limits in the context of calculus, we presume that the variable x was meant to take only real values. The notion of prime number makes no sense for nonintegers. So when we talk about n being a prime number, as in Example 2.1, we presume that only integer values are intended for n. Example 2.16 below is an instance in which we need to determine the context in this way.

Even the choice of notation can indicate the intended scope of a variable. For example the letters i, j, k, ℓ, m, n are most often used to refer to integer-valued variables. Although this is not universal, when we encounter variables with these names it is often a clue that the author intends them to be integer-valued.

Often the scope of variables is explicitly stated in hypotheses that immediately precede the proposition itself. The proposition is only meant to be asserted within the limitations imposed by those hypotheses. In Problem 2.7 for instance the hypotheses that x and y are real and n is a positive integer are stated at the beginning, and are intended as the context for all the individual parts a)–h). In Example 2.8 we state the context explicitly. If the context there were different (x and y complex for instance) then in fact the original statement would be true!

B.2 Quantifiers

To make a valid proposition out of an open statement we must include *quantifiers* which explain how to combine the various statements which result from the different possible values of the variables (within their scope). The next example illustrates what we mean by a quantifier.

Example 2.1. Let $\mathrm{P}(n)$ stand for the statement, "n is a prime number." $\mathrm{P}(n)$ is true for some values of n (like $n = 2, 3, 5, 7$), but is false for others (like $n = 4, 6, 8, 9$). So $\mathrm{P}(n)$ by itself, without any information about what value of n to consider, is ambiguous. But the statement of the previous example,

$\mathrm{P}(n)$ *for all* positive integers n

is a valid proposition, a logically clear statement (but one which is false). The "for all positive integers n" is the *quantifier* that tells us what values of n we want to consider, and how to logically combine all the different statements $\mathrm{P}(n)$. The meaning of the "for all" statement is to make an infinite number of statements all at once, as if we joined them all with "and":

$$\mathrm{P}(1) \text{ and } \mathrm{P}(2) \text{ and } \mathrm{P}(3) \text{ and } \ldots$$

As always, there are many alternate ways to express the same proposition.

- For every positive integer n, $\mathrm{P}(n)$ holds.
- $\mathrm{P}(n)$ holds for every positive integer n.
- $\mathrm{P}(n)$ is true whenever n is a positive integer.
- If n is a positive integer, then $\mathrm{P}(n)$ holds.

Notice that the last two of these are worded as implications (the next-to-last using the "whenever" phrasing of an implication). When working with the negation or converse of an implication involving variables it can be important to understand that statement is really a formulation of a statement involving a quantifier. We will discuss that more below; see page 35.

Example 2.2. Continuing with Example 2.1, we can combine the statements $\mathrm{P}(n)$ in a different way by saying,

for some positive integer n, n is a prime number.

The quantifier here is the "for some." This means the same thing as combining all the statements $\mathrm{P}(n)$ with "or":

$$\mathrm{P}(1) \text{ or } \mathrm{P}(2) \text{ or } \mathrm{P}(3) \text{ or } \ldots$$

This is a true statement. Its meaning is simply that there is at least one positive integer which is prime.

Some alternate expressions of the "for some" quantifier are

- There exists a n so that ...
- There exists a n such that ...
- There exists a n for which ...
- ... for some n.

Open statements involving two or more variables are more complicated, because different quantifiers may be applied to the different variables. Consider the following statement, which we will denote by $\mathrm{R}(r, s)$.

$$r \geq 0 \text{ and } r^2 = s.$$

This is an open statement involving two (real) variables. For some r, s combinations $\mathrm{R}(r, s)$ is true while for others it is false. For $\mathrm{R}(r, s)$ to be true is the definition of $r = \sqrt{s}$. Now consider the following proposition formed using two different quantifiers for r and s.

For every $s \geq 0$ there exists an r so that $\mathrm{R}(r, s)$.

This is just our statement of Proposition 1.3 that $\sqrt{s}$ exists for all $s \geq 0$. It is important to understand that **the order of the quantifiers matters** — we get a *false* statement if we reverse their order:

there exists an r so that for every $s \geq 0$ the statement $\mathrm{R}(r, s)$ holds.

This proposition claims that there is a "universal square root," a single value r which is $r = \sqrt{s}$ for all $s \geq 0$ simultaneously — certainly a false claim.

Example 2.3.

- The proposition "there exists an integer n so that for all integers m we have $m < n$" is false. It claims there is one integer that is strictly larger than all integers (and so strictly larger than itself).
- The proposition "for all integers m there exists an integer n for which $m < n$" is true. It says that given an integer m you can always find another integer n which is larger ($n = m + 1$ for instance).

Problem 2.7 Assume that x and y are real numbers and n is a positive integer. Determine whether each of the following is true or false, and explain your conclusion.

a) For all x there exists a y with $x + y = 0$.

b) There exists y so that for all x, $x + y = 0$.

c) For all x there exists y for which $xy = 0$.

d) There exists y so that for all x, $xy = 0$.

e) For all x there exists y such that $xy = 1$.

f) There exists y so that for all x, $xy = 1$.

g) For all n, n is even or n is odd.

h) Either n is even for all n, or n is odd for all n.

(From [9].)

..... tf

B.3 Subtleties

Negation, Converse, Contrapositive within the Scope

When we form a contrapositive, converse, or negation of a proposition **we do *not* change the context or hypotheses within which the implication is made. We form the new statements *within the same context and hypotheses* as the original.** In particular the intended scope of a variable in a contrapositive, converse, or negation remains the same as for for the original.

So for instance in Problem 2.7 if we were writing the negation of part h) we would *not* change the hypothesis that n is limited to positive integers. The statement of the negation would still be understood subject to the hypothesis that n is a positive integer.

Example 2.4. Suppose the variable x is allowed to take real values. Consider the proposition "$x^2 \geq 0$ for all x." Its negation is the following:

$$x^2 < 0 \text{ for some } x.$$

We read this within the *same context*: the variable x is still understood to be limited to real values. With this understanding the original statement is true. The negation would *not* consider non-real values for x because the scope of x is still that specified by the context.

Example 2.5. Suppose n is an integer. Consider the following statement.

$$\text{If } n \text{ is even then } n^2 \text{ is even.}$$

According to the hypothesis the scope of n is limited to integer values. However we could rephrase this by bundling the hypothesis into the antecedent of the implication to get the following, which we will consider within the context of all *real* values for x:

$$\text{If } x \text{ is an even integer then } x^2 \text{ is an even integer.}$$

For the first version, the contrapositive would be

$$\text{If } n^2 \text{ is odd then } n \text{ is odd,}$$

still understood within the context of n being a positive integer. For the second version the contrapositive would be

$$\text{If } x^2 \text{ is not an even integer even, then either } x \text{ is an odd integer or } n \text{ is not an integer,}$$

considered with the scope of x being all real numbers. This statement of the contrapositive is rather different then before because the context is different. There is no presumption that x is an integer, so that "not even" could include fractions and irrational numbers. In both cases the contrapositive is equivalent to the original, but the contexts are different in the two cases and that affects the statements of the contrapositive.

Just as for converse and contrapositive, negations are formed within the prescribed context; we don't move the hypotheses into the antecedent of an implication and then negate that.

Subtleties of the "For All" Quantifier

Some statements with variables are lacking the words "for all" even though they are part of the intended meaning. So you sometimes need to read between the lines to recognize the presence of a "for all" quantifier. Consider for instance the statement of Problem 1.10.

Suppose that a and b are positive real numbers. Prove that

$$\frac{2ab}{a+b} \leq \cdots \leq \sqrt{(a^2+b^2)/2}.$$

There are variables a and b here, but you don't see the words "for all" in the statement of the proposition, even though the intent is to claim that the inequalities of the problem are true for all possible choices of a and b. Thus the intent was to say

$$\frac{2ab}{a+b} \leq \cdots \leq \sqrt{(a^2+b^2)/2} \ \textit{ for all positive } a \text{ and } b.$$

We encountered other statements like this on page 27, in Proposition 2.1, and in Examples 2.2 and 2.6. In all of these the intent was to make statements that applied to *all* possible values of the variables. We would say there is an unwritten but *implicit* "for all" quantifier. In general, **when you encounter an open statement with an unspecified variable, the intended meaning is most likely that the statement is meant *for all* values of the variables** within their scope. On the other hand, the quantifier "there exists" or "for some" is virtually always stated explicitly.

A "for all" quantifier is sometimes phrased as an implication; we indicated this as the last of the alternate phrasings on page 31. If an implication has variables in the anticedent then a "for all" statement is usually intended. Consider Proposition 1.3 for example.

If s is a nonnegative real number, then $\sqrt{s}$ exists and is unique.

Although worded as an implication, clearly the intent was

For all nonnegative real numbers s, $\sqrt{s}$ exists and is unique.

The meaning of either phrasing is the same: knowing that $s \geq 0$ (regardless of the specific value of s) is all that we need in order to be sure that $\sqrt{s}$ exists. A different statement but with the same ultimate content would be this:

For all real numbers s, if $s \geq 0$ then $\sqrt{s}$ exists and is unique.

Here we are making a statement in the form of an implication within the enlarged scope of *all* real values for s. The correctness of this statement is depends on our understanding that "S1 implies S2" is true when S1 is false. We consider "if $s \geq 0$ then $\sqrt{s}$ exists" to be a true statement *for any* real number s, in particular it is true for negative values of s, in which case the anticedent ("if $s \geq 0$") is false.

Example 2.6. Here are some examples of implications involving a variable x (whose scope we take to be all real numbers), along with their converses.

- The implication "If $x > 3$ then $x^2 > 9$" is true, but it's converse, "If $x^2 > 9$ then $x > 3$," is false. (Just think about $x = -4$.) However it's contrapositive, "If $x^2 \leq 9$ then $x \leq 3$," is true.
- The implication "For $\sin(x) = 0$ it is sufficient that x be an integral multiple of π" is true. So is its converse, which we can express as "For $\sin(x) = 0$ it is necessary that x be an integral multiple of π."
- "If $x < 3$ then $x^2 > 9$" is false and so is its converse: "If $x^2 > 9$ then $x < 3$."

In the first example the converse "If $x^2 > 9$ then $x > 3$" would be true for some specific values of x but it is not true for all such values. That is why the converse is a false statement.

The third of these examples illustrates that when variables are involved it is possible that both the original and converse can be false, contrary to what we saw in the table from page 28 when no variables were involved. If we replace x by a value for which the first implication is false (say $x = 2$) then the converse

would take the form of false implies true (e.g. "if $4 > 9$ then $2 < 3$") which is true. But with the variable the converse means that "$x^2 > 9$ implies $x < 3$" is true *for every* value x, not just those that make the original implication false but also those (like $x = -4$) which make it true. In general, for implications involving variables (which are really "for all" statements) you can't say the converse is true or false based on whether the original is true or false.

Problem 2.8 Determine whether each of the implications below is true or false. For each, write a statement of its converse and its contrapositive and explain whether they are true or false. (Notice that there are variables here: n refers to an arbitrary integer; x and a refer to arbitrary real numbers.)

a) An integer n is divisible by 6 whenever $4n$ is divisible by 6.

b) For $f(x) = \begin{cases} \frac{\sin(x)}{x} & \text{for } x \neq 0 \\ a & \text{for } x = 0 \end{cases}$ to be a continuous function it is necessary that $a = 0$.

c) If x^2 is irrational then x is irrational.

..... mprops

Vacuous Statements

What about a "for all" statement in which there are no possible values for the variable? Consider this statement.

For all real numbers x with $x^2 < 0$ the equation $x = x + 1$ holds.

We are considering the open statement "$x = x + 1$" in which the variable x is allowed to take any real value for which $x^2 < 0$. But there are *no* real values of x for which $x^2 < 0$. So is the above statement true or false? Essentially it is an empty statement; there are no values of x about which the statement has anything to say. Such a statement is considered true; we often say that it is *vacuous* or *vacuously true*. Reworded as an implication, the statment becomes "if $x^2 < 0$ then $x = x + 1$." In that form it becomes another instance of an implication "S1 implies S2" being true when S1 is false. While the above example may seem silly, vacuous statements do sometimes come up in legitimate proofs. For instance you might formulate a proof by cases and find that one of the cases can never hold. The "if (description of vacuous case) then (conclusion)" would still be considered a true implication as part of the proof.

B.4 Negating Quantified Propositions

For compound statements, we need to know how to handle logical connectives when we form their negations. On page 30 we observed that negating a statement exchanges "and" with "or." The negation of an implication "S1 implies S2" can be expressed "S1 and not S2." This can be confirmed with a truth table.

S1, S2	T,T	T,F	F,T	F,F
S1 implies S2	T	F	T	T
S1, not S2	T,F	T,T	F,F	F,T
S1 and not S2	F	T	F	F

Since the first and last lines have the opposite true/false value in all cases, we do have a correct formulation of the negation.

We can think of a statement with a "for all" quantifier as a multiple "and" statement. Since negating an "and" statement produces an "or" statement, we expect the negation of a "for all" statement to be a "for some" statement. Similarly, the negation of a "for some" statement produces a "for all" statement. In brief, **negation interchanges the two types of quantifiers**.

"Not (for all x, P(x))" is equivalent to "for some x, not P(x).
"Not (for some x, P(x))" is equivalent to "for all x, not P(x).

Example 2.7. The negation of

$$\text{there exists a real number } y \text{ with } y^2 = -1$$

is

$$y^2 \neq -1 \text{ for all real numbers } y.$$

The negation of

$$\text{all prime numbers } n \text{ are odd,}$$

is

$$\text{there exists a prime number } n \text{ which is even.}$$

Its simple enough when there is just one quantifier. When there are two or more, form the negation by working from the outside in, as illustrated in the following examples.

Example 2.8. Suppose x and y denote real numbers. The following statement is false.

$$\text{For every } x \text{ there exists a } y \text{ with } x = y^2.$$

Its negation is therefore true. We can develop a natural wording of the negation by first putting a "not" in front of the whole thing, and then progressively moving the "not" from the outside in, producing the following sequence of equivalent statements.

Not (for every x there exists a y with $x = y^2$).
For some x, not (there exists a y with $x = y^2$).
For some x (for all y, not $x = y^2$).
There exists an x so that $x \neq y^2$ for all y.

To see that this negated statement is true, just consider $x = -1$; in the context of the real numbers all y have $y^2 \geq 0$, so $x < y^2$ for all y.

Example 2.9. In the same manner we can now work out the negation of the definition of $\lim_{x\to a} f(x) = L$, which we contemplated at the beginning of the chapter.

Not (for every $\epsilon > 0$ (there exists a $\delta > 0$ so that (for all $0 < |x - a| < \delta$, ($|f(x) - L| < \epsilon$)))).
For some $\epsilon > 0$ not (there exists a $\delta > 0$ so that (for all $0 < |x - a| < \delta$, ($|f(x) - L| < \epsilon$))).
For some $\epsilon > 0$ (for all $\delta > 0$ not (for all $0 < |x - a| < \delta$, ($|f(x) - L| < \epsilon$))).
For some $\epsilon > 0$ (for all $\delta > 0$ (there exists an $0 < |x - a| < \delta$ so that not ($|f(x) - L| < \epsilon$))).
For some $\epsilon > 0$ (for all $\delta > 0$ (there exists an $0 < |x - a| < \delta$ so that $|f(x) - L| \geq \epsilon$)).
There exists an $\epsilon > 0$ so that for every $\delta > 0$ there is an x with $0 < |x - a| < \delta$ and $|f(x) - L| \geq \epsilon$.

The next two example are to make the following point. **When forming negations, it is important to recognize quantifiers which are implicit or phrased with "if."**

Example 2.10. Consider the statement

$$\text{if } n \text{ is prime then } n \text{ is odd.}$$

This is false, because of $n = 2$. So its negation should be true. But if we are careless and ignore the implied "for all" quantifier (taking the negation of "S1 implies S2" to be "S1 and not S2") we might write the following for the negation

$$n \text{ is prime and } n \text{ is even.}$$

Seeing the variable n and references to "prime" and "odd" we would probably presume n was intended to be a positive integer. But since no other qualifications are stated for n we might read an implicit "for all" into this and interpret is as

all positive integers n are both prime and even.

That is not the correct negation of our original statement. If we are careful to include the implicit qualifier, our orginal statement is

for all prime numbers n, n is odd.

Now we get the correct negation by properly handling the quantifier:

for some prime number n, n is even.

Example 2.11. Suppose x is a positive real number. Consider the statement

$$\ln(x) = 0.$$

Although there is a value of $x > 0$ for which this is true, it is not true for all positive x. Since we would understand the statment with an implicit "for all $x > 0$" quantifier, we would properly understand the statement to be false. We want to form the negation, which should be true. But if we neglected the implicit quantifier we might say the negation is simply

$$\ln(x) \neq 0.$$

That would be the correct negation if there were only a single x under consideration. But with the implicit quantifier in the original statement is the correct negation is

There exists a positive x for which $\ln(x) \neq 0$.

This is indeed true; just consider $x = 2$.

Problem 2.9 Consider the proposition

for all real numbers a, if $f(a) = 0$ then $a > 0$.

Write a version of the negation of this proposition.

..... `negpos`

Problem 2.10 Suppose A is a set of real numbers. We say x is an *upper bound* of A when

For all a in A, $a \leq x$.

We say α is the *supremum* of A when α is an upper bound of A and for every upper bound x of A we have $\alpha \leq x$.

a) Write a statement of what it means to say x is not an upper bound of A.

b) Write a statement of what it means to say α is not the supremum of A.

..... `supremum`

C Some Basic Types of Proofs

There is no general prescription for how to write a proof. However, for various standard types of propositions there are some natural ways to arrange a proof. The chart below lists some of the most common. The following subsections will elaborate and cite examples.

Statement	Typical Methods
S (elementary proposition)	• Direct: Assume hypotheses. $\cdots$ Therefore S. • Contradiction
S1 and S2	• Prove S1. Prove S2. (Two separate proofs.)
S1 or S2	• Cases, each leading to one of the possible conclusions. • Prove that (not S1) implies S2.
S1 implies S2	• Assume S1. $\cdots$ Therefore S2
S1 iff S2	• S1 implies S2. S2 implies S1. (Prove two implications.)
Multiple equivalences	• A connecting cycle of implications.
For all x satisfying P(x), S(x)	• Cases. • Assume P(x). $\cdots$ Therefore S(x). (Treat as an implication.)
For all positive integers n, S(n)	• Induction.
There exists x for which ...	• Either construct x or connect to some other existence result.
Uniqueness of x such that ...	• Assume x and $\tilde{x}$ both qualify. Show $x = \tilde{x}$.

C.1 Elementary Propositions and "And" Statements

By a direct proof, we mean an argument which simply starts with the hypotheses and proceeds through a sequence of logical deductions arriving in the end at the desired conclusion. Our proof of Proposition 1.5 falls into this category. Even though the proposition was worded as a "for all" statement, we proved it directly: we assumed the hypotheses (x and y are real numbers) and gave a sequence of deductions (taking advantage of other facts about absolute value and square roots that were already known to us) which led to the conclusion. There were no cases, no contradiction, no induction, no auspicious new objects, just a direct line of reasoning leading from the hypotheses to the conclusion.

Proving a statement "S1 and S2" is nothing other than proving S1 and then proving S2. It is not really any different than preparing proofs of S1 and S2 separately and then writing them down one after the other.

C.2 "Or" Statements

Things can be a little trickier when proving "S1 or S2." One approach is to prove the equivalent statement, "not S1 implies S2;" in other words suppose that S1 is false and show that S2 must be true. Here is an example of such a proof.

Example 2.12. Suppose a and b are real numbers. Either the system of two linear equations

$$\begin{aligned} x + y + z &= a \\ x + 2y + 3z &= b \end{aligned}$$

has has no solution (x, y, z) or it has infinitely many solutions.

Proof. Suppose the system has a solution (x_0, y_0, z_0). Consider

$$x = x_0 + t,\ y = y_0 - 2t,\ z = z_0 + t.$$

Observe that for every real number t

$$\begin{aligned} x + y + z &= (x_0 + t) + (y_0 - 2t) + (z_0 + t) = x_0 + y_0 + z_0 + (t - 2t + t) = x_0 + y_0 + z_0 = a \\ x + 2y + 3z &= (x_0 + t) + 2(y_0 - 2t) + 3(z_0 + t) = x_0 + 2y_0 + 3z_0 + (t - 4t + 3t) = x_0 + 2y_0 + 3z_0 = b, \end{aligned}$$

so that (x, y, z) is a solution. Thus there are infinitely many solutions, one for each t. □

(If you are wondering where the $+t, -2t, +t$ in this proof came from, remember your linear algebra: $(1, -2, 1)$ is a solution of the associated homogeneous linear system.)

Another approach is to consider an exhaustive set of cases in some of which you can conclude S1 while in others you can conclude S2. We didn't see any examples of this in Chapter 1, but here is one with a four-way "or."

Example 2.13. Suppose $q(x) = ax^2 + bx + c$ is a quadratic polynomial with real coefficients a, b, c. The number of solutions to the equation $q(x) = 0$ is 0, 1, 2, or infinite.

Proof. We consider six cases.

Case 1: $a = b = c = 0$. In this case $q(x) = 0$ for all x, so there are infinitely many solutions.

Case 2: $a = b = 0$ and $c \neq 0$. In this case $q(x) = c \neq 0$ for all x, so there are no solutions.

Case 3: $a = 0$ and $b \neq 0$. In this case $q(x) = bx + c$. There is exactly one solution, $x = -c/b$.

In the remaining cases $a \neq 0$. Assuming that, we can rewrite $q(x)$ by completing the square.

$$\begin{aligned} q(x) &= a\left[x^2 + \frac{b}{a}x + \frac{c}{a}\right] \\ &= a\left[x^2 + 2\frac{b}{2a}x + \frac{b^2}{4a^2} - \frac{b^2}{4a^2} + \frac{4ac}{4a^2}\right] \\ &= a\left[\left(x + \frac{b}{2a}\right)^2 - \frac{b^2 - 4ac}{4a^2}\right] \end{aligned} \tag{2.1}$$

We now proceed to the remaining cases using this formulation.

Case 4: $a \neq 0$ and $b^2 - 4ac < 0$. Then $\frac{b^2-4ac}{4a^2} < 0$. For x to be a solution to $q(x) = 0$ would require $(x + \frac{b}{2a})^2 = \frac{b^2-4ac}{4a^2} < 0$, which is impossible. Thus there are no solutions in this case.

Case 5: $a \neq 0$ and $b^2 - 4ac = 0$. Now (2.1) implies that $q(x) = 0$ is equivalent to $(x + \frac{b}{2a})^2 = 0$, for which $x = -\frac{b}{2a}$ is the one and only solution.

Case 6: $a \neq 0$ and $b^2 - 4ac > 0$. In this case $\frac{b^2-4ac}{4a^2} > 0$, so $q(x) = 0$ is equivalent to $(x + \frac{b}{2a})^2 = \frac{b^2-4ac}{4a^2}$, for which there are exactly two solutions: $x = -\frac{b}{2a} \pm \sqrt{\frac{b^2-4ac}{4a^2}}$.

Since the cases exhaust all the possibilities, and in each case there were either 0, 1, 2, or infinitely many solutions, we have proven the result. □

C.3 Implications and "For all . . ." Statements

We explained above that to prove "S1 implies S2" we typically *assume* S1 and use that to deduce S2. Since an implication is equivalent to its contrapositive, another way to prove "S1 implies S2" is to suppose that S2 is false and show S1 is false.

We have observed that when variables are involved a "for all . . ." statement can often be expressed as an implication. It is often proven as an implication as well. We essentially prove it directly as one statement about the quantities referred to by the variables, and for which what we know about them is not their precise values, but just the properties given in the hypotheses. For instance consider the proof of the Pythagorean Theorem in Chapter 1. The theorem is really making an assertion *for all* a, b, c which occur as the sides of a right triangle. But the proof just gives one argument which applies simultaneously to all right triangles, so it is worded as if there is a single right triangle under consideration. Proposition 1.5 is another example. Sometimes you need to break into separate cases for the proof. Lemma 1.1 and the Triangle Inequality are simple examples of that kind of proof.

To *disprove* a "for all x, S(x)" statement we need to show that there exists an x (within the intended context) for which S(x) is false. In other words you need to exhibit an x which is a *counterexample*. Even if the "for all" is expressed as an implication, that is still what you need to do to disprove it.

Example 2.14. Consider the following proposition.

If A and B are 2×2 matrices then $AB = BA$.

The implication is really a "for all" assertion. It is the same as saying "for all 2×2 matrices A and B, the equation $AB = BA$ is true." To disprove this all we need to do is exhibit one *counterexample*, i.e. we want to produce a specific pair of matrices for which $AB \neq BA$. For instance let

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

We find that

$$AB = \begin{bmatrix} -2 & 1 \\ -4 & 3 \end{bmatrix} \neq \begin{bmatrix} 3 & 4 \\ -1 & -2 \end{bmatrix} = BA.$$

This one counterexample is enough to disprove the statement above.

Problem 2.11 Assume that a, b, and c are positive integers. Prove that if a divides b or c then a divides bc.

..... div

Problem 2.12 Prove that if m and n are odd integers, then $n^2 - m^2$ is divisible by 8.

..... D8

Problem 2.13 Assume a and b are real numbers. Prove or disprove the following statement.

$$\text{If } 4ab < (a+b)^2 \text{ then } a < b.$$

..... ineQ

C.4 Equivalence

To prove "S1 if and only if S2" we simply prove both implications, "S1 implies S2" and "S2 implies S1." Proposition 2.1 was an example. Here is another.

Proposition 2.2. *Suppose $a < b < c$ are three consecutive positive integers. There is a right triangle with sides of length a, b, and c if and only if $a = 3$.*

Proof. Suppose $a = 3$. Then $b = 4$ and $c = 5$. The triangle with sides $3, 4, 5$ *is* a right triangle. This proves the "if" assertion.

Suppose $a < b < c$ are consecutive positive integers which form the sides of a right triangle. Then $b = a + 1$, $c = a + 2$, and $a^2 + b^2 = c^2$ and by the Pythagorean Theorem. Therefore we have

$$\begin{aligned} a^2 + (a+1)^2 &= (a+2)^2 \\ a^2 + a^2 + 2a + 1 &= a^2 + 4a + 4 \\ a^2 - 2a - 3 &= 0 \\ (a-3)(a+1) &= 0. \end{aligned}$$

Therefore either $a = -1$ or $a = 3$. Since a is a positive integer, $a \neq -1$. Thus $a = 3$. □

Multiple Equivalences

Sometimes we encounter statements that three or more propositions are equivalent. Here is an example[8].

Proposition 2.3. *Supoose $f(x)$ is a twice differentiable function of a real variable x. The following are equivalent.*

[8] Functions satisfying a) are called "concave up" in typical calculus books. Functions satisfying c) are called "convex" in more advanced analysis texts.

a) $f''(x) \geq 0$ *for all* x.

b) For any three values $a < b < c$ *the following inequality holds:*

$$\frac{f(b) - f(a)}{b - a} \leq \frac{f(c) - f(b)}{c - b}.$$

c) For any two values s *and* t *and any* $0 \leq \lambda \leq 1$ *the following inequality holds:*

$$f(\lambda s + (1 - \lambda)t) \leq \lambda f(s) + (1 - \lambda)f(t).$$

The phrase "the following are equivalent" means that any two of the parts are equivalent to each other. With 3 parts that comes to 3 individual equivalences. We could prove the proposition by proving each equivalence as a pair of implications, which would require 6 different implications. But a proof does not need to include arguments for all 6 implications. If we prove that a) implies b) and that b) implies c) then the truth of a) implies c) follows immediately and does not need to be proven separately. The most common proof of a multiple equivalence like this is to establish a collection of the implications which comprise a cycle through all of the parts. In this case we will prove that a) implies b), b) implies c), and that c) implies a). Nothing more is needed.

Proof. Assume a). That means that f' is an increasing function. By the Mean Value Theorem of calculus there exists a value x between a and b and a value y between b and c for which

$$f'(x) = \frac{f(b) - f(a)}{b - a} \text{ and } f'(y) = \frac{f(c) - f(b)}{c - b}.$$

Because of the assumed ordering $a < b < c$ this implies that $x \leq y$. Since f' is increasing we know that $f'(x) \leq f'(y)$. Therefore

$$\frac{f(b) - f(a)}{b - a} \leq \frac{f(c) - f(b)}{c - b}.$$

Thus a) implies b).

Now assume b). Consider any s, t and $0 \leq \lambda \leq 1$. The cases of $s = t$, $\lambda = 0$ or $\lambda = 1$ are all trivial. Without loss of generality we can assume that $s < t$ and $0 < \lambda < 1$. Let $a = s$, $c = t$ and $b = \lambda s + (1 - \lambda)t$. Observe that $a < b < c$. So from b) we know that

$$\frac{f(b) - f(a)}{b - a} \leq \frac{f(c) - f(b)}{c - b}.$$

This can be rearranged as

$$\begin{aligned}
\left(\frac{1}{b - a} + \frac{1}{c - b}\right) f(b) &\leq \frac{1}{b - a} f(a) + \frac{1}{c - b} f(c) \\
\frac{c - a}{(b - a)(c - b)} f(b) &\leq \frac{1}{b - a} f(a) + \frac{1}{c - b} f(c) \\
f(b) &\leq \frac{c - b}{c - a} f(a) + \frac{b - a}{c - a} f(c).
\end{aligned}$$

Bust since $\lambda = \frac{c-b}{c-a}$ and $1 - \lambda = \frac{b-a}{c-a}$ we see that

$$f(\lambda s + (1 - \lambda)t) \leq \lambda f(s) + (1 - \lambda)f(t).$$

Thus b) implies c).

Finally assume c) and consider any x. For any $h > 0$ take $s = x - h$, $t = s + h$ and $\lambda = \frac{1}{2}$. Then $\lambda s + (1 - \lambda)t = x$ and so we know that

$$f(x) \leq \frac{1}{2} f(x + h) + \frac{1}{2} f(x - h). \tag{2.2}$$

L'Hopital's Rule tells us that

$$\begin{aligned}\lim_{h\to 0^+}\left[\frac{\frac{1}{2}f(x+h)-f(x)+\frac{1}{2}f(x-h)}{\frac{1}{2}h^2}\right] &= \lim_{h\to 0^+}\left[\frac{\frac{1}{2}f'(x+h)-\frac{1}{2}f'(x-h)}{h}\right]\\ &= \lim_{h\to 0^+}\left[\frac{f'(x+h)-f'(x)}{2h}+\frac{f'(x)-f'(x-h)}{2h}\right]\\ &= f''(x).\end{aligned}$$

From (2.2) we know that the limit on the left is nonnegative. Therefore $f''(x) \geq 0$, proving that c) implies a). □

Other examples of multiple equivalences occur in the fourth bullet on page 82 and in Theorem 6.15 below. You will be asked for a proof of a multiple equivalence in Problem 3.15 in the next chapter.

C.5 Existence and Uniqueness

To prove a statement of the form "there exists x such that $\mathrm{P}(x)$" often requires particular creativity. Instead of showing that something is always true or always false, you must somehow find a way to pick or identify *one special* x out of all the possibilities for which $\mathrm{P}(x)$ holds. There is no general pattern to follow for this kind of proof; it depends very much on the specifics of *what* you are proving the existence of.

Problem 1.6 in Chapter 1 is an example of one approach to this kind of problem. The proof you wrote for that worked by connecting the existence of what we want ($x \geq 0$ with $x^2 = s$) to a more general-purpose existence theorem (the Intermediate Value Theorem).

Sometimes you can prove the existence of the thing you want by just finding a way to construct or produce it, apart from using some fancy theorem. Here is an example that can be done that way.

Example 2.15. Prove that for every real number y there exists a real number x for which $y = \frac{1}{2}(e^x - e^{-x})$.

We could do this by appealing to the Intermediate Value Theorem again, but this time we can find the x we want by some algebra (this will be our scratch work) and then just verify that our expression for x works as the proof. Here is the scratch work: given y we are trying to solve for x.

$$\begin{aligned}y &= \frac{1}{2}(e^x - e^{-x})\\ 0 &= e^x - 2y - e^{-x}\\ 0 &= e^{2x} - 2ye^x - 1\\ 0 &= (e^x)^2 - 2ye^x - 1.\end{aligned}$$

So, using the quadratic formula,

$$e^x = \frac{2y \pm \sqrt{4y^2+4}}{2} = y \pm \sqrt{y^2+1}.$$

Since $e^x > 0$, we must use the positive of the two possible roots, and so taking the logarithm we get

$$x = \ln(y + \sqrt{y^2+1}).$$

That's the formula we wanted. Now we prove that it is indeed the x we were looking for. Observe that the proof itself gives the reader no clue about how we came up with the formula for x.

Proof. Given a real number y, consider $x = \ln(y + \sqrt{y^2+1})$. We will confirm that this x works as claimed.

$$\begin{aligned} e^x - e^{-x} &= y + \sqrt{y^2+1} - \frac{1}{y+\sqrt{y^2+1}} \\ &= \frac{(y+\sqrt{y^2+1})^2 - 1}{y+\sqrt{y^2+1}} \\ &= \frac{y^2 + 2y\sqrt{y^2+1} + y^2 + 1 - 1}{y+\sqrt{y^2+1}} \\ &= \frac{2y(y+\sqrt{y^2+1})}{y+\sqrt{y^2+1}} \\ &= 2y \end{aligned}$$

Therefore we find that $\frac{1}{2}(e^x - e^{-x}) = y$, proving the existence of x as claimed. □

A statement of uniqueness says that there is not more than one object satisfying some description or collection of properties. Proposition 1.3 is a good example. We assumed that two numbers both satisfied the definition of $\sqrt{s}$ and then argued from there that those two numbers must be the same. That is the typical pattern of a uniqueness proof.

C.6 Contradiction

A proof by contradiction is essentially the proof that the conclusion can not be false. We simply take all the hypotheses and add to them the assumption that the conclusion is false, and show that from all those together we can deduce something impossible[9]. We saw several examples in Chapter 1: Lemma 1.4, Theorems 1.9 and 1.10.

If you are proving a "for all" statement by contradiction, the negation is a "there exists" statement. To show that the negation is false you must show that in fact no values of the variables do what the negation would say.

Example 2.16. Prove that for all $x > 0$, the inequality $\frac{x}{x+1} < \frac{x+1}{x+2}$ holds.

Before proceeding, what is the intended scope for x here? The statement does not state it outright so we are left to surmise it on our own. Since $x > 0$ only makes sense for real numbers, the context must not extend beyond that. If the variables i, j, n, k, ℓ, or m had been used we might suspect that just integers were intended. However the choice of x suggests that real numbers are intended, so that is how we will understand the statement.

Proof. Suppose the statement is false. Then *there exists* an $x > 0$ with $\frac{x}{x+1} \geq \frac{x+1}{x+2}$. Since both $x+1$ and $x+2$ are positive we can multiply both sides of this inequality by them, obtaining $x(x+2) \geq (x+1)^2$. Multiplying this out we find that

$$x^2 + 2x \geq x^2 + 2x + 1, \quad \text{and therefore } 0 \geq 1.$$

This is clearly false, showing that the existence of such a x is not possible. This completes the proof by contradiction. □

Problem 2.14 Prove that if a, b, c are positive integers with $a^2 + b^2 = c^2$ then either a or b is even.
..... sqeven

Problem 2.15 Prove (by contradiction) that $\sqrt{3} - \sqrt{2}$ is irrational.

[9] The technique is called "reductio ad absurdum," which is Latin for "reduction to the absurd". In brief, the negation of the conclusion leads to an absurdity. I like to say that by assuming the conclusion is false we can produce a mathematical train wreck.

..... r3r2

Problem 2.16 Prove that there does not exist a smallest positive real number.

..... minR

Problem 2.17 Suppose x, y, z are positive real numbers. Prove that $x > z$ and $y^2 = xz$ together imply $x > y > z$. (You can do this by contradiction; negating $x > y > z$ results in an "or" statement, so the proof by contradiction will be in two cases.)

..... ineq3

Problem 2.18 Prove that if f and g are differentiable functions with $x = f(x)g(x)$, then either $f(0) \neq 0$ or $g(0) \neq 0$. Hint: take a derivative. (Adapted from [24].)

..... noxprd

C.7 Induction

Induction is a special technique for proving a proposition of the form

$$P(n) \text{ for all positive integers } n.$$

A proof by induction of $P(n)$ for all $n = 1, 2, 3, \cdots$ consists of two parts.

1. A proof that $P(1)$ is true. (The *base case.*)
2. A proof that $P(n)$ implies $P(n+1)$, for all positive integers n. (The *induction step.*)

The base case is often simple, because it boils down to just checking one case (and usually an easy case at that). The proof of the induction step typically takes the form

$$\text{suppose P}(n). \ \ldots \text{therefore P}(n+1).$$

Newcomers to inductive proofs sometimes object to this, because it looks like we are assuming what we are trying to prove. But as we discussed in Chapter 1, the "suppose P(n)" doesn't mean that we are claiming it is true, but showing how the truth of P($n+1$) will be a consequence of the truth of P(n). We saw several examples of such induction proofs in the last chapter: Propositions 1.11, 1.12, and 1.13.

This proof technique only works for statements in which n ranges over integer values. It doesn't work for statements of the form "...for all real numbers x." Its validity boils down to a basic property of the integers, as we will discuss in Chapter 4. But there are a couple variants of it, which we will describe in this section.

Example 2.17. First, here is a spoof by induction to show that all Canadians were born on the same day. For each positive integer n let $S(n)$ be the statement that every group of n Canadians all were born on the same day. We will show by induction that $S(n)$ is true for each n.

In any group that consists of just one Canadian, everybody in the group has the same age, because after all there is only one person! Thus $S(1)$ is true.

Next suppose $S(n)$ is true and consider any group G of $n+1$ Canadians. Suppose p and q are two of the Canadians in G. We want to show that p and q have the same birthday. Let F be the group of all the people in G excluding p. Then F has n people in it, so since $S(n)$ is true all the people in F have a common birthday. Likewise, let H be the group of people in G excluding q. Then H has n people and so by $S(n)$ all the people in H share a common birthday. Now let r be someone in G other than p or q. Both r and q are in F, so they have the same birthday. Similarly, both r and p are in H so they have the same birthday. But then p and q must have the same birthday, since they both have the same birthday as r. Thus any two people in G have the same birthday, so everyone in G has the same birthday. Since G was any group of $n+1$

Canadians, it follows that $S(n+1)$ is true. This completes the proof by induction that $S(n)$ is true for all n. To finish, let n be the current population of Canada. Since we know $S(n)$ is true, we see that all Canadians have the same birthday.

Can you spot the flaw here[10]? Hint: find the first n for which $\mathrm{S}(n+1)$ is false — the argument for the induction step must be invalid for that n.

Problem 2.19 Let $\mathrm{P}(n)$ be the proposition that if i, j are positive integers with $\max(i, j) = n$, then $i = j$. We spoof by induction that $\mathrm{P}(n)$ is true for all n. First consider $n = 1$. If i and j are positive integers with $\max(i, j) = 1$, then $i = 1 = j$. This shows that $\mathrm{P}(1)$ is true. For the induction step suppose $\mathrm{P}(n)$ is true and i, j are positive integers with $\max(i, j) = n + 1$. Then $\max(i - 1, j - 1) = n$, so by the induction hypothesis we know $i-1 = j-1$. Adding 1 to both sides we deduce that $i = j$. This shows that $\mathrm{P}(n)$ implies $\mathrm{P}(n+1)$, and completes the proof by induction. As a corollary, if i and j are any two positive integers, take $n = \max(i, j)$. Then since $\mathrm{P}(n)$ is true we conclude that $i = j$. Thus all positive integers are equal!

What's wrong here? (Taken from [5].)

..... `fi`

Generalized Induction

The first variant of induction is sometimes called *generalized induction*. It's only difference from ordinary induction is that it proves something for all $n \geq n_0$ where n_0 can be any integer starting value. Ordinary induction is just the case of $n_0 = 1$. For a generalized induction proof you prove the base case of $n = n_0$, and then (assuming $n \geq n_0$) prove the usual induction step. Here is an example.

Example 2.18. Prove that $(1 + \frac{1}{n})^n < n$ for all integers $n \geq 3$. Observe that this inequality is false for $n = 1, 2$. The fact that the proof starts with $n = 3$ (instead of $n = 1$) is what makes this *generalized* induction.

Proof. We use induction on n. First consider $n = 3$.

$$\left(1 + \frac{1}{n}\right)^n = \left(\frac{4}{3}\right)^3 = \frac{64}{27} = 2 + \frac{10}{27} < 3,$$

verifying the assertion for $n = 3$.

Next, suppose that $(1 + \frac{1}{n})^n < n$ and $n \geq 3$. We want to prove that $(1 + \frac{1}{n+1})^{n+1} < n + 1$. We begin by writing

$$\left(1 + \frac{1}{n+1}\right)^{n+1} = \left(1 + \frac{1}{n+1}\right)^n \left(1 + \frac{1}{n+1}\right) = \left(1 + \frac{1}{n+1}\right)^n \frac{n+2}{n+1}.$$

Since $\frac{1}{n+1} < \frac{1}{n}$, it follows that $(1 + \frac{1}{n+1})^n < (1 + \frac{1}{n})^n$. Therefore

$$\begin{aligned}\left(1 + \frac{1}{n+1}\right)^{n+1} &< \left(1 + \frac{1}{n}\right)^n \frac{n+2}{n+1} \\ &< n\frac{n+2}{n+1}, \quad \text{using the induction hypothesis.}\end{aligned}$$

To finish we claim that $n\frac{n+2}{n+1} < n + 1$. We can deduce this from the following inequalities.

$$\begin{aligned} 0 &< 1 \\ n^2 + 2n &< n^2 + 2n + 1 \\ n(n+2) &< (n+1)^2 \\ n\frac{n+2}{n+1} &< n + 1, \end{aligned}$$

[10] If not then on what day do you think all those Canadians were born. April 1 perhaps?

the last line following from the preceding since $n+1$ is positive. So putting the pieces together we have shown that

$$\left(1+\frac{1}{n+1}\right)^{n+1} < n\frac{n+2}{n+1} < n+1.$$

This proves the assertion for $n+1$ and completes the proof by induction. □

Problem 2.20 Prove that for all integers $n \geq 2$, we have $n^3+1 > n^2+n$.

..... cubeineq

Problem 2.21 Prove that $n^2 < 2^n$ for all $n \geq 5$.

..... i3

Problem 2.22 Prove that $n! > 2^n$ for all integers $n \geq 4$.

..... facineq

Problem 2.23 Prove that for all $x > 0$ and all nonnegative integers n,

$$(1+x)^n \geq 1+nx.$$

..... pineq

Strong Induction

A more significant variation is *strong* induction. In the induction step for ordinary (or generalized) induction we assume $\mathrm{P}(n)$ for *just one value* n and use that to prove the next case, $\mathrm{P}(n+1)$. But for strong induction we assume *all* the preceding cases $\mathrm{P}(n_0)$, $\mathrm{P}(n_0+1)$, ..., $\mathrm{P}(n)$ and use them to prove $\mathrm{P}(n+1)$. Sometimes we also prove multiple base cases $\mathrm{P}(n_0)$, $\mathrm{P}(n_0+1)$, ..., $\mathrm{P}(k)$, and then just apply the induction argument for $n > k$. Here is an elementary example.

Example 2.19. Prove that every integer $n \geq 12$ can be written in the form $n = 3m + 7\ell$ where m and ℓ are nonnegative integers.

Let's discuss the proof before writing out the finished version. The basic issue for any induction proof is connecting the next case, $n+1$, to the previous case(s). Here we want to produce nonnegative integers m and ℓ for which

$$n+1 = 3m+7\ell. \tag{2.3}$$

If we were using ordinary induction, we would assume $n = 3m' + 7\ell'$, for some nonnegative m' and ℓ', and try to use this to produce m and ℓ in (2.3). This doesn't work very well;

$$n+1 = 3m'+7\ell'+1, \text{ but how does this become } 3m+7\ell \text{ ?}$$

But we *can* make a connection between $n+1$ and $n-2$ because they are 3 steps apart: $n+1 = (n-2)+3$. If we assume that the $n-2$ case is true, that is that $n-2 = 3m'+7\ell'$, then

$$n+1 = (n-2)+3 = 3m'+7\ell'+3 = 3(m'+1)+7\ell' = 3m+7\ell \text{ where } m = m'+1 \text{ and } \ell = \ell'.$$

So the induction step works out neatly if we are not limited to assuming only the n case but assume *all* the cases up to n, including the case of $n-2$ in particular. Here is this strong induction proof written out.

Proof. We use strong induction. First consider the cases of $n = 12$, $n = 13$, and $n = 14$. Observe that

$$12 = 3 \cdot 4 + 7 \cdot 0, \quad 13 = 3 \cdot 2 + 7 \cdot 1, \quad 14 = 3 \cdot 0 + 7 \cdot 2.$$

Next, consider $14 \leq n$ and suppose that the assertion is true for each $12 \leq k \leq n$. Consider $n + 1$ and let $k = n - 2$. Then $12 \leq k \leq n$ so by our induction hypothesis there exist nonnegative integers m', ℓ' with $n - 2 = 3m' + 7\ell'$ and consequently, $n + 1 = n - 2 + 3 = 3(m' + 1) + 7\ell'$. Since $m = m' + 1$ and $\ell = \ell'$ are nonnegative integers, this proves the assertion for $n + 1$. By induction, this completes the proof. □

Notice that we proved multiple base cases: $n = 12$, 13, and 14. The reason we had to do that is that the argument to prove the case of $n + 1$ uses the case of $n - 2$. Since our proposition is only true for integers at least as large as 12, the induction step is only valid for $n - 2 \geq 12$, which means $n \geq 14$. So we have to check the values $12 \leq n \leq 14$ individually as base cases before we can start using the induction argument.

We will see more examples of strong induction in Proposition 2.4 and Theorem 2.5 below. Additional examples will be found in Chapter 4.

Problem 2.24 Here's a spoof to show that in fact $\frac{d}{dx}x^n = 0$ for all $n \geq 0$. Clearly $\frac{d}{dx}x^0 = 0$. If we make the (strong) induction hypothesis that $\frac{d}{dx}x^k = 0$ for all $k \leq n$, then the product rule implies

$$\begin{aligned}
\frac{d}{dx}x^{n+1} &= \frac{d}{dx}(x \cdot x^n) \\
&= \left(\frac{d}{dx}x^1\right)x^n + x^1\frac{d}{dx}x^n \\
&= 0\,x^n + x^1\,0 \\
&= 0.
\end{aligned}$$

Thus $\frac{d}{dx}x^n = 0$ What's wrong? (See [5] for the source of this one.)

..... ezd

Recursive Definitions

Induction proves a proposition $\mathrm{P}(n)$ for all $n \geq n_0$ by proving it for a first case n_0 and then showing why the truth of each successive case $\mathrm{P}(n + 1)$ follows from the preceding one $\mathrm{P}(n)$. The same idea can be used to *define* some function $f(n)$ for $n \geq n_0$, by specifying a starting value $f(n_0)$ and then giving a formula for $f(n+1)$ which is based on the preceding $f(n)$. Together these two pieces of information determine or define the values $f(n)$ for all $n \geq n_0$. This is what we call a *recursive definition.* For instance, a recursive definition of $n!$ could consist of specifying that

- $0! = 1$, and
- $(n + 1)! = (n + 1) \cdot n!$ for each $n \geq 0$.

This determines the same values for $n!$ as our original definition on page 20. For instance, applying the definition repeatedly (recursively), we find that

$$4! = 4 \cdot 3! = 4 \cdot 3 \cdot 2! = 4 \cdot 3 \cdot 2 \cdot 1! = 4 \cdot 3 \cdot 2 \cdot 1 \cdot 0! = 4 \cdot 3 \cdot 2 \cdot 1 \cdot 1 = 24.$$

We will see another recursive definition in our proof of Theorem 3.8 below.

Here is a famous example that uses a strong induction sort of recursive definition.

Definition. The *Fibonacci numbers*, F_n for $n \geq 1$, are defined recursively by

$$F_1 = 1,\ F_2 = 1, \text{ and } F_{n+1} = F_n + F_{n-1} \text{ for } n \geq 2.$$

The first few Fibonacci numbers are

$$F_1 = 1,\ F_2 = 1,\ F_3 = 2,\ F_4 = 3,\ F_5 = 5,\ F_6 = 8,\ F_7 = 13,\ F_8 = 21, \ldots$$

The Fibonacci numbers have many unusual properties. Since F_{n+1} depends on the previous *two* terms, it is not surprising that strong induction is a natural tool for proving those properties. As an example we offer the following.

Proposition 2.4 (Binet's Formula)**.** *The Fibonacci numbers are given by the following formula:*

$$F_n = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right].$$

It is surprising that this formula always produces integers, but it does!

Proof. We prove Binet's formula by (strong) induction. First consider the cases of $n = 1, 2$. For $n = 1$ we have

$$\frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^1 - \left(\frac{1-\sqrt{5}}{2}\right)^1\right] = \frac{2\sqrt{5}}{2\sqrt{5}} = 1 = F_1.$$

For $n = 2$ we have

$$\frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^2 - \left(\frac{1-\sqrt{5}}{2}\right)^2\right] = \frac{1}{\sqrt{5}}\left[\frac{1+2\sqrt{5}+5}{4} - \frac{1-2\sqrt{5}+5}{4}\right] = \frac{4\sqrt{5}}{4\sqrt{5}} = 1 = F_2.$$

Now suppose the formula is correct for *all* $1 \le k \le n$, where $n \ge 2$. We will show that it must also be true for $n+1$. By definition of the Fibonacci sequence and the induction hypothesis we have

$$\begin{aligned}
F_{n+1} &= F_{n-1} + F_n \\
&= \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{n-1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n-1}\right] + \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{n} - \left(\frac{1-\sqrt{5}}{2}\right)^{n}\right] \\
&= \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{n-1} + \left(\frac{1+\sqrt{5}}{2}\right)^{n}\right] - \frac{1}{\sqrt{5}}\left[\left(\frac{1-\sqrt{5}}{2}\right)^{n-1} + \left(\frac{1-\sqrt{5}}{2}\right)^{n}\right].
\end{aligned}$$

Now observe that

$$\begin{aligned}
\left(\frac{1\pm\sqrt{5}}{2}\right)^{n+1} &= \left[\frac{1\pm 2\sqrt{5}+5}{4}\right]\left(\frac{1\pm\sqrt{5}}{2}\right)^{n-1} \\
&= \left[1 + \frac{2\pm 2\sqrt{5}}{4}\right]\left(\frac{1\pm\sqrt{5}}{2}\right)^{n-1} \\
&= \left[1 + \left(\frac{1\pm 1\sqrt{5}}{2}\right)\right]\left(\frac{1\pm\sqrt{5}}{2}\right)^{n-1} \\
&= \left(\frac{1\pm\sqrt{5}}{2}\right)^{n-1} + \left(\frac{1\pm\sqrt{5}}{2}\right)^{n}.
\end{aligned}$$

Using this we find that

$$F_{n+1} = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right],$$

which verifies Binet's formula for $n+1$ and completes the proof by induction. □

Why is this strong induction? Because the proof for F_{n+1} needs to use the assumption that the formula is true for *both* F_{n-1} and F_n, not just F_n. Why are there two base cases? Because we need *both* $n \geq 1$ and $n-1 \geq 1$ for the induction step to work, i.e. the induction argument is only valid once $n \geq 2$.

Here is another interesting property of the Fibonacci numbers; see [2].

Theorem 2.5 (Zeckendorf's Theorem)**.** *Every positive integer n can be expressed as a sum of distinct Fibonacci numbers:*

$$n = F_{i_1} + F_{i_2} + \cdots F_{i_k},$$

where $1 < i_1 < i_2 < \cdots < i_k$.

The proof is another example of strong induction.

Proof. We prove this by strong induction. First consider $n = 1$. Since $1 = F_2$ the theorem holds for $n = 1$.

Now we make the strong induction hypothesis that the theorem holds for every $1 \leq m \leq n$ and consider $n+1$. If $n+1$ is itself a Fibonacci number, $n+1 = F_i$, then again the theorem's assertion is true. If $n+1$ is not a Fibonacci number then it falls between two Fibonacci numbers:

$$F_\ell < n+1 < F_{\ell+1}, \text{ for some } \ell \geq 4.$$

(F_4, F_5 is the first pair of Fibonacci numbers between which there is a gap.) Let $m = n+1-F_\ell$. It follows that

$$0 < m \leq F_{\ell+1} - F_\ell = F_{\ell-1} < F_\ell < n+1,$$

using the basic recurrence relation for the Fibonacci numbers. By our induction hypothesis, we know that m can be written as a sum of Fibonacci numbers,

$$m = F_{i_1} + F_{i_2} + \cdots F_{i_{k'}},$$

where $1 < i_1 < i_2 < \cdots < i_{k'}$. Since $m < F_\ell$ we know $i_{k'} < \ell$. Now define $k = k'+1$ and $i_k = \ell$. We have

$$\begin{aligned} n+1 &= m + F_\ell \\ &= F_{i_1} + F_{i_2} + \cdots F_{i_{k'}} + F_\ell \\ &= F_{i_1} + F_{i_2} + \cdots F_{i_k}. \end{aligned}$$

This proves the assertion of the theorem for $n+1$, and completes the proof. □

Problem 2.25 Define a sequence a_n recursively by $a_1 = 1$ and $a_{n+1} = \frac{6a_n+5}{a_+2}$ for $n \in \mathbb{N}$. Prove by induction that $0 < a_n < 5$. [9, #17, p.55]

..... `recseq`

Problem 2.26 Here are some more equations involving the Fibonacci numbers. Prove them. (Do *not* use Binet's formula. Ordinary induction will suffice for all of these.)

a) $\sum_{i=1}^n F_i = F_{n+2} - 1$.

b) $\sum_{i=1}^n F_{2i-1} = F_{2n}$.

c) $\sum_{i=1}^n F_{2i} = F_{2n+1} - 1$.

d) $\sum_{i=1}^n F_i^2 = F_n F_{n+1}$.

..... `Fib2`

Problem 2.27 Prove that F_{5n} is divisible by 5 for all positive integers n. (Apply the recursive definition several times to find a formula for F_{5n+5} in terms of F_{5n+1} and F_{5n}. Once you have that an easy induction argument will finish the problem.)

..... Fib5

Problem 2.28 Write a proof that $\lim_{n\to\infty} F_n = \infty$ *without* using Binet's Formula. (One way to do this is prove a simple lower bound, $F_n \geq f(n)$ for some expression $f(n)$ with $f(n) \to \infty$ as $n \to \infty$.)

..... FibInfty

Problem 2.29 We can define the Fibonacci numbers F_n for $n \leq 0$ by using the basic recursion formula

$$F_{n+1} = F_n + F_{n-1}$$

and working *backwards* from $F_1 = F_2 = 1$. For instance, from $F_2 = F_1 + F_0$ we deduce that $F_0 = 0$; and from $F_1 = F_0 + F_{-1}$ we then deduce that $F_{-1} = 1$.

a) Find the values of $F_{-2}, \ldots, F_{-8}$.

b) Prove that for all $n \geq 1$,

$$F_{-n} = (-1)^{n+1} F_n.$$

..... S1

D Some Advice for Writing Proofs

Creating a proof consists of two tasks. The first is to find the mathematical ideas, relationships and arguments that form the proof. The second is writing it down for someone else to read. *Don't try to do both at once.* Start with a piece of scratch paper and just explore the hypotheses and what you are supposed to prove to see if you can discover how they are connected, or some strategy for getting from the hypotheses to the conclusion. There is no prescription or procedure to help you. This is fundamentally a creative process. Play with what you want to prove, try some examples, draw pictures, see if you can work out any possibilities that occur to you — just try things until you find an idea that seems to work. Once you come up with the mathematical idea(s) of your proof, *only then* comes the job of writing it out in careful, logical language. (Don't expect someone else to read your scratch paper. It's for your eyes only — don't turn it in.)

When you read someone else's proof you may wonder how they ever thought of it. For instance in the proof of the Pythagorean Theorem above, how did someone first think of the idea of using a partition of one square into smaller figures and adding up the areas? There is almost always some sort of a basic idea or strategy behind a proof. Sometimes it is obvious. Sometimes it takes some effort to stand back and look past the details to figure out what the guiding idea was, but it is usually there to be found. When studying a proof, try to identify the main idea(s) behind it. Once you've recognized that, the rest of the details will be easier to understand. Learning to think that way will also help you to design proofs of you own.

The writing part is the process of putting your argument on paper, carefully arranged and explained *for someone else to read.* This might involve deciding what symbols and notation to use, what order to say things in. After logical correctness the most important thing is to design the written proof with the reader in mind. Picture one of your classmates who has already read the statement of what is to be proven, but has not thought about how to do the proof, and is even a bit skeptical that it is really true. Your written proof has to lead them through the reasoning, so that by the end they can't help but concede the conclusion. Expect them to be skeptical, looking for holes in your argument. Try to anticipate where in your proof they might raise objections and put in explanation to dispel their doubts. Your written proof is their tour guide through the labyrinth of your argument; you don't want them to wander away from your logical path.

We often mingle mathematical notation with English grammar. For instance, instead of writing "The square root of two is an irrational number," we write "$\sqrt{2}$ is an irrational number." The latter is briefer and easier to read. We also use mathematical notation to squeeze several things we want to say into a shorter sentence. Consider this example:

$$\sqrt{2} \text{ is the real number } x \geq 0 \text{ for which } x^2 = 2.$$

This is briefer than "...the real number x which is greater than or equal to 0 and ..." The goal is clarity for your reader; decide how much notation to use by thinking about what will make it easiest for them to understand.

Even if we are mixing mathematical notation with conventional English, we are still writing in English; there is no substitute for fluency. Proper punctuation is still important. For instance, sentences should end with periods, even when a mathematical symbol is the last "word." In other words we view the use of mathematical notation as an enhancement of our natural English language usage. The language we use is precise, but it still leaves plenty of room for stylistic nuances. Good writing is as important in mathematics as in any other subject. Like any language, fluency is only attained by experience in actually using the language.

Here are a few more specific suggestions that may be helpful as you write out your proofs.

- Be sure you are clear in your own mind about what you are assuming as opposed to what you need to prove. Don't use language that claims things are true before you have actually proven they are true. For instance, you could start a proof of the Pythagorean Theorem with "We want to show that $a^2 + b^2 = c^2$," but *don't* start with "Therefore square of the hypotenuse equals the sum of the squares of the other two sides." (It seems silly to warn you about that, but you would be surprised at how often students write such things.)

- Be sure your argument accounts for all possibilities that the hypotheses allow. Don't add extra assumptions beyond the hypotheses, unless you are considering cases which, when taken together, exhaust all the possibilities that the hypotheses allow.

- Don't confuse an example, which illustrates the assertion in a specific instance, with a proof that accounts for the full scope of the hypotheses.

- Be wary of calculations that start at the conclusion and work backwards. These are useful in discovering a line of reasoning connecting hypotheses and conclusion, but often are logically the converse of what you want to show. Such an argument needs to be rewritten in the correct logical order, so that it proceeds from hypotheses to conclusion, not the other way around. We saw this as we developed our proof of Proposition 1.5.

- Don't just string formulas together. Your proof is not a mathematical diary, in which you write a record of what you did for your own sake. It is an explanation that will guide another reader through a logical argument. Explain and narrate. *Use words, not just formulas.* (And don't include things you thought about but ended up not using.)

- Don't use or talk about objects or variables you haven't introduced to the reader. Introduce them with "let" or "define" so the reader knows what they are when they are first encountered. On the other hand, don't define things long before the reader needs to know about them (or worse, things that aren't needed at all).

- Keep notation consistent, and no more complicated than necessary.

- When you think you are done, reread your proof skeptically, looking for possibilities that your argument overlooked. Perhaps get someone else (who is comfortable with proofs) to read it. Make it your responsibility to find and fix flaws in your reasoning and writing before you turn it in.

- One thing that some find awkward at first is the traditional use of first person plural "we ..." instead of "I" in mathematical writing. The use of "I" is fine for personal communications, talking about yourself. But when you are describing a proof it is not a personal statement or testimony. The "we" refers to you and all your readers — you are leading all of them with you through your reasoning. Write so as to take them along with you through your arguments. By writing "we" you are saying that this is what everyone in your logical tour group should be thinking, not just your personal perspective.

- Separate symbols by words. For instance, "Consider q^2, where $q \in \mathbb{Q}$" is good. "Consider q^2, $q \in \mathbb{Q}$" is vague.

- Don't start a sentence with a symbol.
- A sentence starting with "if" needs to include "then."
- Use words accurately. An inequality is not an equation, for instance.

If you want more discussion of good mathematical writing, you could try *Tips on Writing Mathematics* at the beginning of [17], or the classic little brochure [12].

E Perspective: Proofs and Discovery

Since proofs are often not emphasized in freshman-sophomore calculus courses, students can get the impression that the proofs are merely a sort of formal ritual by which a fact receives the official title of "Theorem." In reality the development of a proof is often part of the process of discovery, i.e. the way we separate what we know to be true from what we do not. If we believe something is true, but can't prove it, then there remains some uncertainty. Perhaps what we are thinking of is only true in some more limited context; until there is a proof we don't know for certain. Here are two famous examples.

Theorem 2.6 (Fermat's Last Theorem)**.** *If $m > 2$ is an integer, there are no triples (a, b, c) of positive integers with*

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

Conjecture (Twin Primes)**.** There are an infinite number of positive integers n for which both n and $n+2$ are prime numbers.

To find a proof of Fermat's Last Theorem was one of the most famous unsolved problems in mathematics since Fermat claimed it was true in 1637. As the years went by many mathematicians[11] tried to prove or disprove it. Even without a valid proof, mathematicians developed an increasing understanding of the problem and its connections to other mathematical subjects. There was a growing consensus that it was probably true, but it remained a strongly held belief or opinion until it was finally proven[12] in 1993 by Andrew Wiles. The point is that Wiles' successful proof is the way we discovered that it really is true.

The Twin Primes conjecture is still on the other side of the fence. So far no one has been able to prove it, so it's not yet a theorem. We don't yet know for sure if it is true or not, even though most number theorists have a pretty strong opinion.

It is important to realize that what makes a proof valid is *not* that it follows some prescribed pattern, or uses authoritative language or some magic phrases like "therefore" or "it follows that". The final criterion is whether the proof presents an argument that is logically complete and exhaustive. There are many examples of proofs that seem convincing at first reading, but which contain flaws on closer inspection. In the history of both Fermat's Last Theorem and the Twin Primes Conjecture there have been people who thought they had found a proof, only to find later that there were logical flaws. It *is* entirely possible to give a false proof of a true conclusion. Often however something is learned from the flawed argument which helps future efforts. A couple interesting examples of things that were thought to have been proven, but later found to be false are given in Ch.11 of [1].

Another very interesting case is the Four Color Problem. Picture a map with borders between various countries (or states, or counties, ...) drawn in. Each country is to be shaded with a color, but two countries that share a common border can not use the same color. The problem is to prove that you can always do this

[11] A prize was offered in 1908 for the first person to prove it, the Wolfskehl prize. It has been estimated that over the years more that 5,000 "solutions" have been submitted to the Göttingen Royal Society of Science, which was responsible for awarding the prize. All of them had flaws; many were submitted by crackpots. The prize of DM 25,000 was awarded to A. Wiles on June 27, 1997. The interesting story of this prize (and the courageous soul who read and replied to each submitted "solution" until his death in 1943) is told in [4].

[12] Wiles announced that he had proved it on June 23, 1993. But shortly after that a flaw in the proof was found. Wiles enlisted the help of Richard Taylor and together they fixed the problem in September 1994. The proof was published as two papers in the May 1995 issue of Annals of Mathematics (vol. 141 no. 3): Andrew Wiles, Modular elliptic curves and Fermat's last theorem, pp. 443–551; Richard Taylor and Andrew Wiles, Ring-theoretic properties of certain Hecke algebras, pp. 553–572. A chapter in [8] provides an overview of Fermat's Last Theorem, its history and final solution by A. Wiles.

with no more than four colors[13] . This problem was posed to the London Mathematical Society in 1878. It resisted a proof until 1976, when Kenneth Appel and Wolfgang Haken of the University of Illinois achieved a proof. But their proof was controversial because it depended in an essential way on a computer program. They were able to reduce the problem to a finite number of cases, but thousands of them. The computer was instrumental both in identifying an exhaustive list of cases and then in checking each of them. It took four years to develop the program and 1200 hours to run. So was this a proof? It wasn't a proof like the ones we have been talking about, which when written down can be read and evaluated by another person. It was a proof based on an algorithm, which (the proof claims) terminates successfully only if the assertion of the four color problem is true, *and* a computer program which (it is claimed) faithfully executes the algorithm, *and* the claim that the algorithm did run and successfully terminate on the authors' computer. By now most of the controversy has died down. Various efforts to test the algorithm and modifications of it, program it differently, run it on different machines, etc... have led most people to accept it, even though some may remain unhappy about it. You can read more about the Four Color Problem in [27].

Additional Problems

Problem 2.30 Prove that for all real numbers a, b, c, the inequality $ab + bc + ca \leq a^2 + b^2 + c^2$ holds.
..... `trineq`

Problem 2.31 If A is a square matrix, the notation A^n refers to the result of multiplying A times itself n times: $A^3 = A \cdot A \cdot A$ for instance. Prove that for every poistive integer n the following formula holds:

$$\begin{bmatrix} a & 1 \\ 0 & a \end{bmatrix}^n = \begin{bmatrix} a^n & na^{n-1} \\ 0 & a^n \end{bmatrix}$$

..... `i2`

Problem 2.32 Suppose that x is a real number and i refers to the complex number described by $i^2 = -1$. For that for all positive integers n the following formula holds.

$$(\cos(x) + i\sin(x))^n = \cos(nx) + i\sin(nx).$$

This is elementary if you use Euler's formula ($e^{ix} = \cos(x) + i\sin(x)$) and assume that the laws of exponents hold for complex numbers. But do *not* use those things to prove it here. Instead use induction and conventional trigonometric identities (which you may need to look up if you have forgotten them).
..... `i5`

Problem 2.33 Prove (by induction) that if we divide the plane into regions by drawing a number of straight lines, then it is possible to "color" the regions with just 2 colors so that no two regions that share an edge have the same color. (From [9].)
..... `color2`

Problem 2.34 Explain Binet's formula in terms of the eigenvalues of the matrix $\left[\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right]$.

[13] Its not hard to see that you need at least four colors. Look at a map of the United States for instance. California, Oregon, and Nevada all share borders with each other, so we will need at least three colors just for them. Suppose we color Oregon green, California blue and Nevada red. If we had only these three colors, then Arizona would have to be different than both California and Nevada, so Arizona would have to be green. Now Utah has to be different from both Arizona and Nevada, so it would have to be blue. Similarly Idaho has to be different from both Oregon and Nevada, so it has to be blue too. But Idaho and Utah share a border and we have colored them both blue. We see that 3 colors alone are not enough.

. . . . . Binet

Problem 2.35 Observe that

$$\begin{aligned} 5^2 &= 3 \cdot 8 + 1 \\ 8^2 &= 5 \cdot 13 - 1 \\ 13^2 &= 8 \cdot 21 + 1 \\ 21^2 &= 13 \cdot 34 - 1 \\ &\vdots \end{aligned}$$

Express the pattern as a formula involving the Fibonacci numbers F_n, and prove it. (Taken from [17].)
. . . . . FibForm

Chapter 3

Sets and Functions

This chapter is devoted to the subject of sets, perhaps the most basic of mathematical objects. The theory of sets is a topic of study in its own right, part of the "foundations of mathematics" which involves deep questions of logic and philosophy. (See Section H.2 below for a hint of this.) For our purposes sets and functions are just another part of the language of mathematics. Sections A–F are intended mainly to familiarize you with this part of the vernacular. We will encounter a lot of notation, of which some may be new to you. This notation too is part of the vocabulary. You need to learn to *use* this language, not be dependent on an example or picture as a way to bypass literacy. Section G introduces some ideas regarding infinite sets. Here we will be going beyond mere terminology and will prove some interesting theorems.

A Notation and Basic Concepts

A *set* is simply a collection of things. The things in the set are the *elements* of the set. For instance, the set of multiples of 13 less than 100 is

$$S = \{13, 26, 39, 52, 65, 78, 91\}.$$

This set has exactly 7 elements. Order does not matter; $S = \{52, 13, 26, 91, 39, 78, 65\}$ describes the same set. The mathematical notation for "is an element of" is $\in$. In our example, $65 \in S$ but $66 \notin S$ (by which we mean 66 is not an element of S). A set is viewed as a single object, distinct from and of a different type than its elements. Our S is not a number; it is the new object formed by putting those specific numbers together as a "package." Even when a set contains only one element, we distinguish between the element and the set containing the element. Thus

$$0 \text{ and } \{0\}$$

are two different objects. The first is a number; the second is a set containing a number.

Simple sets are indicated by listing the elements inside braces $\{\cdots\}$ as we have been doing above. When a pattern is clear we might abbreviate the listing by writing "... " For instance we might identify the set of prime numbers by writing

$$P = \{2, 3, 5, 7, 11, 13, \ldots\}.$$

But to be more precise it is better to use a descriptive specification, such as

$$P = \{n : \ n \text{ is a prime number}\}.$$

This would be read "the set of n such that n is a prime number." The general form for this kind of set specification is[1]

$$\{x : \ \text{criteria}(x)\},$$

where "criteria(x)" is an open sentence specifying the qualifications for membership in the set. For example,

$$T = \{x : x \text{ is a nonnegative real number and } x^2 - 2 = 0\}$$

[1] Many authors prefer to write "$\{n| \ \ldots\}$, using a vertical bar instead of a colon. That's just a different notation for the same thing.

is just a cumbersome way of specifying the set $T = \{\sqrt{2}\}$. In a descriptive set specification we sometimes limit the scope of the variable(s) by indicating something about it before the colon. For instance we might write our example T above as

$$T = \{x \in \mathbb{R} : \ 0 \leq x \text{ and } x^2 - 2 = 0\},$$

or even

$$T = \{x \geq 0 : x^2 - 2 = 0\},$$

if we understand $x \in \mathbb{R}$ to be implicit in $x \geq 0$.

There are special symbols[2] for many of the most common sets of numbers:

$$\begin{aligned}
\text{the } \textit{natural numbers}: &\quad \mathbb{N} = \{1, 2, 3, \ldots\} = \{n : \ n \text{ is a positive integer}\}, \\
\text{the } \textit{integers}: &\quad \mathbb{Z} = \{\ldots - 3, -2, -1, 0, 1, 2, 3, \ldots\} = \{n : \ n \text{ is an integer}\}, \\
\text{the } \textit{rational numbers}: &\quad \mathbb{Q} = \{x : \ x \text{ is a real number expressible as } x = \frac{n}{m}, \text{ for some } n, m \in \mathbb{Z}\} \\
\text{the } \textit{real numbers}: &\quad \mathbb{R} = \{x : \ x \text{ is a real number}\}, \text{ and} \\
\text{the } \textit{complex numbers}: &\quad \mathbb{C} = \{z : \ z = x + iy, \ x, y \in \mathbb{R}\}.
\end{aligned}$$

Intervals are just special types of sets of real numbers, for which we have a special notation.

$$\begin{aligned}
[a, b) &= \{x \in \mathbb{R} : \ a \leq x < b\} \\
(b, +\infty) &= \{x \in \mathbb{R} : \ b < x\}.
\end{aligned}$$

In particular $(-\infty, +\infty)$ is just another notation for $\mathbb{R}$.

Another special set is the *empty set*

$$\emptyset = \{\ \},$$

the set with no elements at all. Don't confuse it with the number 0, or the set containing 0. **All of the following are *different* mathematical objects.**

$$0, \quad \emptyset, \quad \{0\}, \quad \{\emptyset\}.$$

B Basic Operations and Properties

Definition. Suppose A and B are sets. We say A is a *subset* of B, and write $A \subseteq B$, when every element of A is also an element of B. In other words, $A \subseteq B$ means that

$$x \in A \text{ implies } x \in B.$$

For instance, $\mathbb{N} \subseteq \mathbb{Z}$, but $\mathbb{Z} \subseteq \mathbb{N}$ is a false statement. (You will often find "$A \subset B$" written instead of "$A \subseteq B$". They mean the same thing; its just a matter of the author's preference.) No matter what the set A is, it is always true that

$$\emptyset \subseteq A.$$

(This is in keeping with our understanding that vacuous statements are true.) To say $A = B$ means that A and B are the same set, that is they contain precisely the same elements:

$$x \in A \text{ if and only if } x \in B,$$

which means the same as

$$A \subseteq B \text{ and } B \subseteq A.$$

This provides the typical way of proving $A = B$: prove containment both ways. See the proof of Lemma 3.1 part d) below for an example.

Starting with sets A and B there are several operations which form new sets from them.

[2] The choice of $\mathbb{Q}$ for the rational numbers is suggested by Q for "quotient." The choice of $\mathbb{Z}$ for the integers comes from the German term "zahlen" for numbers.

Definition. Suppose A and B are sets. The *intersection* of A and B is the set

$$A \cap B = \{x : \ x \in A \text{ and } x \in B\}.$$

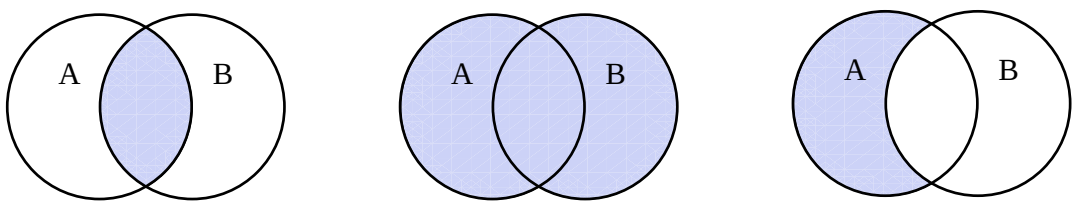
The *union* of A and B is the set

$$A \cup B = \{x : \ x \in A \text{ or } x \in B\}.$$

The set difference, A *remove* B, is the set

$$A \setminus B = \{x : \ x \in A \text{ and } x \notin B\}.$$

We can illustrate these definitions by imagining that A and B are sets of points inside two circles in the plane, and shade the appropriate regions to indicate the newly constructed sets. Such illustrations are called *Venn diagrams*. Here are Venn diagrams for the definitions above.



$A \cap B$ $\qquad$ $A \cup B$ $\qquad$ $A \setminus B$

Don't let yourself become dependent on pictures to work with sets. For one thing, not all sets are geometrical regions in the plane. Instead you should try to work in terms of the definitions, using precise logical language. For instance $x \in A \cap B$ means that $x \in A$ and $x \in B$.

Example 3.1. For $A = \{1, 2, 3\}$ and $B = \{2, 4, 6\}$ we have

$$A \cap B = \{2\}, \quad A \cup B = \{1, 2, 3, 4, 6\}, \quad A \setminus B = \{1, 3\}, \quad \text{and } B \setminus A = \{4, 6\}.$$

When two sets have no elements in common, $A \cap B = \emptyset$ and we say the two sets are *disjoint*. When we have three or more sets we say they are disjoint if every *pair* of them is disjoint. That is *not* the same as saying their combined intersection is empty.

Example 3.2. Consider $A = \{1, 2, 3\}$, $B = \{2, 4, 6\}$, and $C = \{7, 8, 9\}$. These are not disjoint (because $A \cap B \neq \emptyset$), even though $A \cap B \cap C = \emptyset$.

We want to say that the complement of a set A, to be denoted[3] $\tilde{A}$, is the set of all things which are not elements of A. But for this to be meaningful we have to know the allowed scope of all possible elements under consideration. For instance, if $A = \{1, 2, 3\}$, is $\tilde{A}$ to contain all natural numbers that are not in A, or all integers that are not in A, or all real numbers that are not in A, or...? It depends on the context in which we are working. If the context is all natural numbers, then $\tilde{A} = \{4, 5, 6, \ldots\}$. If the context is all integers, then $\tilde{A} = \{\ldots, -3, -2, -1, 0, 4, 5, 6, \ldots\}$. If the context is all real numbers $\mathbb{R}$, then

$$\tilde{A} = (-\infty, 1) \cup (1, 2) \cup (2, 3) \cup (3, +\infty).$$

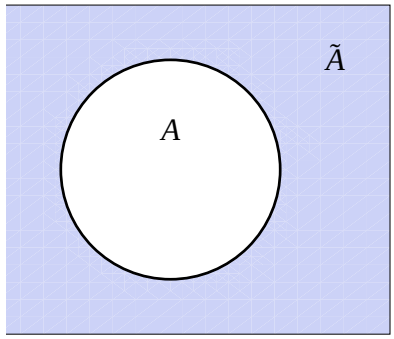
The point is that the complement of a set is always determined *relative to an understood context* of what the scope of all possible elements is.

Definition. Suppose X is the set of all elements which are allowed as elements of sets. The *complement* of a set $A \subseteq X$ is

$$\tilde{A} = X \setminus A.$$

Here is a Venn diagram. The full box illustrates X. The shaded region is $\tilde{A}$.

[3] Other common notations are A^c and A'.



In the context of the real numbers for instance,

$$\widetilde{(a,b)} = (-\infty, a] \cup [b, +\infty).$$

Lemma 3.1. *Suppose X is the set of all elements under consideration, and that A, B, and C are subsets of X. Then the following hold.*

a) $A \cup B = B \cup A$ *and* $A \cap B = B \cap A$.

b) $A \cup (B \cup C) = (A \cup B) \cup C$ *and* $A \cap (B \cap C) = (A \cap B) \cap C$

c) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ *and* $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

d) $(A \cup B)^{\sim} = \tilde{A} \cap \tilde{B}$ *and* $(A \cap B)^{\sim} = \tilde{A} \cup \tilde{B}$.

e) $A \cup \tilde{A} = X$ *and* $A \cap \tilde{A} = \emptyset$.

f) $\widetilde{(\tilde{A})} = A$.

Proofs of these statements consist of just translating the symbols into logical language, working out the implications, and then translating back. We will prove the first part of d) as an example.

Proof. To prove the first part of d), we prove containment both ways.

Suppose $x \in (A \cup B)^{\sim}$. This means $x \in X$ and $x \notin A \cup B$. Since ($x \in A$ or $x \in B$) is false, we know $x \notin A$ and $x \notin B$. So we know that $x \in \tilde{A}$ and $x \in \tilde{B}$, which means that $x \in \tilde{A} \cap \tilde{B}$. This shows that $(A \cup B)^{\sim} \subseteq \tilde{A} \cap \tilde{B}$.

Now assume that $x \in \tilde{A} \cap \tilde{B}$. This means that $x \in \tilde{A}$ and $x \in \tilde{B}$, and so $x \in X$, $x \notin A$, and $x \notin B$. Since x is in neither A nor B, $x \notin A \cup B$. We find therefore that $x \in (A \cup B)^{\sim}$. This shows that $x \in \tilde{A} \cap \tilde{B} \subseteq (A \cup B)^{\sim}$.

Having proven containment both ways we have established that $(A \cup B)^{\sim} = \tilde{A} \cap \tilde{B}$. □

Problem 3.1 If x and y are real numbers, $\min(x, y)$ refers to the smaller of the two numbers. If $a, b \in \mathbb{R}$, prove that

$$\{x \in \mathbb{R} \mid x \leq a\} \cap \{x \in \mathbb{R} \mid \min(x, a) \leq b\} = \{x \in \mathbb{R} \mid x \leq \min(a, b)\}.$$

..... `sets1`

Problem 3.2 Prove part c) of the above lemma.

..... `L21c`

Problem 3.3 Prove that for any two sets, A and B, the three sets

$$A \cap B, \ A \setminus B, \text{ and } B \setminus A$$

are disjoint.

djex

Problem 3.4 The *symmetric difference* of two sets A, B is defined to be

$$A \triangle B = (A \setminus B) \cup (B \setminus A).$$

a) Draw a Venn diagram to illustrate $A \triangle B$.

b) Prove that $\widetilde{A \triangle B} = (A \cap B) \cup (\tilde{A} \cap \tilde{B})$.

c) Is it true that $\widetilde{A \triangle B} = \tilde{A} \triangle \tilde{B}$? Either prove or give a counterexample.

d) Draw a Venn diagram for $(A \triangle B) \triangle C$.

symdif

Indexed Families of Sets

Sometimes we need to work with many sets whose descriptions are similar to each other. Instead of giving them all distinct names, "A, B, C, ..." it may be more convenient to give them the same name but with a subscript which takes different values.

Example 3.3. Let

$$A_1 = (-1, 1), \quad A_2 = (-\frac{1}{2}, \frac{1}{2}), \quad A_3 = (-\frac{1}{3}, \frac{1}{3}) \quad \ldots;$$

in general for $k \in \mathbb{N}$,

$$A_k = (-\frac{1}{k}, \frac{1}{k}).$$

In the above example, the "k" in "A_k" is the *index*. The set of allowed values for the index is called the *index set*. In this example the index set is $\mathbb{N}$. The collection $A_k, k \in \mathbb{N}$ is an example of an *indexed family* of sets. We can form the union of an indexed family, but instead of writing

$$A_1 \cup A_2 \cup A_3 \cup \ldots$$

we write

$$\cup_{k \in \mathbb{N}} A_k \text{ or, for this particular index set, } \cup_{k=1}^{\infty} A_k.$$

Similarly,

$$\cap_{k \in \mathbb{N}} A_k = A_1 \cap A_2 \cap A_3 \cap \ldots,$$

We will see some indexed families of sets in the proof of the Schroeder-Bernstein Theorem below. Here are some simpler examples.

Example 3.4. Continuing with Example 3.3, we have $\cup_{k=1}^{\infty} A_k = (-1, 1)$, and $\cap_{k \in \mathbb{N}} A_k = \{0\}$.

Example 3.5. $\mathbb{R} \setminus \mathbb{Z} = \cup_{k \in \mathbb{Z}} I_k$, where $I_k = (k, k+1)$.

There is no restriction on what can be used for an index set.

Example 3.6. Let $U = (0, \infty)$, the set of positive real numbers. For each $r \in U$ let

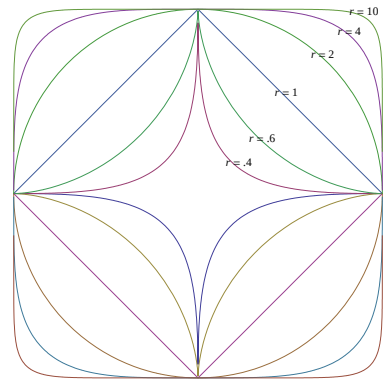
$$S_r = \{(x, y) : \ x, y \text{ are real numbers with } |x|^r + |y|^r = 1\}.$$

The figure at right illustrates some of the sets S_r. The set $\cup_{r\in U} S_r$ is somewhat tricky to describe:

$$\cup_{r\in U} S_r = (B \setminus A) \cup C,$$

where

$$\begin{aligned} B &= \{(x,y) : |x| < 1 \text{ and } |y| < 1\} \\ A &= \{(x,y) : \ x = 0 \text{ or } y = 0\} \\ C &= \{(0,1),(0,-1),(1,0),(-1,0)\}. \end{aligned}$$



Problem 3.5

a) For each $x \in \mathbb{R}$ let $C_x = \{y : x^2 + y^2 \le 1\}$. What is $\cup_{x\in\mathbb{R}} C_x$? What is $\cap_{x\in\mathbb{R}} C_x$?

b) Let $M_n = \{k \in \mathbb{N} : k = nm \text{ for some integer } m > 1\}$. What is $\cup_{n\in\mathbb{N}} M_n$?

c) Suppose $A \subseteq \mathbb{R}$ and for each $\epsilon > 0$ let $I_\epsilon = \{a \in \mathbb{R} : (a-\epsilon, a+\epsilon) \subseteq A\}$. Is it true that $A = \cup_{\epsilon>0} I_\epsilon$? Explain.

d) Let $S_\epsilon = \{n \in \mathbb{N} : \sin(n) > 1 - \epsilon\}$. For $\epsilon > 0$ is S_ϵ finite or infinite? What is $\cap_{\epsilon>0} S_\epsilon$?

..... Idx

Problem 3.6 For each $x \in \mathbb{R}$, let P_x be the set

$$P_x = \{y : \ y = x^n \text{ for some } n \in \mathbb{N}\}.$$

a) There are exactly three values of x for which P_x is a finite set. What are they?

b) Find $\cap_{0<x<1} P_x$ and $\cup_{0<x<1} P_x$.

c) For a positive integer N, find $\cap_{k=1}^{N} P_{2^k}$. Find $\cap_{k=1}^{\infty} P_{2^k}$.

..... Sx

C Product Sets

Sets don't care about the order of their elements. For instance,

$$\{1,2,3\} \text{ and } \{3,1,2\}$$

are the same set. You might think of a set as an unordered list of elements. An ordered list of numbers is a different kind of thing. For instance if we are thinking of $(1,2)$ and $(2,1)$ as the coordinates of points in the plane, then the order matters. When we write (x,y) we mean the *ordered* pair of numbers. The use of parentheses indicates that we mean ordered pair, not set.

If A and B are sets, we can consider ordered pairs (a,b) where $a \in A$ and $b \in B$.

Definition. Suppose A and B are sets. The set of all such ordered pairs (a,b) with $a \in A$ and $b \in B$ is called the *Cartesian product* of A and B:

$$A \times B = \{(a,b) : \ a \in A, \ b \in B\}.$$

Don't let the use of the word "product" and the notation "$\times$" mislead you here — there is no multiplication involved. The elements of $A \times B$ are a different kind of object than the elements of A and B. For instance the elements of $\mathbb{R}$ and $\mathbb{Z}$ are individual numbers (the second more limited that the first), but an element of $\mathbb{R} \times \mathbb{Z}$ is an ordered pair of two numbers (*not* the result of multiplication). For instance $(\pi, -3) \in \mathbb{R} \times \mathbb{Z}$.

Example 3.7. Let $X = \{-1, 0, 1\}$ and $Y = \{\pi, e\}$. Then

$$X \times Y = \{(-1, \pi),\ (0, \pi),\ (1, \pi),\ (-1, e),\ (0, e),\ (1, e)\}\,.$$

We can do the same thing with more than two sets: for a set of ordered triples we would write

$$A \times B \times C = \{(a, b, c) :\ a \in A,\ b \in B, c \in C\}.$$

Example 3.8. If $\Gamma = \{a, b, c, d, e, \ldots, z\}$ is the English alphabet (thought of as a set) then $\Gamma \times \Gamma \times \Gamma \times \Gamma$ is the (notorious) set of four-letter words (including ones of no known meaning).

When we form the Cartesian product of the same set with itself, we often write

$$A^2 \text{ as an alternate notation for } A \times A.$$

So the coordinates of points in the plane make up the set $\mathbb{R}^2$. The set of all possible coordinates for points in three dimensional space is $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$. The set of four-letter words in the example above is Γ^4. The next lemma lists some basic properties of Cartesian products.

Lemma 3.2. *Suppose A, B, C, D are sets. The following hold.*

a) $A \times (B \cup C) = (A \times B) \cup (A \times C)$.

b) $A \times (B \cap C) = (A \times B) \cap (A \times C)$.

c) $(A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D)$.

d) $(A \times B) \cup (C \times D) \subseteq (A \cup C) \times (B \cup D)$.

As an example of how this sort of thing is proven, here is a proof of part b).

Proof. Suppose $(x, y) \in A \times (B \cap C)$. This means $x \in A$ and $y \in B \cap C$. Since $y \in B$ it follows that $(x, y) \in A \times B$. Similarly, since $y \in C$ it follows that $(x, y) \in A \times C$ as well. Since (x, y) is in both $A \times B$ and $A \times C$, we conclude that $(x, y) \in (A \times B) \cap (A \times C)$. This proves that $A \times (B \cap C) \subseteq (A \times B) \cap (A \times C)$.

Suppose $(x, y) \in (A \times B) \cap (A \times C)$. Since $(x, y) \in A \times B$ we know that $x \in A$ and $y \in B$. Since $(x, y) \in A \times C$ we know that $x \in A$ and $y \in C$. Therefore $x \in A$ and $y \in B \cap C$, and so $(x, y) \in A \times (B \cap C)$. This proves that $(A \times B) \cap (A \times C) \subseteq A \times (B \cap C)$.

Having proven containment both ways, this proves b) of the lemma. □

The next example shows why part d) is only a subset relation, not equality, in general.

Example 3.9. Consider $A = \{a\}$, $B = \{b\}$, $C = \{c\}$, $D = \{d\}$ where a, b, c, d are distinct. Observe that (a, d) is an element of $(A \cup C) \times (B \cup D)$, but not of $(A \times B) \cup (C \times D)$.

Problem 3.7 Prove part c) of the above lemma.

. . . . . L12c

D The Power Set of a Set

We can also form sets of sets. For instance if $A = \{1\}$, $B = \{1, 2\}$ and $C = \{1, 2, 3\}$, we can put these together into a new set

$$\mathcal{F} = \{A, B, C\}.$$

This $\mathcal{F}$ is also a set, but of a different type than A, B, or C. The elements of C are integers; the elements of $\mathcal{F}$ are *sets* of integers, which are quite different. We have used a script letter "$\mathcal{F}$" to emphasize the fact that it is a different kind of set than A, B, and C. While it is true that $A = \{1\} \in \mathcal{F}$, it is *not* true that $1 \in \mathcal{F}$.

Starting with any set, A we can form its power set, namely the set of all subsets of A.

Definition. Suppose A is a set. The *power set* of A is

$$\mathcal{P}(A) = \{B : \ B \subseteq A\}.$$

As an example

$$\mathcal{P}(\{1, 2, 3\}) = \big\{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\big\}.$$

Observe that $B \in \mathcal{P}(A)$ means that $B \subseteq A$. We will have something interesting to say about the power set of an infinite set in Section G.2.

We could keep going, and form the power set of the power set, and so forth. But if you carelessly wander too far down that path you will find yourself in a logical quagmire! See Section H.2.

E Relations

There are many ways we can compare two mathematical objects of the same type. Here are some examples.

- inequality for real numbers: $x \leq y$,
- containment for sets: $A \subseteq B$,
- division for integers: $n|m$ (defined on page 15 above),
- equality of integers mod k: $n \equiv_k m$ (defined in Section D of the next chapter).

Each of these is an example of what we call a *relation* between two objects of the same type. If X is a set, a relation on X is really an open sentence $\mathrm{R}(x, y)$ taking two variables $x, y \in X$. For the frequently used relations we have special symbols (like the examples above) that we write between the two arguments instead of in front of them: "$x \leq y$" instead of "$\leq (x, y)$." We will follow this pattern and write "$x \,\mathrm{R}\, y$" instead of "$\mathrm{R}(x, y)$" for our discussion below.

A statement of relation "$x \,\mathrm{R}\, y$" does *not* refer to some calculation we are to carry out using x and y. It is simply a statement which has a well-defined truth value for each pair $(x, y) \in X \times X$. Whether it is true or false depends on the specific choices for $x, y \in X$. For instance consider again the inequality relation, $\leq$. The statement $x \leq y$ is true for $x = 2$ and $y = 3$, but false for $x = 3$ and $y = 2$. (This also shows that order matters; in general $x \,\mathrm{R}\, y$ is not the same statement as $y \,\mathrm{R}\, x$.)

We can make up all kinds of strange relations; you will find several in the problems below. The important ones are those which describe a property which is significant for some purpose, like the examples above. Most useful relations have one or more of the following properties.

Definition. Suppose R is a relation on a set X and $x, y, z \in X$.

- R is called *reflexive* when $x \,\mathrm{R}\, x$ is true for all $x \in X$.
- R is called *symmetric* when $x \,\mathrm{R}\, y$ is equivalent to $y \,\mathrm{R}\, x$.
- R is called *transitive* when $x \,\mathrm{R}\, y$ and $y \,\mathrm{R}\, z$ together imply $x \,\mathrm{R}\, z$.

Example 3.10.

- Inequality ($\leq$) on $\mathbb{R}$ is transitive and reflexive, but not symmetric .

- Strict inequality ($<$) on $\mathbb{R}$ is transitive, but not reflexive or symmetric.
- Define the coprime relation C on $\mathbb{N}$ so that $n \,\mathrm{C}\, m$ means that n and m share no common positive factors other than 1. Then C is symmetric, but not reflexive or transitive.

Example 3.11. Define the lexicographic order relation on $\mathbb{R}^2$ by $(x, y)\,\mathrm{L}\,(u, v)$ when $x < u$ or ($x = u$ and $y \leq v$). Then L is transitive and reflexive, but not symmetric. Let's write out the proof that L is transitive.

Suppose $(x, y)\,\mathrm{L}\,(u, v)$ and $(u, v)\,\mathrm{L}\,(w, z)$, where $x, y, u, v, w, z \in \mathbb{R}$. Our goal is to show that $(x, y)\,\mathrm{L}\,(w, z)$. We know that $x \leq u$ and $u \leq w$. If either of these is a strict inequality, then $x < w$ and therefore $(x, y)\,\mathrm{L}\,(w, z)$. Otherwise $x = u = w$, $y \leq v$, and $v \leq z$. But then $x = w$ and $y \leq z$, which imply $(x, y)\,\mathrm{L}\,(w, z)$. So in either case we come to the desired conclusion.

Problem 3.8 For each of the following relations on $\mathbb{R}$, determine whether or not it is reflexive, symmetric, and transitive and justify your answers.

a) $x \diamondsuit y$ means $xy = 0$.

b) $x \curlyvee y$ means $xy \neq 0$.

c) $x \boxtimes y$ means $|x - y| < 5$.

d) $x \odot y$ means $x^2 + y^2 = 1$.

..... `relations`

Definition. A relation R is called an *equivalence relation* when it is symmetric, reflexive, and transitive.

Equivalence relations are especially important because they describe some notion of "sameness." For instance, suppose we are thinking about angles θ in the plane. We might start by saying an angle is any real number $x \in \mathbb{R}$. But that is not quite what we mean: $\pi/3$ and $7\pi/3$ are different as numbers, but they are *the same* as angles, because they differ by a multiple of 2π. The next example defines an equivalence relation on $\mathbb{R}$ that expresses this idea of sameness as angles.

Example 3.12. Let $\oslash$ be the relation on $\mathbb{R}$ defined by

$$x \oslash y \text{ means that there exists } k \in \mathbb{Z} \text{ so that } x = 2k\pi + y.$$

For instance $\pi/3 \oslash 7\pi/3$, because $\pi/3 = 2(-1)\pi + 7\pi/3$ so that the definition holds with $k = -1$.

This *is* an equivalence relation, as we will now check. Since $x = 2 \cdot 0 \cdot \pi + x$, the relation is reflexive. If $x = 2k\pi + y$, then $y = 2(-k)\pi + x$ and $-k \in \mathbb{Z}$ if k is. This proves symmetry. If $x = 2k\pi + y$ and $y = 2m\pi + z$, then $x = 2(k + m)\pi + z$, showing that the relation is transitive.

Definition. Suppose R is an equivalence relation on X and $x \in X$. The *equivalence class* of x is the set

$$[x]_{\mathrm{R}} = \{y \in X : \; x\mathrm{R}y\}.$$

The $y \in [x]_{\mathrm{R}}$ are called *representatives* of the equivalence class.

When it is clear what equivalence relation is intended, often we leave it out of the notation and just write $[x]$ for the equivalence class. The equivalence classes "partition" X into to the sets of mutually equivalent elements.

Example 3.13. Continuing with Example 3.12, the equivalence classes of $\oslash$ are sets of real numbers which differ from each other by multiples of 2π.

$$[x] = \{\ldots, x - 4\pi, x - 2\pi, x, x + 2\pi, x + 4\pi, \ldots\}.$$

For instance $\pi/3$ and $7\pi/3$ belong to the same equivalence class, which we can refer to as either $[\pi/3]$ or $[7\pi/3]$ — both refer to the same set of numbers. But $[\pi/3]$ and $[\pi]$ are different.

One of the uses of equivalence relations is to define new mathematical objects by disregarding irrelevant properties or information. The equivalence relation $x \oslash y$ of the above examples defines exactly what we mean by saying x and y represent the "same angle." If we have a particular angle in mind, then the set of all the x values corresponding to that angle form one of the equivalence classes of $\oslash$. If we want a precise definition of what an "angle" actually is (as distinct from a real number), the standard way to do it is say that an angle θ is an equivalence class of $\oslash$: $\theta = [x]$. The $x \in \theta$ are the representatives of the angle θ. The angle θ is the set of all x which represent the same angle. In this way we have defined a new kind of object (angle) by basically gluing together all the equivalent representatives of the same object and considering that glued-together lump (the equivalence class) as the new object itself. This will be the basis of our discussion of the integers mod m in Section D of the next chapter.

Problem 3.9 Show that each of the following is an equivalence relation.

a) On $\mathbb{N} \times \mathbb{N}$ define $(j, k) \parallel (m, n)$ to mean $jn = km$. (From [10].)

b) On $\mathbb{R} \times \mathbb{R}$ define $(x, y) \Cup (u, v)$ to mean $x^2 - y = u^2 - v$.

c) On $(0, \infty)$ define $x \eqsim y$ to mean $x/y \in \mathbb{Q}$.

For a) and b), give a geometrical description of the the equivalence classes.

..... er

Problem 3.10 Suppose $\vdash$ is an equivalence relation on X. For $x, y \in X$ let $[x]$ and $[y]$ be their equivalence classes with respect to $\vdash$.

a) Show that $x \in [x]$.

b) Show that either $[x] = [y]$, or $[x]$ and $[y]$ are disjoint.

c) Show that $x \vdash y$ iff $[x] = [y]$.

d) Consider the realtion $\Diamond$ on $\mathbb{R}$ from Problem 3.8. (Recall that this is *not* an equivalence relation.) Define

$$\langle x \rangle = \{y \in \mathbb{R} : \ x \Diamond y\}.$$

Which of a), b) and c) above are true if $[x]$ and $[y]$ are replaced by $\langle x \rangle$ and $\langle y \rangle$ and $\vdash$ is replaced by $\Diamond$?

..... ec

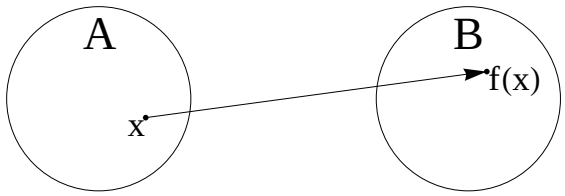
F Functions

The basic notion of a function f is familiar to any calculus student. It is a "rule" which assigns to each allowed "input" value x value an "output" value $f(x)$. In calculus the allowed values for x are usually all real numbers, or some interval of real numbers. For functions of several variables the $x = (x_1, x_2)$ or $x = (x_1, x_2, x_3)$ take values in $\mathbb{R}^2$ or $\mathbb{R}^3$ respectively. In general we can talk about a function whose *domain* A is a set and whose *codomain* B is another. We write

$$f : A \to B. \tag{3.1}$$

This notation means that f is a function[4] for which the allowed input values are the $a \in A$, and for each such a the associated output value $b = f(a) \in B$. Don't confuse "function" with "formula." We can describe a function with a table for instance, even if we can't think of a formula for it. For our purposes, it is helpful to visualize a function with a diagram like the following. The circle A indicates the points in the domain; the circle B indicates the points in the codomain; and the function provides an arrow from each $x \in A$ to its associated value $f(x) \in B$.

[4]We sometimes use the words "map" or "mapping" instead of "function." They mean the same thing.



Example 3.14. Here are some examples to make the point that the elements of the domain and codomain can be more complicated than just numbers.

a) Let $\mathbb{R}_{2\times 2}$ denote the set of 2×2 matrices with real entries, $M = \left[\begin{smallmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{smallmatrix}\right]$ and $\det(M)$ be the usual determinant:
$$\det(M) = m_{11}m_{22} - m_{12}m_{21}.$$
Then $\det : \mathbb{R}_{2\times 2} \to \mathbb{R}$.

b) Let $C([0,1])$ be the set of all continuous functions $f : [0,1] \to \mathbb{R}$. We can view the integral $I(f) = \int_0^1 f(x)\,dx$ as a function $I : C([0,1]) \to \mathbb{R}$.

c) For any $c = (c_1, c_2, c_3) \in \mathbb{R}^3$ we can form a quadratic polynomial $P(c) = c_3x^2 + c_2x + c_1$. We can view P as a function $P : \mathbb{R}^3 \to C([0,1])$ from ordered triples to continuous functions.

Definition. Suppose $f : A \to B$ is a function. The *range* of f is the set
$$\mathrm{Ran}(f) = \{b : \text{ there exists } a \in A \text{ such that } b = f(a)\}.$$

Don't confuse the codomain of a function with its range. The range of f is the set of values in $b \in B$ for which *there actually does exist* an $a \in A$ with $f(a) = b$. The codomain B can be any set containing the range as a subset. In general there is no presumption that $\mathrm{Ran}(f)$ is all of B. What we understand the codomain to be *does* affect how we answer some questions, as we will see shortly. When $\mathrm{Ran}(f) = B$ we say f is onto, or surjective; see the definition below. In particular whether f is onto or not depends on what we understand the codomain to be.

Definition. Suppose $f : A \to B$ is a function.

a) f is said to be *surjective* (or a *surjection*, or *onto*) when for every $b \in B$ there exists an $a \in A$ for which $b = f(a)$.

b) f is said to be *injective* (or an *injection*, or *one-to-one*) when $a = a'$ is necessary for $f(a) = f(a')$.

c) When f is both surjective and injective, we say f is *bijective* (or a *bijection*).

Example 3.15. Suppose we consider the function $f(x) = x^2$ for the following choices of domain A and codomain B: $f : A \to B$.

a) $A = [0, \infty)$ and $B = \mathbb{R}$. This makes f injective but not surjective. To prove that it is injective, suppose $a, a' \in A$ and $f(a) = f(a')$. This means $a, a' \geq 0$ and $a^2 = (a')^2$. It follows that $a = a'$ just as in the proof of Proposition 1.3. This proves that f is injective. To see that f is not surjective, simply observe that $-1 \in B$ but there is no $a \in A$ with $f(a) = -1$.

b) Now take $A = \mathbb{R}$ and $B = \mathbb{R}$. For these choices f is still not surjective, but is not injective either, because $f(1) = f(-1)$.

c) Consider $A = \mathbb{R}$ and $B = [0, \infty)$. Compared to b) all we have done is change what we understand the codomain to be. As in b) f is not injective, but now *is* surjective, because every $b \in B$ does have a square root: $a = \sqrt{b}$ is in A and $f(a) = b$.

When we have two functions, f and g, we can sometimes follow one by the other to obtain the composition of f with g.

Definition. Suppose $f : A \to B$ and $g : C \to D$ are functions, and that $B \subseteq C$. Their *composition* is the function $g \circ f : A \to D$ defined by

$$g \circ f(a) = g\left(f(a)\right) \text{ for all } a \in A.$$

Observe that we must have $B \subseteq C$ in order for $g\left(f(a)\right)$ to be defined.

Example 3.16. Let $f : [0, \infty) \to \mathbb{R}$ be defined by $f(x) = \sqrt{x}$ and $g : \mathbb{R} \to \mathbb{R}$ defined by $g(x) = \frac{x^2}{x^2+1}$. Then $g \circ f(x) = \frac{x}{x+1}$, for $x \geq 0$.

Proposition 3.3. *Suppose $f : A \to B$ and $g : C \to D$ are functions with $B \subseteq C$.*

a) If $g \circ f$ is injective then f is injective.

b) If $g \circ f$ is surjective then g is surjective.

c) If $B = C$ and both f and g are bijective, then $g \circ f$ is bijective.

Problem 3.11 Prove the Proposition.

..... prpr

Problem 3.12

a) For part a) of the Proposition, give an example to show that g can fail to be injective.

b) For part b) of the Proposition, give an example to show that f can fail to be surjective.

c) In part c) of the Proposition, show that it is possible for f and g to fail to be bijections even if $g \circ f$ is.

..... comp

Suppose $f : A \to B$ is a function. We think of f as "sending" each $a \in A$ to a $b = f(a) \in B$: $a \to b$. What happens if we try to reverse all these arrows: $a \leftarrow b$; does that correspond to a function $g : B \to A$? The answer is "yes" precisely when f is a bijection, and the resulting function g is called its inverse.

Definition. Suppose $f : A \to B$ is a function. A function $g : B \to A$ is called an *inverse* function to f when $a = g(f(a))$ for all $a \in A$ and $b = f(g(b))$ for all $b \in B$. When such a g exists we say that f is *invertible* write $g = f^{-1}$.

Example 3.17. The usual exponential function $\exp : \mathbb{R} \to (0, \infty)$ is invertible. Its inverse is the natural logarithm: $\exp^{-1}(x) = \ln(x)$.

Example 3.18. Consider $X = (-2, \infty)$, $Y = (-\infty, 1)$ and $f : X \to Y$ defined by $f(x) = \frac{x}{x+2}$. We claim that f is invertible and its inverse $g : Y \to X$ is given by $g(y) = \frac{2y}{1-y}$. To verify this we need to examine both $g \circ f$ and $f \circ g$. First consider $g \circ f$. For any $x \in X$ we can write

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

Since $x + 2 > 0$ it follows that $1 - \frac{2}{x+2} < 1$. Therefore $f(x) \in Y$ for every $x \in X$, and so $g(f(x))$ is defined. Now we can calculate that for every $x \in X$,

$$g(f(x)) = \frac{2\frac{x}{x+2}}{1 - \frac{x}{x+2}} = \frac{2x}{x+2-x} = \frac{2x}{2} = x.$$

Similar considerations apply to $f(g(y))$. For any $y \in Y$, $1 - y > 0$. Since $2y > 2y - 2 = -2(1 - y)$, we see that $g(y) = \frac{2y}{1-y} > -2$, so that $g(y) \in X$. For each $y \in Y$ we can now check that

$$f(g(y)) = \frac{\frac{2y}{1-y}}{\frac{2y}{1-y} + 2} = \frac{2y}{2y + 2(1-y)} = \frac{y}{1} = y.$$

Problem 3.13 Suppose a, b, c, d are positive real numbers. What is the largest subset $X \subseteq \mathbb{R}$ on which $f(x) = \frac{ax+b}{cx+d}$ is defined? Show that $f : X \to \mathbb{R}$ is injective provided $ad \neq bc$. What is its range Y? Find a formula for $f^{-1} : Y \to X$.

..... domex

Problem 3.14 Prove that if an inverse function exists, then it is unique.

..... invu

Problem 3.15 Suppose that $f : A \to B$ and $g : B \to A$ such that $g(f(x)) = x$ for all $x \in A$.

a) Show by example that f need not be surjective and g need not be injective. Show that $f(g(y)) = y$ for all $y \in B$ fails for your example.

b) Show that the following are equivalent.

   1. f is surjective.
   2. g is injective.
   3. $f(g(y)) = y$ for all $y \in B$.

..... invcex

Theorem 3.4. *A function $f : X \to Y$ is a bijection if and only if it has an inverse function.*

Proof. Suppose f is a bijection. For any $y \in Y$, since f is a surjection there exists $x \in X$ with $f(x) = y$. Since f is injective, this x is unique. Thus each $y \in Y$ determines a unique $x \in X$ for which $f(x) = y$. We can therefore define a function $g : Y \to X$ by

$$g(y) = x \text{ for that } x \in X \text{ with } f(x) = y.$$

We claim that g is an inverse function to f. To see that, consider any $x \in X$ and let $y = f(x)$. Then $y \in Y$ and by definition of g we have $x = g(y)$. In other words $x = g(f(x))$ for all $x \in X$. Next consider any $y \in Y$ and let $x = g(y)$. Then $x \in X$ and by definition of g we know $f(x) = y$. Thus $y = f(g(y))$ for all $y \in Y$. We see that g is an inverse function to f.

Now assume that there does exist a function $g : Y \to X$ which is an inverse to f. We need to show that f is a bijection. To see that it is surjective, consider any $y \in Y$ and take $x = g(y)$. Then $f(x) = f(g(y)) = y$. Thus all of Y is in the range of f. To see that f is injective, consider $x, x' \in X$ with $f(x) = f(x')$. Then $x = g(f(x)) = g(f(x')) = x'$. Hence f is indeed injective. $\square$

Problem 3.16 Let $f : (-1, 1) \to \mathbb{R}$ be the function given by $f(x) = \frac{x}{1-x^2}$. Prove that f is a bijection, and that its inverse $g : \mathbb{R} \to (-1, 1)$ is given by by

$$g(y) = \begin{cases} \frac{1-\sqrt{1+4y^2}}{2y} & \text{for } y \neq 0 \\ 0 & \text{for } y = 0. \end{cases}$$

One way to do this is to show f is injective, surjective, and solve $y = f(x)$ for x and see that you get the formula $x = g(y)$. An alternate way is to verify that $g(f(x)) = x$ for all $x \in (-1, 1)$, that $f(g(y)) = y$ for all $y \in \mathbb{R}$, and then appeal to Theorem 3.4.

..... invex2

Images and Preimages of Sets

We usually think of a function $f : A \to B$ as sending elements a of A to elements $b = f(a)$ of B. But sometimes we want to talk about what happens to all the elements of a subset at once. We can think of f as sending a set $E \subseteq A$ to the set $f(E) \subseteq B$ defined by

$$f(E) = \{b \in B : \ b = f(e) \text{ for some } e \in E\}.$$

We call $f(E)$ the *image of E under f*. We can do the same thing in the backward direction. If $G \subseteq B$, the *preimage of G under f* is the set

$$f^{-1}(G) = \{a \in A : \ f(a) \in G\}.$$

We do not need f to have an inverse function to be able to do this! In other words $f^{-1}(G)$ is defined for sets $G \subseteq B$ even if $f^{-1}(b)$ is not defined for elements $b \in B$. The meaning of "$f^{-1}(\cdot)$" depends on whether what is inside the parentheses is a subset or an element of B.

Example 3.19. Consider $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$.
Then

- $f((-2, 3)) = [0, 9)$. In other words the values $0 \le y < 9$ are the only real numbers which arise as $y = f(x)$ for $-2 < x < 3$.
- $f^{-1}([1, 2]) = [-\sqrt{2}, -1] \cup [1, \sqrt{2}]$. In other words the values of x for which $1 \le f(x) \le 2$ are those for which $1 \le |x| \le \sqrt{2}$.
- $f^{-1}([-2, -1)) = \emptyset$. In other words there are no values of x for which $-2 \le f(x) \le -1$.

Warning: This function is neither injective nor surjective — f^{-1} does not exist as a function. However f^{-1} of *sets* is still defined. $f^{-1}(2)$ is *not* defined, but $f^{-1}(\{2\})$ is $(= \{-\sqrt{2}, \sqrt{2}\})$.

Problem 3.17 Consider $f : \mathbb{Z} \to \mathbb{Z}$ defined by $f(n) = n^2 - 7$. What is $f^{-1}(\{k \in \mathbb{Z} : \ k \le 0\})$?

..... `invex`

Problem 3.18 Suppose $f : X \to Y$ and $A, B \subseteq Y$. Prove that

a) $A \subseteq B$ implies $f^{-1}(A) \subseteq f^{-1}(B)$.

b) $f^{-1}(A \cup B) = f^{-1}(A) \cup f^{-1}(B)$.

c) $f^{-1}(A \cap B) = f^{-1}(A) \cap f^{-1}(B)$.

..... `setinv`

Problem 3.19 Suppose $f : X \to Y$, $A, B \subseteq Y$ and $C, D \subseteq X$.

a) Show that it is *not* necessarily true that $f^{-1}(A) \subseteq f^{-1}(B)$ implies $A \subseteq B$.

b) Show that it is *not* necessarily true that $f(C) \cap f(D) = f(C \cap D)$.

c) If f is injective, show that it *is* true that $f(C) \cap f(D) = f(C \cap D)$.

..... `nsetinv`

G Cardinality of Sets

Now that we have summarized all the standard ideas about sets and functions, we are going to use them to discuss the basic idea of the size of a set. Most of us have no trouble with what it would mean to say that a set "has 12 elements" for instance, or even that a set "is finite."

Example 3.20. Let $A = \{a, b, c, \ldots, z\}$ be our usual alphabet, and $B = \{1, 2, 3, \ldots, 26\}$. There is a simple bijection $f(a) = 1$, $f(b) = 2$, $f(c) = 3$, ... in which f assigns to each letter its position in the usual ordering of the alphabet. This f provides a one-to-one pairing of letters with numbers $1 \leq n \leq 26$:

$$\begin{array}{c} \text{a goes with 1} \\ \text{b goes with 2} \\ \text{c goes with 3} \\ \vdots \\ \text{z goes with 26.} \end{array}$$

Of course this works because both sets have exactly 26 elements.

In general when there is a bijection between two sets A and B, that means that A and B have exactly the same number of elements. The bijection provides a way to pair the elements of A with the elements of B, a "one-to-one correspondence" between the two sets. Here is the formal definition for this idea.

Definition. Two sets, A and B, are called *equipotent*[5] if there exists a bijection $f : A \to B$. We will write $A \simeq B$ to indicate that A and B are equipotent.

Two equipotent sets are often said to have *the same cardinality*. The idea of equipotence gives a precise meaning to "have the same number of elements" even when the number of elements is infinite. This agrees with our intuitive idea of "same size" for finite sets, as the example above illustrated. Things becomes more interesting for infinite sets.

Example 3.21. Let $A = (0, 1)$ and $B = (0, 2)$. The function $f : A \to B$ defined by $f(x) = 2x$ is a bijection, as is simple to check. So f provides a one-to-one correspondence between A and B, showing that these two intervals are equipotent.

The same reasoning applied to Problem 3.16 says that $(-1, 1)$ is equipotent to $\mathbb{R}$. In both these examples we have two sets A and B which are "of the same size" in terms of equipotence, even though in terms of geometrical size (length) one is larger than the other.

G.1 Finite Sets

What exactly do we mean by saying as set A is a finite set? One way to express that idea is to say we can count its elements, "this is element #1, this is element #2, ..." and get done at some point, "...this is element #n, and that accounts for all of them." That counting process consists of picking a function $f : \{1, 2, 3, \ldots, n\} \to A$; $f(1)$ is the element of A we called #1, $f(2)$ is the element we called #2, and so on. The fact that we accounted for all of them means that f is surjective. The fact that we didn't count the same element more than once means that f is injective. We can turn our counting idea of finiteness into the following definition.

Definition. A set A is said to be *finite* if it is either empty or equipotent to $\{1, 2, 3, \ldots, n\}$ for some $n \in \mathbb{N}$. A set which is not finite is called *infinite*.

Could it happen that a set is equipotent to $\{1, 2, 3 \ldots, n\}$ as well as to $\{1, 2, 3, \ldots, m\}$ for two *different* integers n and m? We all know that this is *not* possible.

Lemma 3.5 (Pigeon Hole Principle[6])**.** *Suppose $n, m \in \mathbb{N}$ and $f : \{1, 2, 3, \ldots, n\} \to \{1, 2, 3, \ldots, m\}$ is a function.*

[5] There are many different words people use for this: equipollent, equinumerous, congruent, equivalent. Many use the symbol "$\approx$," but since that suggests approximately in some sense, we have chosen "$\simeq$" instead.

[6] Although this name sounds like something out of Winnie the Pooh, its what everyone calls it.

a) If f is injective then $n \leq m$.

b) If f is surjective then $n \geq m$.

We could write a proof of this (using the Well-Ordering Principle of the integers from the next chapter), but it is tricky because what we are proving seems so obvious. Most of the effort would go toward sorting out exactly what we can and cannot assume about subsets of $\mathbb{Z}$. Instead we will take it for granted.

Corollary 3.6. *If $f : \{1, 2, 3, \ldots, n\} \to \{1, 2, 3, \ldots, m\}$ is a bijection ($n, m \in \mathbb{N}$), then $n = m$.*

Problem 3.20 Prove that if there is an injection $f : A \to A$ which is not surjective, then A is not a finite set.

..... `mapfin`

G.2 Countable and Uncountable Sets

To say that "infinite" means "not finite" is simple enough. But now the big question: are all infinite sets equipotent to each other? This is where things get interesting: the answer is "no"! In fact there are infinitely many nonfinite cardinalities, as we will see. But a point to make first is that we are beyond our intuition here. We depend on proofs and counterexamples to know what is true and what is not. Few people can guess their way through this material.

Example 3.22. We will exhibit bijections for the following in class.

- $\{2k : \ k \in \mathbb{N}\} \simeq \{2, 3, 4, \ldots\} \simeq \mathbb{N} \simeq \mathbb{Z}$.
- $\mathbb{R} \simeq (0, 1) \simeq (0, 1]$.

Problem 3.21 Find an injective function $f : \mathbb{N} \to [0, 1)$ (the half-open unit interval).

..... `injint`

The next example shows that it is possible for two infinite sets to *fail* to be equipotent.

Example 3.23. $\mathbb{N}$ and $[0, 1)$ are *not* equipotent! To see this, consider any function $f : \mathbb{N} \to [0, 1)$. We will show that f is not surjective, by identifying a $y \in [0, 1)$ which is not in $\mathrm{Ran}(f)$. We will do this by specifying the decimal representation $y = .d_1 d_2 d_3 \ldots$ where each digit $d_i \in \{0, 1, 2, \ldots, 9\}$. Consider the decimal representation of $f(1)$. We want to choose d_1 to be *different* than the first digit in the decimal representation of $f(1)$. For instance, if $f(1) = .352011\ldots$ we could choose $d_1 = 7$, since $7 \neq 3$. To be systematic about it, if $f(1) = .3\ldots$ take $d_1 = 7$, and otherwise take $d_1 = 3$. If $f(2) = .*3\ldots$ take $d_2 = 7$, and otherwise take $d_2 = 3$. If $f(k) = .*\cdots*3\ldots$ (a 3 in the kth position) take $d_k = 7$, and $d_k = 3$ otherwise. This identifies a specific sequence of digits d_i which we use to form the decimal representation of y. Then $y \in [0, 1)$, and for every k we know $y \neq f(k)$ because their decimal expansions differ in the kth position. Thus y is not in the range of f. Hence f is not surjective.

Definition. A set A is called *countably infinite* when $A \simeq \mathbb{N}$. We say A is *countable* when A is either finite or countably infinite. A set which is not countable is called *uncountable*.

Theorem 3.7 (Cantor's Theorem)**.** *If A is a set, there is no surjection $F : A \to \mathcal{P}(A)$.*

(Note that we are using an upper case "F" for this function to help us remember that $F(a)$ is a subset, not an element, of A.)

Its easy to write down an injection: $F(a) = \{a\}$. Essentially Cantor's Theorem says that $\mathcal{P}(A)$ always has greater cardinality than A itself. So the sets $\mathbb{N}$, $\mathcal{P}(\mathbb{N})$, $\mathcal{P}(\mathcal{P}(\mathbb{N}))$, ... will be an unending list of sets, each with greater cardinality than the one before it! This is why we said above there are infinitely many nonfinite cardinalities. Here is the proof of Cantor's Theorem.

Proof. Suppose $F : A \to \mathcal{P}(A)$ is a function. In other words, for each $a \in A$, $F(a)$ is a subset of A. Consider

$$C = \{a \in A : \ a \notin F(a)\}.$$

Clearly $C \subseteq A$, so $C \in \mathcal{P}(A)$. We claim that there is no $b \in A$ for which $F(b) = C$. If such a b existed then either $b \in C$ or $b \in \tilde{C}$. But $b \in C$ means that $b \notin F(b) = C$, which is contradictory. And $b \in \tilde{C}$ would mean that $b \in F(b) = C$, again a contradiction. Thus $C = F(b)$ leads to a contradiction either way. Hence no such b can exist. □

Problem 3.22 Let $F : \mathbb{R} \to \mathcal{P}(\mathbb{R})$ be defined by

$$F(x) = \begin{cases} \emptyset & \text{if } -1 \le x \le 0, \\ (-x, x^2) & \text{otherwise.} \end{cases}$$

Find the set C described in the proof of Theorem 3.7.

..... Cset

G.3 The Schroeder-Bernstein Theorem

The definition of equipotence gives a precise meaning for saying two sets have the same size. It is pretty easy to express what we mean by a set A being at least as large as B, the existence of an injection $f : B \to A$. This implies that B is equipotent to a subset of A, which seems to agree with our idea of A being at least as large as B. Now, if A is at least as large as B *and* B is at least as large as A, we might naturally expect that A and B must be equipotent. That seems natural, but is it true? Yes, it is – this is the Schroeder-Bernstein Theorem. But to prove it is not a simple task.

Theorem 3.8 (Schroeder-Bernstein Theorem)**.** *Suppose X and Y are sets and there exist functions $f : X \to Y$ and $g : Y \to X$ which are both injective. Then X and Y are equipotent.*

Example 3.24. To illustrate how useful this theorem can be, lets use it to show that $\mathbb{Z} \simeq \mathbb{Q}$. Its easy to exhibit an injection $f : \mathbb{Z} \to \mathbb{Q}$; just use $f(n) = n$. Its also not too hard to find an injection $g : \mathbb{Q} \to \mathbb{Z}$. Given $q \in \mathbb{Q}$, start by writing it as $q = \pm\frac{n}{m}$, where n, m are nonnegative integers with no common factors. Using this representation, define $g(q) = \pm 2^n 3^m$. It is clear that this is also an injection. Theorem 3.8 now implies that $\mathbb{Z} \simeq \mathbb{Q}$. By Example 3.22, it follows that $\mathbb{Q}$ is countable!

Problem 3.23 Use Theorem 3.8 to show $[0, 1] \simeq (0, 1)$.

..... SBappl

Proof. Suppose $f : X \to Y$ and $g : Y \to X$ are both injective. Let $Y_0 = f(X)$, the range of f. Then f is a bijection from X to Y_0, so $X \simeq Y_0$. Although Y_0 is only a subset of Y we will show that there is a bijection $Y_0 \simeq Y$.

Let $B_0 = Y \setminus Y_0$, the part of Y that f misses. Then recursively define

$$B_1 = f(g(B_0)), \ B_2 = f(g(B_1)), \ \ldots, B_{n+1} = f(g(B_n)), \ldots$$

For $n \ge 1$ note that B_n is a subset of Y_0 and therefore disjoint from B_0. Define $h : Y \to Y_0$ by

$$h(y) = \begin{cases} f(g(y)) & \text{if } y \in \cup_0^\infty B_n \\ y & \text{otherwise.} \end{cases}$$

To see that h is surjective consider any $y_0 \in Y_0$. If y_0 does not belong to $\cup_0^\infty B_n$ then $y_0 = h(y_0)$. Suppose y_0 *does* belong to $\cup_0^\infty B_n$. Since $y_0 \in Y_0$ which is disjoint from B_0 it must be that $y_0 \in B_n$ for some $n \ge 1$. By definition of B_n this means $y_0 = f(g(y))$ for some $y \in B_{n-1}$. Thus $y_0 = h(y)$ for some $y \in Y$.

To see that h is injective suppose that $h(y) = h(y')$ for some $y, y' \in Y$. Observe that $h(\cup_0^\infty B_n) = \cup_1^\infty B_n$. Suppose $y \notin \cup_0^\infty B_n$ then $h(y') = h(y) = y$, and therefore $y' \notin \cup_0^\infty B_n$. That implies that $y' = h(y') = h(y) = y$. Next suppose $y \in \cup_0^\infty B_n$. Then it must be that $y' \in \cup_0^\infty B_n$ as well, else $y' = h(y') = h(y) \in \cup_1^\infty B_n$ would be a contradiction. But then $f(g(y)) = h(y) = h(y') = f(g(y'))$. Since $f \circ g$ is injective it follows that $y = y'$.

Having shown that h is a bijection, we know it has an inverse $h^{-1} : Y_0 \to Y$. The composition $h^{-1} \circ f$ is a bijection from X to Y, proving the theorem. $\square$

H Perspective: The Strange World at the Foundations of Mathematics

We close this chapter with a brief look at a couple of issues from the foundations of mathematics.

H.1 The Continuum Hypothesis

The collection of possible cardinalities of sets form what are called the *cardinal numbers*, a number system that includes the nonnegative integers as well as the infinite cardinals. The possible cardinalities of finite sets are just the nonnegative integers $0, 1, 2, 3, \ldots$. After that come the cardinalities of infinite sets — we have seen that they are not all the same. We know that both $\mathcal{P}(\mathbb{N})$ and $\mathbb{R}$ have strictly larger cardinality than $\mathbb{N}$; both are uncountable. (In fact it can be shown that $\mathcal{P}(\mathbb{N})$ and $\mathbb{R}$ are equipotent.) A few infinite cardinal numbers have symbols associated with them. For instance the cardinality of $\mathbb{N}$ (or any other countably infinite set) is usually denoted $\aleph_0$, and the cardinality of $\mathbb{R}$ is sometimes denoted $\mathfrak{c}$. We can say with confidence that $\aleph_0 < \mathfrak{c}$, because there is an injection from $\mathbb{N}$ into $\mathbb{R}$ but we know that $\mathbb{R}$ is uncountable. During the latter nineteenth and early twentieth centuries mathematicians pondered the question of whether there could be a set K with cardinality strictly between these two: $\aleph_0 <$ cardinality of $K < \mathfrak{c}$. *The Continuum Hypothesis* is the assertion that there is no such K, that $\mathfrak{c}$ is the next largest cardinality after $\aleph_0$. Is the Continuum Hypothesis true? The answer (if you can call it that) is more amazing than you would ever guess. We will leave it hanging over a cliff until the end of the next chapter!

H.2 Russell's Paradox

We have mentioned that a logical swamp awaits those who wander too far along the path of forming sets of sets of sets of ... This is a real mind-bender — take a deep breath, and read on.

> Let $\mathcal{S}$ be the set of *all* sets. It contains sets, and sets of sets, and sets of sets of sets and so on. If A and B are sets then $A, B \in \mathcal{S}$, and it is possible that $A \in B$, e.g. $B = \{A, \ldots\}$. In particular we can ask whether $A \in A$. Consider then the set
>
> $$\mathcal{R} = \{A \in \mathcal{S} : \ A \notin A\}.$$
>
> Now we ask the "killer" question: is $\mathcal{R} \in \mathcal{R}$? If the answer is "yes," then the definition of $\mathcal{R}$ says that $\mathcal{R} \notin \mathcal{R}$, meaning the answer is "no." And if the answer is "no," the definition of $\mathcal{R}$ says that $\mathcal{R} \in \mathcal{R}$, meaning the answer is "yes." Either answer to our question is self-contradictory.

What is going on here? We seem to have tied ourselves in a logical knot. This conundrum is called *Russell's Paradox*. A paradox is not quite the same as a logical contradiction or impossibility. Rather it is an *apparent* contradiction, which typically indicates something wrong or inappropriate about our reasoning. Sometimes a paradox is based on a subtle/clever misuse of words. Here we are allowing ourselves to mingle the ideas of set and element too freely, opening the door to the above paradoxical discussion that Bertrand Russell brewed up. Historically, Russell's Paradox showed the mathematical world that they had not yet fully worked out what set theory actually consisted of, and sent them back to the task of trying to decide which statements about sets are legitimate and which are not. This led to the development of axioms which govern formal set theory (the Zermelo-Fraenkel axioms) which prevent the misuse of language that Russell's paradox illustrates. (In the next chapter we will see what we mean by "axioms" in the context of the integers.) This is a difficult

topic in mathematical logic which we cannot pursue further here. Our point here is only that careless or sloppy use of words can tie us in logical knots, and a full description of what constitutes appropriate usage is a major task.

These things do *not* mean that there is something wrong with set theory, and that we are just choosing to go ahead in denial. Rather they mean that we have to be careful not to be too cavalier with the language of sets. For virtually all purposes, if we limit ourselves to two or three levels of elements, sets, sets of sets, and sets of sets of sets (but stop at some point) we will be able to sleep safely, without fear of these dragons from the logical abyss. See Chapter 4, Section E for a bit more on these topics.

Chapter 4

The Integers

The number systems we use most frequently are the natural numbers ($\mathbb{N}$), the integers ($\mathbb{Z}$), the real numbers ($\mathbb{R}$), and the complex numbers ($\mathbb{C}$). Each of these has properties that the others don't. In this chapter we are going to focus on the essential properties of the integers. But as we work through that we will observe the contrasts with other number systems. In the final section we will introduce the integers mod m ($\mathbb{Z}_m$), a finite number system that can be remarkably useful.

A Properties of the Integers

We begin in this section by stating some of the most basic properties of the integers. These properties are of two types, algebraic properties and order properties.

A.1 Algebraic Properties

All of the number systems mentioned above have two basic algebraic operations: addition $+$ and multiplication $\cdot$. The algebraic properties of the integers are shared by the other number systems we mentioned. We describe them below for $\mathbb{Z}$ but you should observe that all of these hold if $\mathbb{Z}$ is replaced by $\mathbb{R}$ or $\mathbb{C}$ as well. The standard terminology describing each property is given in parentheses following the statement of the property itself.

(A1) $a + b \in \mathbb{Z}$ for all $a, b \in \mathbb{Z}$ (closure under addition).

(A2) $a + b = b + a$ for all $a, b \in \mathbb{Z}$ (commutative law of addition).

(A3) $a + (b + c) = (a + b) + c$ for all $a, b, c \in \mathbb{Z}$ (associative law of addition).

(A4) There exists an element $0 \in \mathbb{Z}$ with the property that $0 + a = a$ for all $a \in \mathbb{Z}$ (existence of additive identity).

(A5) For each $a \in \mathbb{Z}$ there exists an element $-a$ with the property that $a + (-a) = 0$ (existence of additive inverses).

(M1) $a \cdot b \in \mathbb{Z}$ for all $a, b \in \mathbb{Z}$ (closure under multiplication).

(M2) $a \cdot b = b \cdot a$ for all $a, b \in \mathbb{Z}$ (commutative law of multiplication).

(M3) $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ for all $a, b, c \in \mathbb{Z}$ (associative law of multiplication).

(M4) There exists an element $1 \in \mathbb{Z}$ with the property that $1 \cdot a = a$ for all $a \in \mathbb{Z}$ (existence of multiplicative identity).

(D) $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$ for all $a, b, c \in \mathbb{Z}$ (distributative law).

Notice first of all that the properties make no reference to an operation of subtraction of two integers: $b-a$. The negation $-a$ of a single integer is described in (A5). We understand subtraction as shorthand for the combined operation of first taking the element $-a$, and then adding it to b:

$$\text{“}b-a\text{”} = b+(-a).$$

In that way the properties of subtraction can be deduced from the above properties.

Notice also that the properties make no mention of division: a/b. This time the reason is different: *there is no fully-defined operation of division for the integers.* We can always take two inegers a, b and add them to get *another integer* $a+b$; that's what closure under addition, property (A1), is about; the operation of addition is always defined and always produces *another integer.* The same holds for multiplication; as property (M1) says. But we can *not* always divide any pair of integers. For instance if $a=3$ and $b=2$, there is no integer $3/2$. You may say "but wait, $3/2$ is defined; it's the number we call 1.5." Well yes, but that is not another integer. You can only make sense of $3/2$ by moving into a different number system like $\mathbb{Q}$ or $\mathbb{R}$ which *does* have operations of division. You can't form $3/2$ and *stay within the integers.* The word "closure" in (A1) and (M1) means that those operations always produce results within the integers. You can't do that with division; it would force you outside the integers in general. So the integers are *not* closed under division. We can talk about doing something like division within the integers if we allow a remainder; we will come to that in Theorem 4.5 below. But the usual operation of division is *undefined* within the integers proper.

Lets consider the properties above with the natural numbers $\mathbb{N}$ instead of $\mathbb{Z}$. We see that they do not satisfy property (A4); the additive identity 0 is not a natural number, and there is no other natural number with the property of (A4). Since 0 does not exist in $\mathbb{N}$ we can't even consider property (A5). We are starting to see how different number systems have different properties. The natural numbers do not have properties (A4) and (A5), while the integers do. The integers have no operation of division, while $\mathbb{R}$ and $\mathbb{C}$ do. Although the real numbers and complex numbers (as well as intergers mod m) do share all the properties of $\mathbb{Z}$ we have listed so far. There is a property of the integers that they do *not* share, and that has do do with the order relation $<$.

A.2 Properties of Order

You know from the last chapter what we mean by a relation on a set $\mathbb{Z}$. Here are the fundamental properties of the order relation $<$.

(O1) For any $a, b \in \mathbb{Z}$ one and only one of the following is true: $a<b$, $a=b$, $b<a$ (total ordering property).

(O2) If $a<b$ and $b<c$ then $a<c$ (order is transitive).

(O3) If $a<b$ and $c \in \mathbb{Z}$ then $a+c<b+c$ (translation invariance).

(O4) If $a<b$ and $0<c$ then $c \cdot a < c \cdot b$ (multiplicative invariance).

Again, these are not unique to $\mathbb{Z}$. They also hold for $\mathbb{R}$. But now we come to the special property of $\mathbb{Z}$ which distinguishes it from $\mathbb{R}$ and other "higher" number systems: the *Well-Ordering Principle.*

(W) If S is a nonempty subset of $\{a \in \mathbb{Z} : \ 0<a\}$, then there exists $s_* \in S$ with the property that $s_* < s$ for all $s \in S$ other than $s = s_*$.

First, notice that $\{a \in \mathbb{Z} : \ 0<a\}$ is just a description of the natural numbers, so we could view (W) as a property of $\mathbb{N}$ which is inherited by $\mathbb{Z}$. What (W) says is that any set S of positive integers that has at least one element always has a smallest element s_*. For instance there is a smallest prime number (2), a smallest perfect square (1), a smallest prime p such that the next prime is $p+10$ ($p=139$). We will see some of the consequences of the Well-Ordering Principle in Section A.5, and in the next section we will observe that neither $\mathbb{R}$ nor $\mathbb{C}$ satisfy this property.

A.3 Comparison with Other Number Systems

We have noted that $\mathbb{N}$ does not satisfy properties (A4) or (A5), but does satisfy all the other properties of the integers. There are other settings in which both multiplication and addition are defined but *fail* to satisfy all the algebraic properties above. For example, consider the set $\mathbb{M}_{2\times 2}$ of all 2×2 matricies. We can add, multiply and negate such matricies: if

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \text{ and } B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix},$$

then

$$A + B = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix}, \quad AB = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}, \quad -A = \begin{bmatrix} -a_{11} & -a_{12} \\ -a_{21} & -a_{22} \end{bmatrix}.$$

The role of 0 is played by the matrix of all zeros $\left[\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right]$, and the role of 1 is played by the identity matrix $I = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$. All the algebraic properties listed above hold for matricies, *except* (M2); matrix multiplication is not commutative. Problem 4.1 will point out another familiar setting in which our algebraic properties are not all satisfied. So as seemingly obvious as all those properties are, they are not universal!

Now lets consider what the Well-Odering Principle would say in the context of the real numbers $\mathbb{R}$. Consider $S = (1, +\infty) = \{x \in \mathbb{R} : \ 1 < x\}$. This is certainly a nonempty subset of $\{a \in \mathbb{R} : \ 0 < a\}$. So if the Well-Ordering Principle were true for $\mathbb{R}$, There would exist a smallest number in $S = (1, +\infty)$. But that is not true; for any number $s \in S$ we can always find an even smaller one, like $(1 + s)/2$ which is halfway between 1 and s. There are numbers like $s_* = 1/2$ or $s_* = 1$ with the property that $s_* < x$ for all $x \in S$, but none of them belong to S themselves, as the Well-Ordering Principle would require. So we see that (W) is not true for $\mathbb{R}$. If you take a course in advanced calculus or analysis, you will learn that $\mathbb{R}$ has a different property in its place, called "completeness," which is the foundation for limits and the other fundamental constructions of calculus.

What about the complex numbers? Problem 4.3 explains that there does not exist an order relation $<$ for the complex numbers satisfying (O1)—(O4), so it's pointless to even consider (W) in the case of $\mathbb{C}$!

Problem 4.1 You are familiar with adding and subtracting vectors: if $\vec{x} = (x_1, \ldots, x_n)$ and $\vec{y} = (y_1, \ldots, y_n)$, then

$$\vec{x} \pm \vec{y} = (x_1 \pm y_1, \ldots, x_n \pm y_n).$$

In general there is no way to multiply vectors, except in the special case of $n = 3$. In that case we have the cross product:

$$\vec{x} \times \vec{y} = (x_2y_3 - x_3y_2, x_3y_1 - x_1y_3, x_1y_2 - x_2y_1).$$

(It is usually defined in terms of $\vec{i} = (1, 0, 0)$, $\vec{j} = (0, 1, 0)$ and $\vec{k} = (0, 0, 1)$, but if you write out whatever description of the cross product you are familiar with you will see that it is equivalent to what we wrote above.) Several of the axioms fail for the cross product.

a) Show (by example) that axioms (M2) and (M3) are false for $\times$.

b) Show that (M4) is false for $\times$. I.e. there is no vector $\vec{e}$ that has the property of the number 1, namely that $\vec{e} \times \vec{x} = \vec{x}$ for all $\vec{x}$.

c) Show (by example) that it is possible for $x \times y = (0, 0, 0)$ even if $x \neq (0, 0, 0)$ and $y \neq (0, 0, 0)$.

..... cross

A.4 Further Properties of the Integers

We have not offered proofs for the properties of the integers we have listed so far. We have just been describing the integers as we know them from our past experience, not trying to explain *why* these properties must be true. There are many additional properties that we have not mentioned yet, like the following.

- $0 < 1$.
- $0 \cdot 1 = 0$
- $(-1) \cdot (-1) = 1$.
- 1 is the smallest positive integer.

At what point can we start giving reasons (proofs!) for additional properties based on those we have already described? Could you prove the algebraic or order proerties we listed in the preceeding sections? Can you give reasons for the properties listed as bullets above? When are we done describing the integers and able to start proving things about it? To answer the questions we need to make a distinction between those properties which *define* the integers and so are assumed without proof, and those which are consequences of the defining properties and so ought to be proven. Those properties which comprise the defintion of the integers, and so are assumed, not proven, are called the *axioms* of the integers. In other words axioms for the integers consists of a collection of basic properties from which all other properties can be derived (proven) logically. There is more than one way to identify a set of such basic properties for the integers. One such set of axiomatic properties consists of those we have listed in the preceeding sections: (A1)—(A5), (M1)—(M4), (D), (O1)—(O4), (W), and one other: the axiom of *nontriviality*:

(N) $0 \neq 1$

This seems like a rather silly, obvious observation. But in fact it cannot be proven from the other axioms, and without it things like $0 < 1$ are unprovable. (It will be used in the proof of part 6) of Proposition 4.1 below.) So it needs to be assumed!. All other properties of the integers (like our bullets just above) should, in principle[1], be provable from the axioms. We don't intend to go through the (long) endeavor of writing proofs of all the well-known properties of the intergers. But we will look at a few, collected as the following proposition, just to get an idea of how this kind of proof would go. Observe that parts 6), 2), 4), and 8) (respectively) provide the justification for our bullets above.

Proposition 4.1.

1) 0 *and* 1 *are unique. (In other words,* 0 *is the only integer satisfying property (A4) and* 1 *is the only integer satisfying (M4).)*

2) For all $a \in \mathbb{Z}$, $0 \cdot a = 0$.

3) For all $a \in \mathbb{Z}$, $-(-a) = a$.

4) For all $a, b \in \mathbb{Z}$, $(-a) \cdot b = -(a \cdot b)$.

5) $a < 0$ *if and only if* $0 < -a$.

6) $0 < 1$.

7) If $c \cdot a = c \cdot b$ *and* $c \neq 0$, *then* $a = b$.

8) If $0 < a$ *then either* $1 < a$ *or* $1 = a$.

We might call 7) the "law of cancellation." The reason it is true is *not* because we can divide both sides by c; there is no fully defined notion of division in $\mathbb{Z}$, as we have already pointed out. Rather the reason is that it is a consequence of 2), as we will see in the proof.

Proof. For 1), to prove that 0 is unique, we suppose $z \in \mathbb{Z}$ also has the property that $z + a = a$ for all $a \in \mathbb{Z}$. Using this we need to show why $0 = z$ must be true. Consider $0 + z$. By (A4) with $a = z$ we find that $0 + z = z$. On the other hand using (A2) we know that $0 + z = z + 0$. Our hypothesis regarding z, with

[1] Actually this is not quite true; see the discussion in Section E for a glimpse of what goes wrong. But for all the properties that we will use, it's true.

$a = 0$, implies that $z + 0 = 0$. Thus $0 + z$ must equal both 0 and z, and therefore $0 = z$. Thus there is only one 0 satisfying (A4). The uniqueness of 1 in (M4) is proved the same way.

For 2), consider any $a \in \mathbb{Z}$. We want to prove that $0 \cdot a = 0$.

$$\begin{aligned}
0 &= a + (-a) && \text{by (A5)}\\
&= 1 \cdot a + (-a) && \text{by (M4)}\\
&= a \cdot 1 + (-a) && \text{by (M2)}\\
&= a \cdot (0 + 1) + (-a) && \text{by (A4)}\\
&= [(a \cdot 0) + (a \cdot 1)] + (-a) && \text{by (D)}\\
&= [(0 \cdot a) + (1 \cdot a)] + (-a) && \text{by (M2)}\\
&= [(0 \cdot a) + a] + (-a) && \text{by (M4)}\\
&= (0 \cdot a) + [a + (-a)] && \text{by (A3)}\\
&= (0 \cdot a) + 0 && \text{by (A5)}\\
&= 0 + (0 \cdot a) && \text{by (A2)}\\
&= 0 \cdot a && \text{by (A3)}
\end{aligned}$$

Thus $0 \cdot a = 0$ as claimed.

Proofs of 3)–5) are left as problems. We continue with proofs of 6)–8) under the assumption that 1)—5) have been established and so can be used in the proofs for the subsequent parts.

For 6), by virtue of (O1) we need to show that neither $1 < 0$ nor $0 = 1$ is possible. We know by the nontrivality axiom that $0 \neq 1$. So suppose $1 < 0$. Since $1 = -(-1)$ by 3), we deduce from 5) that $0 < -1$. Using this in (O4) we have that

$$(-1) \cdot 0 < (-1) \cdot (-1). \tag{4.1}$$

By (M2) and 2) $(-1) \cdot 0 = 0 \cdot (-1) = 0$; and by 4)

$$\begin{aligned}
(-1) \cdot (-1) &= -[1 \cdot (-1)]\\
&= -[-1] && \text{by (M4)}\\
&= 1 && \text{by 4).}
\end{aligned}$$

Making these substitutions in (4.1) we find that $0 < 1$. But according to (O1) both $1 < 0$ and $0 < 1$ cannot be true simultaneously. This shows that $1 < 0$ is not possible.

For 7), since $c \neq 0$ we know from (O1) that either $0 < c$ or $c < 0$. Suppose first that $0 < c$. From (O1) again we know that one of the following must be true: $a < b$, $a = b$, or $b < a$. If $a < b$, then multiplying both sides by c and using (O4), we deduce that $c \cdot a < c \cdot b$, contrary to our hypothesis that $c \cdot a = c \cdot b$. If $b < a$, then we deduce in the same way that $c \cdot b < c \cdot a$, again contrary to our hypotheses. This proves 7) in the case that $0 < c$. If $c < 0$, then we know by 5) that $0 < -c$. By 4) it follows that

$$(-c) \cdot a = -(c \cdot a) = -(c \cdot b) = (-c) \cdot b.$$

Since $-c > 0$ the same argument we just gave implies that $a = b$. This completes the proof of 7).

For 8), consider

$$S = \{a \in \mathbb{Z} : \ 0 < a\}$$

and let s_* be the smallest element of S, as guaranteed by (W). Since $1 \in S$ by 6), we just need to show that $s_* < 1$ is not possible. To prove this by contradiction, suppose $s_* < 1$. Since we also know $0 < s_*$, we can use (O4) and 2) to deduce that $0 < s_* \cdot s_*$ and $s_* \cdot s_* < s_*$. Therefore $s_* \cdot s_* \in S$. But the property of s_* from (W) does not allow $s_* \cdot s_* < s_*$. This contradiction shows that $s_* < 1$ is not possible, and therefore $s_* = 1$. This completes the proof. $\square$

Problem 4.2 Write proofs for parts 3)–5) of the proposition. (You may use any of the earlier parts in your proofs of later parts.)

..... basics

Problem 4.3 Consider $\mathbb{C}$ and suppose that there *was* a relation $<$ satisfying (O1)—(O4). Since $i \neq 0$ (else $i^2 = 0$, not -1) property (O1) would require that either $0 < i$ or $i < 0$. Show how using either of these, (O4) leads to a contradiction to $-1 < 0 < 1$. Proposition 4.1 showed that $-1 < 0 < 1$ for the integers, but the same proof applies in *any* number system satisfying all the algebraic and order properties above. So even if there were some bizare order realtion $<$ on $\mathbb{C}$ that is incompatible with the usual order on $\mathbb{Z}$, it still would have to be true that $-1 < 0 < 1$. You can therefore take it for granted that $-1 < 0 < 1$, and anything that violates that is indeed a contradiction.

..... nonord

Problem 4.4 Use the fact that 1 is the smallest positive integer to prove that if $m, n \in \mathbb{N}$ and $n|m$ then $n \leq m$.

..... small1

A.5 A Closer Look at the Well-Ordering Principle

The Well-Ordering Principle is probably the least familiar of the axioms for $\mathbb{Z}$. In this subsection we are going to develop two of its important consequences, neither of which will be big surprizes to you. However they are properties that do *not* follow from the other axioms alone, but really depend on well-ordering. That makes them peculiar to the integers. In particular neither the real numbers nor the complex numbers have properties like those we prove below.

The Principle of Mathematical Induction

Recall the technique of proof by induction described in earlier chapters. We said that in order to prove a statement of the form

$$\mathrm{P}(n) \text{ for all } n \in \mathbb{N},$$

we could a) prove that $\mathrm{P}(1)$ is true, and b) prove that whenever $\mathrm{P}(n)$ is true then $\mathrm{P}(n+1)$ is true. The validity of this proof technique rests on the Well-Ordering Principle. To see the connection consider the set

$$T = \{n \in \mathbb{N} : \ \mathrm{P}(n) \text{ is true}\}.$$

The method of proof by induction consists of showing that T has these properties: a) $1 \in T$, and b) if $n \in T$ then $n+1 \in T$. From these we are supposed to be able to conclude that $T = \mathbb{N}$. So at the heart of proof by induction is the following fact.

Theorem 4.2 (Induction Principle)**.** *Suppose $T \subseteq \mathbb{N}$ contains 1 and has the property that whenever $n \in T$, then $n+1 \in T$ as well. Then $T = \mathbb{N}$.*

The proof is by contradiction; if $T \neq \mathbb{N}$ then $S = \mathbb{N} \setminus T$ would be a nonempty subset of $\mathbb{N}$. The Well-Ordering Principle would say that S has a smallest element, s_*. Let $n = s_* - 1$. Since $n < s_*$ $n \notin S$ and therefore n *is* in T. The picture is like this:

$$1 \in T, \ 2 \in T, \cdots, n \in T, \quad n+1 = s_* \notin T, \ \cdots$$

But this is contrary to the hypothesis that $n \in T$ implies $n+1 \in T$. Here is the proof written out with the rest of the details.

Proof. Consider $S = \mathbb{N} \setminus T$. Our goal is to show that $S = \emptyset$. Suppose on the contrary that S is not empty. Since $S \subseteq \mathbb{N}$ we know that $0 < n$ for all $n \in S$. Thus the Well-Ordering Principle applies to S, and so there is a smallest element s_* of S. In particular, $s_* \notin T$. Since $S \subseteq \mathbb{N}$ we know $1 \leq s_*$. But since $1 \in T$, we know that $1 \notin S$ and therefore $2 \leq s_*$. Consider $n = s_* - 1$. Then $1 \leq n$ and since $n < s_*$, we know $n \notin S$. Therefore $n \in T$. By the hypotheses on T, it follows that $n+1 \in T$ as well. But $n+1 = s_*$ which we know is in S, which is contrary to $n + 1 \in T$. This contradiction proves that S must in fact be empty, which means that $T = \mathbb{N}$. □

The Well-Ordering Principle can be used to establish a more general version in which the set S can include negative integers, provided it has a lower bound. A second generalization considers sets with an upper bound, and guarantees the existence of a largest element.

Theorem 4.3 (Generalized Well-Ordering Principle)**.** *Suppose S is a nonempty subset of $\mathbb{Z}$.*

a) If S has a lower bound (i.e. there exists $k \in \mathbb{Z}$ with $k \leq s$ for all $s \in S$) then S has a least element s_*

b) If S has an upper bound (i.e. there exists $m \in \mathbb{Z}$ with $s \leq m$ for all $s \in S$) then S has a largest element.

We will write a proof of a) and leave the proof of b) as a problem. The idea for a) is to reduce the situation to one for which standard well-ordering applies. Our S is a subset of $\{a \in \mathbb{Z} : \ k \leq a\}$, but to apply standard well-ordering it needs to be a subset of $\mathbb{N}$. So what we will do is construct a new set S' by adding $1 - k$ to every element of S, shifting everything up so that S' is a subset of $\{a \in \mathbb{Z} : \ 1 \leq a\} = \{a \in \mathbb{Z} : \ 0 < a\}$. Standard well-ordering will then apply to S'. We take the least element s'_* of S' and subtract $1 - k$ back off. The result should be the least element of S.

Proof of a). Let S be nonempty with a lower bound k. Define

$$S' = \{s + 1 - k : \ s \in S\}.$$

Since S is nonempty, S' is also nonempty. For any $s' \in S'$ there is some $s \in S$ such that $s' = s + 1 - k$. Since $s \geq k$ it follows that $s' \geq 1 > 0$. Therefore $S' \subseteq \{a \in \mathbb{Z} : \ 0 < a\}$. The Well-Ordering Principle applies to S', so there exists $s'_* \in S'$ with $s'_* \leq s'$ for all $s' \in S'$. Since $s'_* \in S'$ it follows that there is $s_* \in S$ with $s'_* = s_* + 1 - k$. For any $s \in S$, we know that $s + 1 - k \in S'$ and therefore $s'_* \leq s + 1 - k$. This implies that

$$s_* = s'_* + k - 1 \leq s.$$

Thus s_* is the least element of S. □

Problem 4.5 Write a proof of part b) of Theorem 4.3. You will need to do something similar to the proof of a), building a new set S' so that the upper bound of S turns into a lower bound of S'. Apply the Well-Ordering Principle to S' and convert back to a statement about S, like what we did for part a).
..... `ub`

Theorem 4.4 (Generalized Induction Principle)**.** *Suppose $T \subseteq \mathbb{Z}$ contains n_0, and that $n \in T$ implies $n + 1 \in T$. Then T contains all $n \in \mathbb{Z}$ with $n \geq n_0$.*

Problem 4.6 Write a proof of the Generalized Induction Principle.
..... `GIPf`

Problem 4.7 Formulate and prove a version of the induction principle for strong induction.
..... `strongind`

The Division Theorem

Although we can't always divide one integer we *can* do something like division if we also allow a remainder.

Theorem 4.5. *Suppose $m, n \in \mathbb{Z}$ and $0 < n$. There exist unique $q, r \in \mathbb{Z}$ so that*

$$m = q \cdot n + r \text{ and } 0 \leq r < n.$$

Everyone knows this; n "goes into" m some maximum number of times q (the *quotient*), leaving a *remainder* r. The theorem says you can always do this, and that q and r are unique, provided we insist that $0 \leq r < n$. For instance if $m = 17$ and $n = 5$, then $q = 3$ and $r = 2$:

$$17 = 3 \cdot 5 + 2 \text{ and } 0 \leq 2 < 5.$$

The theorem is sometimes called the "division algorithm," but that is really a misnomer. An algorithm is a procedure or sequence of steps that accomplishes a specific task. The theorem above offers no algorithm for finding q and r (although you can probably think of a procedure without too much trouble).

You don't need to see a proof of the division theorem to know it is true. But we *are* going to write out a proof, because we want to exhibit its connection to the Well-Ordering Principle. (It can also be proven by strong induction. But, as we just saw, the Induction Principle is a manifestation of well-ordering, so however you prove it it comes down to well-ordering in the end.)

Proof. Define the set

$$R = \{b \in \mathbb{Z} : \ b \geq 0 \text{ and there exists an integer } k \text{ for which } m = k \cdot n + b\}.$$

Our goal is to apply the Generalized Well-Ordering Principle to show that R has a smallest element, which will be the r that we want.

We need to verify that R satisfies the hypotheses of the Generalized Well-Ordering Principle. First, by definition every $b \in R$ satisfies $0 \leq b$. So 0 is a lower bound for R. Second, we need to show that R is nonempty. We do that by considering two cases: $0 \leq m$ or $m < 0$. If $0 \leq m$, then $m = 0 \cdot n + m$ and so $m \in R$. Now consider the case of $m < 0$. Since $m = mn + (n-1)(-m)$ and $(n-1)(-m) \geq 0$, we see that $(n-1)(-m) \in R$. Thus R is nonempty in either case.

We can now appeal to the Generalized Well-Ordering Principle to deduce that there is an $r \in R$ which is smallest possible. By definition of R we know that $0 \leq r$ and there is $q \in \mathbb{Z}$ for which $m = q \cdot n + r$. To finish we need to show that $r < n$. We do this by contradiction. Suppose to the contrary that $n \leq r$. Then consider $b = r - n$. Since $n \leq r$ it follows that $0 \leq b$. Since $m = (q+1) \cdot n + (r - n)$, we see that $b = r - n \in R$. But $r - n < r$, contradicting the fact that r is the smallest element of R. This contradiction proves that $r < n$, finishing the proof of existence.

Now we need to prove uniqueness. Assume that there is a second pair of integers q', r' with $m = q' \cdot n + r'$ and $0 \leq r' < n$. We want to show that $q = q'$ and $r = r'$. Since $qn + r = q'n + r'$, it follows that $(q - q')n = r' - r$. But $0 \leq r, r' < n$ implies that $-n < r' - r < n$, and the only multiple of n in that interval is $0 \cdot n = 0$. Thus $r = r'$, and therefore $(q - q')n = 0$. Since $0 < n$ we are forced to conclude that $q = q'$. $\square$

As elementary as the division theorem seems it is an important tool in writing proofs about $\mathbb{Z}$. We will see that in several places in the rest of this chapter.

Problem 4.8 Find the quotient and remainder of the division theorem for the following

a) $m = 2297$, $n = 172$.

b) $m = 44633$, $n = 211$.

c) $m = 64016$, $n = 253$.

..... divex

B Greatest Common Divisors

If in the Division Theorem the remainder turns out to be $r = 0$, then $m = qn$ for some integer q, which means that n divides m as defined on page 15: $n|m$. Some elementary properties of divisibility are the following.

- $1|n$ for all $n \in \mathbb{Z}$.
- $n|0$ for all $n \in \mathbb{Z}$.
- $0|n$ if and only if $n = 0$.
- The following are equivalent:
  - i) $n|m$,
  - ii) $n|(-m)$,
  - iii) $(-n)|m$.

Problem 4.9 Suppose $k, m, n \in \mathbb{Z}$. Prove that if k divides m and m divides n, then k divides n.
..... `divtrans`

Problem 4.10

a) Suppose $m, n \in \mathbb{N}$. Explain why $n|m$ impliies $n \leq m$

b) Suppose $n, m \in \mathbb{Z}$ and that both $n|m$ and $m|n$. Prove that either $m = n$ or $m = -n$.

..... `divpm`

Problem 4.11 Prove that if m divides a and b, then m divides $a \pm b$ and ac (for any integer c).
..... `division`

Many of the most interesting properties of the integers, such as those involving prime numbers, are due to the fact that division (with zero reminder) is not always possible. The notion of the greatest common divisor of two integers m, n is an important tool for discussing various divisibility related issues. Because divisibility does not care about $\pm$ signs, we will state the definition just for natural numbers.

Definition. Suppose $a, b \in \mathbb{N}$. We say $g \in \mathbb{N}$ is the *greatest common divisor* of a and b, and write $g = \gcd(a, b)$, if both the following hold:

a) g divides both a and b,

b) whenever $d \in \mathbb{N}$ divides both a and b then d also divides g.

This definition is unexpected in a couple ways. First, it does not refer to factorizations of n and m, even though that is probably how you are used to thinking of greatest common divisors. (See Example 4.2 below.) Second, part b) of the definition doesn't interpret "greatest" as $d \leq g$ but rather as $d|g$. Perhaps that seems strange to you. However in other algebraic systems we won't always have an order relation. This division interpretation of "greatest" turns out to be the way to generalize the definition in those settings. For instance, when we replace n and m by polynomials $p(x), q(x)$ in the next chapter, the interpretation of "greater" as $p(x) \leq q(x)$ would be inadequate, because for polynomials it is possible that neither $p(x) \leq q(x)$ nor $q(x) \leq p(x)$ is true — not all polynomials are comparable in the sense of inequality. But the definition based on divisors dividing other divisors will work just fine. We are using the more general definition here for integers as well. Just remember that when you are asked to prove something about $\gcd(a, b)$, don't revert to a naive interpretation of "greatest," but use the definition above.

Notice that the definition said "the" rather than "a" greatest common divisor. The next theorem justifies this presumption of uniqueness. The proof in fact also yields an unexpected fact!

Theorem 4.6. *For $a, b \in \mathbb{N}$ the greatest common divisor of a and b exists and is unique. Moreover, there exist $\alpha, \beta \in \mathbb{Z}$ for which* $\gcd(a, b) = \alpha a + \beta b$.

Example 4.1. Looking ahead to Example 4.2, $\gcd(8613, 2178) = 99$. Observe that

$$99 = (-1) \cdot 8613 + 4 \cdot 2178,$$

showing that in this case the theorem does hold, with $\alpha = -1$ and $\beta = 4$.

In this example we have given no hint as to how we found the values of α and β. The proof of the theorem establishes theoretically that they exist, but that is not a very practical way to actually find them. Problem 4.14 below suggests an algorithm for calculating the α and β.

Proof. We first prove uniqueness. Suppose both g and g' satisfy the definition of $\gcd(a, b)$. Then g divides g' and g' divides g. Since they are both positive, Problem 4.10 implies that $g = g'$ or $g = -g'$. But $g, g' \in \mathbb{N}$ implies that $g \neq -g'$. Therefore $g = g'$, proving uniqueness.

The existence will be a consequence of the proof of the last part. For that, define

$$D = \{k \in \mathbb{N} : \ k = \alpha a + \beta b \text{ for some } \alpha, \beta \in \mathbb{Z}\}.$$

Since $a, b \in D$ we know that D is nonempty, so the Well-Ordering Principle implies that there is a smallest element d of D. In particular there are integers α_0, β_0 for which

$$d = \alpha_0 a + \beta_0 b. \tag{4.2}$$

We claim that d divides both a and b. Suppose d did not divide a. Then there would exist a quotient q and remainder $0 < r < d$ with

$$a = qd + r.$$

We can rearrange this as

$$r = a - qd = (1 - q\alpha_0)a + (-q\beta_0)b.$$

But this implies that $r \in D$ and is smaller than d, contrary to the minimality of d. Thus it must be that d divides a. A similar argument shows that d also divides b.

It follows from (4.2) that every common divisor of a and b must divide d. Since $d \in \mathbb{N}$ and is itself a common divisor of a and b, we see that d satisfies the definition of $d = \gcd(a, b)$ proving existence of the greatest common divisor. Now (4.2) establishes the final claim of the theorem. □

Problem 4.12 Suppose $n, m \in \mathbb{N}$ and $n|m$. Prove that $n = \gcd(m, n)$.

. . . . . `dgcd`

B.1 The Euclidean Algorithm

Suppose someone asks us to determine a particular greatest common divisor $\gcd(a, b)$. How might we proceed? A natural approach is to find the prime factorizations of both a and b, and then take all the common factors. Although this is based only on an intuitive understanding of "greatest common divisor," rather than the definition above, it does produce the correct answer.

Example 4.2. Find $\gcd(8613, 2178)$.

$$\begin{aligned} 8613 &= 3^3 \cdot 11 \cdot 29 \\ 2178 &= 2 \cdot 3^2 \cdot 11^2 \\ \gcd(8613, 2178) &= 3^2 \cdot 11 = 99. \end{aligned}$$

This approach depends on being able to work out the prime factorization of an integer. For large integers that can be a very difficult task[2]. It turns out there is a much better way, known as the Euclidean Algorithm[3]. It doesn't require you to know anything about primes. You only need to be able to work out quotients and remainders as in the Division Theorem.

Here is how the Euclidean Algorithm works. Suppose we want to calculate $\gcd(a, b)$ $(a, b \in \mathbb{N})$. The algorithm will recursively produce a sequence $a_0, a_1, a_2, \ldots$ of integers, starting with $a_0 = a$ and $a_1 = b$. If the algorithm has produced $a_0, \ldots, a_k$ so far then we get the next term by dividing a_{k-1} by a_k and forming the remainder r_k as in the Division Theorem,

$$a_{k-1} = q_k a_k + r_k, \quad 0 \le r_k < a_k.$$

If the remainder is positive $r_k > 0$, we make the remainder the next term in our sequence, $a_{k+1} = r_k$, and keep going. But the first time the remainder is $r_k = 0$, we stop and take a_k as the final value of the algorithm.

Example 4.3. Here is the algorithm applied to Example 4.2.

$$\begin{aligned} & & a_0 &= 8613 \\ & & a_1 &= 2178 \\ 8613 &= 2178 \cdot 3 + 2079 & a_2 &= 2079 \\ 2178 &= 2079 \cdot 1 + 99 & a_3 &= 99 \\ 2079 &= 99 \cdot 21 + 0 & \gcd &= 99. \end{aligned}$$

Example 4.4. Here is a longer example: $\gcd(1953, 2982)$.

$$\begin{aligned} & & a_0 &= 2982 \\ & & a_1 &= 1953 \\ 2982 &= 1 \cdot 1953 + 1029 & a_2 &= 1029 \\ 1953 &= 1 \cdot 1029 + 924 & a_3 &= 924 \\ 1029 &= 1 \cdot 924 + 105 & a_4 &= 105 \\ 924 &= 8 \cdot 105 + 84 & a_5 &= 84 \\ 105 &= 1 \cdot 84 + 21 & a_6 &= 21 \\ 84 &= 4 \cdot 21 + 0 & \gcd &= 21. \end{aligned}$$

To prove that the algorithm really works we will need a little more notation. We will let k be the "counter" which keeps track of the different stages of the algorithm; the kth stage will be the one in which we divide by a_k. We will let k^* denote the value of k at which we stop: $r_{k^*} = 0$, $\gcd = a_{k^*}$. (In Example 4.4, $k^* = 6$.)

Theorem 4.7. *Suppose $a, b \in \mathbb{N}$. The Euclidean algorithm with $a_1 = a$ and $a_2 = b$ terminates after a finite number k^* of steps with $r_{k^*} = 0$ and $a_{k^*} = \gcd(a, b)$.*

We have to prove two things, 1) that the algorithm does terminate, $k^* < \infty$, and 2) that the resulting value a_{k^*} is actually $\gcd(a, b)$. The key to the latter is the following.

Lemma 4.8. *Suppose $m, n \in \mathbb{N}$. Let q and r be the quotient and remainder as in the division theorem: $m = qn + r$, $0 \le r < n$. If $g = \gcd(n, r)$ then $g = \gcd(m, n)$.*

Proof of Lemma. From $m = qn + r$ it follows that any common divisor of n and r is also a common divisor of m and n. From $r = m - qn$ it follows that any common divisor of n and m is also a common divisor of n and r. Suppose $g = \gcd(n, r)$, then g is also a common divisor of m and n. If d is any other common divisor of m and n, then it is also a common divisor of n and r. By definition of $g = \gcd(n, r)$ we know d divides g. Thus g satisfies the definition of $g = \gcd(m, n)$. □

[2]Many schemes for computer security and data encryption depend on the difficulty of finding prime factors of very large integers.

[3]It is one of the oldest mathematical algorithms in the world. It was known to Euclid, and may have been known as early as 500 BC.

Informally, the reason the algorithm must terminate is that at each stage $a_{k+1} = r_k < a_k = r_{k-1}$, so the sequence of remainders is strictly decreasing and must reach $r_{k^*} = 0$ in a finitely many steps. The lemma says that

$$\gcd(a, b) = \gcd(a_0, a_1) = \gcd(a_1, a_2) = \cdots = \gcd(a_{k^*-1}, a_{k^*}) = a_{k^*},$$

the last equality by Problem 4.12 because a_{k^*} divides a_{k^*-1}.

The "$\cdots$" in the above indicates that there really is an induction argument here. But the things we are claiming are only for $k \leq k^*$. To write this as an induction proof of something which holds for *all* k we can make the statement being proved an implication with $k \leq k^*$ as the anticedent. This makes the proof more cumbersome to read than the informal explanation above, but does make the logical structure of the argument explicit.

Proof of Theorem. We will prove by induction that the following implication is true for every positive integer k:

$$\text{if } k \leq k^* \text{ then both } r_k \leq b - k \text{ and } \gcd(a, b) = \gcd(a_{k-1}, a_k).$$

First consider $k = 1$. Since $a_0 = a$ and $a_1 = b$ the first step of the algorithm is to find q_1 and $0 \leq r_1 < b$ with $a = q_1 b + r_1$. In particular, $r_1 \leq b - 1$. And since $a_0 = a$ and $a_1 = b$ we certainly know that $\gcd(a, b) = \gcd(a_{k-1}, a_k)$. This verifies the base case of $k = 1$. (Since $k^* \geq 1$ the anticedent $k \leq k^*$ is true for the base case, but since the consequent of the implication is true we don't need to say anything about the anticedent to verify the implication.)

Now for the induction step we assume it *is* true that $k \leq k^*$ implies both $r_k \leq b - k$ and $\gcd(a, b) = \gcd(a_{k-1}, a_k)$. We want to prove that this implication is also true for $k + 1$. So we assume the anticedent, namely that $k + 1 \leq k^*$. Then $k \leq k^*$ follows and so the consequent of the assumed implication is true: $r_k \leq b - k$ and $\gcd(a, b) = \gcd(a_{k-1}, a_k)$. Since $k < k^*$ the algorithm continues at least one more step:

$$a_k = q_{k+1} a_{k+1} + r_{k+1}, \quad 0 \leq r_{k+1} < a_{k+1}.$$

Since $a_{k+1} = r_k$ it follows that $r_{k+1} \leq a_{k+1} - 1 = r_k - 1 \leq b - k - 1 = b - (k + 1)$. By the induction hypothesis $\gcd(a, b) = \gcd(a_{k-1}, a_k)$, and by the lemma (applied to $a_{k-1} = q_{k-1} a_k + r_k$) it follows that $\gcd(a_{k-1}, a_k) = \gcd(a_k, r_k) = \gcd(a_k, a_{k+1})$. Thus both $r_{k+1} \leq b - (k + 1)$ and $\gcd(a, b) = \gcd(a_k, a_{k+1})$ hold. This completes the induction proof of the claimed implication for all $k \in \mathbb{N}$.

Now $k \leq k^*$ implies that $0 \leq b - k$ and consequently $k \leq b$. This means that for $k = b + 1$ it cannot be that $k \leq k^*$, so $k^* < k = b + 1$, proving termination of the algorithm in no more than b steps. And for k^* itself we have

$$\gcd(a, b) = \gcd(a_{k^*-1}, a_{k^*}) = a_{k^*},$$

the last equality by Problem 4.12 because a_{k^*} divides a_{k^*-1}. □

In the next section we are going to prove the Fundamental Theorem of Arithmetic, the most basic result about prime factorizations of integers. The notion of relatively prime integers is a key concept for the proofs to come.

Definition. Two integers $a, b \in \mathbb{N}$, are called *relatively prime* if $\gcd(a, b) = 1$.

Example 4.5. Observe that 6 divides $4 \cdot 15 = 60$, because $60 = 10 \cdot 6$. But 6 does *not* divide 4, and 6 does *not* divide 15.

In general if a divides a product bk, we can *not* conclude that a divides one of the two factors, b or k. If, however, a and b are relatively prime, the story is different.

Lemma 4.9. *Suppose a and b are relatively prime and a divides bk. Then a divides k.*

Proof. Since $\gcd(a, b) = 1$ there exist integers α and β for which

$$1 = \alpha a + \beta b.$$

Therefore

$$k = \alpha a k + \beta b k.$$

By hypothesis, a divides both terms on the right, so it must divide k. □

Lemma 4.10. *Suppose $a, b \in \mathbb{N}$ are relatively prime, and that both of them divide k. Then ab divides k.*

Problem 4.13 Prove Lemma 4.10.

. . . . . `prodpf`

Problem 4.14 In this problem you are going to work out a method to calculate the α and β of Theorem 4.6. Suppose $a_0, a_1, \ldots$ is the sequence of values produced by the Euclidean Algorithm. At each stage we have

$$a_{k-1} = q_k a_k + r_k, \quad \text{and } a_{k+1} = r_k \text{ provided } r_k \neq 0. \tag{4.3}$$

We want to find integers $\alpha_1, \ldots, , \ldots$ and $\beta_1, \ldots, \beta_k, \ldots$ so that at each stage

$$a_k = \alpha_k \cdot a + \beta_k \cdot b.$$

We are going to keep track of our work in a table, like this.

k	a_k	q_k	α_k	β_k
0	a		α_0 (=?)	β_0 (=?)
1	b	q_1	α_1 (=?)	β_1 (=?)
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$k-1$	a_{k-1}	q_{k-1}	α_{k-1}	β_{k-1}
k	a_k	q_k	α_k	β_k
$k+1$	a_{k+1} $(= r_k)$	q_{k+1}	α_{k+1} (=?)	β_{k+1} (=?)
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k_*	a_{k_*}	q_{k_*}	α_{k_*}	β_{k_*}

Suppose we have filled in the correct values for rows 0 through k. We need to work out a procedure for filling in row $k+1$. The a_{k+1} and q_{k+1} values are just what the Euclidean Algorithm prescribes: set $a_{k+1} = r_k$ from the previous row (k), calculate the quotient and remainder, $a_k = q_{k+1}a_{k+1} + r_{k+1}$ and then fill in the values in the q column. What we need are formulas to tell us what to fill in for α_{k+1} and β_{k+1}. Since $a_{k+1} = r_k$ we know from (4.3) that

$$a_{k+1} = a_{k-1} - q_k a_k.$$

Since we know how to write a_{k-1} and a_k in terms of a and b we can substitute the appropriate expressions into this to find values $\alpha_{k+1}, \beta_{k+1}$ for which

$$a_{k+1} = \alpha_{k+1} \cdot a + \beta_{k+1} \cdot b.$$

Work this out to find formulae for α_{k+1} and β_{k+1} in terms of values from rows $k-1$ and k of the table. All we need now are values for $\alpha_0, \beta_0, \alpha_1, \beta_1$ to get the calculations started, and that is trivial. When you get to the bottom row of the table, you will have the desired values as $\alpha = \alpha_{k_*}$ and $\beta = \beta_{k_*}$.

Using this method, find the following gcds and the corresponding α and β values. Turn in your explaination of the formulae for α_{k+1} and β_{k+1}, and what values to use in the first two rows, and your filled in table for each of the examples.

a) $\gcd(357, 290)$.

b) $\gcd(2047, 1633)$.

c) $\gcd(912, 345)$.

. . . . . `EAplus`

Problem 4.15 Suppose $a, b, c \in \mathbb{N}$.

a) Write a definition of $\gcd(a, b, c)$ which is analogous to the definition of $\gcd(a, b)$ we gave above.

b) Prove that $\gcd(a, b, c) = \gcd(\gcd(a, b), c)$.

..... gcd3

Problem 4.16 Show that $a, b \in \mathbb{N}$ are relatively prime if and only if there exist integers α, β such that $1 = \alpha \cdot a + \beta \cdot b$.

..... lc

Problem 4.17 Suppose that $a, b \in \mathbb{N}$ are relatively prime. Use the preceding problem to prove that, for any $k \in \mathbb{N}$, a^k and b are relatively prime. (Hint: use induction.)

..... power

Problem 4.18 Suppose that a and b are relatively prime, and that c and d are relatively prime. Prove that $ac = bd$ implies $a = d$ and $b = c$. ([9, #17.6 page 215])

..... coprime

C Primes and the Fundamental Theorem

We gave the definition of prime number on page 15. In this section we want to prove some further properties of prime numbers. Here is the theorem which guarantees that all natural numbers can be factored (uniquely) into primes.

Theorem 4.11 (Fundamental Theorem of Arithmetic)**.** *Every natural number $n > 1$ has a unique factorization into a product of prime numbers, $n = p_1 p_2 \cdots p_s$.*

For this to be true when n is prime, we need to consider an individual prime number to be a "product of primes;" i.e. one prime all by itself will be consdered to be a "product" of one prime. We all know this theorem is true. The proof of existence is a strong induction argument. We leave it to Problem 4.19 below and concentrate here on proving uniqueness. First we must be clear about what we *mean* in saying the factorization is unique. For instance, we do not consider $12 = 2 \cdot 2 \cdot 3 = 2 \cdot 3 \cdot 2 = 3 \cdot 2 \cdot 2$ to be different factorizations. To eliminate the possible reorderings of a factorization we can insist on using a standard ordering, $p_1 \leq p_2 \leq \ldots \leq p_k$.

The following lemma is the key to the proof.

Lemma 4.12. *Suppose p is prime and divides a product of positive integers $a_1 \cdots a_m$. Then p divides a_i for some $1 \leq i \leq m$.*

Proof. We will prove the lemma by induction on m. First consider $m = 1$. Then by hypothesis p divides a_1, which is what we needed to show.

Next suppose the lemma is true for m and suppose p divides $a_1 \cdots a_{m+1}$. If p divides a_{m+1} then we are done. So suppose p does not divide a_{m+1}. Then, since the only (positive) divisors of p are 1 and p, it must be that $\gcd(p, a_{m+1}) = 1$. By applying Lemma 4.9 we conclude that p divides $a_1 \cdots a_m$. By the induction hypothesis it follows that p divides a_i for some $1 \leq i \leq m$. Thus the lemma holds for $m + 1$. This completes the proof. □

Now we can write a proof of uniqueness of prime factorizations.

Proof (Uniqueness in Theorem 4.11). Suppose there exists a natural number n with two different prime factorizations

$$p_1 \cdots p_s = n = q_1 \cdots q_r, \tag{4.4}$$

where $p_1 \leq \cdots \leq p_s$ are primes numbered in order and $q_1 \leq \cdots \leq q_r$ are also primes numbered in order. Since these two factorizations are different, either $s \neq r$ or $p_i \neq q_i$ for some i. Starting with (4.4) we can cancel all common factors and renumber the primes to obtain an equality

$$p_1 \cdots p_k = q_1 \cdots q_m, \tag{4.5}$$

in which none of the p_i appear among the q_i. Since the two factorizations in (4.4) were assumed different, there is at least one prime on each side of (4.5). In particular p_1 is a prime which divides $q_1 \cdots q_m$. By the lemma above, p_1 must divide one of the q_i. Since $p_1 \neq 1$ and $p_1 \neq q_i$ this contradicts the primality of q_i. This contradiction proves that different factorizations do not exist. □

Notice that we have used an "expository shortcut" by referring to a process of cancellation and renumbering but without writing it out explicitly. We are trusting that the reader can understand what we are referring to without needing to see it all in explicit notation. Just describing this in words is clearer than what we would get if we worked out notation to describe the cancellation and renumbering process explicitly.

Problem 4.19 Write a proof of the existence part of Theorem 4.11, namely that a prime factorization exists for each $n > 1$. [Hint: use strong induction, starting with $n = 2$. For the induction step, observe that either $n + 1$ is prime or $n + 1 = mk$ where both $2 \leq m, k \leq n$.]

..... FAexist

D The Integers Mod m

All our usual number systems are infinite, but there are finite number systems too! The most basic are the integers mod m, which we introduce in this section. We said "are" because for different choices of $m \in \mathbb{N}$ we will get different number systems. So bear in mind throughout this section that m is allowed to be any given positive integer.

Definition. We say $a, b \in \mathbb{Z}$ are *congruent modulo* m, and write $a \equiv_m b$ (or $a \equiv b \bmod m$) when $b - a$ is divisible by m.

Example 4.6. $3 \equiv 27 \bmod 8$, because $27 - 3 = 3 \cdot 8$. But $3 \not\equiv 27 \bmod 10$, because $27 - 3 = 24$ is not divisible by 10.

Lemma 4.13. *Congruence modulo m is an equivalence relation on $\mathbb{Z}$.*

Proof. For any $a \in \mathbb{Z}$, since $a - a = 0 = 0 \cdot m$ we see that $a \equiv_m a$, showing that $\equiv_m$ is reflexive. If $a \equiv_m b$, then $b - a$ is divisible by m. But then $a - b = -(b - a)$ is also divisible by m, so that $b \equiv_m a$. This shows that $\equiv_m$ is symmetric. For transitivity, suppose $a \equiv_m b$ and $b \equiv_m c$. Then $a - b$ and $b - c$ are both divisible by m. It follows that $a - c = (a - b) + (b - c)$ is also divisible by m, implying $a \equiv_m c$. □

Since $\equiv_m$ is an equivalence relation, we can define it's equivalence classes according to Definition 3.8 on page 63. We abbreviate the notation for an equivalence class, writing $[n]_m$ rather than $[n]_{\equiv_m}$, and will refer to $[n]_m$ as a *congruence class* mod m.

Definition. Suppose m is a positive integer. The *integers modulo* m is the set $\mathbb{Z}_m$ of equivalence classes modulo m:

$$\mathbb{Z}_m = \{[n]_m : \ n \in \mathbb{Z}\}.$$

Back on page 64 we talked about the idea of defining new mathematical objects to be equivalence classes with respect to some equivalence relation. There we talked about considering an angle to be the set all real numbers which were "equivalent to each other as angles," i.e. an equivalence class of the relation $\oslash$ of Example 3.12. We are doing the same thing here using the relation $\equiv_m$: we take the set of all integers which are congruent to each other mod m and put them together as a set (congruence class); that set is a single element of $\mathbb{Z}_m$.

Example 4.7. A typical element of $\mathbb{Z}_8$ is

$$[27]_8 = \{\ldots, -13, -5, 3, 11, 19, 27, \ldots\}.$$

We can indicate the same equivalence class several ways, for instance $[27]_8 = [3]_8$. (We have several different ways of referring to the same real number as well, for instance $\frac{1}{2} = .5$.) We would say that 27 and 3 are both representatives of the equivalence class $[27]_8$. We can choose any representative of an equivalence class to identify it. But we often use the smallest nonnegative representative, which would be 3 in this example.

Whether we refer to it as $[27]_8$ or $[3]_8$ it is just one element of $\mathbb{Z}_8$. There are a grand total of 8 elements in $\mathbb{Z}_8$:

$$\mathbb{Z}_8 = \{[0]_8, [1]_8, [2]_8, [3]_8, [4]_8, [5]_8, [6]_8, [7]_8\}.$$

Every congruence class mod 8 is the same as one of these.

We have been saying that $\mathbb{Z}_m$ is a number system. That must mean there is a way to define addition and multiplication on the elements of $\mathbb{Z}_m$, i.e. there is a way to add and mutiply congruence classes. The next example begins to explain.

Example 4.8. $3 \equiv_8 27$ and $5 \equiv_8 45$. Observe that

$$3 \cdot 5 \equiv_8 27 \cdot 45 \text{ and } 3 + 5 \equiv_8 27 + 45.$$

This example illustrates the fact that $\equiv_m$ "respects" the operations of multiplication and addition. The next lemma states this precisely.

Lemma 4.14. *Suppose $a \equiv_m a$ and $b \equiv_m b'$. Then $a + b \equiv_m a' + b'$, $a \cdot b \equiv_m a' \cdot b'$, and $a - b \equiv_m a' - b'$.*

Proof. By hypothesis there exist $k, \ell \in \mathbb{Z}$ for which $a' = a + km$ and $b' = b + \ell m$. Then

$$a'b' = (a + km)(b + \ell m) = ab + (a\ell + bk + k\ell m)m,$$

which implies that $ab \equiv_m a'b'$. The proofs for addition and subtraction are similar. □

Here is how you should understand this. Suppose A and B are any two elements of $\mathbb{Z}_m$. (For example, $A = [3]_8$ and $B = [5]_8$.) We can add A and B in the following way: pick any element a of A and any element b of B. (For instance $a = 3$ and $b = 5$.) Form $a + b$ using ordinary arithmetic, and then take C to be the equivalence class of the result: $C = [a + b]_m$. (In our example, $C = [3 + 5]_8 = [0]_8$.) Then C is what we mean by $A + B$. What the lemma says is that the a and b that you picked don't matter; you will arrive at the same result C regardless. (For instance if we picked $a' = 27$ and $b' = 45$ instead, we would still get $C = [27 + 45]_8 = [72]_8 = [0]_8$.) The same procedue works for multiplication: $D = A \cdot B$ is $D = [a \cdot b]_m$.

Definition. Addition, multiplication, and negation are defined on $\mathbb{Z}_m$ by

$$\begin{aligned} [a]_m + [b]_m &= [a + b]_m, \\ [a]_m \cdot [b]_m &= [a \cdot b]_m, \\ -[a]_m &= [-a]_m. \end{aligned}$$

With this definition $\mathbb{Z}_m$ is a *finite* number system, and satisfies all the algebraic properties we listed in Section A.1:(A1)–(A5), (M1)–(M4), and (D). (There is no order realation, however.)

Example 4.9. Here are the addition and multiplication tables for $\mathbb{Z}_6$. (All the entries should really be surrounded by "$[\cdot]_6$" but we have left all these brackets out to spare our eyes from the strain.)

+	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

*	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

Notice that $[2]_6 \neq [0]_6$ and $[3]_6 \neq [0]_6$, but $[2]_6 \cdot [3]_6 = [0]_6$. In other words in $\mathbb{Z}_6$ two nonzero numbers can have zero as their product! (We have seen this happen before; see Problems 4.1 and the 2×2 matricies of Section A.3.)

There are many clever and creative things we can use modular arithmetic for.

Example 4.10. There do not exist positive integers a, b for which $a^2 + b^2 = 1234567$. A long, tedious approach would be to examine all possible pairs a, b with $1 \leq a, b < 1234567$. A faster way is to consider the implications modulo 4. If $a^2 + b^2 = 1234567$ were true then (mod 4),

$$[a]^2 + [b]^2 = [a^2 + b^2] = [1234567] = [3].$$

(For the last equality, observe that $1234567 = 1234500 + 64 + 3$, which makes it clear that $12345467 \equiv_4 3$.) Now in $\mathbb{Z}_4$, $[n]^2$ is always either $[0]$ or $[1]$. So there are four cases: $[a]^2 = [0]$ or $[1]$ and $[b]^2 = [0]$ or $[1]$. Checking the four cases, we find

$[a]^2$	$[b]^2$	$[a]^2 + [b]^2$
$[0]$	$[0]$	$[0]$
$[0]$	$[1]$	$[1]$
$[1]$	$[0]$	$[1]$
$[1]$	$[1]$	$[2]$

In no case do we find $[a]^2 + [b]^2 = [3]$. Thus $a^2 + b^2 = 1234567$ is not possible, no matter what the values of a and b.

In fact, we can turn this idea into a proposition. The proof is essentially the solution of the above example, so we won't write it out again.

Proposition 4.15. *If $c \equiv_4 3$, there do not exist integers a, b for which $a^2 + b^2 = c$.*

Problem 4.20 A natural question to ask about Example 4.10 is why we choose to use mod 4; why not some other m?

a) Show that in $\mathbb{Z}_6$ every $[n]$ occurs as $[a]^2 + [b]^2$ for some a and b. What happens if we try to repeat the argument of Example 4.10 in $\mathbb{Z}_6$ — can we conclude that $a^2 + b^2 = 1234567$ is impossible in that way?

b) For the argument of Example 4.10 to work in $\mathbb{Z}_m$, we need to use an m for which

$$\{[a]_m^2 + [b]_m^2 : \ a, b \in \mathbb{Z}\} \neq \mathbb{Z}_m.$$

This happens for $m = 4$ but not for $m = 6$. Can you find some values of m other than 4 for which this happens? [Hint: there are two values $m < 10$ other than $m = 4$ for which it works.]

..... expyth

Problem 4.21 Find values of $a, b, m \in \mathbb{N}$ so that $a^2 \equiv_m b^2$ but $a \not\equiv_m b$.

..... powne

Problem 4.22 Suppose $n \in \mathbb{N}$ is expressed in the usual decimal notation $n = d_k d_{k-1} \cdots d_1 d_0$, where each d_i is one of the digits $0, \ldots, 9$. You probably know that n is divisible by 3 if and only if $d_k + d_{k-1} + \cdots + d_1 + d_0$ is divisible by 3. Use $\mathbb{Z}_3$ to prove why this is correct. [Hint: The notation we use for the number one hundred twenty three, "n=123," does *not* mean $n = 1 \cdot 2 \cdot 3$. What *does* it mean? More generally what does "$n = d_k d_{k-1} \cdots d_1 d_0$" mean?] Explain why the same thing works for divisibility by 9.

..... div39

Problem 4.23 Along the same lines as the preceding problem, show that n is divisible by 11 if and only if the *alternating* sum of its digits $d_0 - d_1 + d_2 \cdots + (-1)^k d_k$ is divisible by 11.

..... div11

Problem 4.24 What is the remainder when $1^{99} + 2^{99} + 3^{99} + 4^{99} + 5^{99}$ is divided by 5? (From [17].)

..... pow99

Problem 4.25 What is the last digit of $2^{1000000}$? (Based on [9, #7 page 272])

..... TwoK

E Axioms and Beyond: Gödel Crashes the Party

We introduced a set of axioms for the integers in Section A.4. Axioms have been developed for many of the most basic mathematical systems, such as the natural numbers, the real numbers, set theory. (Russell's paradox showed that we need to be careful about what kinds of statements about sets we allow. To resolve this this requires developing a system of axioms for set theory.) If you take a modern algebra class you will see definitions of other types of algebraic systems (such as groups, rings and fields) in terms of axioms. In any of these settings, a set of axioms is a collection of basic properties from which all other properties can be derived and proven logically.

In the 1920s David Hilbert proposed that all of mathematics might be reduced to an appropriate list of axioms, from which everything mathematical could be then be derived in an orderly, logical way. This system of axioms should be complete, i.e. all true statements should be provable from it. It should also be consistent, i.e. there should be no contradictions that follow logically from the axioms. This would put all of mathematics on a neat and tidy foundation. By developing formal rules that govern logical arguments and deductions, so that proofs could be carried out mechanically, we would in principle be able to turn over all of mathematics to computers which would then determine all mathematical truth for us. In 1931 Kurt Gödel pulled the plug on that possibility. He showed that in *any* axiomatic system (provided it is at least elaborate enough to include $\mathbb{N}$) there are statements that can be neither proven nor disproven, i.e. whose truth or falsity *cannot* be logically established based on the axioms. (A good discussion of Gödel's brilliant proof is given in [21].) Gödel's result tells us that we can not pin our hopes on some ultimate set of axioms. There will always be questions which the axioms are not adequate to answer.

For instance suppose we consider the axioms for the integers as listed in Section A.4, but leave out the well-ordering principle. Now we ask if the well-ordering principle is true or false based on the other axioms. We know that $\mathbb{Z}$ satisfies the axioms and the well-ordering principle is true for $\mathbb{Z}$. That means it is impossible to prove that the well-ordering property is false from the other axioms. But the axioms also are true for $\mathbb{R}$, and the well-ordering property is false in the context of $\mathbb{R}$. That means there is no way to prove the well-ordering property is true from the other axioms. So whether the well-ordering property is true or false cannot be decided based on the other axioms alone, it is *undecidable* from the other axioms. For more difficult propositions it can take years before we can tell if the proposition is undecidable as opposed to just really hard. For many years people wondered if Euclid's fifth postulate (axiom) was provable from his other four. Eventually (2000 years after Euclid) other "geometries" were discovered which satisfied Euclid's other axioms but not the fifth, while standard plane geometry does satisfy the fifth. That made it clear; the fifth axiom is undecidable based on the first four. You could assume it (and get standard plane geometry) or replace it by something different (and get various non-Euclidean geometries). It took 300 years before we knew that Fermat's Last Theorem *was* provable from the axioms of the integers. We still don't know if the Twin Primes Conjecture is provable. Maybe it is unprovable — we just don't know (yet). Another example is the Continuum Hypothesis, long considered one of the leading unsolved problems of mathematics. In 1938 Gödel showed that it was consistent with set theory, i.e. could not be disproved. Then in 1963 Paul Cohen showed that it's negation was also consistent. Thus it can not be proven true and it cannot be proven false!

It is undecidable based on "standard" set theory. (This of course requires a set of axioms to specify exactly what "set theory" consists of.)

Please don't leave this discussion thinking that Gödel's result makes the study of axioms useless. Identifying a set of axioms remains one of the best ways to delineate exactly what a specific mathematical system consists of and what we do and don't know about it. However, the axioms of a system are something that distills what we have learned about a system after years of study. It is *not* typically where we begin the study of a new mathematical system.

Additional Problems

Problem 4.26 Prove that each row of the multiplication table for $\mathbb{Z}_m$ contains 1 if and only if it contains 0 only once. (See [9, #13 page 272].)

..... `row1`

Problem 4.27 The following quotation appeared in Barry A. Cipra's article *Sublinear Computing* in SIAM News **37** (no. 3) April 2004: "Suppose that someone removes a card from a deck of 52 and proceeds to show you the rest, one at a time, and then asks you to name the missing card. In principle, you could mentally put check marks in a 4×13 array and then scan the array for the unchecked entry, but very few people can do that, even with lots of practice. It's a lot easier to use some simple arithmetic: convert each card into a three-digit number, the hundreds digit specifies the suit — say 1 for clubs, 2 for diamonds, 3 for hearts, and 4 for spades — and the other two digits specify the value, from 1(ace) to 13 (king); then simply keep a running sum, subtracting 13 whenever the two digit part exceeds 13 and dropping and thousands digit (e.g. adding the jack of hearts — 311 — to the running sum 807 gives 1118, which reduces to 105). The missing card is simply what would have to be added to the final sum to get 1013 (so that for a final sum of 807, the missing card would be the 6 of diamonds)." Explain why this works!

..... `trick`

Problem 4.28 Prove the following.

Theorem 4.16 (Fermat's Little Theorem)**.** *If p is prime and a is an integer, then $a^p = a \bmod p$.*

(Hint: You can do this by induction on a. For the induction step use the Binomial Theorem to work out $(a+1)^p$.)

..... `FLT`

Problem 4.29 Suppose that p is prime. Then $\mathbb{Z}_p$ has additional properties. Prove the following.

a) If $[a] \cdot [b] = [0]$, then either $[a] = [0]$ or $[b] = [0]$.

b) If $[a] \neq [0]$ then the function $f_a : \mathbb{Z}_p \to \mathbb{Z}_p$ defined by $f_a([b]) = [a] \cdot [b]$ is injective. (Hint: suppose not and use a).)

c) If $[a] \neq [0]$ the function f_a above is surjective. (Hint: use the Pigeon Hole Principle.)

d) If $[a] \neq [0]$ there exists $[b] \in \mathbb{Z}_p$ for which $[a] \cdot [b] = [1]$.

e) The $[b]$ of d) is unique.

In other words, in $\mathbb{Z}_p$ we can divide by nonzero numbers!

..... `Zp`

Problem 4.30 Prove that infinitely many prime numbers are congruent to 3 mod 4.

. . . . . `inf3primes`

Problem 4.31 Not every equivalence relation will produce equivalence classes that respect addition and multiplication, as $\equiv_m$ did. For instance define an equivalence realtion $\sqcup$ on $\mathbb{Z}$ so that $n \sqcup m$ means that one of the following holds:

i) both $-10 \leq n \leq 10$ and $-10 \leq m \leq 10$,

ii) both $n < -10$ and $m < -10$,

iii) both $10 < n$ and $10 < m$.

It is not hard to check that $\sqcup$ is indeed an equivalence relation (but you don't need to do it), and it has three equivalence classes:

$$\begin{aligned} [-11]_\sqcup &= \{\ldots\ ,-13,\ -12,\ -11\} \\ [0]_\sqcup &= \{-10,\ -9,\ \ldots\ ,\ 9,\ 10\} \\ [11]_\sqcup &= \{11,\ 12,\ 13,\ \ldots\}. \end{aligned}$$

Now, the point we want to make in this problem is that we cannot define addition (or multiplication) on this set of equivalence classes in the same way that we did for $\mathbb{Z}_m$. Show (by example) that Lemma 4.14 is false for $\sqcup$ and its the equivlaence classes.

. . . . . `notnum`

Problem 4.32 A 3-tuple of positive integers (a, b, c) is called a *Pythagorean triple* if $a^2 + b^2 = c^2$. For instance $(3, 4, 5)$ is the most familiar example. The next best known is $(5, 12, 13)$. Prove that Euclid's formula

$$a = m^2 - n^2, \quad b = 2mn, \quad c = m^2 + n^2$$

produces a Pythagoran triple for any two integers $0 < n < m$. Show that if m and n are relatively prime and one of them is even, then a, b, c have no common factor (other than 1). Use this to prove that there are infinitely many distinct Pythagorean triples, no two of which can be obtained as integer multiples of each other..

. . . . . `PyTrip`

Chapter 5

Polynomials

A *polynomial* is a function of the form

$$\begin{aligned} p(x) &= c_n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0 \\ &= \sum_0^n c_k x^k. \end{aligned} \tag{5.1}$$

In other words it is a function that can be expressed as a (finite) sum of *coefficients* c_k times nonnegative integer powers of the variable x. (We understand $x^0 = 1$ for all x, and we customarily write the higher powers first and lower powers last.) Polynomials arise in almost all branches of mathematics. Usually the roots of a polynomial are particularly important. There are many important theorems about the roots of polynomials. In this chapter we will look at proofs of three such theorems. The first is the Rational Root Theorem, whose proof is not too hard using what we know about prime numbers. The other two, Descartes' Rule of Signs and the Fundamental Theorem of Algebra, are difficult. Their proofs are more substantial than any we have looked at previously, and are good examples of how a proof is built around a creative idea, not just methodically applying definitions. The purpose for studying these more difficult proofs is *not* that you will be expected to produce equally difficult proofs on your own. Rather the purpose is for you to gain experience reading and understanding difficult proofs written by someone else. But before we can study those proofs we need first to go through some preliminary material about polynomials.

A Preliminaries

To start, we need to make the distinction between a function and the expression we use to describe it.

Example 5.1. Consider the following descriptions of a function $f : \mathbb{R} \to \mathbb{R}$:

$$\begin{aligned} f(x) &= x^2 + 1 \\ f(x) &= \sqrt{x^4 + 2x^2 + 1} \\ f(x) &= (x+3)^2 - 6(x+3) + 10 \\ f(x) &= 2 + \int_1^x 2t\,dt. \end{aligned}$$

All four of these expressions describe the *same* function.

By an "expression" we mean a combination of symbols and notation, essentially what we write or type, like the righthand sides above. Clearly each of the four expressions above are different in terms of what is printed on the page. But when we substitute a real value in for x, all four expressions produce the same real value for $f(x)$. For instance, $x = -1/3$ produces the value $f(-1/3) = 10/9$ in all four expressions for f above, and you will likewise find agreement for every possible value of $x \in \mathbb{R}$ that you try. So the four *different* expressions above all describe the *same* function.

When we say $p(x)$ is a polynomial, we mean that it is a function which can be described using an expression of the particular form (5.1). I.e. it is a polynomial function if *there exists* a polynomial expression which describes $p(x)$ for all x. Our example $f(x)$ above is a polynomial because the first description[1] we gave for it is the required polynomial expression.

Example 5.2. The following functions are *not* polynomials.

$$\begin{aligned} g(x) &= \frac{x}{x^2+1}, \\ h(x) &= e^x, \\ s(x) &= \sin(x), \\ r(x) &= \sqrt[3]{x}. \end{aligned}$$

We are claiming that for each of these no polynomial expression exists[2].

We will be talking about polynomials using three different number fields: $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$. The number field specifies the domain and codomain of the polynomial function, as well as what values are allowed for the coefficients. When we say that $p(x)$ is a polynomial *over* $\mathbb{R}$, we mean that it is a function $p : \mathbb{R} \to \mathbb{R}$ expressible in the form (5.1) using coefficients $c_k \in \mathbb{R}$. The set of all polynomials over $\mathbb{R}$ is denoted $\mathbb{R}[x]$. The set of polynomials over $\mathbb{Q}$ is denoted $\mathbb{Q}[x]$; when we say $p(x) \in \mathbb{Q}[x]$ that means the coefficients are all rational numbers and we view $\mathbb{Q}$ as its domain. For functions of complex numbers it is customary to use z for the variable, so $\mathbb{C}[z]$ will denote the set of all polynomials over $\mathbb{C}$. A $p(z) \in \mathbb{C}[z]$ may have any complex numbers as its coefficients, and all complex z are in its domain. To discuss all three cases at once we will write $\mathbb{F}$ to stand for any of the number fields $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$ with the understanding that everything we say about $\mathbb{F}[x]$ is meant to refer to all three cases of $\mathbb{Q}[x]$, $\mathbb{R}[x]$, and $\mathbb{C}[z]$ simultaneously.

Definition. Let $\mathbb{F}$ be either $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$. A *polynomial* over $\mathbb{F}$ is a function $p : \mathbb{F} \to \mathbb{F}$ such that there exist $c_0, c_1, \ldots, c_n \in \mathbb{F}$ (for some integer $0 \le n$) for which $p(x)$ is given by

$$p(x) = \sum_{k=0}^{n} c_k x^k.$$

Suppose $p(x) \in \mathbb{F}[x]$. This means that it can be represented in the form (5.1) for some n and choice of coefficients $c_k \in \mathbb{F}$, $k = 0, \ldots, n$. Now we come to the first important question about polynomials: might it be possible for there to be a second such representation of the *same* function but using *different* coefficients? In other words, could there be a different choice of coefficients $a_k \in \mathbb{F}$ so that

$$\sum_{0}^{n} c_k x^k = \sum_{0}^{n} a_k x^k \text{ for all } x \in \mathbb{F}\,?$$

This is an important question — if two different representations are possible, then it will be meaningless to talk about "the" coefficients of $p(x)$ unless we also specify which particular expression for it we have in mind. Fortunately, the answer is no.

Lemma 5.1. *Suppose two polynomial expressions, with coefficients in $\mathbb{F}$, agree for all all $x \in \mathbb{F}$:*

$$\sum_{0}^{n} c_k x^k = \sum_{0}^{n} a_k x^k.$$

Then $c_k = a_k$ for all $0 \le k \le n$.

[1] The third description would be called a polynomial "in $(x+3)$," as opposed to a polynomial "in x" as we intend here.

[2] The reasons, in brief, are as follows. There are infinitely many roots of $s(x)$, but a polynomial has only finitely many roots (see Lemma 5.5). For $g(x)$ observe that $\lim_{x\to\infty} g(x) = 0$. But the only polynomial with this property is the zero polynomial. For the others observe that for any nonzero polynomial there is a nonnegative integer n (its degree) with the property that $\lim_{x\to\infty} p(x)/x^n$ exists but is not 0. For $h(x)$ and $r(x)$ that does not hold, so they can't be polynomials.

Proof. By subtracting the right side from the left, the hypothesis says that

$$\sum_0^n (c_k - a_k)x^k = 0 \text{ for all } x \in \mathbb{F}.$$

We want to show that this implies $c_k - a_k = 0$ for all $0 \le k \le n$. Thus what we need to show is that if $\alpha_k \in \mathbb{F}$ and

$$\sum_0^n \alpha_k x^k = 0 \text{ for all } x \in \mathbb{F}, \tag{5.2}$$

then the coefficients are $\alpha_k = 0$ for $k = 0, \ldots, n$. We will prove this by induction on n. For $n = 0$ (5.2) simply says that $\alpha_0 = 0$, which is what we want to show.

Now we make the induction hypothesis that (5.2) for n implies $\alpha_k = 0$ for all $0 \le k \le n$, and suppose

$$\sum_0^{n+1} \alpha_k x^k = 0 \text{ for all } x \in \mathbb{F}. \tag{5.3}$$

Define a new polynomial by

$$\begin{aligned} p(x) &= 2^{n+1} \sum_0^{n+1} \alpha_k x^k - \sum_0^{n+1} \alpha_k (2x)^k \\ &= \sum_0^{n+1} (2^{n+1} - 2^k)\alpha_k x^k \\ &= \sum_0^{n} (2^{n+1} - 2^k)\alpha_k x^k. \end{aligned}$$

It follows from our hypothesis (5.3) that $p(x) = 0$ for all x. But observe that the coefficients of x^{n+1} cancel so that $p(x)$ is in fact an expression of the form (5.2) with nonzero terms only for $0 \le k \le n$. By our induction hypothesis, we know that $(2^{n+1} - 2^k)\alpha_k = 0$ for all $0 \le k \le n$. But for these k we know $2^{n+1} - 2^k \ne 0$ and so $\alpha_k = 0$ for all $0 \le k \le n$. Only α_{n+1} remains. The hypothesis (5.3) now reduces to

$$\alpha_{n+1} x^{n+1} = 0 \text{ for all } x \in \mathbb{F}.$$

Plugging in $x = 1$ we conclude that In particular $\alpha_{n+1} = 0$ as well. This completes the induction step. □

Notice that in this proof we used the induction hypothesis twice! It might occur to you that an easier way to do the induction step would be to first plug in $x = 0$ to deduce that $\alpha_0 = 0$. Then we could say that for all $x \in \mathbb{F}$

$$0 = \sum_1^{N+1} \alpha_k x^k = x \sum_0^N \alpha_{k+1} x^k.$$

Now we might divide out the extra x, and then appeal to the induction hypothesis to conclude the $\alpha_{k+1} = 0$ for $k = 0, \ldots, N$. But there is one problem with that argument: after dividing by x we can only say that the resulting equation ($\sum_0^N \alpha_{k+1} x^k = 0$) holds for $x \ne 0$. We can't say it is true for *all* x, and so can't appeal to the induction hypothesis. One way to remedy this would be to take $\lim_{x \to 0}$ to see that in fact the equation must hold for $x = 0$ as well. But that would require appealing to properties of limits, which (especially in the case of $\mathbb{F} = \mathbb{C}$) would take us beyond what you know about limits from calculus. Another approach would be to use the derivative to reduce a degree $N + 1$ polynomial to one of degree N — an induction proof can be based on that, but again we would need to justify the calculus operations when $\mathbb{F}$ is $\mathbb{Q}$ or $\mathbb{C}$. Although the proof we gave above is somewhat more complicated, it is purely algebraic and does not rely on limits.

The significance of Lemma 5.1 is that for a given a polynomial $p(x)$ there is only one way to express it in the form (5.1). In brief, two polynomial expressions are equal (for all x) if and only if they have the same coefficients. Knowing that, we can now make the following definitions, which depend on this uniqueness of coefficients.

Definition. Suppose $p(x) = \sum_{k=0}^{n} c_k x^k$ is a polynomial in $\mathbb{F}[x]$. The *degree* of $p(x)$, denoted $\deg(p)$, is the largest k for which $c_k \neq 0$, and that c_k is the *leading coefficient* of $p(x)$. The *zero polynomial* is the constant function $p(x) = 0$ for all x, and is considered[3] to have degree 0 (even though it has no nonzero coefficient). A *root* (or *zero*) of $p(x)$ is a value $r \in \mathbb{F}$ for which $p(r) = 0$.

In the expression $p(x) = \sum_{k=0}^{n} c_k x^k$, the degree of $p(x)$ is n provided $c_n \neq 0$. A polynomial of degree 0 is just a constant function, $p(x) = c_0$ for all x. We only consider values in the appropriate $\mathbb{F}$ as possible roots, so for $p(x) \in \mathbb{Q}[x]$, a root is a *rational* number r with $p(r) = 0$; for $p(x) \in \mathbb{R}[x]$ a root is a *real* number r with $p(r) = 0$; for $p(z) \in \mathbb{C}[z]$ a root is a *complex* number r with $p(r) = 0$

We can add and multiply two polynomials and get polynomials as the results. The coefficients of $p(x) + q(x)$ are just the sums of the coefficients of $p(x)$ and $q(x)$, but the coefficients of $p(x)q(x)$ are related to the coefficients of $p(x)$ and $q(x)$ in a more complicated way. Part c) of the next lemma says that when working with polynomial equations, it is valid to cancel common factors from both sides. (While we might call this "dividing out $s(x)$" that is not quite right; $s(x)$ may have roots so for those values of x we cannot consider cancellation to be the same as division. We encountered the same thing in 7) of Proposition4.1.)

Lemma 5.2. *Suppose $p(x), q(x), s(x) \in \mathbb{F}[x]$.*

a) If neither $p(x)$ nor $q(x)$ is the zero polynomial, then $\deg(p(x)q(x)) = \deg(p(x)) + \deg(q(x))$.

b) If $p(x)q(x)$ is the zero polynomial, then either $p(x)$ is the zero polynomial or $q(x)$ is the zero polynomial.

c) If $s(x)$ is not the zero polynomial and $s(x)p(x) = s(x)q(x)$ (for all x), then $p(x) = q(x)$.

Proof. We first prove a). Suppose $\deg(p) = n$ and $\deg(q) = m$. The hypotheses imply that for some choice of coefficients a_i, b_j, with $a_n \neq 0$ and $b_m \neq 0$,

$$p(x) = \sum_{i=0}^{n} a_i x^i, \quad q(x) = \sum_{j=0}^{m} b_j x^j.$$

When we multiply $p(x)q(x)$ out, the highest power of x appearing will be x^{n+m} and its coefficient will be $a_n b_m$. Since $a_n b_m \neq 0$, we see that $\deg(pq) = n + m$. This proves a).

We prove b) by contradiction. If neither $p(x)$ nor $q(x)$ were the zero polynomial, then, as in a), $p(x)q(x)$ would have a nonzero leading coefficient $a_n b_m$. But that would mean that $p(x)q(x)$ is *not* the zero polynomial, a contrary to the hypotheses.

For c), the hypothesis implies that $s(x)\,[p(x) - q(x)] = 0$ for all x. By b), since $s(x)$ is not the zero polynomial it follows that $p(x) - q(x)$ is the zero polynomial and therefore $p(x) = q(x)$, proving c). $\square$

Problem 5.1 Find a formula for the coefficients of $p(x)q(x)$ in terms of the coefficients of $p(x)$ and $q(x)$.
..... conv

Sometimes we can divide one polynomial by another, *but not always*. So, like the integers, we have a meaningful notion of divisibility for polynomials.

Definition. Suppose $p(x), d(x)$ are polynomials in $\mathbb{F}[x]$. We say $d(x)$ *divides* $p(x)$ (or $p(x)$ is *divisible by* $d(x)$) when there exists $q(x) \in \mathbb{F}[x]$ so that

$$p(x) = q(x) \cdot d(x).$$

Theorem 5.3 (Division Theorem for Polynomials)**.** *Suppose $p(x), d(x)$ are polynomials in $\mathbb{F}[x]$ and that $d(x)$ has degree at least 1. There exist unique polynomials $q(x)$ (the quotient) and $s(x)$ (the remainder) in $\mathbb{F}[x]$ such that $s(x)$ has lower degree than $d(x)$ and*

$$p(x) = q(x) \cdot d(x) + s(x).$$

[3]Some authors consider the zero polynomial to have degree $-\infty$. With that definition, part a) of Lemma 5.2 below holds without the restriction to nonzero polynomial.

Example 5.3. Let $p(x) = x^5 + 2x^4 + 6x^2 + 18x - 1$ and $d(x) = x^3 - 2x + 9$. We want to work out the quotient and remainder as in the Division Theorem. We essentially carryout a long division process, finding $q(x)$ one term at a time, working from the highest power to the smallest.

$$(x^5 + 2x^4 + 6x^2 + 18x - 1) - x^2(x^3 - 2x + 9) = 2x^4 + 2x^3 - 3x^2 + 18x - 1.$$

The x^2 is just right to make the x^5 terms cancel. Now we do the same to get rid of the $2x^4$ by subtracting just the right multiple of $d(x)$.

$$(2x^4 + 2x^3 - 3x^2 + 18x - 1) - 2x(x^3 - 2x + 9) = 2x^3 + x^2 - 1.$$

Next we eliminate the $2x^3$.

$$2x^3 + x^2 - 1 - 2(x^3 - 2x + 9) = x^2 + 4x - 19.$$

We can't reduce that any further by subtracting multiples of $d(x)$, so that must be the remainder.

$$q(x) = x^2 + 2x + 2, \quad s(x) = x^2 + 4x - 19.$$

We want to prove the Division Theorem. We proved the integer version using the well-ordering principle. But there is no well-ordering principle for polynomials[4]. So we will need a different proof. The simplest approach is to use induction on the degree of $p(x)$. The induction step is essentially the idea used in the example above.

Proof. Let $k \geq 1$ be the degree of $d(x)$ and n be the degree of $p(x)$. If $n < k$, we can just take $q(x) \equiv 0$ and $s(x) = p(x)$. This proves the theorem for $0 \leq n \leq k - 1$. So we need to prove the theorem for $n \geq k$. We use (strong) induction on n. Suppose $n \geq k - 1$ and that the theorem holds whenever the degree of $p(x)$ is at most n. We want to prove it when the degree of $p(x)$ is $n + 1$. Let $a_{n+1} \neq 0$ be the leading coefficient of $p(x)$ and $b_k \neq 0$ the leading coefficient of $d(x)$. Consider

$$\tilde{p}(x) = p(x) - \frac{a_{n+1}}{b_k} x^{n+1-k} d(x).$$

The right side is a polynomial of degree at most $n + 1$, but the coefficients of x^{n+1} cancel, so in fact the degree of $\tilde{p}(x)$ is at most n. By the induction hypothesis, there exist $\tilde{q}(x), \tilde{s}(x)$ with degree of $\tilde{s}(x)$ less than k so that

$$\tilde{p}(x) = \tilde{q}(x) \cdot d(x) + \tilde{s}(x).$$

Therefore

$$\begin{aligned} p(x) &= \frac{a_{n+1}}{b_k} x^{n+1-k} d(x) + \tilde{p}(x) \\ &= \frac{a_{n+1}}{b_k} x^{n+1-k} d(x) + \tilde{q}(x) \cdot d(x) + \tilde{s}(x) \\ &= \left[\frac{a_{n+1}}{b_k} x^{n+1-k} + \tilde{q}(x) \right] \cdot d(x) + \tilde{s}(x) \\ &= q(x) \cdot d(x) + s(x), \end{aligned}$$

where $q(x) = \frac{a_{n+1}}{b_k} x^{n+1-k} + \tilde{q}(x)$ and $s(x) = \tilde{s}(x)$. This proves the existence of $q(x)$ and $s(x)$ for $p(x)$ by induction. You will verify uniqueness in the following homework problem. □

Problem 5.2 Show that the $q(x)$ and $s(x)$ of the above theorem are unique.

. . . . . `divuniq`

[4]There is no reasonable sense of order for polynomials. For instance neither $x < 1 - x$ nor $1 - x < x$ is correct — it depends on which x you consider!

Corollary 5.4. *Suppose $p(x) \in \mathbb{F}[x]$ and $r \in \mathbb{F}$. Then $p(x)$ is divisible by $x - r$ if and only if r is a root of $p(x)$.*

Proof. By the division theorem, there is a constant c (polynomial of degree less than 1) and a polynomial $q(x)$ so that

$$p(x) = q(x) \cdot (x - r) + c.$$

Evaluating both sides at $x = r$, we see that $p(r) = c$. So $(x - r)$ divides $p(x)$ iff $c = 0$, and $c = 0$ iff $p(r) = 0$. □

Lemma 5.5. *A nonzero polynomial of degree n has at most n distinct roots.*

Proof. We use induction. Consider $n = 0$. In that case $p(x) = c_0$ for some $c_0 \neq 0$. For such a polynomial there are no roots at all; zero roots. Since zero is indeed at most $0 = \deg(p)$, this proves the $n = 0$ case.

Suppose the lemma is true for n and consider a polynomial $p(x)$ with $\deg(p) = n + 1$. If $p(x)$ has no roots, then we are done because $0 \leq n + 1$. So suppose $p(x)$ has a root r. Then $p(x) = (x - r)q(x)$ where $\deg(q) = n$. By the induction hypothesis $q(x)$ has at most n roots. Any root of $p(x)$ is either r or one of the roots of $q(x)$, so there are at most $1 + n$ roots of $p(x)$. This completes the induction step. □

Lemma 5.6. *Suppose $p(x)$ and $q(x)$ are polynomials with $\deg(p), \deg(q) \leq n$ and $p(x) = q(x)$ for $n + 1$ distinct x values. Then $p(x) = q(x)$ for all x.*

Problem 5.3 Prove the lemma.

..... lprove

We could continue to develop the notions of greatest common divisor, the Euclidean algorithm, and unique factorization results, all analogous to what we discussed for the integers. But instead we will proceed to the theorems about roots of polynomials that we mentioned in the introduction.

B $\mathbb{Q}[x]$ and the Rational Root Theorem

Suppose $p(x)$ is a polynomial with rational coefficients, and we are trying to find its roots. We can multiply $p(x)$ by the least common multiple N of the denominators of the coefficients to obtain a new polynomial $\tilde{p}(x) = Np(x)$ all of whose coefficients will be integers. The roots of $p(x)$ and $\tilde{p}(x)$ are the same, so we will proceed assuming that all the coefficients of $p(x)$ are integers.

Theorem 5.7 (Rational Root Theorem)**.** *Suppose*

$$p(x) = c_n x^n + \cdots + c_1 x + c_0$$

is a polynomial with coefficients $c_k \in \mathbb{Z}$ and that $r = \frac{\ell}{m}$ is a rational root, expressed in lowest terms. Then m divides c_n and ℓ divides c_0.

Example 5.4. Consider the polynomial $p(x) = .6x^4 + 1.3x^3 + .1x^2 + 1.3x - .5$. We want to find the roots. All the coefficients are rational numbers. Multiplying by 10 produces a polynomial with integer coefficients:

$$\tilde{p}(x) = 10p(x) = 6x^4 + 13x^3 + x^2 + 13x - 5.$$

The divisors of 6 are 1, 2, 3, 6 and their negatives. The divisors of -5 are 1, 5, and their negatives. So the possible rational roots are

$$\pm 1,\ \pm\frac{1}{2},\ \pm\frac{1}{3},\ \pm\frac{1}{6},\ \pm\frac{5}{1},\ \pm\frac{5}{2},\ \pm\frac{5}{3},\ \pm\frac{5}{6}.$$

Checking them all, we find that $\frac{1}{3}$ and$-\frac{5}{2}$ are the only ones which are actually roots. By Theorem 5.7 we conclude that these are the *only* rational roots.

Proof. Assume that $r = \frac{\ell}{m}$ is a root expressed in lowest terms (i.e. m and ℓ are relatively prime). We know $p(r) = 0$. Multiplying this by m^n we find that

$$0 = c_n\ell^n + c_{n-1}\ell^{n-1}m + \cdots + c_1\ell m^{n-1} + c_0 m^n.$$

Since all the terms here are integers, it follows that m divides $c_n\ell^n$. By hypothesis m and ℓ are relatively prime, and so m and ℓ^n are relatively prime, by Problem 4.17. By Lemma 4.9 it follows that m divides c_n. Similarly, ℓ divides $c_0 m^n$. Since ℓ and m^n are relatively prime, we conclude that ℓ divides c_0. □

Problem 5.4 Use Theorem 5.7 to prove that $\sqrt{2}$ is irrational, by considering $p(x) = x^2 - 2$.
..... root2

Problem 5.5 Show that for each $n \geq 2$ there is a $p(x) \in \mathbb{Q}[x]$ of degree n with no (rational) roots. (Do this by exhibiting such a $p(x)$. You may want to consider cases depending of the value of n.)
..... norQ

C $\mathbb{R}[x]$ and Descartes' Rule of Signs

Next we consider a result about the roots of polynomials over $\mathbb{R}$,

$$p(x) = c_n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0, \quad c_k \in \mathbb{R}.$$

Such a polynomial is a function $p : \mathbb{R} \to \mathbb{R}$. The standard results from calculus apply in this setting, and we will use them in our proofs. The theorem of this section does not tell us what *what* the roots are, but it will tell us *how many* there can be (at most). Here is how it works: convert each coefficient c_k to a $\pm$ sign according to whether $c_k > 0$ or $c_k < 0$ (you can just skip any $c_k = 0$) and write the signs out in order.

Example 5.5. If

$$p(x) = 2x^6 + x^5 - 4.7x^4 + \frac{1}{2}x^2 + 2x - 3$$

we would get (skipping $c_3 = 0$)

$$+ \ + \ - \ + \ + \ - \ .$$

Now count the number of *sign changes* in this sequence, i.e. occurrences of $-+$ or $+-$. The rule of signs says that the number of *positive* roots of $p(x)$ is no greater than the number of sign changes. In our example, this says there can't be more than 3 positive roots. (In fact there is just one.)

Rene Descartes stated his "rule" in 1637, but it was almost 100 years before the first proofs were given. A number of different proofs have been developed since then. We are going to consider a very recent proof given by Komornik [15]. This an interesting example of a proof which works by proving a more general statement than the original — sometimes working in a more general setting makes possible techniques that would not have worked in the original setting. Here is Kormornik's generalization of Descartes' rule.

Theorem 5.8 (Rule of Signs)**.** *Suppose $p(x) = a_m x^{b_m} + a_{m-1}x^{b_{m-1}} + \cdots + a_1 x^{b_1} + a_0 x^{b_0}$ is a function with nonzero real coefficients $a_m, \ldots, a_0$ and real exponents $b_m > b_{m-1} > \cdots > b_0$. Then $p(x)$ cannot have more positive roots (counted with multiplicity) than the number of sign changes in the sequence $a_m \ldots a_0$.*

We need to carefully consider the statement of this theorem before we can prove it. There are several things to observe.

The coefficients a_k are assumed to be nonzero. This means any terms $0x^b$ are simply not included in the expression for $p(x)$. This also means that the exponents are not simply $m, m-1, \ldots$ but a more general sequence $b_m > b_{m-1} > \cdots$. In our example, $b_5 = 6$, $b_4 = 5$, $b_3 = 4$, $b_2 = 2$, $b_1 = 1$, $b_0 = 0$.

The exponents are *not* required to be nonnegative integers. They can be *any real numbers*, including fractional and negative ones. Thus the $p(x)$ of the theorem are more general functions than polynomials. We will call such functions generalized polynomials.

Definition. A *generalized polynomial* is a function $p(x) : (0, \infty) \to \mathbb{R}$ of the form

$$p(x) = a_m x^{b_m} + a_{m-1} x^{b_{m-1}} + \cdots + a_1 x^{b_1} + a_0 x^{b_0},$$

for some sequence of exponents $b_m > b_{m-1} > \cdots b_1 > b_0$ and coefficients $a_k \in \mathbb{R}$.

Example 5.6. The function

$$p(x) = x^{3/2} - \sqrt{2} x^{\sqrt{2}} + 4 + 2x^{-1/5} - \pi x^{-2}$$

is a generalized polynomial, but not a polynomial.

The domain of $p(x)$ is only $(0, \infty)$. Since the exponents of a generalized polynomial can be fractional we can't allow negative x. For instance $x^{1/2}$ is undefined if $x < 0$. And since exponents are allowed to be negative we can't allow $x = 0$; x^{-1} is undefined for $x = 0$. *The theorem only talks about positive roots of $p(x)$; it says nothing about the possibility of negative roots!*

At this point you may be wondering why we have allowed such troublesome functions to be included in the theorem. Why haven't we just stayed with nice polynomials? The reason is that Kormornik's proof doesn't work if we confine ourselves to conventional polynomials.

But there is still one more thing we need to explain about the theorem statement. **The phrase "counted with multiplicity" needs to be defined.** Suppose $p(x)$ is a polynomial in the usual sense (nonnegative integral exponents). If r is a root of $p(x)$, that means $(x - r)$ is a factor of $p(x)$. It might be a repeated factor $p(x) = (x - r)^2(\cdots)$ or $p(x) = (x - r)^3(\cdots)$, or maybe $p(x) = (x - r)^k(\cdots)$ for some higher power k. The highest power k for which $(x - r)^k$ divides $p(x)$ is called the *multiplicity* of the root r. The number of roots "counted with multiplicity" means that a root of multiplicity 2 is counted twice, a root of multiplicity 3 is counted three times, and so on. To count the roots with multiplicity we add up the multiplicities of the different roots.

Example 5.7. Consider $p(x) = (x - 1)(x + 1)^3(x - 2)^5$. This has 3 roots, but 9 counted with multiplicity. It has 2 positive roots, 6 counted with multiplicity. If we multiply this polynomial out, we get

$$p(x) = x^9 - 8x^8 + 20x^7 - 2x^6 - 61x^5 + 58x^4 + 56x^3 - 80x^2 - 16x + 32.$$

This has exactly 6 sign changes in the coefficients, confirming Descartes' rule for this example.

However this definition of multiplicity does not work for generalized polynomials.

Example 5.8. Consider[5] $p(x) = x^\pi - 1$. Clearly $r = 1$ is a root. But that does *not* mean we can write $p(x) = (x - 1)q(x)$ for some other generalized polynomial $q(x)$. In fact if you try to work out what $q(x)$ would need to be you get an infinite series:

$$q(x) = x^{\pi-1} + x^{\pi-2} + x^{\pi-3} + \cdots$$

So for generalized polynomials we need a different definition of multiplicity, one that does not appeal to $(x - r)^k$ as a factor of $p(x)$, and yet which reduces to the usual definition when $p(x)$ is a conventional polynomial. The key to this generalized definition is to observe that if $p(x)$ is a generalized polynomial, then the derivative $p'(x)$ exists and is also a generalized polynomial. That is simply because $\frac{d}{dx} x^b = bx^{b-1}$ is valid for any $b \in \mathbb{R}$, for $x > 0$. Now suppose $p(x)$ is a conventional polynomial and that r is a root of multiplicity 1: $p(x) = (x - r)q(x)$ where $q(r) \neq 0$. Then observe that

$$p'(x) = q(x) + (x - r)q'(x); \quad p'(r) = q(r) + 0 \neq 0.$$

If r is a root of multiplicity 2, $p(x) = (x - r)^2 q(x)$ with $q(r) \neq 0$. Then looking at the derivatives we have

$$\begin{aligned} p'(x) &= 2(x - r)q(x) + (x - r)^2 q'(x); & p'(r) &= 0 \\ p''(x) &= 2q(x) + 4(x - r)q'(x) + (x - r)^2 q''(x); & p''(r) &\neq 0. \end{aligned}$$

If we continue in this way we find that if r is a root of multiplicity k then k is the smallest integer such that $p^{(k)}(r) \neq 0$. (We use the customary notation $p^{(k)}(x)$ for the kth derivative of $p(x)$.) If we make that the definition of multiplicity, then we have a definition that makes sense for generalized polynomials as well!

[5] Thanks to Randy Cone for this example.

Definition. Suppose $p(x)$ is a generalized polynomial with root $r > 0$ and k is a positive integer. We say the root r has *multiplicity* k when

$$0 = p(r) = p'(r) = \ldots = p^{(k-1)}(r), \quad \text{and } p^{(k)}(r) \neq 0.$$

So when the theorem says "roots counted with multiplicity" it means multiplicity as defined in terms of derivatives.

We are ready now to turn our attention to the proof of Theorem 5.8. Some notation will help. For a generalized polynomial $p(x)$ we will use $\zeta(p)$ to denote the number of positive roots of $p(x)$, counted according to multiplicity, and $\sigma(p)$ to denote the number of sign changes in the coefficients. The assertion of Theorem 5.8 is simply

$$\zeta(p) \leq \sigma(p).$$

The proof will be by induction, but *not* on m which counts the number of terms in $p(x)$. Rather the induction will be on the number $n = \sigma(p)$ of sign changes in the coefficients. That's Komornik's first clever idea. In other words we will prove that $\mathrm{P}(n)$ is true for all $n = 0, 1, 2, \ldots$ where $\mathrm{P}(n)$ is the statement

For every generalized polynomial $p(x)$ with n sign changes in the coefficients,
there are at most n positive real roots (counted according to multiplicity).

The induction step for such a proof needs a way of taking a $p(x)$ with $n+1$ sign changes and connecting to some other generalized polynomial $q(x)$ with n sign changes. The induction hypothesis will tell us that $q(x)$ has at most n roots. Then we will need to use that fact to get back to our desired conclusion that $p(x)$ has at most $n+1$ roots. Komornik's second clever idea was a way to do this: find one of the sign changes in the coefficients: say $a_{j+1} < 0 < a_j$, marked by the $\curlywedge$ in the $\pm$ pattern below:

$$+ \ + \ - \curlywedge + \ + \ - \ .$$

Let b_{j+1} and b_j be the exponents on either side of the sign change and pick a value β between them: $b_{j+1} > \beta > b_j$. Form a new generalized polynomial $\tilde{p}(x) = x^{-\beta}p(x)$. This will have the same number of sign changes and the same number of roots (with the same multiplicities) as $p(x)$, but will have the property that all the exponents to the left of the sign change position will be positive: $b_k - \beta > 0$ for $k \geq j+1$; and all the exponents to the right of the sign change position will be negative: $b_k - \beta < 0$ for $k \leq j$. Now let $q(x) = \tilde{p}'(x)$, the derivative of $\tilde{p}(x)$. The negative exponents to the right of the sign change position will reverse the signs of those coefficients, while the positive exponents on the left will leave the signs of those coefficients unchanged. The $\pm$ pattern for the coefficients of $q(x)$ will be

$$+ \ + \ - \curlywedge - \ - + \ ,$$

which has exactly one fewer sign change than the original $p(x)$. So the induction hypothesis will tell us that q has at most n roots (counted with multiplicity), $\zeta(q) \leq \sigma(q) = n$ in our notation. Now we need to make a connection between $\zeta(q)$ and $\zeta(\tilde{p}) = \zeta(p)$. That, it turns out, is not so hard. All the facts we need to make the proof work are collected in the next lemma.

Lemma 5.9. *Suppose $p(x)$ is a generalized polynomial.*

a) If $r > 0$ is a root of $p(x)$ with multiplicity $k > 1$, then $r > 0$ is a root of $p'(x)$ with multiplicity $k - 1$.

b) $\zeta(p) \leq 1 + \zeta(p')$.

c) If $\beta \in \mathbb{R}$ and $\tilde{p}(x) = x^{\beta}p(x)$, then $\sigma(\tilde{p}) = \sigma(p)$ and $\zeta(\tilde{p}) = \zeta(p)$.

Proof. Part a) is elementary, because the definition of multiplicity k for $p(x)$ says that

$$0 = p'(r) = p''(r) \ldots = (p')^{(k-2)}(r), \quad \text{and } (p')^{(k-1)}(r) \neq 0,$$

which is the definition of multiplicity $k-1$ for $p'(x)$.

To prove b), suppose there are ℓ roots of $p(x)$, numbered in order: $r_1 < r_2 < \ldots < r_\ell$. Let their multiplicities be $m_1, m_2, \ldots, m_\ell$. For each successive pair of roots $p(r_k) = 0 = p(r_{k+1})$. (Here $k = 1, \ldots, \ell -$

1.) Now Rolle's Theorem (from calculus) implies that there exists a t_k in (r_k, r_{k+1}) for which $p'(t_k) = 0$. These t_k give us $\ell - 1$ new roots of $p'(x)$ in addition to the r_k. As roots of p' the r_k have multiplicities $m_k - 1$ (by part a)), and the t_k have multiplicities at least 1. Thus, counting multiplicities, we have that $\zeta(p')$ is at least

$$(m_1 - 1) + (m_2 - 1) + \cdots + (m_\ell - 1) + \ell - 1 = (m_1 + \cdots m_\ell) - 1.$$

(If $m_k = 1$ then r_k is not a root of $p'(x)$ at all. But then $m_k - 1 = 0$ so including it in the sum above is not incorrect.) This shows that

$$\zeta(p') \geq \zeta(p) - 1.$$

Rearranging, this is b).

For c), the coefficients of $p(x)$ and $\tilde{p}(x) = x^\beta p(x)$ are the same, so $\sigma(p) = \sigma(\tilde{p})$. Since $r^\beta \neq 0$, it is clear from $\tilde{p}(r) = r^\beta p(r)$ that r is a root of $p(x)$ if and only if it is a root of $\tilde{p}(x)$. We need to show that if r has multiplicity m for $p(x)$ then it has multiplicity m for $\tilde{p}(x)$. We do this by induction on m. First consider $m = 1$. This means $p(r) = 0$ and $p'(r) \neq 0$. We already know $\tilde{p}(r) = 0$. The product rule tells us

$$\tilde{p}'(r) = \beta r^{\beta-1} p(r) + r^\beta p'(r) = 0 + r^\beta p'(r) \neq 0.$$

So the multiplicity of r for $\tilde{p}(x)$ is indeed $m = 1$. Next we make the induction hypothesis:

If r is a root of $p(x)$ of multiplicity m then it is a root of $\tilde{p}(x) = x^\beta p(x)$ of multiplicity m.

Suppose that r has multiplicity $m + 1$ with respect to $p(x)$. The chain rule tells us that

$$\tilde{p}'(x) = x^{\beta-1} t(x), \quad \text{where } t(x) = \beta p(x) + x p'(x).$$

Now $t(x)$ is another generalized polynomial. We know from a) that r is a root of multiplicity m for $p'(x)$. By our induction hypothesis r is also a root of multiplicity m for $xp'(x)$: all its derivatives through the $(m-1)^{\text{st}}$ are $= 0$ at $x = r$, but its mth derivative is $\neq 0$ there. The other term, $\beta p(x)$, has all derivatives through *and including* the mth derivative $= 0$ at r. It follows then that r is a root of $t(x)$ of multiplicity m. Our induction hypothesis implies that it also has multiplicity m as a root of $\tilde{p}'(x)$. This means that as a root of $\tilde{p}(x)$ it has multiplicity $m + 1$. This completes the induction argument, proving that roots of $p(x)$ and $\tilde{p}(x)$ are the same with the same multiplicities. Now the conclusion of c) is clear. □

We are ready now to write the proof of Theorem 5.8.

Proof. The proof is by induction on the value of $\sigma(p)$. First consider $\sigma(p) = 0$. This means all the coefficients a_k of $p(x)$ are positive (or all negative), and therefore $p(x) > 0$ for all $x > 0$ (or $p(x) < 0$ for all $x > 0$). Therefore $p(x)$ has no positive roots, so $\zeta(p) = 0$, confirming $\zeta(p) \leq \sigma(p)$.

Next, we assume the theorem is true for any generalized polynomial with $\sigma(p) = n$. Suppose $\sigma(p) = n+1$. Among the sequence of coefficients a_k there is at least one sign change; choose a specific j where one of the sign changes occurs: $a_{j+1} a_j < 0$. Choose $b_{j+1} > \beta > b_j$, and define a new generalized polynomial $\tilde{p}(x) = x^{-\beta} p(x)$. Then

$$\begin{aligned} \tilde{p}(x) &= a_m x^{b_m - \beta} + a_{m-1} x^{b_{m-1} - \beta} + \cdots + a_1 x^{b_1 - \beta} + a_0 x^{b_0 - \beta}, \\ &= a_m x^{\bar{b}_m} + a_{m-1} x^{\bar{b}_{m-1}} + \cdots + a_1 x^{\bar{b}_1} + a_0 x^{\bar{b}_0}, \end{aligned}$$

where $\bar{b}_k = b_k - \beta$. The coefficients of $\tilde{p}$ are identical with those of p. By the lemma,

$$\zeta(\tilde{p}) = \zeta(p), \quad \sigma(\tilde{p}) = \sigma(p) = n + 1. \tag{5.4}$$

Now consider $q(x) = \tilde{p}'(x)$:

$$q(x) = \alpha_m x^{\bar{b}_m - 1} + \cdots \alpha_1 x^{\bar{b}_1 - 1} + \alpha_0 x^{\bar{b}_0 - 1},$$

where $\alpha_k = a_k \bar{b}_k$. Now for $k \geq j + 1$ we have $\bar{b}_k = b_k - \beta > 0$, so that the sign of α_k is the same as the sign of a_k. But for $j \geq k$, we have $\bar{b}_k = b_k - \beta < 0$ so that the signs of α_k are opposite the signs of a_k. It follows that q has exactly one fewer sign changes than $\tilde{p}$:

$$\sigma(q) = \sigma(\tilde{p}) - 1 = n. \tag{5.5}$$

By our induction hypothesis,

$$\zeta(q) \le \sigma(q). \tag{5.6}$$

We can conclude that

$$\begin{aligned}
\zeta(p) &= \zeta(\tilde{p}) \quad \text{by (5.4)}, \\
&\le \zeta(\tilde{p}') + 1 \quad \text{by Lemma 5.9}, \\
&= \zeta(q) + 1 \quad \text{since } q = \tilde{p}', \\
&\le \sigma(q) + 1 \quad \text{by (5.6)}, \\
&= \sigma(\tilde{p}) \quad \text{by (5.5)}, \\
&= \sigma(p) \quad \text{by (5.4)}.
\end{aligned}$$

This completes the induction step, and the proof of the theorem. □

Problem 5.6 Explain why it is important for the above proof that the exponents b_k are *not* required to be nonnegative integers.

..... `needgp`

Problem 5.7 Explain how the rule of signs can be applied to $p(-x)$ to say something about the number of negative roots (counted according to multiplicity) of a polynomial $p(x)$. Apply this to the polynomial of Example 5.5, and compare the result to the actual number of negative roots counted according to multiplicity.

..... `negroot`

D $\mathbb{C}[z]$ and The Fundamental Theorem of Algebra

D.1 Some Properties of the Complex Numbers

Next we turn to properties of polynomials with complex coefficients. Since complex numbers may be unfamiliar to some readers, this section will give a quick presentation of some of their basic properties.

The complex numbers $\mathbb{C}$ consist of all numbers z which can be written in the form

$$z = x + iy,$$

where $x, y \in \mathbb{R}$ and i is a new number with the property that $i^2 = -1$. We refer to x as the *real part* of z and y as the *imaginary part* of z. The usual notation is

$$x = \mathrm{Re}(z), \quad y = \mathrm{Im}(z).$$

Arithmetic ($+$ and $\cdot$) on complex numbers is carried out by just applying the usual properties (axioms (A1)–(M1) for the integers), just remembering that $i^2 = -1$. So if $z = x + iy$ and $w = u + iv$ ($x, y, u, v \in \mathbb{R}$) then

$$\begin{aligned}
zw &= (x + iy) \cdot (u + iv) \\
&= xu + x\,iv \;\; + iy\,u + iy\,iv \\
&= xu + ixv + iyu + i^2 yv \\
&= xu + i\,xv \;\; + i\,yu - yv \\
&= (xu - yv) + i(xv + yu).
\end{aligned} \tag{5.7}$$

In $\mathbb{C}$ the zero element (additive identity) is $0 = 0 + i\,0$, and the multiplicative identity is $1 = 1 + 0i$.

The *conjugate* of $z = x + iy$ is

$$\bar{z} = x - yi.$$

and the *modulus* of z is

$$|z| = \sqrt{x^2 + y^2}.$$

Observe that the modulus is an extension of absolute value from $\mathbb{R}$ to $\mathbb{C}$, and continues to use the same notation "$|\cdot|$." Also note that $|z|$ is always a nonnegative real number, even though z is complex. The next problem brings out some important properties.

Problem 5.8 Prove the following for all $z, w \in \mathbb{C}$.

a) $z = 0$ if and only if $|z| = 0$.

b) $z\bar{z} = |z|^2$.

c) $\overline{zw} = \bar{z}\bar{w}$.

d) $|zw| = |z|\,|w|$.

e) $|z^n| = |z|^n$ for all $n \in \mathbb{N}$.

f) $|z + w| \leq |z| + |w|$.

g) $|z + w| \geq |z| - |w|$.

..... `props`

Problem 5.9 Suppose $z = x + iy \in \mathbb{C}$ and $z \neq 0$. (That means x and y are not both 0.) Find formulas for $u, v \in \mathbb{R}$ in terms of x and y so that $w = u + iv$ has the property that $wz = 1$. I.e. find the formula for $w = z^{-1}$.

..... `prd`

Problem 5.10 Show that in $\mathbb{C}$ if $wz = 0$ then either z or w mult be 0. (Hint: the properties in the preceding problems might be useful.)

..... `nozd`

A complex number $z = x + iy$ corresponds to the pair $(x, y) \in \mathbb{R}^2$ of its real and imaginary parts, which we can plot as a point in the plane. Viewed this way the plane is often called the *complex plane*. It is important to recognize that there is no order relation in $\mathbb{C}$. We can compare real numbers with $\leq$, $>$ and so forth, but inequalities are meaningless for complex numbers in general. See Problem 4.3.

Problem 5.11 Show that if $z = x + iy$ is not 0, then by using the standard polar coordinates for the point (x, y) in the plane z can be written as

$$z = |z|(\cos(\alpha) + i\sin(\alpha))$$

for some real number α. This is called the *polar form* of z. Write both -1 and i in this form. Show that if $w = |w|(\cos(\beta) + i\sin(\beta))$ is a second complex number written in polar form, then

$$zw = |zw|(\cos(\alpha + \beta) + i\sin(\alpha + \beta)).$$

(This is just an exercise in trig identities.) The complex number $\cos(\alpha) + i\sin(\alpha)$ is usually denoted $e^{i\alpha}$, for several very good reasons. One is that the above multiplication formula is just the usual law of exponents: $e^{i\alpha}e^{i\beta} = e^{i(\alpha+\beta)}$.

polar

Problem 5.12 If $z = r(\cos(\theta) + i\sin(\theta))$, what are the polar forms of $\bar{z}$ and $1/z$?

polars

Problem 5.13 Suppose n is a positive integer. Find n different solutions to $z^n = 1$ and explain why they are all different. If $w \neq 0$ is nonzero, find n different solutions to $z^n = w$. (Hint: Find the polar form of z.) In particular let $\pm\xi$ be the roots of $z^2 = 2 + i\sqrt{3}$. Determine $|\xi|$ and $\mathrm{Re}(\xi)$ exactly. (The trigonometric identity connecting $\cos(2\theta)$ and $\cos(\theta)$ will be useful.)

cplxroots

D.2 The Fundamental Theorem

Now we turn to a very important result about $\mathbb{C}$: every polynomial in $\mathbb{C}[z]$ can be factored all the way down to a product of linear factors!

Theorem 5.10 (Fundamental Theorem of Algebra)**.** *Suppose*

$$p(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$$

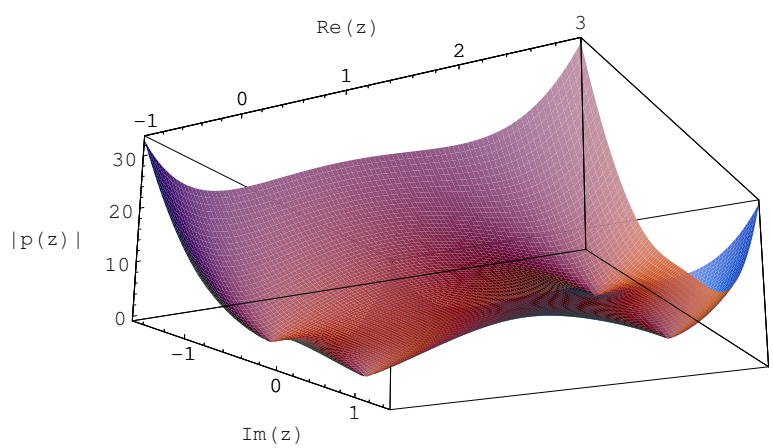
is a polynomial of degree $n \geq 1$ with complex coefficients $a_k \in \mathbb{C}$. Then $p(z)$ can be completely factored over $\mathbb{C}$. In other words, there exist $\zeta_k \in \mathbb{C}$ so that

$$p(z) = a_n(z - \zeta_1)(z - \zeta_2)\cdots(z - \zeta_n).$$

Example 5.9. Consider the polynomial[6]

$$p(z) = z^4 - 4z^3 + 2z^2 + 4z + 4.$$

In an effort to see the locations of the roots of $p(z)$, here is the graph of $|p(z)|$ viewed from its underside. (To plot the graph of $p(z)$ would require a four-dimensional picture. Since $|p(z)| \in \mathbb{R}$ its graph is only three-dimensional, which we are much better at visualizing.)



The locations of the roots are visible as the four "dimples" where $|p(z)|$ comes down to touch 0.

[6]This example is significant in the history of the fundamental theorem. Nikolas Bernouli believed it could not be factored at all, but in 1742 L. Euler wrote to him with a factorization into a pair of quadratics.

The first correct proof of the Fundamental Theorem was due to Argand in 1806, although there had been numerous flawed attempts previously. There are now many different proofs. We will discuss the elementary proof given in E. Landau's classic calculus book [16]. The essence of the proof is to show that given $p(z)$ there always exists a root $p(\zeta) = 0$. By Corollary 5.4 that means that $p(z) = (z - \zeta)q(z)$. Now apply the same reasoning to $q(z)$ and keep going. So the proof focuses on on proving that one root of $p(z)$ exists. This is done in two parts. First is to argue that $|p(z)|$ has a minimum point: a z_* such that $|p(z_*)| \leq |p(z)|$ for all $z \in \mathbb{C}$. To do that we are going to ask you to accept on faith an extension of the Extreme Value Theorem from calculus. The Extreme Value Theorem says that if $f(x)$ is a continuous function of $x \in [a, b]$, then there is a minimum point $c \in [a, b]$: $f(c) \leq f(x)$ all $x \in [a, b]$. Here is the fact we will need.

Proposition 5.11. *If $p(z)$ is a polynomial with complex coefficients, and $B \subseteq \mathbb{C}$ is a closed rectangle*

$$B = \{z = x + iy : \ a \leq x \leq b, \ c \leq y \leq d\}$$

for some real numbers $a \leq b$, $c \leq d$, then $|p(z)|$ has a minimum point z_* over B: $|p(z_*)| \leq |p(z)|$ for all $z \in B$.

This proposition is true if $|p(z)|$ is replaced by any continuous real-valued function $f(z)$ defined on B. But we don't want to make the diversion to explain continuity for functions of a complex variable (or of two real variables: $z = x + iy$). That is something you are likely to discuss in an advanced calculus course. We have simply stated it for the $f(z) = |p(z)|$ that we care about in our proof. If you want to see a proof of this Proposition, you can find one in [16].

The second part of the proof is to show that $p(z_*) \neq 0$ leads to a contradiction. This part of the argument will illustrate a technique that is sometimes used to simplify a complicated proof: reducing the argument to a special case. This is usually introduced with the phrase "without loss of generality, we can assume..." Although we didn't use exactly those words, we used this device back in the proof of Theorem 1.10 of Chapter 1. We will use it more than once below.

Proof. Suppose $p(z)$ is a complex polynomial of degree at least 1.

$$p(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$$

with $n \geq 1$ and $a_n \neq 0$. It is sufficient to prove that there exists a root $p(\zeta) = 0$. Indeed, this implies $p(z) = (z - \zeta)q(z)$ for some polynomial of one lower degree. The theorem follows by induction from this. So, by way of contradiction, suppose $p(z)$ has no roots. Then $|p(z)| > 0$ for all $z \in \mathbb{C}$. We will show that this leads to a contradiction. The argument is in several steps. First observe that the case of $n = 1$ is trivial because any polynomial of degree 1 does have a root. So we will assume that $1 < n$ for the rest of the proof.

Step 1. We claim that there exists a square

$$B = \{z = x + iy : \ |x| \leq b, \ |y| \leq b\}$$

with the property that $|p(z)| \geq |p(0)|$ for all $z \in \mathbb{C} \setminus B$. First observe that all such z have $|z| \geq b$. We want to estimate $|p(z)|$ from below, and show that if b is big enough then $|p(z)| > |a_0|$. Using the various parts of Problem 5.8,

$$\begin{aligned} |p(z)| &\geq |a_n||z|^n - \sum_{k=0}^{n-1} |a_k||z|^k \\ &= |a_n||z|^n \left[1 - \sum_{k=0}^{n-1} \frac{|a_k|}{|a_n|} |z|^{k-n} \right] \\ &\geq |a_n||z|^n \left[1 - \sum_{k=0}^{n-1} \frac{|a_k|}{|a_n|} b^{k-n} \right]. \end{aligned} \tag{5.8}$$

The last inequality follows from our assumption that $|z| \geq b > 0$. Now suppose that $b \geq \left(\frac{2n|a_k|}{|a_n|}\right)^{\frac{1}{n-k}}$ for each k in this sum. That implies that $\frac{|a_k|}{|a_n|}b^{k-n} \leq \frac{1}{2n}$, and therefore

$$1 - \sum_{k=0}^{n-1} \frac{|a_k|}{|a_n|} b^{k-n} \geq \frac{1}{2}. \tag{5.9}$$

This gives the inequality

$$|p(z)| \geq \frac{1}{2}|a_n||z|^n \geq \frac{1}{2}|a_n|b^n,$$

From here, if $b \geq \left(1 + \frac{2|a_0|}{|a_n|}\right)^{\frac{1}{n}}$ it will follow that $b^n > \frac{2|a_0|}{|a_n|}$, and therefore

$$|p(z)| > |a_0| = |p(0)|. \tag{5.10}$$

In summary, if we choose b to be

$$b = \max\left(\left(\frac{2n|a_0|}{|a_n|}\right)^{\frac{1}{n-0}}, \ldots, \left(\frac{2n|a_{n-1}|}{|a_n|}\right)^{1}, \left(1 + \frac{2|a_0|}{|a_n|}\right)^{\frac{1}{n}}\right)$$

then for all $z \in \mathbb{C} \setminus B$, we have

$$|p(z)| > |p(0)|,$$

as desired. The significance of this is that the minimum value of $|p(z)|$ over all $z \in \mathbb{C}$ will be found in B; we don't need to look outside of B for the minimum.

Step 2. Let B be the square of Step 1. By the Proposition 5.11 above there exists a minimizing point z_* for $|p(z)|$ over B. By Step 1, this implies that z_* is a minimum point of $|p(z)|$ over *all* of $\mathbb{C}$:

$$|p(z_*)| \leq |p(z)|, \quad \text{for all } z \in \mathbb{C}. \tag{5.11}$$

By our hypothesis that $p(z)$ has no roots, we must have that $|p(z_*)| > 0$. The rest of the proof is devoted to deriving a contradiction from this.

Step 3. Without loss of generality, we can assume $z_* = 0$ and $p(0) = 1$. To see this consider $\tilde{p}(z) = p(z + z_*)/p(z_*)$. This is again a polynomial, and from (5.11) has the property that

$$|\tilde{p}(z)| = \frac{|p(z + z_*)|}{|p(z_*)|} \geq \frac{|p(z_*)|}{|p(z_*)|} = 1 = \tilde{p}(0).$$

In other words $|\tilde{p}(z)|$ has its minimum at $z = 0$, and $|\tilde{p}(0)| = 1$. Considering $\tilde{p}(z)$ in place of $p(z)$ is the same as assuming $z_* = 0$ and $p(0) = 1$. We proceed under this assumption.

Step 4. Since $p(0) = 1$, we can write

$$p(z) = 1 + a_1 z + \cdots a_n z^n.$$

Let $1 < m \leq n$ be index of the first nonzero coefficient after the constant term (it exists since $\deg(p) > 1$):

$$p(z) = 1 + a_m z^m + \cdots a_n z^n,$$

with $a_m \neq 0$. We know from Problem 5.13 that there exists a $w \in \mathbb{C}$ with $w^m = -a_m$. Consider $\tilde{p}(z) = p(z/w)$. This is another complex polynomial with the property that $\tilde{p}(0) = 1 \leq |\tilde{p}(z)|$ for all $z \in \mathbb{C}$. Moreover $\tilde{p}(z)$ has the form $\tilde{p}(z) = 1 - z^m + \tilde{a}_{m+1}z^{m+1} + \cdots + \tilde{a}_n z^n$. In other words, without loss of generality we can assume $a_m = -1$ and write

$$p(z) = 1 - z^m + (a_{m+1}z^{m+1} + \cdots a_n z^n) = 1 - z^m + \sum_{k=m+1}^{n} a_k z^k.$$

Note that it is not possible that $m = n$, else we would have $p(1) = 0$, contrary to the properties of $p(z)$.

Step 5. Now suppose $z = x$ is a real number with $0 < x \leq 1$. Then, using the properties from Problem 5.8 again, we have

$$\begin{aligned}
|p(x)| &\leq |1 - x^m| + \sum_{k=m+1}^{n} |a_k x^k| \\
&= 1 - x^m + \sum_{k=m+1}^{n} |a_k| x^k \\
&\leq 1 - x^m + x^{m+1} \sum_{k=m+1}^{n} |a_k| \\
&= 1 - x^m + x^{m+1} c \\
&= 1 - x^m (1 - cx),
\end{aligned}$$

where c is the nonnegative real number

$$c = \sum_{k=m+1}^{n} |a_k|.$$

Consider the specific value

$$x = \frac{1}{1+c}.$$

Since $c \geq 0$ we see that $0 < x \leq 1$, so the above inequalities do apply. Moreover,

$$1 - cx = 1 - \frac{c}{1+c} = \frac{1}{1+c} > 0,$$

and therefore $x^m(1 - cx) > 0$. We find that

$$|p(x)| \leq 1 - x^m(1 - cx) < 1.$$

This is a contradiction to our hypothesis (Step 3) that $1 \leq |p(z)|$ for all z. The contradiction shows that in fact our original $p(z)$ must have a root, completing the proof. □

Notice that there is really an induction argument in this proof, but the reader is trusted to be able to fill that in for themselves so that the proof can focus on the argument for the induction step: the existence of a root ζ. When you read a difficult proof like this, you often find that the author has left a number of things for you to check for yourself.

Problem 5.14 Write out in detail the justifications of (5.8) and (5.9).

..... `details`

Problem 5.15 There is only one place in the proof where we used a property of $\mathbb{C}$ that is not true for $\mathbb{R}$ — where is that?

..... `diagnose`

Problem 5.16 Suppose $p(z)$ is a polynomial all of whose coefficients are real numbers, and $\zeta \in \mathbb{C}$ is a complex root. Show that the conjugate $\bar{\zeta}$ is also a root.

..... `conjroot`

Corollary 5.12. *Every polynomial $p(x) \in \mathbb{R}[x]$ of degree $n \geq 3$ can be factored into a product of lower degree polynomials with real coefficients.*

Proof. By keeping the same real coefficients but replacing the real variable x by a complex variable z we can consider $p(z) = \sum_0^n a_k z^k$ as a polynomial in $\mathbb{C}[z]$ whose coefficients are all real numbers. We know that in that setting it factors as

$$p(z) = a_n(z - \zeta_1) \cdots (z - \zeta_n),$$

where $\zeta_k \in \mathbb{C}$ are the roots.

According to Problem 5.16, for any root ζ_i that is not a real number, one of the other roots ζ_j must be its conjugate: $\zeta_j = \bar{\zeta}_i$. If we pair up these two factors we get

$$(z - \zeta_i)(z - \zeta_j) = (z - \zeta_i)(z - \bar{\zeta}_i) = z^2 - (\zeta_i + \bar{\zeta}_i)z + \zeta_i\bar{\zeta}_i = z^2 - 2\mathrm{Re}(\zeta_i)z + |\zeta_i|^2,$$

which is a quadratic polynomial with *real* coefficients. By pairing each factor with a nonreal root with its conjugate counterpart, we obtain a factorization of $p(z)$ into linear factors (the $(z - \zeta_i)$ with $\zeta_i \in \mathbb{R}$) and quadratic factors, all with real coefficients. In this factorization, simply replace z by the original real variable x to obtain a factorization of $p(x)$ into a product of linear and quadratic factors, all with real coefficients. □

Problem 5.17 Consider the polynomial

$$p(x) = x^4 - 4x^3 + 2x^2 + 4x + 4$$

of Example 5.9.

a) Find all rational roots.

b) What does Descarte's Rule of Signs say about the number of positive roots? What can you deduce from the rule of signs about *negative* roots?

c) Verify that

$$p(x) = \left[(x-1)^2 - 2\right]^2 + 3.$$

From here find the full factorization of $p(z)$ into the product of first order terms and identify all the complex roots. (By writing $3 = -(i\sqrt{3})^2$ this is the difference of two squares: $A^2 - B^2 = (A+B)(A-B)$. Using the values of ξ and $\bar{\xi}$ from Problem 5.13 each of the two factors are themselves the difference of squares!)

d) Find a way to write $p(x)$ as a product of two quadratic polynomials each with real coefficients, as in Corollary 5.12.

..... Ber

Problem 5.18 As above, let $P_m(n)$ denote the summation formula

$$P_m(n) = \sum_{k=1}^{n} k^m,$$

Back in Chapter 1 (Proposition 1.11 specifically) we saw that $P_1(n) = \frac{n(n+1)}{2}$, $P_2(n) = \frac{n(n+1)(2n+1)}{6}$, and $P_3(n) = \frac{n^2(n+1)^2}{4}$. We observe that each of these is a polynomial in n. Does the following constitute a proof that $\sum_{k=1}^n k^m$ is always a polynomial in n?

Since

$$\sum_{k=1}^{n} k^m = 1 + \ldots + n^m,$$

and every term on the right is a polynomial, the sum is a polynomial.

(See [5] for the source of this one.)

..... poly

Chapter 6

Determinants and Linear Algebra in $\mathbb{R}^n$

A matrix is a rectangular array of numbers. We will assume that all the entries are real numbers. We typically use upper case letters to denote matrices, and the same letter (lower case) with subscripts to denotes its entries. In other words, if A is a matrix, we use $a_{i\,j}$ to denote the entry in row i and column j. We sometimes write $A = [a_{i\,j}]$ to indicate this notation. We will only be working with square matrices in this chapter, meaning there are the same number of rows as columns. We say A is an $n \times n$ matrix if there are n rows and n columns. So when we say A is the 5×5 matrix $[a_{i\,j}]$ we mean

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix}.$$

If we have two (square) matrices $A = [a_{i\,j}]$ and $B = [b_{i\,j}]$ of the same size $(n \times n)$ then we can add them, multiply them by each other, and multiply them by numbers $\alpha \in \mathbb{R}$: $A + B = [a_{i\,j} + b_{i\,j}]$, $\alpha A = [\alpha a_{i\,j}]$, and $AB = [c_{i\,j}]$ where the entries of the product are given by

$$c_{i\,j} = \sum_{k=1}^{n} a_{i\,k} b_{k\,j}. \tag{6.1}$$

The subject of this chapter is determinants of square matrices. We are going define the determinant of an $n \times n$ square matrix $A = [a_{ij}]$ directly using the following formula.

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i\,\sigma(i)}. \tag{6.2}$$

But before formally stating this as the defintion we need to understand what all the pieces of this formula are. The $\sigma \in S_n$ refer to permutations of $\{1, 2, 3, \ldots n\}$, which we will describe in §A. In §B we will discuss the sign of a permutation, $\operatorname{sgn}(\sigma)$. Then in §C we will state the above as the definition and begin using it to prove various properties of deteminants in the rest of the chapter.

Although the formula (6.2) may look formidable, it will reduce to the familiar formula for 2×2 matrices:

$$\det\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right) = a_{11}a_{22} - a_{21}a_{12}.$$

Maybe you have also seen the formula for 3×3 matrices; the determinant of

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

is given by

$$\det(A) = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} - a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}. \tag{6.3}$$

This likewise will just be formula (6.2) in the special case of $n = 3$. For 4×4 matricies the formula has 24 terms, more than we really want to write out[1]. For larger matrices you were probably never told exactly how its determinant is actually defined, but just given some rules or procedures for calculating it (without being told why they are valid). Our formula (6.2) above is the general defintion that you were never told. The various rules and procedures that you may have learned are valid because they are provable from the formula, as we will see below. This chapter is about working all this out. Many of the proofs we encounter here will be different than those we have considered previously; they will be proofs that are based on careful manipulation of a complicated formula. Often we will be able to "see" why what we want to prove is true by working out examples, and being convinced that what happened in the examples will always work. The challenge in writing a proof is to find ways to express in explicit and precise notation what we could see in an example.

A different approach to the study of determinants, typical of advanced algebra texts, is to start by prescribing a list of properties for $\det(A)$, and then show that these properties completely determine what $\det(A)$ must be, leading to our formula above. This is the axiomatic approach, which we leave to the algebra texts.

Problem 6.1 Use the formula (6.1) to show that for any two $n \times n$ matrices A and B and any scalar α, $(\alpha A)B = A(\alpha B)$. [Hint: to do this let S be the matrix αA. In other words $s_{i\,j} = \alpha a_{i\,j}$ for every i and j. The defintion of matrix multiplication says the $i\,j$ entry of $(\alpha A)B$ is

$$\sum_{k=1}^{n} s_{i\,k} b_{k\,j} = \sum_{k=1}^{n} (\alpha a_{i\,k}) b_{k\,j}.$$

In a similar manner work out the $i\,j$ entry of $A(\alpha B)$, and then explain why it is the same thing.]

..... sm

Problem 6.2 Suppose that A, B, and C are $n \times n$ matricies. Show that matrix multiplication is associative, i.e. that $(AB)C = A(BC)$. [Hint: follow the same approach as Problem 6.1.]

..... assoc

A Permutations: $\sigma \in S_n$

Each σ in our formula (6.2) stands for a permutation of $\{1, 2, 3, \ldots n\}$.

Definition. A *permutation* of $\{1, \ldots, n\}$ is a bijection from $\{1, 2, 3, \ldots n\}$ to itself. The set of all permutations of $\{1, \ldots, n\}$ is denoted S_n.

It is traditional to use lower case Greek letters (σ, π, τ, ϵ, $\ldots$) for the names of permutations. There are several different notations for identifying a specific permutation. One is to simply list the values in the form of a table:

$$\sigma = \begin{pmatrix} 1 & 2 & \ldots & n \\ \sigma(1) & \sigma(2) & \ldots & \sigma(n) \end{pmatrix}.$$

[1] In general there are $n!$ terms in (6.2). It was once pointed out to me by a famous mathematician (S. Ulam) that 52! is greater than the number of molecules in the known universe. That was in 1973; today's estimates might come out different. But in any case 52! is an extremely large number, larger than 10^{67}. It is the number of different ways to shuffle a deck of playing cards. It is also the number of terms in the formula (6.2) for a 52×52 matrix. So it is laughable to think about using (6.2) to actually evaluate a determinant of any significant size. As a mathematical expression however it is still perfectly valid. Fortunately the properties that follow from the formula lead to much better ways to compute determinants of big matrices (it only takes on the order of n^3 operations if you do it sensibly). That's fortunate, since in many applications matrices with n in the thousands are common.

For instance, the permutation of $\{1,2,3,4,5\}$ defined by $\sigma(1)=5,\ \sigma(2)=2,\ \sigma(3)=1,\ \sigma(4)=4,\ \sigma(5)=3$ would be written

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 2 & 1 & 4 & 3 \end{pmatrix}.$$

Another notation is the cycle notation which we will talk about on the next page.

The set of all permutations of $\{1,2,\ldots n\}$ is denoted S_n. Thus the outside sum $\sum_{\sigma\in S_n}$ in (6.2) refers to what we get by taking each different permutation σ of $\{1,2,3,\ldots n\}$, compute the expression $\operatorname{sgn}(\sigma)\prod_{i=1}^{n} a_{i\,\sigma(i)}$, and then add up the results. For a given permutation, the inside product $\prod_{i=1}^{n} a_{i\,\sigma(i)}$ refers to

$$\prod_{i=1}^{n} a_{i\,\sigma(i)} = a_{1\,\sigma(1)}a_{2\,\sigma(2)}\cdots a_{n\,\sigma(n)}.$$

The values $a_{i\,\sigma(i)}$ are a selection of n entries from the matrix with exactly one from each row and exactly one from each column. For instance, if $n=5$ and σ is our example permutation above, then the product involves the boxed entries below:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \boxed{a_{15}} \\ a_{21} & \boxed{a_{22}} & a_{23} & a_{24} & a_{25} \\ \boxed{a_{31}} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & \boxed{a_{44}} & a_{45} \\ a_{51} & a_{52} & \boxed{a_{53}} & a_{54} & a_{55} \end{bmatrix}.$$

So for this σ,

$$\prod_{i=1}^{n} a_{i\,\sigma(i)} = a_{15}a_{22}a_{31}a_{44}a_{53}.$$

In the next section we will talk about $\operatorname{sgn}(\sigma)$. For the moment let's just point out that $\operatorname{sgn}(\sigma)$ will always be ± 1. So (6.2) stands for an expression consisting of the products of all possible combinations of n terms from A, one per row and one per column, either added or subtracted depending on $\operatorname{sgn}(\sigma)$. At this point you can already see that (6.2) refers to a formula of the same general type as of the determinant expressions for 2×2 and 3×3 matrices that we wrote down above. When $n=2$ there are only two permutations in S_2, the inversion τ and the identity map ϵ:

$$\tau = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}; \quad \epsilon = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}.$$

If you will take for granted that $\operatorname{sgn}(\tau)=-1$ and $\operatorname{sgn}(\epsilon)=1$, then we find see that (6.2) reduces to the familiar formula for 2×2 determinants.

$$\begin{aligned} \prod_{i=1}^{n} a_{i\,\epsilon(i)} &= a_{11}a_{22}, \text{ using } \sigma=\epsilon \\ \prod_{i=1}^{n} a_{i\,\tau(i)} &= a_{12}a_{21}, \text{ using } \sigma=\tau \\ \det(A) &= 1\cdot(a_{11}a_{22}) + (-1)\cdot(a_{12}a_{21}). \end{aligned}$$

A more concise notation for permutations is the *cycle notation*. Our example

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 2 & 1 & 4 & 3 \end{pmatrix}$$

would be written[2] in cycle notation as

$$\sigma = \langle 1, 5, 3 \rangle.$$

The meaning is that the first value gets sent to the second, the second gets sent to the third, ..., and the last gets sent back to the first. In our example we might indicate this by $1 \to 5 \to 3 \to 1$. Since 2 and 4 don't appear, the understanding is that σ leaves them unchanged: $\sigma(2) = 2$ and $\sigma(4) = 4$.

Definition. A permutation $\sigma \in S_n$ is called a *k-cycle* if there is a set of k elements, $\{i_1,\ i_2, \ldots, i_k\} \subseteq \{1, \ldots, n\}$ so that $\sigma(i_1) = i_2$, $\sigma(i_2) = i_3$, ..., $\sigma(i_k) = i_1$, and $\sigma(j) = j$ for all other j. We use

$$\sigma = \langle i_i,\ i_2, \ldots, i_k \rangle$$

to indicate such a permutation. A 2-cycle is also called a *transposition*.

The σ of our example is a 3-cycle. $\langle 1, 3, 5, 2 \rangle$ would be a 4-cycle. The cycle notation is more concise since we only need to write one row, and don't need to write anything for the $\sigma(i) = i$ terms. However not every permutation is a single cycle. For instance

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 1 & 2 & 4 \end{pmatrix}.$$

consists of two "disjoint" cycles. We write

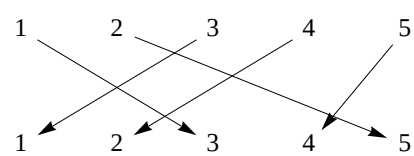
$$\pi = \langle 1, 3 \rangle \langle 2, 5, 4 \rangle.$$

In general when two or more permutations are written next to each other, as if they were being multiplied, we mean that they are to be composed as functions.

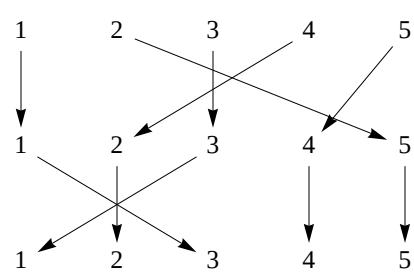
$$\sigma\pi = \sigma \circ \pi.$$

Thus $\sigma\pi$ applied to i is $(\sigma\pi)(i) = \sigma(\pi(i))$. With that understanding you can check that $\tau = \langle 1, 3 \rangle \langle 2, 5, 4 \rangle$ is indeed correct in our example.

A useful way to illustrate a permutation σ is write out $1, 2, \ldots, n$ in two rows and then draw arrows from i in the top row to $\sigma(i)$ in the bottom row. For instance our example π above would be illustrated as follows.



We will refer to this as the *diagram* of the permutation. Such diagrams are convenient for composing two permutations; we can just stack the two diagrams on top of each other and follow the arrows through both layers to get to find the composition. Here is the illustration for our example of $\pi = \langle 1, 3 \rangle \langle 2, 5, 4 \rangle$.



[2]It is common to simply use parentheses to denote a cycle, "$(1, 5, 3)$" instead of "$\langle 1, 5, 3 \rangle$." We however are choosing to use angle brackets to more easily distinguish the notation for the permutation from the parentheses used to demark its argument. Thus $\langle 1, 5, 3 \rangle(5) = 3$ seems a little clearer than $(1, 5, 3)(5) = 3$.

Note that the rightmost permutation is applied first, in keeping with our usual function notation $\sigma\pi(i) = \sigma(\pi(i))$.

If σ, π, and α are three permutations, then $(\sigma\pi)\alpha = \sigma(\pi\alpha)$ since both refer to the permutation that takes i to $\sigma(\pi(\alpha(i)))$. Thus the parentheses are unnecessary; we can just write $\sigma\pi\alpha$ with no ambiguity.

Problem 6.3

a) Let $\mathcal{I} : S_n \to S_n$ be defined by $\mathcal{I}(\sigma) = \sigma^{-1}$. Explain why $\mathcal{I}$ is a bijection.

b) Suppose $\pi \in S_n$, and for this π define $\mathcal{C}_\pi : S_n \to S_n$ be defined by $\mathcal{C}_\pi(\sigma) = \pi\sigma$. Explain why $\mathcal{C}_\pi$: is a bijection.

c) Show that $(\sigma\tau)^{-1} = \tau^{-1}\sigma^{-1}$ for all $\sigma, \tau \in S_n$. (Note that the order on the right is reversed!)

d) Show that if τ is a transpositon, then $\tau^{-1} = \tau$. Is the converse true?

e) Suppose $\sigma \in S_n$ is any permutation and $i_1, \ldots, i_k$ are k distinct elements of $\{1, \ldots, n\}$. Show that

$$\sigma\langle i_1, i_2, \ldots, i_k\rangle\sigma^{-1} = \langle \sigma(i_1), \sigma(i_2), \ldots, \sigma(i_k)\rangle.$$

[Hint: Consider any $j \in \{1, \ldots, n\}$. You want to show that both sides of the above send j to the same thing. Consider cases: (1) j is one of the values $i_1, \ldots i_{k-1}$, (2) $j = i_k$, (3) j is not one of the values $i_1, \ldots, i_k$.]

..... d1

Lemma 6.1. *Every permutation $\sigma \in S_n$ which is not the identity can be written as the composition of some number $k \leq n-1$ of transpositions τ_i:*

$$\sigma = \tau_1\tau_2 \ldots \tau_k.$$

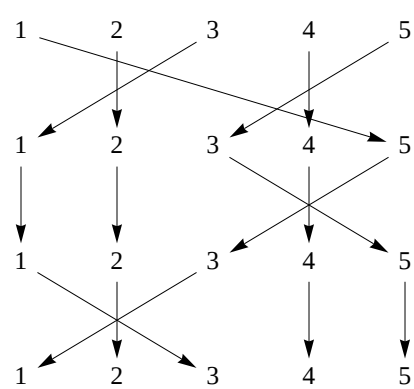
Using parts 3 and 4 of Problem 6.3, we see that the conclusion of the lemma is equivalent to

$$\tau_k \ldots \tau_2\tau_1\sigma = \epsilon \quad \text{(the identity permutation).}$$

So what we want to do is show that σ can be followed by a sequence of transpositions so as to "untangle" the diagram and get every i back to its starting position. For instance consider our example from page 112.

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 2 & 1 & 4 & 3 \end{pmatrix}.$$

We start by drawing the diagram for σ. We can first get 5 back to its starting position using $\tau_1 = \langle 5, 3\rangle$. Then using $\tau_2 = \langle 1, 3\rangle$ will finish restoring all i to there initial positions. Here is the combined picture.



This shows that $\tau_2\tau_1\sigma = \epsilon$, which means that $\sigma = \tau_1\tau_2$. Our proof of the lemma consists simply of describing a process for choosing the transpositions which will always accomplish this.

Proof. If σ is not the identity there is some value of i which is not fixed: $\sigma(i) \neq i$. Let ℓ be the *largest* value which is not fixed by σ: $\sigma(\ell) \neq \ell$. This means that for $\ell < m \leq n$ we have $\sigma(m) = m$. It must be that $\sigma(\ell) < \ell$, because if $\sigma(\ell) > \ell$, then $\sigma(\sigma(\ell)) = \sigma(\ell)$, contrary to the fact that σ is injective. Now let $\tau_1 = \langle \sigma(\ell), \ell \rangle$. Then

$$\sigma' = \tau_1\sigma$$

not only leaves $\ell + 1, \ldots, n$ fixed, it leaves ℓ fixed as well.

If $\sigma' = \epsilon$ we are done. Otherwise repeat the process above for σ': let ℓ' be the largest value not fixed by σ'. We know that $\ell' \leq \ell - 1$. Let $\tau_2 = \langle \sigma'(\ell'), \ell' \rangle$. Then

$$\sigma'' = \tau_2\sigma' = \tau_2\tau_1\sigma$$

must leave all the values from ℓ' up to n fixed. If $\sigma'' = \epsilon$ we are done. Otherwise we repeat the process again, and continue to we reach the identity permutation ϵ.

Since the value of ℓ goes down by at least one at each step, and can never be less than 2, the process must terminate after some number $k \leq n - 1$ steps, resulting in

$$\epsilon = \tau_k \ldots \tau_2\tau_1\sigma.$$

This implies that

$$\sigma = \tau_1\tau_2 \ldots \tau_k,$$

completing the proof. □

This is a somewhat informal proof, because of it's reliance on "repeat the process." It really is the description of an algorithm to produce the $\tau_1, \ldots, \tau_k$. We could present it more formally as a strong induction argument on the value of ℓ, or a proof by contradiction, but those would probably be less clear than the informal description above.

Problem 6.4

a) Write the permutation

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 5 & 6 & 3 & 1 \end{pmatrix}$$

as the compotion of cycles.

b) Write $\pi = \langle 3, 2, 5 \rangle \langle 2, 5, 4 \rangle$ in "table" notation: $\pi = \begin{pmatrix} 1 & 2 & \cdots & 5 \\ \cdot & \cdot & \cdots & \cdot \end{pmatrix}$.

c) Write $\gamma = \langle 3, 1, 2, 5 \rangle$ as the composition of transpositions.

..... `permrep`

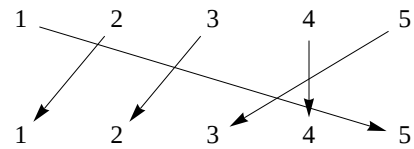
B The Sign of a Permutation: $\mathrm{sgn}(\sigma)$

Next we need to define the sign of a permutation. As we have already said, $\mathrm{sgn}(\sigma)$ will always be ± 1, but we need to explain how the choice of $\pm$ is determined.

As an example consider the following permutation.

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 1 & 2 & 4 & 3 \end{pmatrix}.$$

Here is it's diagram.



Now consider a pair of input values from the top row, 2 and 5 for example. The order of this particular pair is preserved by σ: $2 < 5$ and $\sigma(2) < \sigma(5)$. But for other input pairs the order of the corresponding outputs is the opposite of the inputs: $1 < 3$ while $\sigma(1) > \sigma(3)$. We see this in the picture because the lines from 2 and 5 do not cross each other, while the lines from 1 and 3 do cross each other. We say that such pairs are *inverted* by σ. In general we want to count the number of pairs[3] $\{i, j\}$ (of distinct inputs: $i \neq j$) which are inverted by σ. In a picture like the one above, this is the number of line crossings. In notation, an inversion occurs when $i < j$ and $\sigma(j) < \sigma(i)$.

Definition. For $\sigma \in S_n$, we define the *inversion count* of σ to be

$$r(\sigma) = |\{\{i, j\} : \ 1 \leq i < j \leq n \text{ and } \sigma(j) < \sigma(i)\}| \, .$$

The *sign* of the permutation σ is defined to be

$$\operatorname{sgn}(\sigma) = (-1)^{r(\sigma)}.$$

If $\operatorname{sgn}(\sigma) = +1$ we say σ is an *even* permutation; if $\operatorname{sgn}(\sigma) = -1$ we say σ is an *odd* permutation.

("$|A|$" is the notation from Chapter 3 for the number of elements in a finite set A.) For the σ above, the pairs that are reversed are $\{1, 2\}, \{1, 3\}, \{1, 4\}, \{1, 5\}, \{4, 5\}$ and no others. So we find $r(\sigma) = 5$, $\operatorname{sgn}(\sigma) = -1$, and thus in our example σ is odd.

Problem 6.5 Show that $r(\langle i, j \rangle) = 2|j - i| - 1$. Thus transpositions are always odd. (Although looking at examples may help you understand the formula, don't just offer an example as a proof! To write a proof, suppose $i < j$. Now describe exactly which pairs $\{k, m\}$ (with $k < m$) are reversed. It will depend on whether k and m are less than i equal to i between i and j, equal to j, or greater than j. With an accurate description of this type, you can now count exactly how many pairs there are which will be reversed.)

..... d5

Problem 6.6 Let $\rho \in S_n$ be

$$\rho = \begin{pmatrix} 1 & 2 & \dots & n \\ n & n-1 & \dots & 1 \end{pmatrix}.$$

What is $r(\rho)$?

..... d6

Here is the essential fact about $\operatorname{sgn}(\sigma)$.

Theorem 6.2. *For all $\sigma, \pi \in S_n$, $\operatorname{sgn}(\sigma\pi) = \operatorname{sgn}(\sigma)\operatorname{sgn}(\pi)$.*

Proof. We divide the distinct pairs $\{i, j\}$ into four different types.

$$\begin{aligned}
A &= \{\{i, j\} \ : \ \pi \text{ reverses } \{i, j\} \text{ but } \sigma \text{ does not reverse} \{\pi(i), \pi(j)\}\} \\
B &= \{\{i, j\} \ : \ \pi \text{ does not reverse } \{i, j\} \text{ but } \sigma \text{ does reverse } \{\pi(i), \pi(j)\}\} \\
C &= \{\{i, j\} \ : \ \pi \text{ reverses } \{i, j\} \text{ and } \sigma \text{ reverses } \{\pi(i), \pi(j)\}\} \\
D &= \{\{i, j\} \ : \ \pi \text{ does not reverse } \{i, j\} \text{ and } \sigma \text{ does not reverse } \{\pi(i), \pi(j)\}\}.
\end{aligned}$$

[3] We have used "$\{i, j\}$" instead of "(i, j)" to refer to a pair because we don't want to distinguish between (i, j) and (j, i), as the use of parentheses would imply. It's just the set of two (distinct) vaules $i \neq j$ that we are trying to refer to.

Let $a = |A|$, $b = |B|$, $c = |C|$. Clearly

$$r(\pi) = a + c.$$

Also,

$$r(\sigma) = b + c.$$

This is because in counting the pairs $\{k, l\}$ that σ reverses, we can identify them as $k = \pi(i)$, $l = \pi(j)$. Thirdly

$$r(\sigma\pi) = a + b,$$

because the pairs in C are reversed by σ but then reversed back to their original order by π so that the combined effect of $\sigma\pi$ is to preserve their order. We now have

$$\operatorname{sgn}(\sigma)\operatorname{sgn}(\pi) = (-1)^{a+c}(-1)^{b+c} = (-1)^{a+b}(-1)^{2c} = (-1)^{a+b} = \operatorname{sgn}(\sigma\pi).$$

□

Corollary 6.3. *For any permutation σ,* $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\sigma^{-1})$.

Problem 6.7 Prove Corollary 6.3.

..... `sgninv`

Problem 6.8 Show that σ is an odd permutation if and only if it can be written as the composition of an odd number of transpositions. (Don't confuse the term "transposition" with the term "reversal.")

..... `d8`

C Definition and Basic Properties

Now that we have explained its components, we can state the definition of the determinant of a matrix.

Definition. If $A = [a_{ij}]$ is an $n \times n$ matrix, its *determinant* is defined to be

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i\,\sigma(i)}. \tag{6.4}$$

Problem 6.9 Show that for $n = 3$ the definition agrees with (6.3).

..... `dim3`

Our goal in the rest of this section is to use the definition to prove basic properties of the determinants. We begin with triangular matricies, for which the determinant is easy to calculate. A matrix $T = [t_{ij}]$ is called *lower triangular* if $t_{ij} = 0$ whenever $i < j$. T is called *upper triangular* if $t_{ij} = 0$ whenever $i > j$.

Lemma 6.4. *If $T = [t_{ij}]$ is either lower triangular or upper triangular, then*

$$\det(T) = \prod_{i=1}^{n} t_{ii}.$$

In other words the determinant of a triangular matrix is simply the product of its diagonal entries. The proof of this lemma is our first example of a proof based on the defintion (6.4). We start by writing down what the defintion of $\det(T)$ is, and then use the assumptions about t_{ij} to manipulate it.

Proof. Suppose T is lower triangular. By defintion,

$$\det(T) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} t_{i\,\sigma(i)}.$$

Observe that if $\sigma \in S_n$ is any permutation other than the identity permutation, then there must exist some i' with $i' < \sigma(i')$. (Indeed let i' be the smallest value such that $\sigma(i') \neq i'$. It cannot be that $\sigma(i') < i'$ since that would imply that $\sigma(\sigma(i')) = \sigma(i')$ contrary to the injectivity of σ. Therefore $i' < \sigma(i')$.) But that means $t_{i'\,\sigma(i')} = 0$, so that

$$\prod_{i=1}^{n} t_{i\,\sigma(i)} = 0.$$

Thus in the formula for $\det(T)$ there is only one $\sigma \in S_n$ for which the product is nonzero, namely the identity permutation $\sigma = \epsilon$. Since $\operatorname{sgn}(\epsilon) = 1$, the expression defining the determinant collapses to

$$\det(T) = 1 \prod_{i=1}^{n} t_{i\,\epsilon(i)} = \prod_{i=1}^{n} t_{i\,i}.$$

The upper triangular case is proven in a similar way. □

Definition. The $n \times n$ identity matrix is the matrix $I = [\delta_{ij}]$ where $\delta_{ij} = 1$ if $i = j$ and 0 if $i \neq j$:

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Corollary 6.5. *If I is the $n \times n$ identity matrix, then* $\det(I) = 1$.

The *transpose* of a matrix $A = [a_{ij}]$ is the matrix $A^T = [a'_{ij}]$ where $a'_{ij} = a_{ji}$. The rows of A become the columns of A^T.

Lemma 6.6. $\det(A^T) = \det(A)$.

Proof. Using the notation just introduced, $a'_{i\sigma(i)} = a_{\sigma(i)i} = a_{j\sigma^{-1}(j)}$, where $j = \sigma(i)$. As the values of i run through $1, 2, \ldots, n$ the corresponding values of j will also run through $1, 2, \ldots, n$, just in a different order (since σ is a bijection). It follows that

$$\prod_{i=1}^{n} a'_{i\sigma(i)} = \prod_{j=1}^{n} a_{j\sigma^{-1}(j)}.$$

Also observe that as a consequence of Corollary 6.3 we know that $\operatorname{sgn}(\sigma^{-1}) = \operatorname{sgn}(\sigma)$. So we have

$$\begin{aligned} \det(A^T) &= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a'_{i\sigma(i)} \\ &= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma^{-1}) \prod_{j=1}^{n} a_{j\sigma^{-1}(j)} \\ &= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{j=1}^{n} a_{j\sigma(j)} \\ &= \det(A). \end{aligned}$$

The equality between the second and third line is due to the fact that the set of all possible σ^{-1} is the same as the set of all possible σ. □

You have probably learned techniques for computing determinants based on manipulating the rows or columns of a matrix. To justify those techniques is our next goal. We will work in terms of columns because that is most natural for the discussion of Cramer's Rule in the next section. It will be helpful to have a standard notation for the columns of a matrix $A = [a_{ij}]$. The jth column of A is the element of $\mathbb{R}^n$ which we will denote by

$$\hat{a}_j = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{bmatrix}.$$

(Think of the "hat" ˆ as a little pointer reminding us to write it in an up-down orientation.) To save space we will often type a column vector horizontally with a transpose symbol to remind us to stand it back up as a column. For instance we will write $\hat{v} = (-1, 5, 7)^T$ instead of $\hat{v} = \left[\begin{smallmatrix} -1 \\ 5 \\ 7 \end{smallmatrix}\right]$.)

We want to view the determinant as a function D of the n columns of the matrix.

$$\det(A) = D(\hat{a}_1, \hat{a}_2, \dots \hat{a}_n).$$

Here are the properties of the determinant as a function of the columns.

Theorem 6.7. *Consider any collection of column vectors $\hat{a}_i, \hat{b}_i \in \mathbb{R}^n$.*

a) $D(\hat{a}_1, \cdots, \gamma\hat{a}_k, \cdots \hat{a}_n) = \gamma D(\hat{a}_1, \cdots, \hat{a}_k, \cdots \hat{a}_n)$*, for any scalar* $\gamma \in \mathbb{R}$.

b) $D(\hat{a}_1, \cdots, \hat{a}_k + \hat{b}_k, \cdots \hat{a}_n) = D(\hat{a}_1, \cdots, \hat{a}_k, \cdots \hat{a}_n) + D(\hat{a}_1, \cdots, \hat{b}_k, \cdots \hat{a}_n)$.

c) $\operatorname{sgn}(\pi) D(\hat{a}_{\pi(1)}, \cdots, \hat{a}_{\pi(i)}, \cdots \hat{a}_{\pi(n)}) = D(\hat{a}_1, \cdots, \hat{a}_i, \cdots \hat{a}_n)$*, for any permutation* $\pi \in S_n$.

Corollary 6.8. *Suppose A is an $n \times n$ matrix with columns $\hat{a}_i$ and α is any real number.*

a) If two columns duplicate each other, in other words $\hat{a}_i = \hat{a}_j$ for some $i \neq j$, then $D(\hat{a}_1, \cdots \hat{a}_n) = 0$.

b) For any $i \neq j$ if we add α times $\hat{a}_i$ to $\hat{a}_j$ the determinant does not change:

$$D(\hat{a}_1, \cdots, \hat{a}_i, \cdots, \hat{a}_j + \alpha\hat{a}_i, \dots, \hat{a}_n) = D(\hat{a}_1, \cdots, \hat{a}_i, \cdots, \hat{a}_j, \dots, \hat{a}_n).$$

c) $\det(\alpha A) = \alpha^n \det(A)$.

Example 6.1. Before writing a proof of the theorem, here is an example of how these column operations can be used to evaluate a determinant. Suppose we want to calculate the following determinant.

$$\det \begin{pmatrix} 0 & 3 & 2 \\ -1 & -6 & 6 \\ 5 & 9 & 1 \end{pmatrix}.$$

The strategy is apply the properties of the theorem and corollary to manipulate the matrix into one which

is triangular, so that determinant will be easy to evaluate.

$$\begin{aligned}
& \det\begin{pmatrix} 0 & 3 & 2 \\ -1 & -6 & 6 \\ 5 & 9 & 1 \end{pmatrix} \\
\text{by permuting the columns using } \pi = \langle 1,2\rangle \text{ this } &= -\det\begin{pmatrix} 3 & 0 & 2 \\ -6 & -1 & 6 \\ 9 & 5 & 1 \end{pmatrix} \\
\text{now by taking a factor of 3 out of the first column } &= -3\det\begin{pmatrix} 1 & 0 & 2 \\ -2 & -1 & 6 \\ 3 & 5 & 1 \end{pmatrix} \\
\text{by adding } -2 \text{ times the first column to the third column } &= -3\det\begin{pmatrix} 1 & 0 & 0 \\ -2 & -1 & 10 \\ 3 & 5 & -5 \end{pmatrix} \\
\text{by adding 10 times the second column to the third column } &= -3\det\begin{pmatrix} 1 & 0 & 0 \\ -2 & -1 & 0 \\ 3 & 5 & 45 \end{pmatrix} \\
&= (-3)\cdot 1\cdot(-1)\cdot(45) \\
&= 135.
\end{aligned}$$

We turn now to the proof. This proof is a good example of what we said in the introduction. We could convince ourselves of each property by looking at examples, but to write a proof we need to find a way to write demonstrate the property in general, based on the definition of determinant. To do that it can be helpful to introduce new notation for some of the other matrices involved in the calculation, like you did in Problem 6.1.

Proof of Theorem. For part a), let $G = [g_{ij}]$ be the matrix with

$$g_{ij} = \begin{cases} a_{ij} & \text{if } j \neq k \\ \gamma a_{ij} & \text{if } j = k \end{cases}$$

Part a) claims that $\det(G) = \gamma\det(A)$. To see this just observe that for any permutation,

$$\prod_{i=1}^{n} g_{i\sigma(i)} = \gamma \prod_{i=1}^{n} a_{i\sigma(i)}.$$

This is because $\sigma(i) = k$ occurs exactly once in the product. Multiplying this by $\operatorname{sgn}(\sigma)$ and summing over all σ proves a).

You will write the proof of b) as Problem 6.10 below.

Turning to c), let $P = [a_{i\,\pi(j)}]$. The left side in c) is $\operatorname{sgn}(\pi)\det(P)$ because $\hat{p}_j = \hat{a}_{\pi(j)}$. For any $\sigma \in S_n$ we have

$$\prod_{i=1}^{n} p_{i\sigma(i)} = \prod_{i=1}^{n} a_{i\,\pi(\sigma(i))}.$$

So we have

$$\operatorname{sgn}(\pi)\operatorname{sgn}(\sigma)\prod_{i=1}^{n} p_{i\sigma(i)} = \operatorname{sgn}(\pi\sigma)\prod_{i=1}^{n} a_{i\,\pi\sigma(i)}.$$

Next observe that $\sigma \mapsto \pi\sigma$ is a bijection of S_n, so that summing over $\sigma \in S_n$ is equivalent to summing over

$\pi\sigma \in S_n$. Therefore

$$\begin{aligned}
\operatorname{sgn}(\pi) D(\hat{a}_{\pi(1)}, \cdots, \hat{a}_{\pi(i)}, \cdots \hat{a}_{\pi(i)}) &= \operatorname{sgn}(\pi) \det(P) \\
&= \operatorname{sgn}(\pi) \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} p_{i\,\sigma(i)} \\
&= \sum_{\sigma \in S_n} \operatorname{sgn}(\pi\sigma) \prod_{i=1}^{n} a_{i\,\pi\sigma(i)} \\
&= \sum_{\pi\sigma \in S_n} \operatorname{sgn}(\pi\sigma) \prod_{i=1}^{n} a_{i\,\pi\sigma(i)} \\
&= \sum_{\beta \in S_n} \operatorname{sgn}(\beta) \prod_{i=1}^{n} a_{i\,\beta(i)} \\
&= D(\hat{a}_1, \cdots \hat{a}_i \cdots \hat{a}_n).
\end{aligned}$$

□

This proof depends on the notation a lot! You can't just gloss over the mathematical symbols and only read the words. You really need to scrutinize the notation to understand exactly what it is saying, and be sure that you agree.

Proof of Corollary, part a). Suppose $\hat{a}_i = \hat{a}_j$ and let $\pi = \langle i, j \rangle$. Then the two determinants in c) of the theorem are the same. Since $\operatorname{sgn}(\pi) = -1$, we have

$$\det(A) = -\det(A)$$

which implies that $\det(A) = 0$. □

Problem 6.10 Write a proof for part b) of Theorem 6.7. To do this define the matricies $C = [c_{ij}]$ and $H = [h_{ij}]$ by

$$c_{ij} = \begin{cases} a_{ij} & \text{if } j \neq k \\ a_{ij} + b_{ij} & \text{if } j = k \end{cases}, \quad h_{ij} = \begin{cases} a_{ij} & \text{if } j \neq k \\ b_{ij} & \text{if } j = k \end{cases}.$$

The task is to show $\det(C) = \det(A) + \det(H)$. Start by writing down the defintion of $\det(C)$. The essential step is to explain why

$$\prod_{i=1}^{n} c_{i\sigma(i)} = \prod_{i=1}^{n} a_{i\sigma(i)} + \prod_{i=1}^{n} h_{i\sigma(i)},$$

and then use that to finish showing that $\det(C) = \det(A) + \det(H)$.

..... `colsum`

Problem 6.11 Prove parts b) and c) of Corollary 6.8.

..... `corbc`

Problem 6.12 Suppose $C = AB$ where A, B and C are square matricies of the same size. Explain why the jth column of C is given by the following formula involving the columns $\hat{a}_k$ of A and the entries of the jth column of B.

$$\hat{c}_j = \sum_k b_{kj} \hat{a}_k.$$

[Hint: What is the ith component of each side?]

colprod

Problem 6.13 Find the determinants of the following matrices using the using the manipulations discussed above.

$$A = \begin{bmatrix} 1 & -2 & 0 & 0 & 0 \\ 3 & -7 & 0 & 0 & 0 \\ 1 & 0 & 1 & 2 & 0 \\ 5 & -4 & 0 & 1 & 1 \\ 3 & 2 & 1 & 1 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 3 & 0 & -1 & 0 & 0 \\ 5 & 2 & 0 & 0 & 1 \\ -3 & 0 & 1 & 4 & 0 \\ 6 & 0 & 4 & 2 & 1 \end{bmatrix}$$

d10

The next theorem is very important, but it is just a consequence of Theorem 6.7 above.

Theorem 6.9. $\det(AB) = \det(A)\det(B)$.

Proof. Let $C = AB$. Observe that the multiplication formula $c_{ij} = \sum_k a_{ik}b_{kj}$ can be interpreted as saying that the columns $\hat{c}_j$ of C are obtained from the columns of A as

$$\hat{c}_j = \sum_k \hat{a}_k b_{kj}.$$

(This is Problem 6.12 above.) Using this for each entry in

$$\det(AB) = D(\hat{c}_1, \dots \hat{c}_n),$$

and using parts a) and b) of Theorem 6.7 we obtain

$$\det(AB) = \sum_{k_1=1}^{n} \cdots \sum_{k_n=1}^{n} D(\hat{a}_{k_1}, \dots, \hat{a}_{k_n}) \prod_{i=1}^{n} b_{k_i\, i}.$$

Now if any two of the k_i are the same, then $D(\hat{a}_{k_1}, \dots, \hat{a}_{k_n}) = 0$ by the corollary. So the only $(k_1, \dots, k_n)$ that we need to consider are those which correspond to a permutation: $k_i = \sigma(i)$ for some $\sigma \in S_n$. Thus by part c) of the theorem we have

$$\begin{aligned} \det(AB) &= \sum_{\sigma \in S_n} D(\hat{a}_{\sigma(1)}, \dots, \hat{a}_{\sigma(n)}) \prod_{i=1}^{n} b_{\sigma(i)i} \\ &= \sum_{\sigma \in S_n} D(\hat{a}_1, \dots, \hat{a}_n) \operatorname{sgn}(\sigma) \prod_{i=1}^{n} b_{\sigma(i)i} \\ &= \det(A) \left[\sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} b_{\sigma(i)i} \right] \\ &= \det(A)\det(B^T) = \det(A)\det(B), \end{aligned}$$

which completes the proof. □

D Cofactors and Cramer's Rule

In this section we will develop some of the interesting properties of determinants related to cofactors. You may have learned how to calculate a determinant using the "cofactor expansion" along some row or column. You also may have learned a method called "Cramer's Rule" for solving a system of linear equations using determinants. This too is related to cofactors.

If we were to write out all the terms of $\det(A)$ and collect all of them which include a factor of a_{ij}, the *ij-cofactor of A* is the quantity C_{ij} which multiplies a_{ij}:

$$\det(A) = a_{ij}C_{ij} + (\cdots \text{ terms without } a_{ij} \cdots)$$

To find a more explicit expression, let $\hat{e}_k$ denote the column vector with a 1 in the kth position and 0 in all other positions:

$$\hat{e}_k = (0 \cdots 0 \overset{k}{1} 0 \cdots 0)^T. \tag{6.5}$$

The jth column of A is

$$\hat{a}_j = \sum_i a_{ij}\hat{e}_i.$$

So

$$\begin{aligned} \det(A) &= D(\hat{a}_1, \ldots \hat{a}_j, \ldots \hat{a}_n) \\ &= D(\hat{a}_1, \ldots \sum_i a_{ij}\hat{e}_i, \ldots \hat{a}_n) \\ &= \sum_i a_{ij} D(\hat{a}_1, \ldots \overset{j}{\hat{e}_i}, \ldots \hat{a}_n). \end{aligned} \tag{6.6}$$

Using this, we see that

$$C_{ij} = D(\hat{a}_1, \ldots, \overset{j}{\hat{e}_i}, \ldots \hat{a}_n),$$

the determinant of the matrix that agrees with A except that its jth column has been replaced by $\hat{e}_i$. Moreover, the calculation (6.6) proves the following theorem. This is called the *cofactor expansion of* $\det(A)$ *along the jth column.*

Theorem 6.10 (Cofactor Expansion)**.** *If A is an $n \times n$ matrix then, for any choice of column $j = 1, \ldots, n$,*

$$\det(A) = \sum_i a_{ij}C_{ij},$$

where C_{ij} are the cofactors of A.

Problem 6.14

a) Let C'_{ji} be the j, i cofactor of C^T. Explain why $C'_{ji} = C_{ij}$.

b) Show that the cofactor expansion along any row is valid also: for any i,

$$\det(A) = \sum_j a_{ij}C_{ij}.$$

..... `cofalt`

You may have learned to calculate cofactors using the minors of A. If A is $n \times n$, form an $(n-1) \times (n-1)$ matrix by deleting the ith row and j^{th} column from A, and take its determinant. The result is called the *ij minor* of A, denoted M_{ij}. The proof of the following lemma is Problem 6.33 below.

Lemma 6.11. $C_{ij} = (-1)^{i+j}M_{ij}$.

Example 6.2. We illustrate cofactor expansions with the same determinant as Eample 6.1.

$$A = \begin{bmatrix} 0 & 3 & 2 \\ -1 & -6 & 6 \\ 5 & 9 & 1 \end{bmatrix}.$$

We can choose which column to use; we choose to use the first column. The cofactors from the first column (computed using minors) are

$$\begin{aligned} C_{1\,1} &= +[-6-54] = -60, \\ C_{2\,1} &= -[3-18] = 15, \\ C_{3\,1} &= +[18-(-12)] = 30. \end{aligned}$$

Now we use the formula from Theorem 6.10 for $j = 1$.

$$\det(A) = \sum_{i=1}^{3} a_{i1}C_{i1} = 0C_{1\,1} + (-1)C_{2\,1} + 5C_{3\,1} = 0 \cdot (-60) - 1 \cdot 15 + 5 \cdot 30 = 135.$$

We could have used any column. The first column is easiest because we really didn't need to work out $C_{1\,1}$.

Problem 6.15 Use cofactor expansions to calculate the determinants from Problem 6.13.
..... `cofcalcs`

Definition. Suppose A is an $n \times n$ matrix. The *adjoint* of A is the matrix obtained by forming the transpose of the matrix of cofactors:

$$\mathrm{Adj}(A) = [C_{ij}]^T.$$

In other words $\mathrm{Adj}(A) = [\tilde{a}_{ij}]$ where $\tilde{a}_{ij} = C_{ji}$.

Example 6.3. To illustrate we return again to the matrix A of Examples 6.1 and 6.2. We already worked out the cofactors from the first column. The others are (you can check)

$$\begin{aligned} C_{1\,2} &= 29 & \qquad C_{1\,3} &= 39 \\ C_{2\,2} &= -10 & \qquad C_{2\,3} &= 15 \\ C_{3\,2} &= 2 & \qquad C_{3\,3} &= -3. \end{aligned}$$

So the matrix of cofactors is

$$[C_{ij}] = \begin{bmatrix} -60 & 31 & 21 \\ 15 & -10 & 15 \\ 30 & -2 & -3 \end{bmatrix}.$$

The adjoint is the transpose of this:

$$\mathrm{Adj}(A) = \begin{bmatrix} -60 & 15 & 30 \\ 31 & -10 & -2 \\ 21 & 15 & -3 \end{bmatrix}.$$

The adjoint has a remarkable property involving the identity matrix.

Theorem 6.12. *For any square matrix A,*

$$\mathrm{Adj}(A)A = \det(A)I,$$

where I is the $n \times n$ identity matrix.

Proof. Consider any pair i, j; we will calculate the i, j entry of $\mathrm{Adj}(A)A$. Using the notation $\mathrm{Adj}(A) = [\tilde{a}_{ij}]$, the i, j entry of the product is

$$\sum_{k=1}^{n} \tilde{a}_{ik} a_{kj} = \sum_{k=1}^{n} a_{kj} C_{ki}. \tag{6.7}$$

If $i = j$ this is the cofactor expansion of $\det(A)$ along the jth column, so yields $\det(A)$. Suppose that $i \neq j$. Let B denote the matrix that agrees with A except that its ith column is a copy of the jth column. Observe that the k, i cofactors of B are the same as for A because the differing column is replaced by $\hat{e}_k$ in calculating the cofactor. Therefore the cofactor expansion of $\det(B)$ along the ith column is

$$\det(B) = \sum_{k=1}^{n} a_{kj} C_{ki}.$$

But, according to the corollary to Theorem 6.7, $\det(B) = 0$. This means that for $i \neq j$ the calculation in (6.7) produces 0. Thus in all cases the i, j entries of $\mathrm{Adj}(A)A$ and $\det(A)I$ agree, proving the theorem. □

Definition. A square matrix A is called *invertible* if there exists a matrix B so that $BA = I$ (the identity matrix). The matrix B is called the *inverse* of A, and denoted $B = A^{-1}$.

The determinant of A tells us whether A is invertible or not.

Theorem 6.13. *A square matrix A is invertible if and only if* $\det(A) \neq 0$*, in which case* $A^{-1} = \frac{1}{\det(A)}\mathrm{Adj}(A)$.

Proof. Suppose A is invertible. There exists B with $BA = I$. It follows from the product rule that $\det(B)\det(A) = 1$, so that $\det(A) \neq 0$. Conversely, suppose $\det(A) \neq 0$. Let $B = \frac{1}{\det(A)}\mathrm{Adj}(A)$. Theorem 6.12 tells us that $BA = I$, so that A is indeed invertible. □

Notice that our defintion of A^{-1} only says that $A^{-1}A = I$. The defintion does *not* say that $AA^{-1} = I$. But that *is* true, as the following problem shows.

Problem 6.16 If $BA = I$, prove that $AB = I$ also. (Hint: $\det(B) \neq 0$ implies that there is C with $CB = I$. But then explain why CBA must equal both A and C.)

..... d13

Cramer's Rule concerns the solution of systems of linear equations,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots \qquad\qquad\qquad \vdots \quad &= \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned}$$

As a single matrix equation, this takes the form

$$A\hat{x} = \hat{b},$$

where $\hat{x}$ and $\hat{b}$ are the column vectors with x_i and b_i as entries.

Lemma 6.14 (Cramer's Rule)**.** *Suppose A is an invertible $n \times n$ matrix and $\hat{b} \in \mathbb{R}^n$. The equation $A\hat{x} = \hat{b}$ has a unique solution given by $\hat{x} = (x_1, \ldots, x_n)^T$, where*

$$x_i = \frac{D(\hat{a}_1, \ldots \overset{i}{\hat{b}}, \ldots \hat{a}_n)}{\det(A)}.$$

Proof. Since $\det(A) \neq 0$ we know A^{-1} exists. The equation $A\hat{x} = \hat{b}$ is equivalent to

$$\hat{x} = A^{-1}\hat{b},$$

proving the existence of one and only one solution. Using the formula for A^{-1} from Theorem 6.13 we have

$$\hat{x} = \frac{1}{\det(A)}\mathrm{Adj}(A)\hat{b}.$$

Using our notation $\mathrm{Adj}(A) = [\tilde{a}_{ij}]$ again, we have

$$\begin{aligned} x_i &= \frac{1}{\det(A)}\sum_{k=1}^{n}\tilde{a}_{ik}b_k \\ &= \frac{1}{\det(A)}\sum_{k=1}^{n}b_k C_{ki} \\ &= \frac{1}{\det(A)}D(\hat{a}_1, \ldots \overset{i}{\hat{b}}, \ldots \hat{a}_n), \end{aligned}$$

a cofactor expansion along the ith column justifying the last equality. □

Problem 6.17

a) Show that for any invertible $n \times n$ matrix A, $\det(\mathrm{Adj}(A)) = \det(A)^{n-1}$.

b) Show that the formula of part a) is correct even if A is *not* invertible. [Hint: if $\det(\mathrm{Adj}(A)) \neq 0$, then $\mathrm{Adj}(A)$ would be invertible. From that you can conclude that A is the matrix $[0]$ of all 0s. Write this out as a proof by contradiction.]

..... d14

Problem 6.18 If A is invertible, find a formula for $\mathrm{Adj}(\mathrm{Adj}(A))$ in terms of A, A^{-1}, and $\det(A)$.

..... d15

E Linear Independence and Bases

Any vector $\hat{u} \in \mathbb{R}^n$ can be written as a *linear combination* of the standard unit vectors of (6.5):

$$\hat{u} = u_1\hat{e}_1 + \cdots + u_n\hat{e}_n.$$

In this section we want to talk about doing the same thing with some other set $V = \{\hat{v}_1, \ldots, \hat{v}_m\}$ of vectors in $\mathbb{R}^n$. In other words we want $\hat{v}_1, \ldots, \hat{v}_m$ to have the property that for any vector $\hat{u} \in \mathbb{R}^n$ it is possible to obtain $\hat{u}$ as a linear combinations of our $\hat{v}_j$:

$$\hat{u} = c_1\hat{v}_1 + \cdots c_m\hat{v}_m$$

by using some choice of scalars $c_i \in \mathbb{R}$. Given the set of vectors V there are two basic issues to consider. The first is whether we really can get all $\hat{u} \in \mathbb{R}^n$ in this way. The second is whether we actually need all the $\hat{v}_j$ we started with, or if we might be able to discard some of the $\hat{v}_j$ but still be able to recover all $\hat{u}$ using a reduced set of $\hat{v}_j$. In other words we want a set V of $\hat{v}_j$ which is as small as possible, but still adequate to reconstruct all other vectors $\hat{u}$. The next example will illustrate this.

Example 6.4. Consider $\mathbb{R}^4$, using

$$\hat{v}_1 = (4,0,0,-1)^T, \quad \hat{v}_2 = (-3,0,1,0)^T, \quad \hat{v}_3 = (1,1,-1,0)^T, \quad \hat{v}_4 = (1,-1,-1,1)^T$$

Observe that it is impossible to write

$$(-1,0,1,0)^T = c_1\hat{v}_1 + c_1\hat{v}_3 + c_3\hat{v}_3 + c_3\hat{v}_4.$$

We can see that using the standard "dot" product $\langle\cdot,\cdot\rangle$:

$$\langle (1,2,3,4)^T, (-1,0,1,0)^T\rangle = 2,$$

while all the $\hat{v}_i$ have

$$\langle (1,2,3,4)^T, \hat{v}_i\rangle = 0.$$

On the other hand, if we add an additional vector,

$$\hat{v}_5 = (-1,0,1,0)^T,$$

then we *will* be able to reconstruct every vector, and we can always do it without using v_4. That follows from Cramer's Rule, since

$$D(\hat{v}_1, \hat{v}_2, \hat{v}_3, \hat{v}_5) = 2.$$

Thus $\{\hat{v}_1, \hat{v}_2, \hat{v}_3, \hat{v}_4, \hat{v}_5\}$ is bigger than needed.

Here are definitions of the concepts we are talking about.

Definition. Suppose $\hat{v}_1, \ldots, \hat{v}_m \in \mathbb{R}^n$. Let

- We say $\hat{v}_1, \ldots, \hat{v}_m$ *span* $\mathbb{R}^n$ if for every $\hat{u} \in \mathbb{R}^n$ there exist scalars $c_j \in \mathbb{R}$ so that

$$\hat{u} = c_1\hat{v}_1 + \cdots c_n\hat{v}_m.$$

- We say $\hat{v}_1, \ldots, \hat{v}_m$ are *linearly independent* if whenever $c_j \in \mathbb{R}$ are such that

$$\hat{0} = c_1\hat{v}_1 + \cdots c_n\hat{v}_m$$

  then $c_1 = \ldots = c_m = 0$.

Problem 6.19 By negating the defintion of linear independence, write what it means to say that $\hat{v}_1, \ldots, \hat{v}_m$ is *not* linearly independent. [Hint: You will need to write in the implicit "for all" quantifier before forming the negation to get it right.]

..... `lindep`

Notice that the definition does *not* assume that the number m of vectors being considered is the same as the number of components n in a vector. Secondly, although it may not be apparent, the idea of linear independence is equivalent to our idea of not being able to do without any of the $\hat{v}_i$. To see this lets return to our example.

Example 6.5. With the same $\hat{v}_1, \ldots, \hat{v}_5$ as above observe that

$$(0,0,0,0)^T = 1\hat{v}_1 + 2\hat{v}_2 + 1\hat{v}_3 + 1\hat{v}_4 + 0\hat{v}_5.$$

This means that $\{\hat{v}_1, \hat{v}_2, \hat{v}_3, \hat{v}_4, \hat{v}_5\}$ is *not* linearly independent, in accord with the definition. In particular we see that $\hat{v}_4$ can be written in terms of the other $\hat{v}_i$:

$$\hat{v}_4 = -1\hat{v}_1 - 2\hat{v}_2 - 1\hat{v}_3 + 0\hat{v}_5.$$

We can replace $\hat{v}_4$ by this expression, converting any linear combination of all five $\hat{v}_i$ to a linear combination of just $\hat{v}_1, \hat{v}_2, \hat{v}_3, \hat{v}_5$:

$$\begin{aligned} c_1\hat{v}_1 + c_2\hat{v}_2 + c_3\hat{v}_3 + c_4\hat{v}_4 + c_5\hat{v}_5 &= c_1\hat{v}_1 + c_2\hat{v}_2 + c_3\hat{v}_3 + c_4(-\hat{v}_1 - 2\hat{v}_2 - \hat{v}_3 + 0\hat{v}_5) + c_5\hat{v}_5 \\ &= (c_1 - c_4)\hat{v}_1 + (c_2 - 2c_4)\hat{v}_2 + (c_3 - c_4)\hat{v}_3 + c_5\hat{v}_5. \end{aligned}$$

Problem 6.20 Show that if $\{\hat{u}, \hat{v}, \hat{w}\}$ is linearly independent, then $\{\hat{u}, \hat{u}+\hat{v}, \hat{u}+\hat{v}+\hat{w}\}$ is linearly independent. (From [14].)

..... H1

Problem 6.21 Show that $\{\hat{u}-\hat{v}, \hat{v}-\hat{w}, \hat{w}-\hat{u}\}$ is never linearly independent. (From [14].)

..... H2

The main theorem of this section says that we can test for *both* of these properties using determinants, but only when the number m of vectors is the same as the number n of coordinates in each vector in $\mathbb{R}^n$. Thus when $m = n$ to span is equivalent to being linearly independent. (But when $m \neq n$ they are *not* equivalent.)

Theorem 6.15. *Suppose $\hat{v}_1, \ldots, \hat{v}_n$ are n vectors in $\mathbb{R}^n$. The following are equivalent.*

a) $\hat{v}_1, \ldots, \hat{v}_n$ are linearly independent,

b) $D(\hat{v}_1, \cdots, \hat{v}_n) \neq 0$

c) $\hat{v}_1, \ldots, \hat{v}_n$ span $\mathbb{R}^n$

Proof. We first prove that a) and b) are equivalent. Suppose the $\hat{v}_i$ are not linearly independent. That means there exist scalars c_i, not all 0, for which

$$\hat{0} = c_1\hat{v}_1 + \cdots + c_n\hat{v}_n.$$

If we let V be the matrix with $\hat{v}_i$ as its columns, and $\hat{c} = (c_1, \ldots, c_n)^T$, we can rephrase this as saying $\hat{c}$ solves

$$V\hat{c} = \hat{0}.$$

But a second (different) solution is the vector $\hat{0}$ of all 0s. According to Cramer's Rule, this can only be if $\det(V) = D(\hat{v}_1, \cdots, \hat{v}_n) = 0$.

Conversely assume the determinant in b) is zero: $\det(V) = 0$ where V is the matrix with $\hat{v}_i$ as its columns. From this we want to produce some c_i not all 0 for which $\sum_1^n c_i\hat{v}_i = \hat{0}$. To do this start by taking $W = V^T$ and $\hat{w}_i$ the columns of W (i.e. the rows of V). By Lemma 6.6

$$D(\hat{w}_1, \ldots, \hat{w}_n) = \det(W) = \det(V) = 0.$$

Pick a maximal subset of the $\hat{w}_i$ such that they can be complemented with some additional vectors $\hat{u}_j$ to produce a nonzero determinant: after renumbering and rearranging,

$$D(\hat{w}_1, \ldots, \hat{w}_k, \hat{u}_{k+1}, \ldots, \hat{u}_n) \neq 0.$$

By hypothesis, $k < n$. Let $c_i = C_{in}$ be the cofactors of this determinant along the last column. The above determinant tells us that $0 \neq \sum_i c_i u_{in}$. Therefore cofactors c_i cannot be all 0. These will be the c_i that we seek, but we still need to explain why they do what we want.

If we replace $\hat{u}_n$ (last column) in the above determinant by *any* of the $\hat{w}_i$ the determinant is 0; for $i = 1, \ldots, k$ this is because of a repeated column; for $i = k+1, \ldots, n$ this is because of the maximality of $\{\hat{w}_1, \ldots \hat{w}_k\}$. Thus, for each j we have

$$\sum_1^n c_i w_{ij} = 0.$$

This is the same as saying

$$\sum_1^n c_i v_{ji} = 0 \text{ for all } j,$$

which in turn is equivalent to

$$\sum_1^n c_i \hat{v}_i = \hat{0}.$$

Since the c_i are not all 0, this means that the $\hat{v}_i$ are not independent. This completes the proof of the equivalence of a) and b).

To prove the equivalence of b) and c), observe that Cramer's Rule says that b) implies that for any $\hat{u}$ there exist c_i for which

$$\sum_1^n c_i \hat{v}_i = \hat{u}.$$

In other words, b) implies c).

Finally assume c), namely that the $\hat{v}_i$ span $\mathbb{R}^n$. In particular, for each j there are some coefficients c_{ij} so that

$$\hat{e}_j = c_{1j}\hat{v}_1 + c_{2j}\hat{v}_2 + \cdots c_{nj}\hat{v}_n. \tag{6.8}$$

The left side is the jth column of the identity matrix I, and the right side is the jth column of the product VC, where $C = [c_{ij}]$ is the matrix assembled from all the c_{ij} values. In other words (6.8) can be restated as

$$I = VC.$$

According to Theorem 6.14, this implies that $\det(V) \neq 0$. Since $\det(V) = D(\hat{v}_1, \cdots, \hat{v}_n)$ this proves that c) implies b). □

Problem 6.22 Use the above theorem to prove the following.

a) If $m > n$ then $v_1, \ldots v_m \in \mathbb{R}^n$ are *not* linearly independent.

b) if $m < n$ then $v_1, \ldots v_m \in \mathbb{R}^n$ do *not* span $\mathbb{R}^n$

[Hint: add some extra 0s so that the theorem can be applied, but be careful to explain why adding the 0s does not change the spanning or linear independence that you are trying to prove.]

..... LI

F The Cayley-Hamilton Theorem

In this last section we will use Theorem 6.12 one more time to prove another famous theorem, the Cayley-Hamilton Theorem, which involves determinants, matrices, and polynomials.

The most familiar way to associate a polynomial with a matrix A is to form its characteristic polynomial. (You have probably encountered the characteristic polynomial of a matrix before; it is used to compute the eigenvalues of A for instance.)

Definition. Suppose A is an $n \times n$ matrix. Its *characteristic polynomial* is

$$p(x) = \det(xI - A).$$

In other words, we form the matrix $xI - A$ where x is a variable, and then compute its determinant, which will be an expression involving that variable.

Example 6.6. For

$$A = \begin{bmatrix} 3 & 5 \\ 2 & 4 \end{bmatrix},$$

the characteristic polynomial is

$$p(x) = \det\left(\begin{bmatrix} x-3 & -5 \\ -2 & x-4 \end{bmatrix}\right) = (x-3)(x-4) - (-2)(-5) = x^2 - 7x + 2.$$

Problem 6.23 Explain why $p(x)$ is *always* a polynomial of degree n (same size as A) with leading coefficient of 1:

$$p(x) = x^n + \text{lower order powers.}$$

..... CPdeg

The Cayley-Hamilton Theorem involves substituting $x = A$ in the characteristic polynomial. It is pretty clear what we should mean by powers of a matrix: $A^2 = A \cdot A$, $A^3 = A \cdot A \cdot A$, and so forth. What we want to do is substitute the powers of A for the corresponding powers of x in the characteristic polynomial. We interpret the constant term as a constant times $x^0 = 1$, and replace the x^0 by the matrix $A^0 = I$ (the identity matrix).

Example 6.7. Continuing with Example 6.6,

$$\begin{aligned} p(x) &= x^2 - 7x^1 + 2x^0 \\ p(A) &= A^2 - 7A + 2I \\ &= \begin{bmatrix} 19 & 35 \\ 14 & 26 \end{bmatrix} - 7 \begin{bmatrix} 3 & 5 \\ 2 & 4 \end{bmatrix} + 2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

Our theorem says that what happened in this example is not a coincidence!

Theorem 6.16 (Cayley-Hamilton)**.** *If A is an $n \times n$ matrix and its characteristic polynomial is*

$$p(x) = \det(xI - A) = \sum_{i=0}^{n} b_i x^i,$$

then

$$p(A) = \sum_{i=0}^{n} b_i A^i = [0], \text{ the zero matrix.}$$

Problem 6.24 Let A be the matrix

$$A = \begin{bmatrix} 1 & 3 & 5 \\ 1 & 1 & 2 \\ 1 & 0 & -1 \end{bmatrix}.$$

a) Calculate $\mathrm{Adj}(A)$ and check that the identity of Theorem 6.13 does hold by calculating both sides.

b) Calculate the characteristic polynomial $p(x)$ for A, and verify the conclusion of the Cayley-Hamilton Theorem by calculating $p(A)$.

..... adjex

The Cayley-Hamilton Theorem is often proved using what is called the Jordan canonical form of a matrix, which you probably have not seen before (and will not see here either). However there is a really nice proof based on Theorem 6.12. Here is the idea. Let $p(x) = \det(xI - A)$ be the characteristic polynomial of A. We know from Theorem 6.12 that

$$\mathrm{Adj}(xI - A) \cdot (xI - A) = p(x)I, \tag{6.8}$$

holding for each value of $x \in \mathbb{R}$. If we just formally[4] plug A in for x on both sides, the right side is $p(A)I = p(A)$ and the left side is $[0]$, because of the term $(xI - A)$. So this formal manipulation seems to

[4]By "formally" we mean just manipulate the symbols without thinking about whether what we are doing really makes any sense.

tell us right away that $[0] = p(A)$, which is exactly what we want to prove! But we need to be careful. We know what we mean by by substituting A for x in $p(x)$. But the $\mathrm{Adj}(xI - A)$ on the left is something more complicated — it is a matrix with the variable x and its powers appearing in the various entries. Does it make any sense to plug $x = A$ into that? Our job is to see if we can find a careful way to get from (6.8) to our desired conclusion that $p(A) = [0]$.

There are two ways we can think of $\mathrm{Adj}(xI - A)$. The most natural is to think of it as a matrix with polynomials as its entries. But there is a second point of view: if we collect up the common powers of x we can think of it as a polynomial with matrix coefficients.

Example 6.8. Suppose

$$A = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & -1 \\ 1 & 0 & 1 \end{bmatrix}.$$

With a little work we find that

$$\mathrm{Adj}(xI - A) = \begin{bmatrix} x^2 - 5x + 4 & 3x - 3 & 5x - 23 \\ 2x - 3 & x^2 - 2x - 4 & -x + 11 \\ x - 4 & 3 & x^2 - 5x - 2 \end{bmatrix}.$$

This is what we mean by a matrix with polynomial entries. But by collecting powers of x we can write it as

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} x^2 + \begin{bmatrix} -5 & 3 & 5 \\ 2 & -2 & -1 \\ 1 & 0 & -5 \end{bmatrix} x + \begin{bmatrix} 4 & -3 & -23 \\ -3 & -4 & 11 \\ -4 & 3 & -2 \end{bmatrix}.$$

This is what we mean by a polynomial with matrix coefficients. (We have written the matrix in front of the powers of x, as is customary for coefficients.)

No we see that both sides of (6.8) can be expressed as polynomials with matrix coefficients. If we read (6.8) that way then at least we know how we would go about plugging A in for x. But if (6.8) is true for each real number x why should it be true when x is replaced by something other than a real number, like A? That is the issue the proof ultimately must answer.

Our proof of the Cayley-Hamilton Theorem is based on a generalization of Lemma 5.1: if two such polynomials produce the same (matrix) value for each value of the variable x (scalar), then in fact the coefficient matricies must agree. In case we are uncertain about that, we present it as a lemma.

Lemma 6.17. *Suppose B_k, C_k are $n \times n$ are matrices for each $k = 0, 1, 2, \ldots, m$, and that for every $x \in \mathbb{R}$*

$$\sum_0^m C_k x^k = \sum_0^m B_k x^k \text{ (as matricies).}$$

Then $C_k = B_k$ for each $k = 0, \ldots, m$.

Proof. Express both sides as matrices of polynomials:

$$\sum_0^n C_k x^k = [c_{ij}(x)], \quad \sum_0^n B_k x^k = [b_{ij}(x)],$$

where each $c_{ij}(x)$ and $b_{ij}(x)$ is a polynomial in x (of degree at most n). The hypothesis is that for each value of x, the matrices $[c_{ij}(x)]$ and $[b_{ij}(x)]$ are equal, and therefore

$$c_{ij}(x) = b_{ij}(x) \text{ for all pairs } i, j.$$

This means the (scalar) coefficients of $c_{ij}(x)$ and $b_{ij}(x)$ agree. The coefficients of x^k in these two polynomials are the i, j entries of C_k and B_k respectively. It follows then that for each k, all entries of C_k and B_k agree. Thus $C_k = B_k$. □

We are ready now for our proof, taken from [19].

Proof of Cayley-Hamilton Theorem. Let the characteristic polynomial be

$$p(x) = \det(xI - A) = \sum_{k=0}^{n} b_k x^k.$$

So the coefficients of $p(x)I$ from the right side of (6.8), viewed as a polynomial with matrix coefficients, are $b_k I$. Let C_i be the matrix coefficients of $\mathrm{Adj}(xI - A)$:

$$\mathrm{Adj}(xI - A) = \sum_{i=0}^{n-1} C_i x^i$$

(Do you see why it has degree at most $n - 1$?) We can now work out the coefficients of the full left side of (6.8).

$$\begin{aligned}
\mathrm{Adj}(xI - A) \cdot (xI - A) &= (C_0 x^0 + \ldots + C_{n-2} x^{n-2} + C_{n-1} x^{n-1})(xI - A) \\
&= \qquad C_0 x^1 + \ldots + C_{n-2} x^{n-1} + C_{n-1} x^n \\
&\quad - (C_0 + C_1 x + \ldots + C_{n-1} x^{n-1}) A \\
&= -C_0 A + (C_0 - C_1 A) x + \ldots + (C_{n-2} - C_{n-1} A) x^{n-1} + C_{n-1} x^n
\end{aligned}$$

Now that we have worked out the matrix coefficients of both sides we can apply Lemma 6.17 to (6.8), to deduce that

$$\begin{aligned}
b_0 I &= -C_0 A \\
b_1 I &= C_0 - C_1 A \\
&\vdots \\
b_{n-1} I &= C_{n-2} - C_{n-1} A \\
b_n I &= C_{n-1}.
\end{aligned}$$

We now evaluate $p(A)$, writing $b_i A^i = (b_i I) A^i$ and using the preceding formulas.

$$\begin{aligned}
p(A) &= b_0 I + b_1 A^1 + \ldots + b_{n-1} A^{n-1} + b_n A^n \\
&= b_0 I + b_1 I A^1 + \ldots + b_{n-1} I A^{n-1} + b_n I A^n \\
&= -C_0 A + (C_0 - C_1 A) A + \ldots + (C_{n-2} - C_{n-1} A) A^{n-1} + C_{n-1} A^n \\
&= [0],
\end{aligned}$$

because all terms on the right cancel. □

Problem 6.25 Suppose that the characteristic polynomial factors as $p(x) = \prod_{i=1}^{n} (x - \lambda_i)$. (By Theorem 5.10 it *always* does if we allow complex λ_i. The λ_i are called the *eigenvalues* of the matrix A.) Prove that $\det(A) = \prod_{i=1}^{n} \lambda_i$.

. . . . . ev

Problem 6.26 We can define what we mean for a collection of $n \times n$ matricies A, B, ... to be linearly independent by analogy with the defintion for vectors on page 128: if

$$c_A A + c_B B + \cdots = [0] \text{ (the matrix of all 0s)}$$

is only true for $c_A = c_B = \cdots = 0$. Use the Cayley-Hamilton Theorem to explain why I, A, A^2, ..., A^n will never be linearly independent.

. . . . . pow

Additional Problems

Problem 6.27 Our definition of $\operatorname{sgn}(\sigma)$ used the natural order relation "$<$" on the integers. Suppose we used a different ordering of the integers. Would that lead to a different notion of $\operatorname{sgn}(\sigma)$? If $\triangleleft$ is a different order relation there is a permutation γ connecting it to the natural order in the sense that

$$i \triangleleft j \text{ if and only if } \gamma(i) < \gamma(j).$$

(Take that for granted.) Suppose we defined the inversion count of σ using $\triangleleft$ instead of $<$:

$$r_{\triangleleft}(\sigma) = \left|\{\{i,j\} : \ 1 \le i \triangleleft j \le n \text{ and } \sigma(j) \triangleleft \sigma(i)\}\right|.$$

Prove that we would still get the same result for $\operatorname{sgn}(\sigma) = (-1)^{r_{\triangleleft}(\sigma)}$ as before. Thus the definition of $\operatorname{sgn}(\sigma)$ is actually independent of the choice of order relation. (Hint: Relate this to $\gamma\sigma\gamma^{-1}$ and use Theorem 6.2.)

..... `ordind`

Problem 6.28 For a permutation $\pi \in S_n$ define the *permutation matrix* to be $P_\pi = [p_{i,j}]$ where

$$p_{ij} = \begin{cases} 1 & \text{if } i = \pi(j) \\ 0 & \text{otherwise.} \end{cases}$$

Show that permutation matricies have the following properties.

a) $P_\pi P_\sigma = P_{\pi\sigma}$, for any two permutations π and σ.

b) $P_\pi^{-1} = P_{\pi^{-1}} = P_\pi^T$.

c) $\det(P_\pi) = \operatorname{sgn}(\pi)$.

d) If $A = [\hat{a}_1, \ldots, \hat{a}_j, \ldots, \hat{a}_n]$, then $AP = [\hat{a}_{\pi(1)}, \ldots, \hat{a}_{\pi(j)}, \ldots, \hat{a}_{\pi(n)}]$. In other words, right multiplication by P_π has the effect of reordering the columns of A according to the permutation π.

..... `permmat`

Problem 6.29 Suppose $A^\sharp$ denotes the matrix resulting from rotating A one quarter turn in the counter-clockwise direction. For instance, if

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \quad \text{then } A^\sharp = \begin{bmatrix} 3 & 6 & 9 \\ 2 & 5 & 8 \\ 1 & 4 & 7 \end{bmatrix}.$$

What is the relationship between $\det(A)$ and $\det(A^\sharp)$ (in the general $n \times n$ case)? [Hint: Problem 6.6 might be useful.]

..... `d11`

Problem 6.30 Let $x_1, x_2, \ldots, x_n$ be real numbers. The following determinant is called the *Vandermonde* determinant:

$$V_n = \det \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \ldots & x_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^{n-1} \end{pmatrix}.$$

Show that its value is given by the following formula:

$$V_n = \prod_{1 \le i < j \le n} (x_j - x_i).$$

First check the formula directly for $n = 2$ and $n = 3$, then prove the general case by induction. Hint: For the induction, multiply each column by x_1 and subtract it from the column to its right, starting from the right side and working your way to the left. Taking the determinant that way should give you something like

$$V_n = (x_n - x_1) \dots (x_2 - x_1) V_{n-1}.$$

..... vdm

Problem 6.31 Explain what Problem 6.30 has to do with the fact that if we are given $n+1$ points (x_i, y_i) in the plane with distinct x_i then there is only one polynomial $p(x)$ of degree n with $y_i = p(x_i)$.

..... vdm2

Problem 6.32 Find and prove a formula (of the same type as in Problem 6.30 above) for this determinant:

$$\det \begin{pmatrix} x_1 & x_1 & x_1 & \dots & x_1 \\ x_1 & x_2 & x_2 & \dots & x_2 \\ x_1 & x_2 & x_3 & \dots & x_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1 & x_2 & x_3 & \dots & x_n \end{pmatrix}.$$

Under what circumstances is the determinant nonzero?

..... d12

Problem 6.33 Prove Lemma 6.11. Like many proofs in this chapter, this mostly requires finding a clear way to write down the steps involved. To get you started, $C_{k\ell} = \det(B)$, where $B = [b_{ij}]$ with

$$b_{ij} = \begin{cases} 0 & \text{if } i \ne k,\ j = \ell \\ 1 & \text{if } i = k,\ j = \ell \\ a_{ij} & \text{if } j \ne \ell \end{cases}$$

and $M_{k\ell} = \det F$ where $F = [f_{ij}]$ is the $(n-1) \times (n-1)$ matrix with

$$f_{ij} = \begin{cases} a_{i\,j} & \text{if } i < k \text{ and } j < \ell \\ a_{i\,j+1} & \text{if } i < k \text{ and } \ell \le j \le n-1 \\ a_{i+1\,j} & \text{if } k \le i \le n-1 \text{ and } j < \ell \\ a_{i+1\,j+1} & \text{if } k \le i \le n-1 \text{ and } \le\le j \le n-1. \end{cases}$$

Finally, let $H = [h_{ij}]$ where $h_{ij} = b_{\alpha(i)\,\beta(j)}$, where α and β are the permutations

$$\alpha = \langle k, k+1, \dots, n \rangle,\ \beta = \langle \ell, \ell+1, \dots, n \rangle.$$

F and A are related to each other by means of the same permutations α and β. Put these pieces together, showing that

$$C_{k\ell} = \det(B) = \operatorname{sgn}(\alpha)\operatorname{sgn}(\beta)\det(H) = \operatorname{sgn}(\alpha)\operatorname{sgn}(\beta)\det(F) = \operatorname{sgn}(\alpha)\operatorname{sgn}(\beta) M_{k\ell}.$$

. . . . . `minors`

Problem 6.34 The Cayley-Hamilton Theorem tells us that the characteristic polynomial of an $n \times n$ matrix A, $p(x) = \det(xI - A)$, has the property that $p(A) = 0$. But there are other polynomials with this property. Prove the following.

1. If $q(x)$ is another polynomial and $p(x)$ divides $q(x)$, then $q(A) = 0$.
2. Any two polynomials of smallest possible degree with $p(A) = [0]$ are constant multiples of each other. So there exists a unique polynomial $m(x)$ of smallest possible degree with the property that $m(A) = 0$ and with leading coefficient 1. This $m(x)$ is called the *minimal polynomial of* A.
3. Find both the characteristic and minimal polynomials for each of the following matricies.
   (a) $A = I$, the $n \times n$ identity matrix.
   (b) $A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.
   (c) $A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$.
4. Let $m(x)$ be the minimal polynomial of A and $q(x)$ any other polynomial. Prove that $q(A) = 0$ iff $m(x)$ divides $q(x)$. [Hint: Use the division theorem.]
5. The minimal polynomial must divide the characteristic polynomial.

. . . . . `minpoly`

Appendix: Mathematical Words

Here is a list of some words and phrases that are common in mathematical writing but which have a somewhat different meaning than in everyday usage. They are ordered so that those with similar or related meanings are close to each other.

theorem — A theorem is a statement of fact that the author wants to highlight as one of the most important conclusions of his whole discussion, something he intends his readers will remember and refer back to even after they have forgotten the details of the proof. Its the way we record major landmarks of mathematical learning. Examples are the Pythagorean Theorem, the Fundamental Theorem of Algebra. Sometimes they end up with people's names attached to them, like the Cantor's Theorem, Rolle's Theorem or Fermat's Last Theorem.

lemma — A lemma is like a little theorem, which is important for the paper in which it appears, but probably won't have much importance beyond that. Often it is just a tool developed in preparation for proving the big theorem the author has in mind.

proposition — We have used "proposition" to stand for any statement with an unambiguous mathematical meaning. A stated Proposition (with a capital P, like Proposition 1.3) is about the same as a lemma, perhaps even of slightly less lasting importance.

corollary — A corollary is an easily obtained consequence of a theorem or lemma that precedes it. See for example the corollary to Lemma 1.8.

claim — In a proof we sometimes announce an assertion that we are going to prove before we prove it. This can help the reader know where we are headed. We might write "We claim that P . To see this ..." It means the same as if we said "We are about to prove P. The proof of P is as follows." See for example the proof of Theorem 3.7.

conjecture — A conjecture is a statement of something the author thinks is probably true, but doesn't know how to prove. A published conjecture is often taken as a challenge to the mathematical community to come up with a proof (or disprove it). An example is the Twin Primes Conjecture, described in Section E.

thus, therefore, hence — These words mean that what came before provides the reason for what comes next. For instance, "$\ldots x = (\ldots)^2$, therefore $x \geq 0$."

it follows that — Sort of like "thus" and "therefore" except that the reasons for what comes next may not be what immediately precedes "it follows that." The way most people use it, it means "Somewhere in what we know and have assumed are reasons for what comes next, but we are not going to spell them all out. You can find them for yourself."

we deduce — This is about the same as "it follows that."

we have — This means "we know" or "we have established the truth of." It sometimes serves as a reminder of things from earlier in an argument.

obviously — This means the author thinks that what is said next is so clear that no explanation is needed. Sometimes it means that the next assertion is tedious to prove and the author just doesn't want to be bothered with it. "Clearly," and "it is easy to see" are used the same way. The phrase "as the reader can check" is pretty close as well, but with that phrase the author is acknowledging that there is something to be worked out, but which is deliberately left out. For veteran mathematicians writing for each other, these are often used to shorten a proof by leaving out parts with the confidence that the reader can fill them in on his/her own. As students be very careful resorting to these. Are you positive you could prove what you are leaving out? If not, put it in!

since — This typically announces the reason for what comes next: "Since $x = (\ldots)^2$ we know $x \geq 0$." Sometimes we reverse the order: "We know $x \geq 0$, since $x = (\ldots)^2$."

let — This is used several ways. Sometimes it announces notation, as in "let A be a 3×3 matrix." Sometimes is announces a hypothesis for the next phase of the argument. For instance "let n be even $\ldots$;" and then later "let n be odd." It might specify the value of a variable to be used in the next part of a calculation, "let $x = 3$, then $\ldots$"

given — This is sometimes used to refer to something we know from hypotheses or an assumption. We used it this way in the paragraph just after the proof of Theorem 1.9, "given our understanding of infinite $\ldots$"

suppose — To suppose something is to assume it. We often say suppose as a way of announcing that we are invoking the hypotheses. See the first sentence of the proof of Theorem 1.2 for instance.

provided — This is a way of identifying some additional assumption that is needed for what was just said, or is about to be. For instance, "x^2 is always nonnegative, provided x is a real number."

prove, show — These mean roughly the same thing. Many would say that to show something, means to describe the main idea of a proof without being quite as complete as a full-dressed proof, a sort of informal or lazy-man's proof.

w.l.o.g — An abbreviation for "without loss of generality." Sometimes something we wish to prove can be reduced to a special case. We might say "without loss of generality" to announce the special case, meaning the reader to understand that the general case can be reduced to the special case actually being considered. Our proof of the Fundamental Theorem of Algebra uses this technique.

such that — This means "with the property that;" it announces a property that is required of the object being described. For example, "$\sqrt{2}$ refers to the positive real number y such that $y^2 = 2$."

a, the — When referring to some mathematical object we use "the" when there is only one such object, and "a" when there could be several to choose from. If you wrote "let r be *the* root of $x^2 - 4 = 0$," the reader would take that as a claim that there is only one such root (and complain because in fact there are more than one). If you wrote "let r be *a* root of $x^2 - 4 = 0$," then I would understand that I could take my choice between the two possibilities $r = 2, -2$ and proceed with either one. (What you say after better work for both of them!)

moreover — This is used to continue with some additional conclusions in a new sentence. See Step 4 of the proof of the Fundamental Theorem of Algebra in Chapter 5.

indeed — This is often used to give reasons for an assertion that was just made with no justification. See the beginning of the proof of the Fundamental Theorem of Algebra.

Appendix: The Greek Alphabet and Other Notation

We use symbols and special notation *a lot* in mathematics. There are many special symbols, like $\int$, ∂ $\leq$, $\div$, ∞, ... that have no meaning outside mathematics. But we also use conventional letters. Since the standard Latin alphabet $(a, b, \ldots, z)$ is not enough, we also use capitals, and sometimes other typefaces $(A, \mathcal{A}, \mathbb{A})$. We use a lot of Greek characters, and a few from other alphabets (The symbol $\aleph$ (aleph) is Hebrew for instance). At right is a table of those Greek letters that are used in mathematics. (Those in the table are those supported by the standard mathematical typesetting language LaTeX, with which this book is written.) Some Greek letters are not used because they look too much like Latin characters. For instance, an uppercase alpha is indistinguishable from A, the lower case upsilon is very close to the italic Latin v, and both cases of omicron look just like our o (or O). To further extend the list of available symbols we add accents, like $a', \tilde{\alpha}, \bar{x}, \hat{f}$.

Some symbols are customary for certain purposes. For instance the letters i, j, k, l, m, n are often used when only integer values are intended. Similarly, z is sometimes used for complex numbers, as opposed to real numbers. Such conventions are another way we limit the possible scope of what we say, so that the meaning is unambiguous. However these conventions are never universal.

spelling	**l. case**	**u. case**
alpha	α	
beta	β	
gamma	γ	Γ
delta	δ	Δ
epsilon	ϵ	
zeta	ζ	
eta	η	
theta	θ	Θ
iota	ι	
kappa	κ	
lambda	λ	Λ
mu	μ	
nu	ν	
xi	ξ	Ξ
omicron		
pi	π	Π
rho	ρ	
sigma	σ	Σ
tau	τ	
upsilon		Υ
phi	ϕ	Φ
psi	ψ	Ψ
chi	χ	
omega	ω	Ω

Bibliography

[1] D. Acheson, 1089 and All That: A Journey into Mathematics, Oxford Univ. Press, NY, 2002.

[2] J. L. Brown Jr., *Zeckendorf's Theorem and Some Applications*, Fibonacci Quarterly vol. 2 (1964), pp. 163–168.

[3] Margherita, Barile, *Curry Triangle*, from MathWorld–A Wolfram Web Resource, created by Eric W. Weisstein, `http://mathworld.wolfram.com/CurryTriangle.html`.

[4] Klaus Barner, *Paul Wolfskehl and the Wolfskehl prize*, Notices of the AMS vol. 44 no. 10 (1997), pp. 1294–1303.

[5] Edward J. Barbeau, Mathematical Fallacies, Flaws and Flimflam, The Mathematical Association of America, 2000.

[6] A. T. Benjamin and J. J. Quinn, Proofs That Really Count: The Art of Combinatorial Proof, The Mathematical Association of America, 2003.

[7] R. H. Cox, *A proof of the Schroeder-Bernstein Theorem*, The American Mathematical Monthly v. 75 (1968), p. 508.

[8] Keith Devlin, Mathematics, The New Golden Age, Columbia Univ. Press, NY, 1999.

[9] P. J. Eccles, An Introduction to Mathematical Reasoning: Numbers, Sets and Functions, Cambridge Univ. Press, Cambridge, UK, 1997.

[10] P. Fletcher and C. W. Patty, Foundations of Higher Mathematics, Brooks/Cole Publishing, Pacific Grove, CA, 1996.

[11] P. R. Halmos, Naive Set Theory, Springer-Verlag, New York, 1974.

[12] Leonard Gillman, Writing Mathematics Well: a Manual for Authors, Mathematical Association of America, 1987.

[13] G. H. Hardy, A mathematician's Apology, Cambridge University Press, London, 1969.

[14] Jim Hefferon, Linear Algebra, `http://joshua.smcvt.edu/linearalgebra/linalg.html`.

[15] Vilmos Komornik, *Another Short Proof of Descartess Rule of Signs*, The American Mathematical Monthly v. 113 (2006), pp. 829–830.

[16] E. Landau, Differential and Integral Calculus, Chelsa, NY, 1951.

[17] A. Levine, Discovering Higher Mathematics, Four Habits of Highly Effective Mathematicians, Academic Press, San Diego, 2000.

[18] P. D. Magnus, Forall x: An Introduction to Formal Logic, version 1.22, 2005, `http://www.fecundity.com/logic/`

[19] M. Marcus and H. Minc, A Survey of Matrix Theory and Matrix Inequalities, Dover Publications, NY, 1992.

[20] T. Nagell, Introduction to Number Theory, Wiley, New Your, 1951.

[21] E. Nagel and J. R. Newman, Gödel's Proof, New York University Press, 1958.

[22] Roger B. Nelson, Proofs Without Words, MAA, Washington, DC, 1993.
House,

[23] J. Nunemacher and R. M. Young, *On the sum of consecutive Kth Powers*, Mathematics Magazine v. 60 (1987), pp. 237–238.

[24] Michael Spivak, Calculus (third edition), Publish or Perish, Inc., Houston, TX, 1994.

[25] J. Stewart, Calculus: Early Transcendentals (fourth edition), Brooks/Cole, Pacific Grove, CA, 1999.

[26] D. Veljan, *The 2500-year-old Pythagorean Theorem*, Mathematics Magazine v. 73 (no. 4, Oct. 2000), pp. 259—272.

[27] Robin J. Wilson, Four Colors Suffice: How the Map Problem was Solved, Princeton University Press, Princeton, NJ, 2002.

Index

Please let me know of any additional items you think ought to be included in the index.