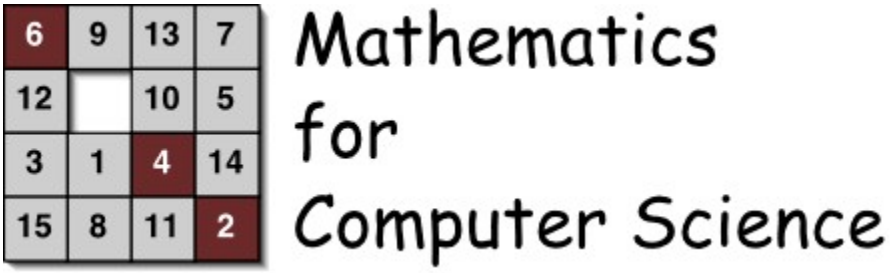
6.042, Spring 2007



Lecture notes, class problems, problem sets, miniquizzes

Contents

Chapter 1

Proofs

1.1 What is a Proof?

A proof is a method of establishing truth. What constitutes a proof differs among fields.

- *Legal* truth is ascertained by a jury based on allowable evidence presented at trial.
- *Authoritative* truth is ascertained by a trusted person or organization.
- *Scientific* truth[1] is ascertained by experiment.
- *Probable* truth is obtained from statistical analysis of sample data. For example, public opinion is ascertained by polling a small random sample of people.
- *Philosophical* proof involves careful exposition and persuasion based on consistency and plausibility. The best example is "Cogito ergo sum," a Latin sentence that translates as "I think, therefore I am." It comes from the beginning of a 17th century essay by the Mathematician/Philospher, René Descartes, and it is one of the most famous quotes in the world: do a web search on the phrase and you will be flooded with hits.

  Deducing your existence from the fact that you're thinking about your existence is a pretty cool and persuasive-sounding first axiom. However, with just a few more lines of proof in this vein, Descartes goes on to conclude that there is an infinitely beneficent God. This ain't Math.

Mathematics also has a specific notion of "proof."

Definition. A *formal proof* of a *proposition* is a chain of *logical deductions* leading to the proposition from a base set of *axioms*.

The three key ideas in this definition are highlighted: proposition, logical deduction, and axiom. In the next sections, we'll discuss these three ideas along with some basic ways of organizing proofs.

[1] Actually, only scientific *falsehood* can be demonstrated by an experiment, when the experiment fails to behave as predicted. But no amount of experiment can confirm that the *next* experiment won't fail. For this reason, scientists rarely speak of truth, but rather of *theories* that accurately predict past, and anticipated future, experiments.

1.2 Propositions

Definition. A *proposition* is a statement that is either true or false.

This definition sounds very general, but it does exclude sentences such as, "Wherefore art thou Romeo?" and "Give me an A!".

But not all propositions are mathematical. For example, "Albert's wife's name is 'Irene' " happens to be true, and could be proved with legal documents and testimony of their children, but it's not a mathematical statement.

Mathematically meaningful propositions must be about well-defined mathematical objects like numbers, sets, functions, relations, *etc.*, and they must be stated using mathematically meaningful terminology, like 'AND' and 'FORALL'. It's best to illustrate this with a few examples about numbers and planar maps that are all mathematically meaningful.

Proposition 1.2.1. *2 + 3 = 5.*

This proposition is true.

Proposition 1.2.2. *Let* $p(n) ::= n^2 + n + 41$. [2]

$$\forall n \in \mathbb{N}.\ p(n) \text{ is a prime number}.$$

The symbol $\forall$ is read "for all". The symbol $\mathbb{N}$ stands for the set of *nonnegative integers*, which are 0, 1, 2, 3, ...(ask your TA for the complete list). The symbol "$\in$" is read as "is a member of" or simply as "is in". The period after the $\mathbb{N}$ is just a separator between phrases.

A *prime* is a nonnegative integer greater than one that is not divisible by any other nonnegative integer other than 1 and itself, for example, 2, 3, 5, 7, 11, ...

Let's try some numerical experimentation to check this proposition: $p(0) = 41$ which is prime. $p(1) = 43$ which is prime. $p(2) = 47$ which is prime. $p(3) = 53$ which is prime. $\ldots p(20) = 461$ which is prime. Hmmm, starts to look like a plausible claim. In fact we can keep checking through $n = 39$ and confirm that $p(39) = 1601$ is prime.

But if $n = 40$, then $p(n) = 40^2 + 40 + 41 = 41 \cdot 41$, which is not prime. So it's not true that the expression is prime *for all* n, the proposition is false! In fact, it's not hard to show that *no* nonconstant polynomial can map all nonnegative integers into prime numbers. The point is that in general you can't check a claim about an infinite set by checking a finite set of its elements, no matter how large the finite set. Here are two even more extreme examples:

Proposition 1.2.3. $a^4 + b^4 + c^4 = d^4$ *has no solution when* a, b, c, d *are positive integers. In logical notation, letting* $\mathbb{Z}^+$ *denote the positive integers, we have*

$$\forall a \in \mathbb{Z}^+ \forall b \in \mathbb{Z}^+ \forall c \in \mathbb{Z}^+ \forall d \in \mathbb{Z}^+.\ a^4 + b^4 + c^4 \neq d^4.$$

[2] The symbol ::= means "equal by definition." It's always ok to simply write "=" instead of ::=, but reminding the reader that an equality holds by definition can be helpful.

Strings of $\forall$'s like this are usually abbreviated for easier reading:

$$\forall a, b, c, d \in \mathbb{Z}^+.\ a^4 + b^4 + c^4 \neq d^4.$$

Euler (pronounced "oiler") conjectured this 1769. But the proposition was proven false 218 years later by Noam Elkies at a liberal arts school up Mass Ave. He found the solution $a = 95800, b = 217519, c = 414560, d = 422481$.

Proposition 1.2.4. *$313(x^3 + y^3) = z^3$ has no solution when $x, y, z \in \mathbb{N}$.*

This proposition is also false, but the smallest counterexample has more than 1000 digits!

Proposition 1.2.5. *Every map can be colored with 4 colors so that adjacent[3] regions have different colors.*

This proposition is true and is known as the "four-color theorem". However, there have been many incorrect proofs, including one that stood for 10 years in the late 19th century before the mistake was found. An extremely laborious proof was finally found in 1976 by mathematicians Appel and Haken who used a complex computer program to categorize the four-colorable maps; the program left a couple of thousand maps uncategorized, and these were checked by hand by Haken and his assistants—including his 15-year-old daughter. There was a lot of debate about whether this was a legitimate proof: the proof was too big to be checked without a computer, and no one could guarantee that the computer calculated correctly, nor did anyone have the energy to recheck the four-colorings of thousands of maps that was done by hand. Finally, about five years ago, a humanly intelligible proof of the four color theorem was found (see `http://www.math.gatech.edu/ thomas/FC/fourcolor.html`). [4]

Proposition 1.2.6 (Goldbach). *Every even integer greater than 2 is the sum of two primes.*

No one knows whether this proposition is true or false. This is the "Goldbach Conjecture," which dates back to 1742.

Problem 1.2.1. Show that no nonconstant polynomial can map all nonnegative integers into prime numbers.

1.3 The Axiomatic Method

The standard procedure for establishing truth in mathematics was invented by Euclid, a mathematician working in Alexadria, Egypt around 300 BC. His idea was to begin with five *assumptions* about geometry, which seemed undeniable based on direct experience. (For example, "There is a straight line segment between every pair of points.) Propositions like these that are simply accepted as true are called ***axioms***.

Starting from these axioms, Euclid established the truth of many additional propositions by providing "proofs". A ***proof*** is a sequence of logical deductions from axioms and previously-proved

[3] Two regions are adjacent only when they share a boundary segment of positive length. They are not considered to be adjacent if their boundaries meet only at a few points.

[4] The story of the four-color proof is told in a well-reviewed recent popular (non-technical) book: "Four Colors Suffice. How the Map Problem was Solved." *Robin Wilson*. Princeton Univ. Press, 2003, 276pp. ISBN 0-691-11533-8.

statements that concludes with the proposition in question. You probably wrote many proofs in high school geometry class, and you'll see a lot more in this course.

There are several common terms for a proposition that has been proved. The different terms hint at the role of the proposition within a larger body of work.

- Important propositions are called ***theorems***.
- A ***lemma*** is a preliminary proposition useful for proving later propositions.
- A ***corollary*** is a proposition that follows in just a few logical steps from a theorem.

The definitions are not precise. In fact, sometimes a good lemma turns out to be far more important than the theorem it was originally used to prove.

Euclid's axiom-and-proof approach, now called the ***axiomatic method***, is the foundation for mathematics today. In fact, there are just a handful of axioms, called the axioms Zermel-Frankel with Choice (ZFC), which, together with a few logical deduction rules, appear to be sufficient to derive essentially all of mathematics.

1.3.1 Our Axioms

The ZFC axioms are important in studying and justifying the foundations of Mathematics. But for practical purposes, they are much too primitive— by one reckoning, proving that $2 + 2 = 4$ requires more than 20,000 steps! So instead of starting with ZFC, we're going to take a *huge* set of axioms as our foundation: we'll accept all familiar facts from high school math!

This will give us a quick launch, but you *will* find this imprecise specification of the axioms troubling at times. For example, in the midst of a proof, you may find yourself wondering, "Must I prove this little fact or can I take it as an axiom?" Feel free to ask for guidance, but really there is no absolute answer. Just be upfront about what you're assuming, and don't try to evade homework and exam problems by declaring everything an axiom!

1.3.2 Proofs in Practice

In principle, a proof can be *any* sequence of logical deductions from axioms and previously-proved statements that concludes with the proposition in question. This freedom in constructing a proof can seem overwhelming at first. How do you even *start* a proof?

Here's the good news: many proofs follow one of a handful of standard templates. Proofs all differ in the details, of course, but these templates at least provide you with an outline to fill in. We'll go through several of these standard patterns, pointing out the basic idea and common pitfalls and giving some examples. Many of these templates fit together; one may give you a top-level outline while others help you at the next level of detail. And we'll show you other, more sophisticated proof techniques later on.

The recipes below are very specific at times, telling you exactly which words to write down on your piece of paper. You're certainly free to say things your own way instead; we're just giving you something you *could* say so that you're never at a complete loss.

The ZFC Axioms

We're *not* going to be working with the Axioms of Zermelo-Frankel Set Theory with Choice (ZFC) in this course, but we thought you might like to see them. Essentially all of mathematics can be derived from these axioms together with a few logical deduction rules.

Extensionality. Two sets are equal if they have the same members. In formal logical notation, this would be stated as:

$$(\forall z.\ (z \in x \longleftrightarrow z \in y)) \longrightarrow x = y.$$

Pairing. For any two sets x and y, there is a set, $\{x, y\}$, with x and y as its only elements.

Union. The union of a collection, z, of sets is also a set.

$$\exists u \forall x.\ (\exists y.\ x \in y \wedge y \in z) \longleftrightarrow x \in u.$$

(The symbol $\exists$ is read "there exists".)

Infinity. There is an infinite set; specifically, a nonempty set, x, such that for any set $y \in x$, the set $\{y\}$ is also a member of x

Subset. Given any set, x, and any proposition $P(y)$, there is a set containing precisely those elements $y \in x$ for which $P(y)$ holds.

Power Set. All the subsets of a set form another set.

Replacement. The image of a set under a function is a set.

Foundation. For every non-empty set, x, there is a set $y \in x$ such that x and y are disjoint. (This most technical of the axioms aims to capture the idea that sets are built up successively from simpler sets. In particular, this axiom prevents a set from being a member of itself.)

Choice. We can choose one element from each set in a collection of nonempty sets. More precisely, if x is a set, and every element of x is itself a set that is nonempty, then there is a "choice" function, g, such that $g(y) \in y$ for every $y \in x$.

1.4 Proving an Implication

Propositions of the form "If P, then Q" are called ***implications***. This implication is often rephrased as "P implies Q."

Here are some examples:

- (Quadratic Formula) If $ax^2 + bx + c = 0$ and $a \neq 0$, then $x = \left(-b \pm \sqrt{b^2 - 4ac}\right)/2a$.
- (Goldbach's Conjecture) If n is an even integer greater than 2, then n is a sum of two primes.
- If $0 \leq x \leq 2$, then $-x^3 + 4x + 1 > 0$.

There are a couple standard methods for proving an implication.

1.4.1 Method #1

In order to prove that P implies Q:

1. Write, "Assume P."
2. Show that Q logically follows.

Example

Theorem 1.4.1. *If $0 \leq x \leq 2$, then $-x^3 + 4x + 1 > 0$.*

Before we write a proof of this theorem, we have to do some scratchwork to figure out why it is true.

The inequality certainly holds for $x = 0$; then the left side is equal to 1 and $1 > 0$. As x grows, the $4x$ term (which is positive) initially seems to have greater magnitude than $-x^3$ (which is negative). For example, when $x = 1$, we have $4x = 4$, but $-x^3 = -1$ only. In fact, it looks like $-x^3$ doesn't begin to dominate until $x > 2$. So it seems the $-x^3 + 4x$ part should be nonnegative for all x between 0 and 2, which would imply that $-x^3 + 4x + 1$ is positive.

So far, so good. But we still have to replace all those "seems like" phrases with solid, logical arguments. We can get a better handle on the critical $-x^3 + 4x$ part by factoring it, which is not too hard:

$$-x^3 + 4x = x(2 - x)(2 + x)$$

Aha! For x between 0 and 2, all of the terms on the right side are nonnegative. And a product of nonnegative terms is also nonnegative. Let's organize this blizzard of observations into a clean proof.

Proof. Assume $0 \leq x \leq 2$. Then x, $2 - x$, and $2 + x$ are all nonnegative. Therefore, the product of these terms is also nonnegative. Adding 1 to this product gives a positive number, so:

$$x(2 - x)(2 + x) + 1 > 0$$

Multiplying out on the left side proves that

$$-x^3 + 4x + 1 > 0$$

as claimed. □

There are a couple points here that apply to all proofs:

- You'll often need to do some scratchwork while you're trying to figure out the logical steps of a proof. Your scratchwork can be as disorganized as you like— full of dead-ends, strange diagrams, obscene words, whatever. But keep your scratchwork separate from your final proof, which should be clear and concise.
- Proofs typically begin with the word "Proof" and end with some sort of doohickey like □ or "q.e.d". The only purpose for these conventions is to clarify where proofs begin and end.

1.4.2 Method #2 - Prove the Contrapositive

An implication ("P implies Q") is logically equivalent to its *contrapositive* "not Q implies not P"; proving one is as good as proving the other, and proving the contrapositive is sometimes easier than proving the original statement. If so, then you can proceed as follows:

1. Write, "We prove the contrapositive:" and then state the contrapositive.
2. Proceed as in Method #1.

Example

Theorem 1.4.2. *If r is irrational, then $\sqrt{r}$ is also irrational.*

Recall that rational numbers are equal to a ratio of integers and irrational numbers are not. So we must show that if r is *not* a ratio of integers, then $\sqrt{r}$ is also *not* a ratio of integers. That's pretty convoluted! We can eliminate both "not"'s and make the proof straightforward by considering the contrapositive instead.

Proof. We prove the contrapositive: if $\sqrt{r}$ is rational, then r is rational.

Assume that $\sqrt{r}$ is rational. Then there exists integers a and b such that:

$$\sqrt{r} = \frac{a}{b}$$

Squaring both sides gives:

$$r = \frac{a^2}{b^2}$$

Since a^2 and b^2 are integers, r is also rational. □

1.5 Proving an "If and Only If"

Many mathematical theorems assert that two statements are logically equivalent; that is, one holds if and only if the other does. Here are three clever examples:

- An integer is a multiple of 3 if and only if the sum of its digits is a multiple of 3.
- Two triangles have the same side lengths if and only if two sidelengths and an angle are the same.
- A positive integer $p \geq 2$ is prime if and only if $1 + (p-1) \cdot (p-2) \cdots 3 \cdot 2 \cdot 1$ is a multiple of p.

1.5.1 Method #1: Prove Each Statement Implies the Other

The statement "P if and only if Q" is equivalent to the two statements "P implies Q" and "Q implies P". So you can prove an "if and only if" by proving *two* implications:

1. Write, "We prove P implies Q and vice-versa."
2. Write, "First, we show P implies Q." Do this by one of the methods in Section 1.4.
3. Write, "Now, we show Q implies P." Again, do this by one of the methods in Section 1.4.

1.5.2 Method #2: Construct a Chain of Iffs

In order to prove that P is true if and only if Q is true:

1. Write, "We construct a chain of if-and-only-if implications."
2. Prove P is equivalent to a second statement which is equivalent to a third staement and so forth until you reach Q.

This method is generally more difficult than the first, but the result can be a short, elegant proof.

Example

The *standard deviation* of a sequence of values $x_1, x_2, \ldots, x_n$ is defined to be:

$$\sqrt{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}$$

where μ is the average of the values:

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Theorem 1.5.1. *The standard deviation of a sequence of values $x_1, \ldots, x_n$ is zero if and only if all the values are equal to the mean.*

For example, the standard deviation of test scores is zero if and only if everyone scored exactly the class average.

Proof. We construct a chain of "if and only if" implications. The standard deviation of $x_1, \ldots, x_n$ is zero if and only if:

$$\sqrt{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2} = 0$$

where μ is the average of $x_1, \ldots, x_n$. This equation holds if and only if

$$(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2 = 0$$

since zero is the only number whose square root is zero. Every term in this equation is nonnegative, so this equation holds if and only if every term is actually 0. But this is true if and only if every value x_i is equal to the mean, μ. □

1.6 How to Write *Good* Proofs

One purpose of a proof is to establish the truth of an assertion with absolute certainty. Mechanically checkable proofs of enormous length or complexity can accomplish this. But the most desirable proofs are humanly intelligible ones that help the reader understand the subject matter. The teachers of 6.042 share the view held nearly universally among Mathematicians that important mathematical results can't be fully understood until their proofs are understood, and that is why proofs are an important part of the 6.042 curriculum.

To be understandable and helpful, more is required of a proof than just logical correctness: a good proof must also be clear. Correctness and clarity are usually complementary; a well-written proof is more likely to be a correct proof, since mistakes are harder to hide.

In practice, the notion of proof is a moving target. Proofs in a professional research journal are generally unintelligible to all but a few experts who know all the terminology and prior results used, often without explicit mention, in the proof. Conversely, proofs in the first weeks of a beginning course like 6.042 would be regarded as tediously long-winded by a professional mathematician. In fact, what we accept as a good proof later in the term will be different from what we consider good proofs in the first couple of weeks of 6.042. But even so, we can offer some general tips on writing good proofs:

State your game plan. A good proof begins by explaining the general line of reasoning, e.g. "We use case analysis" or "We argue by contradiction". This creates a rough mental picture into which the reader can fit the subsequent details.

Keep a linear flow. We sometimes see proofs that are like mathematical mosaics, with juicy tidbits of reasoning sprinkled across the page. This is not good. The steps of your argument should follow one another in a sequential order.

A proof is an essay, not a calculation. Many students initially write proofs the way they compute integrals. The result is a long sequence of expressions without explanation, making it very hard to follow. This is bad. A good proof usually looks like an essay with some equations thrown in. Use complete sentences.

Avoid excessive symbolism. Your reader is probably good at understanding words, but much less skilled at reading arcane mathematical symbols. So use words where you reasonably can.

Revise and simplify. Your readers will be grateful.

Introduce notation thoughtfully. Sometimes an argument can be greatly simplified by introducing a variable, devising a special notation, or defining a new term. But do this sparingly since you're requiring the reader to remember all that new stuff. And remember to actually *define* the meanings of new variables, terms, or notations; don't just start using them!

Structure long proofs. Long programs are usually broken into a hierarchy of smaller procedures. Long proofs are much the same. Facts needed in your proof that are easily stated, but not readily proved are best pulled out and proved in preliminary lemmas. Also, if you are repeating essentially the same argument over and over, try to capture that argument in a general lemma, which you can cite repeatedly instead.

Be wary of the "obvious". When familiar or truly obvious facts are needed in a proof, it's OK to label them as such and to not prove them. But remember that what's obvious to you, may not (and typically is not) obvious to your reader.

Most especially, don't use phrases like "clearly" or "obviously" in an attempt to bully the reader into accepting something which you're having trouble proving. Also, go on the alert whenever you see one of these phrases in someone else's proof.

Finish. At some point in a proof, you'll have established all the essential facts you need. Resist the temptation to quit and leave the reader to draw the "obvious" conclusion. Instead, tie everything together yourself and explain why the original claim follows.

The analogy between good proofs and good programs extends beyond structure. The same rigorous thinking needed for proofs is essential in the design of critical computer systems. When algorithms and protocols only "mostly work" due to reliance on hand-waving arguments, the results can range from problematic to catastrophic. An early example was the Therac 25, a machine that provided radiation therapy to cancer victims, but occasionally killed them with massive overdoses due to a software race condition. More recently, in August 2004, a single faulty command to a computer system used by United and American Airlines grounded the entire fleet of both companies— and all their passengers!

It is a certainty that we'll all one day be at the mercy of critical computer systems designed by you and your classmates. So we really hope that you'll develop the ability to formulate rock-solid logical arguments that a system actually does what you think it does!

1.7 Propositional Formulas

It's really sort of amazing that people manage to communicate in the English language. Here are some typical sentences:

1. "You may have cake or you may have ice cream."

2. "If pigs can fly, then you can understand the Chebyshev bound."
3. "If you can solve any problem we come up with, then you get an A for the course."
4. "Every American has a dream."

What *precisely* do these sentences mean? Can you have both cake and ice cream or must you choose just one dessert? If the second sentence is true, then is the Chebyshev bound incomprehensible? If you can solve some problems we come up with but not all, then do you get an A for the course? And can you still get an A even if you can't solve any of the problems? Does the last sentence imply that all Americans have the same dream or might some of them have different dreams?

Some uncertainty is tolerable in normal conversation. But when we need to formulate ideas precisely— as in mathematics and programming— the ambiguities inherent in everyday language can be a real problem. We can't hope to make an exact argument if we're not sure exactly what the individual words mean. So before we start into mathematics, we need to investigate the problem of how to talk about mathematics.

To get around the ambiguity of English, mathematicians have devised a special mini-language for talking about logical relationships. This language mostly uses ordinary English words and phrases such as "or", "implies", and "for all". But mathematicians endow these words with definitions more precise than those found in an ordinary dictionary. Without knowing these definitions, you could sort of read this language, but you would miss all the subtleties and sometimes have trouble following along.

Surprisingly, in the midst of learning the language of logic, we'll come across the most important open problem in computer science— a problem whose solution could change the world.

1.7.1 Combining Propositions

In English, we can modify, combine, and relate propositions with words such as "not", "and", "or", "implies", and "if-then". For example, we can combine three propositions into one like this:

If all humans are mortal **and** all Greeks are human, **then** all Greeks are mortal.

For the next while, we won't be much concerned with the internals of propositions— whether they involve mathematics or Greek mortality— but rather with how propositions are combined and related. So we'll frequently use variables such as P and Q in place of specific propositions such as "All humans are mortal" and "$2 + 3 = 5$". The understanding is that these variables, like propositions, can take on only the values **T**(true) and **F**(false). Such true/false variables are sometimes called ***Boolean variables*** after their inventor, George— you guessed it— Boole.

"Not", "And" and "Or"

We can precisely define these special words using ***truth tables***. For example, if P denotes an arbitrary proposition, then the truth of the proposition "not P" is defined by the following truth table:

P	not P
T	**F**
F	**T**

The first row of the table indicates that when proposition P is true, the proposition "not P" is false (F). The second line indicates that when P is false, "not P" is true. This is probably what you would expect.

In general, a truth table indicates the true/false value of a proposition for each possible setting of the variables. For example, the truth table for the proposition "P and Q" has four lines, since the two variables can be set in four different ways:

P	Q	P and Q
T	**T**	**T**
T	**F**	**F**
F	**T**	**F**
F	**F**	**F**

According to this table, the proposition "P and Q" is true only when P and Q are both true. This is probably reflects the way you think about the word "and".

There is a subtlety in the truth table for "P or Q":

P	Q	P or Q
T	**T**	**T**
T	**F**	**T**
F	**T**	**T**
F	**F**	**F**

This says that "P or Q" is true when P is true, Q is true, or *both* are true. This isn't always the intended meaning of "or" in everyday speech, but this is the standard definition in mathematical writing. So if a mathematician says, "You may have cake or you may have ice cream", then you *could* have both.

"Implies"

The least intuitive connecting word is "implies". Here is its truth table, with the lines labelled so we can refer to the them later.

P	Q	P implies Q	
T	**T**	**T**	(tt)
T	**F**	**F**	(tf)
F	**T**	**T**	(ft)
F	**F**	**T**	(ff)

Let's experiment with this definition. For example, is the following proposition true or false?

"If Goldbach's Conjecture is true, then $x^2 \geq 0$ for every real number x."

Now, we told you before that no one knows whether Goldbach's Conjecture is true or false. But that doesn't prevent you from answering the question! This proposition has the form $P \longrightarrow Q$ where P is "Goldbach's Conjecture is true" and Q is "$x^2 \geq 0$ for every real number x". Since Q is definitely true, we're on either line (tt) or line (ft) of the truth table. Either way, the proposition as a whole is *true*!

One of our original examples demonstrates an even stranger side of implications.

"If pigs fly, then you can understand the Chebyshev bound."

Don't take this as an insult; we just need to figure out whether this proposition is true or false. Curiously, the answer has *nothing* to do with whether or not you can understand the Chebyshev bound. Pigs do not fly, so we're on either line (ft) or line (ff) of the truth table. In both cases, the proposition is *true*!

In contrast, here's an example of a false implication:

"If the moon shines white, then the moon is made of white cheddar."

Yes, the moon shines white. But, no, the moon is not made of white cheddar cheese. So we're on line (tf) of the truth table, and the proposition is false.

The truth table for implications can be summarized in words as follows:

An implication is true exactly when the if-part is false or the then-part is true.

This sentence is worth remembering; a large fraction of all mathematical statements are of the if-then form!

"If and Only If"

Mathematicians commonly join propositions in one additional way that doesn't arise in ordinary speech. The proposition "P if and only if Q" asserts that P and Q are logically equivalent; that is, either both are true or both are false.

P	Q	P if and only if Q
T	T	T
T	F	F
F	T	F
F	F	T

The following if-and-only-if statement is true for every real number x:

"$x^2 - 4 \geq 0$ if and only if $|x| \geq 2$"

For some values of x, *both* inequalities are true. For other values of x, *neither* inequality is true . In every case, however, the proposition as a whole is true.

The phrase "if and only if" comes up so often that it is often abbreviated "iff".

1.7.2 Propositional Logic in Computer Programs

Propositions and logical connectives arise all the time in computer programs. For example, consider the following snippet, which could be either C, C++, or Java:

```
if ( x > 0 || (x <= 0 && y > 100) )
    ⋮
    (further instructions)
```

The symbol `||` denotes "or", and the symbol `&&` denotes "and". The *further instructions* are carried out only if the proposition following the word `if` is true. On closer inspection, this big expression is built from two simpler propositions. Let A be the proposition that `x > 0`, and let B be the proposition that `y > 100`. Then we can rewrite the condition this way:

$$A \text{ or } ((\text{not } A) \text{ and } B)$$

A truth table reveals that this complicated expression is logically equivalent to "A or B".

A	B	A or $((\text{not } A)$ and $B)$	A or B
T	**T**	**T**	**T**
T	**F**	**T**	**T**
F	**T**	**T**	**T**
F	**F**	**F**	**F**

This means that we can simplify the code snippet without changing the program's behavior:

```
if ( x > 0 || y > 100 )
    (further instructions)
```

Rewriting a logical expression involving many variables in the simplest form is both difficult and important. Simplifying expressions in software might slightly increase the speed of your program. But, more significantly, chip designers face essentially the same challenge. However, instead of minimizing && and || symbols in a program, their job is to minimize the number of analogous physical devices on a chip. The payoff is potentially enormous: a chip with fewer devices is smaller, consumes less power, has a lower defect rate, and is cheaper to manufacture.

1.7.3 A Cryptic Notation

Programming languages use symbols like && and ! in place of words like "and" and "not". Mathematicians have devised their own cryptic symbols to represent these words, which are summarized in the table below.

English	**Cryptic Notation**
"not P"	$\neg P$ (alternatively, $\overline{P}$)
"P and Q"	$P \wedge Q$
"P or Q"	$P \vee Q$
"P implies Q" or "if P then Q"	$P \longrightarrow Q$
"P if and only if Q"	$P \longleftrightarrow Q$

For example, using this notation, "If P and not Q, then R" would be written:

$$(P \wedge \neg Q) \longrightarrow R$$

This symbolic language is helpful for writing complicated logical expressions compactly. But in most contexts ordinary words such as "or" and "implies" are much easier to understand than symbols such as $\vee$ and $\longrightarrow$. So we'll use this symbolic language sparingly, and we advise you to do the same.

1.7.4 Logically Equivalent Implications

Are these two sentences saying the same thing?

If I am hungry, then I am grumpy.
If I am not grumpy, then I am not hungry.

We can settle the issue by recasting both sentences in terms of propositional logic. Let P be the proposition "I am hungry", and let Q be "I am grumpy". The first sentence says "P implies Q" and the second says "(not Q) implies (not P)". We can compare these two statements in a truth table:

P	Q	P implies Q	(not Q) implies (not P)
T	T	T	T
T	F	F	F
F	T	T	T
F	F	T	T

Sure enough, the two statements are precisely equivalent. In general, "(not Q) implies (not P)" is called the ***contrapositive*** of "P implies Q". And, as we've just shown, the two are just different ways of saying the same thing.

In contrast, the ***converse*** of "P implies Q" is the statement "Q implies P". In terms of our example, the converse is:

If I am grumpy, then I am hungry.

This sounds like a rather different contention, and a truth table confirms this suspicion:

P	Q	P implies Q	Q implies P
T	T	T	T
T	F	F	T
F	T	T	F
F	F	T	T

Thus, an implication *is* logically equivalent to its contrapositive, but is *not* equivalent to its converse.

One final relationship: an implication and its converse together are equivalent to an if and only if statement, specifically, to these two statements together. For example,

If I am grumpy, then I am hungry.
If I am hungry, then I am grumpy.

are equivalent to the single statement:

I am grumpy if and only if I am hungry.

Once again, we can verify this with a truth table:

P	Q	(P implies Q) and (Q implies P)	Q if and only if P
T	T	T	T
T	F	F	F
F	T	F	F
F	F	T	T

SAT

A proposition is **satisfiable** if some setting of the variables makes the proposition true. For example, $P \wedge \neg Q$ is satisfiable because the expression is true when P is true and Q is false. On the other hand, $P \wedge \neg P$ is not satisfiable because the expression as a whole is false for both settings of P. But determining whether or not a more complicated proposition is satisfiable is not so easy. How about this one?

$$(P \vee Q \vee R) \wedge (\neg P \vee \neg Q) \wedge (\neg P \vee \neg R) \wedge (\neg R \vee \neg Q)$$

The general problem of deciding whether a proposition is satisfiable is called ***SAT***. One approach to SAT is to construct a truth table and check whether or not a **T** ever appears. But this approach is not very efficient; a proposition with n variables has a truth table with 2^n lines. For a proposition with just 30 variables, that's already over a billion!

Is there an *efficient* solution to SAT? Is there some ingenious procedure that *quickly* determines whether any given proposition is satifiable or not? No one knows. And an awful lot hangs on the answer. An efficient solution to SAT would immediately imply efficient solutions to many, many other important problems involving packing, scheduling, routing, and circuit verification. This sounds fantastic, but there would also be worldwide chaos. Decrypting coded messages would also become an easy task (for most codes). Online financial transactions would be insecure and secret communications could be read by everyone.

At present, though, researchers are completely stuck. No one has a good idea how to either solve SAT more efficiently or to prove that no efficient solution exists. This is the outstanding unanswered question in theoretical Computer Science.

1.8 Logical Deductions

Logical deductions or *inference rules* are used to prove new propositions using previously proved ones.

A fundamental inference rule is *modus ponens*. This rule says that a proof of P together with a proof of $P \longrightarrow Q$ is a proof of Q.

Inference rules are sometimes written in a funny notation. For example, *modus ponens* is written:

Rule.

$$\frac{P, \quad P \longrightarrow Q}{Q}$$

When the statements above the line, called the *antecedents*, are proved, then we can consider the statement below the line, called the *conclusion* or *consequent*, to also be proved.

A key requirement of an inference rule is that it must be *sound*: any assignment of truth values that makes all the antecedents true must also make the consequent true. So it we start off with true axioms and apply sound inference rules, everything we prove will also be true.

There are many other natural, sound inference rules, for example:

Rule.

$$\frac{P \longrightarrow Q, \quad Q \longrightarrow R}{P \longrightarrow R}$$

Rule.

$$\frac{\neg P \longrightarrow Q, \quad \neg Q}{P}$$

Rule.

$$\frac{\neg P \longrightarrow \neg Q}{Q \longrightarrow P}$$

On the other hand,

Rule.

$$\frac{\neg P \longrightarrow \neg Q}{P \longrightarrow Q}$$

is not sound: if P is assigned **T** and Q is assigned **F**, then the antecedent is true and the consequent is not.

Problem 1.8.1. Prove that a propositional inference rule is sound iff the conjunction (AND) of all its antecedents implies its consequent.

As with axioms, we will not be too formal about the set of legal inference rules. Each step in a proof should be clear and "logical"; in particular, you should state what previously proved facts are used to derive each new conclusion.

1.9 In-Class Problems Week 1, Wed.

Problem 1.9.1. Identify exactly where the bugs are in each of the following bogus proofs.[5]

(a) $1/8 > 1/4$.

Bogus proof.

$$\begin{aligned} 3 &> 2 \\ 3\log_{10}(1/2) &> 2\log_{10}(1/2) \\ \log_{10}(1/2)^3 &> \log_{10}(1/2)^2 \\ (1/2)^3 &> (1/2)^2, \end{aligned}$$

and the claim now follows by the rules for multiplying fractions. □

Solution. $\log x < 0$, for $0 < x < 1$, so since both sides of the inequality "$3 > 2$" are being multiplied by the negative quantity $\log_{10}(1/2)$, the "$>$" in the second line should have been "$<$." ■

(b) $1¢ = \$0.01 = (\$0.1)^2 = (10¢)^2 = 100¢ = \$1$.

Solution. $\$0.01 = \$(0.1)^2 \neq (\$0.1)^2$ because the units $\$^2$ and $\$ don't match (just as in physics the difference between sec^2 and sec indicates the difference between acceleration and velocity). Similarly, $(10¢)^2 \neq 100¢$. ■

Problem 1.9.2.

Proposition (Arithmetic-Geometric Mean Inequality). *For all nonnegative real numbers a and b*

$$\frac{a+b}{2} \geq \sqrt{ab}.$$

What is wrong with the following proof of this proposition?

Bogus proof.

$$\begin{aligned} \frac{a+b}{2} &\stackrel{?}{\geq} \sqrt{ab} \\ a+b &\stackrel{?}{\geq} 2\sqrt{ab} \\ a^2+2ab+b^2 &\stackrel{?}{\geq} 4ab \\ a^2-2ab+b^2 &\stackrel{?}{\geq} 0 \\ (a-b)^2 &\geq 0 \end{aligned}$$

[5] From Stueben, Michael and Diane Sandford. *Twenty Years Before the Blackboard*, Math. Assoc America, ©1998.

The last statement is true because $a - b$ is a real number, and the square of a real number is never negative. This proves the claim. □

Solution. In this argument, we started with what we wanted to prove and then reasoned until we reached a statement that is surely true. The little question marks presumably are supposed to indicate that we're not quite certain that the inequalities are valid until we get down to the last step. At that step, the inequality checks out, *but that doesn't prove the claim*. All we have proved is that **if** $(a + b)/2 \geq \sqrt{ab}$, **then** $(a - b)^2 \geq 0$, which is not very interesting, since we already knew that the square of any nonnegative number is nonnegative.

To be fair, this bogus proof is pretty good: if it was written in reverse order – or if "is implied by" was simply inserted after each line – it would actually prove the Arithmetic-Geometric Mean Inequality:

Proof.

$$\begin{aligned} \frac{a+b}{2} &\geq \sqrt{ab} && \text{is implied by} \\ a + b &\geq 2\sqrt{ab}, && \text{which is implied by} \\ a^2 + 2ab + b^2 &\geq 4ab, && \text{which is implied by} \\ a^2 - 2ab + b^2 &\geq 0, && \text{which is implied by} \\ (a - b)^2 &\geq 0. \end{aligned}$$

The last statement is true because $a - b$ is a real number, and the square of a real number is never negative. This proves the claim. □

But the problem with the bogus proof as written is that it reasons backward, beginning with the proposition in question and reasoning to a true conclusion. This kind of backward reasoning can easily "prove" false statements. Here's an example:

False Claim.

$$0 = 1.$$

Bogus proof.

$$\begin{aligned} 0 &\stackrel{?}{=} 1 \\ 1 &\stackrel{?}{=} 0 \\ 0 + 1 &\stackrel{?}{=} 1 + 0 \\ 1 &= 1 \end{aligned}$$

and the last equality is trivially true. □

We can also come up with very easy "proofs" of true theorems, for example, here's an easy "proof" of the Arithmetic-Geometric Mean Inequality:

Bogus proof.

$$\frac{a+b}{2} \stackrel{?}{\geq} \sqrt{ab}$$
$$0 \cdot \frac{a+b}{2} \stackrel{?}{\geq} 0 \cdot \sqrt{ab}$$
$$0 \geq 0 \qquad \square$$

So watch out for backward proofs! ■

1.10 In-Class Problems Week 1, Fri.

Problem 1.10.1. Albert announces that he plans a surprise 6.042 quiz next week. His students wonder if the quiz could be next Friday. The students realize that it obviously cannot, because if it hadn't been given before Friday, everyone would know that there was only Friday left on which to give it, so it wouldn't be a surprise any more.

So the students ask whether Albert could give the surprise quiz Thursday? They observe that if the quiz wasn't given *before* Thursday, it would have to be given *on* the Thursday, since they already know it can't be given on Friday. But having figured that out, it wouldn't be a surprise if the quiz was on Thursday either. Similarly, the students reason that the quiz can't be on Wednesday, Tuesday, or Monday. Namely, it's impossible for Albert to give a surprise quiz next week. All the students now relax having concluded that Albert must have been bluffing.

And since no one expects the quiz, that's why, when Albert gives it on Tuesday next week, it really is a surprise!

What do you think is wrong with the students' reasoning?

Solution. The basic problem is that "surprise" is not a mathematical concept, nor is there any generally accepted way to give it a mathematical definition. The "proof" above assumes some plausible axioms about surprise, without defining it. The paradox is that these axioms are inconsistent. But that's no surprise `:-)`, since—mathematically speaking—we don't know what we're talking about.

Mathematicians and philosophers have had a lot more to say about what might be wrong with the students' reasoning, (see Chow, Timothy Y. *The surprise examination or unexpected hanging paradox*, American Math. Monthly (January 1998), pp.41–51.) ■

Problem 1.10.2. Identify the antecedents and conclusions of each of the following deductions and translate them into propositional logic notation using logical operators:

$$\begin{aligned} \wedge &::= \text{AND}, \\ \vee &::= \text{OR}, \\ \neg &::= \text{NOT}, \\ \longrightarrow &::= \text{IMPLIES}, \\ \longleftrightarrow &::= \text{IFF (if and only if)} \end{aligned}$$

This may require that you "pin down" a statement that could be interpreted in more than one way.

Identify which of the deductions are sound ones.

(a) Jane and Pete won't both win the math prize. Pete will win either the math prize or the chemistry prize. Jane will win the math prize. Thus, Pete will win the chemistry prize.

Solution. The deduction is:

$$\frac{\neg(J \wedge M), \quad M \vee C, \quad J}{C}$$

where

$$\begin{aligned} J &::= \text{“Jane will win the math prize.”} \\ M &::= \text{“Pete will win the math prize.”} \\ C &::= \text{“Pete will win the chemistry prize.”} \end{aligned}$$

This deduction is sound. ■

(b) The main course will be beef or fish. The vegetable will be peas or corn. We will not have both fish as a main course and corn as a vegetable. Therefore, we will not have both beef as a main course and peas as a vegetable.

Solution. The deduction is:

$$\frac{B \vee F, \quad C \vee P, \quad \neg(F \wedge C)}{\neg(B \wedge P)}$$

where

$$\begin{aligned} B &::= \text{“The main course will be beef.”} \\ F &::= \text{“The main course will be fish.”} \\ C &::= \text{“The vegetable will be corn.”} \\ P &::= \text{“The vegetable will be peas.”} \end{aligned}$$

This deduction is not sound. For example, $B \wedge \neg F \wedge C \wedge P$ is consistent with the antecedents but not with the conclusion. Note that as formalized, there need not be only one main course and only one vegetable; it is possible, for example, for the vegetable to be both corn and peas, as in the scenario given.

If we wished to exclude the possibility of multiple courses we could have used *exclusive-or* instead of inclusive-or. So our antecedent about the main course would then read $B \oplus F$ or, equivalently, $(B \vee F) \wedge \neg(B \wedge F)$. The antecedent about the vegetable could be changed similarly. The deduction is still unsound in this formalization. ■

(c) Either John or Bill is telling the truth. Either Sam or Bill is lying. Thus, either John is telling the truth or Sam is lying.

Solution. We interpret “John is lying,” to be the negation of “John is telling the truth.” Similarly for the corresponding propositions involving Bill and Sam. The deduction is:

$$\frac{J \vee B, \quad \neg S \vee \neg B}{J \vee \neg S}$$

where

$$\begin{aligned} J &::= \text{“John is telling the truth.”} \\ B &::= \text{“Bill is telling the truth.”} \\ S &::= \text{“Sam is telling the truth.”} \end{aligned}$$

This deduction is sound. It is an example of a common “cancellation” or *cut* rule that lets us get rid of the proposition B in the conclusion. ■

(d) Either sales will go up and the boss will be happy, or expenses will go up and the boss won't be happy. Therefore, sales and expenses will not both go up.

Solution. The deduction is:

$$\frac{(S \wedge H) \vee (E \wedge \neg H)}{\neg(S \wedge E)}$$

where

$$\begin{aligned} S &::= \text{“Sales will go up.”} \\ H &::= \text{“The boss will be happy.”} \\ E &::= \text{“Expenses will go up.”} \end{aligned}$$

This deduction is not sound. For example, $S \wedge E \wedge H$ is consistent with the antecedent but not with the conclusion. ■

Problem 1.10.3. (* *No group really got to this problem in class, so we promise it won't show up on a mini-quiz. However, if you want, you're welcome to check out the solutions!*)

Boolean logic comes up in digital circuit design using the convention that **T** corresponds to 1 and **F** to 0. For example, suppose we want to describe a circuit with $n+1$ inputs $a_n, a_{n-1}, \ldots, a_1, a_0$ which are the $n+1$ bits of the binary representation of an integer, k, between 0 and $2^{n+1}-1$. We want outputs $o_{n+1}, o_n, \ldots, o_1, o_0$ to be the bits of $k+b$ where b is a single bit.

For example, for $n = 1$, the formulas

$$\begin{aligned} o_0 &::= a_0 \oplus b \\ c_1 &::= a_0 \wedge b && \text{the carry into column 1} \\ o_1 &::= c_1 \oplus a_1 \\ c_2 &::= c_1 \wedge a_1 && \text{the carry into column 2} \\ o_2 &::= c_2 \end{aligned}$$

do the job. Here $\oplus$ is the "mod 2 sum" operator: $a \oplus b$ is 1 iff $a + b$ is odd.

(a) Generalize the example above for any $n \geq 0$. That is, give simple formulas for o_i and c_i for $0 \leq i \leq n+1$.

Solution. Define

$$\begin{aligned} o_0 &::= a_0 \oplus b, \\ c_1 &::= a_0 \wedge b, \\ c_{i+1} &::= a_i \wedge c_i && \text{for } 1 \leq i \leq n, \\ o_{i+1} &::= c_{i+1} \oplus a_{i+1} && \text{for } 0 \leq i < n, \\ o_{n+1} &::= c_{n+1}. \end{aligned}$$

■

(b) Write similar definitions for the $n+1$ bits of the sum of two binary numbers $a_n, a_{n-1}, \ldots, a_1, a_0$ and $b_n, b_{n-1}, \ldots, b_1, b_0$.

Solution. Define

$$\begin{aligned}
o_0 &::= a_0 \oplus b_0, \\
c_1 &::= a_0 \wedge b_0, \\
c_{i+1} &::= (a_i \wedge b_i) \vee (a_i \wedge c_i) \vee (b_i \wedge c_i) && \text{for } 1 \leq i \leq n, \\
o_{i+1} &::= c_{i+1} \oplus a_{i+1} \oplus b_{i+1} && \text{for } 0 \leq i < n, \\
o_{n+1} &::= c_{n+1}.
\end{aligned}$$

■

(c) How many Boolean operations does your system use to calculate the sum?

Solution. The scheme above uses $3(n+1)$ AND'S, $2n+1$ MOD-2-SUMS and $2(n+1)$ OR's for a total of $7n+5$ operations. ■

Chapter 2

Predicates & Sets

2.1 More Proof Techniques

2.1.1 Proof by Cases

In Week1 Notes we verified by truth table that the two expressions $A \vee (\overline{A} \wedge B)$ and $A \vee B$ were equivalent. Another way to prove this would be to reason *by cases*:

A is **T**. Then $A \vee$ anything will have truth value **T**. Since both expressions are of this form, in this case, both have the same truth value, namely, **T**.

A is **F**. Then $A \vee P$ will have the same truth value as P for any proposition, P. So the second expression has the same truth value as B. Similarly, the first expression has the same truth value as $\overline{\mathbf{F}} \wedge B$ which also has the same value as B. So in this case, both expressions will have the same truth value, namely, the value of B.

Here's a slightly more interesting example. Let's agree that given any two people, either they have met or not. If every pair of people in a group has met, we'll call the group a ***club***. If every pair of people in a group has not met, we'll call it a group of ***strangers***.

Theorem. *Every collection of 6 people includes a club of 3 people or a group of 3 strangers.*

Proof. The proof is by case analysis[1]. Let x denote one of the six people. There are two cases:

1. Among the remaining 5 people, at least 3 have met x.
2. Among the remaining 5 people, at least 3 have not met x.

We first argue that at least one of these two cases must hold.[2] We'll prove this itself by cases. Namely, let m be the number of the five remaining people that have met x and n the number that

[1] Describing your approach at the outset helps orient the reader.

[2] Part of a case analysis argument is showing that you've covered all the cases. Often this is trivial, because the two cases are of the form "P" and "not P". However, the situation above is not quite so simple.

have not met x. So $m + n = 5$. Now one case is that $m \geq 3$, in which case 1 holds. The other case is that $m < 3$, so $n = 5 - m > 5 - 3 = 2$, that is, $n \geq 3$ which means case 2 holds. So at least one of the cases 1 and 2 must hold.

Case 1: Suppose that at least 3 people did meet x.

This case splits into two subcases:

> **Case 1.1:** no pair among those people met each other. Then these people are a group of at least 3 strangers. So the Theorem holds in this subcase.
>
> **Case 1.2:** some pair among those people have met each other. Then that pair, together with x, form a club of 3 people. So the Theorem holds in this subcase.

This implies that the Theorem holds in Case 1.

Case 2: Suppose that at least 3 people did not meet x.

This case also splits into two subcases:

> **Case 2.1**: every pair among those people met each other. Then these people are a club of at least 3 people. So the Theorem holds in this subcase.
>
> **Case 2.2:** some pair among those people have not met each other. Then that pair, together with x, form a group of at least 3 strangers. So the Theorem holds in this subcase.

This implies that the Theorem also holds in Case 2, and therefore holds in all cases. □

2.1.2 Proof by Contradiction

In a ***proof by contradiction*** or ***indirect proof***, you show that if a proposition were false, then some logical contradiction or absurdity would follow. Thus, the proposition must be true. So proof by contradiction would be described by the inference rule

Rule.

$$\frac{\neg P \longrightarrow \mathbf{F}}{P}$$

Proof by contradiction is *always* a viable approach. However, as the name suggests, indirect proofs can be a little convoluted. So direct proofs are generally preferable as a matter of clarity.

2.1.3 Method

In order to prove a proposition P by contradiction:

1. Write, "We use proof by contradiction."
2. Write, "Suppose P is false."
3. Deduce something known to be false (a logical contradiction).
4. Write, "This is a contradiction. Therefore, P must be true."

Example

Remember that a number is *rational* if it is equal to a ratio of integers. For example, $3.5 = 7/2$ and $0.1111\cdots = 1/9$ are rational numbers. On the other hand, we'll prove by contradiction that $\sqrt{2}$ is irrational.

Theorem 2.1.1. *$\sqrt{2}$ is irrational.*

Proof. We use proof by contradiction. Suppose the claim is false; that is, $\sqrt{2}$ is rational. Then we can write $\sqrt{2}$ as a fraction a/b in *lowest terms*.

Squaring both sides gives $2 = a^2/b^2$ and so $2b^2 = a^2$. This implies that a is even; that is, a is a multiple of 2. Therefore, a^2 must be a multiple of 4. Because of the equality $2b^2 = a^2$, we know $2b^2$ must also be a multiple of 4. This implies that b^2 is even and so b must be even. But since a and b are both even, the fraction a/b is not in lowest terms.

This is a contradiction. Therefore, $\sqrt{2}$ must be irrational. □

2.2 Predicates

A ***predicate*** is a proposition whose truth depends on the value of one or more variables. For example,

"n is a perfect square"

is a predicate whose truth depends on the value of n. The predicate is true for $n = 4$ since 4 is a prefect square, but false for $n = 5$ since 5 is not a perfect square.

Like other propositions, predicates are often named with a letter. Furthermore, a function-like notation is used to denote a predicate supplied with specific variable values. For example, we might name our earlier predicate P:

$$P(n) ::= \text{"}n \text{ is a perfect square"}$$

Now $P(4)$ is true, and $P(5)$ is false.[3]

This notation for predicates is confusingly similar to ordinary function notation. If P is a predicate, then $P(n)$ is either *true* or *false*, depending on the value of n. On the other hand, if p is an ordinary function, like $n^2 + 1$, then $p(n)$ is a *numerical quantity*. Students frequently confuse these two.

2.2.1 Quantifying a Predicate

There are a couple of assertions one commonly makes about a predicate: that it is *sometimes* true and that it is *always* true. For example, the predicate

"$x^2 \geq 0$"

[3]The symbol ::= means "equal by definition." It's always ok to simply write "=" instead of ::=, but reminding the reader that an equality holds by definition can be helpful.

is always true when x is a real number. On the other hand, the predicate

$$\text{“}5x^2 - 7 = 0\text{”}$$

is only sometimes true; specifically, when $x = \pm\sqrt{7/5}$.

There are several ways to express the notions of “always true” and “sometimes true” in English. The table below gives some general formats on the left and specific examples using those formats on the right. You can expect to see such phrases hundreds of times in mathematical writing!

Always True

For all n, $P(n)$ is true.	For all x, $x^2 \geq 0$.
$P(n)$ is true for every n.	$x^2 \geq 0$ for every x.

Sometimes True

There exists an n such that $P(n)$ is true.	There exists an x such that $5x^2 - 7 = 0$.
$P(n)$ is true for some n.	$5x^2 - 7 = 0$ for some x.
$P(n)$ is true for at least one n.	$5x^2 - 7 = 0$ for at least one x.

All these sentences quantify how often the predicate is true. Specifically, an assertion that a predicate is always true is called a ***universal*** quantification, and an assertion that a predicate is sometimes true is an ***existential*** quantification. Sometimes the English sentences are unclear with respect to quantification:

> “If you can solve any problem we come up with, then you get an A for the course.”

The phrase “you can solve any problem we can come up with” could reasonably be interpreted as either a universal or existential quantification:

> “you can solve *every* problem we come up with,”

or maybe

> “you can solve *at least one* problem we come up with.”

In any case, notice that this quantified phrase appears inside a larger if-then statement. This is quite normal; quantified statements are themselves propositions and can be combined with and, or, implies, etc., just like any other proposition.

2.2.2 More Cryptic Notation

There are symbols to represent universal and existential quantification, just as there are symbols for “and” ($\wedge$), “implies” ($\longrightarrow$), and so forth. In particular, to say that a predicate, P, is true for all values of x in some set, D, one writes:

$$\forall x \in D.\ P(x)$$

The symbol $\forall$ is read "for all", so this whole expression is read "for all x in D, $P(x)$ is true". To say that a predicate $P(x)$ is true for at least one value of x in D, one writes:

$$\exists x \in D.\ P(x)$$

The backward-E is read "there exists". So this expression would be read, "There exists an x in D such that $P(x)$ is true." The symbols $\forall$ and $\exists$ are always followed by a variable and then a predicate, as in the two examples above.

As an example, let Probs be the set of problems we come up with, Solves(x) be the predicate "You can solve problem x", and A be the proposition, "You get an A for the course." Then the two different interpretations of

"If you can solve any problem we come up with, then you get an A for the course."

can be written as follows:

$$(\forall x \in \text{Probs. Solves}(x)) \longrightarrow A,$$

or maybe

$$(\exists x \in \text{Probs. Solves}(x)) \longrightarrow A.$$

2.2.3 Mixing Quantifiers

Many mathematical statements involve several quantifiers. For example, Goldbach's Conjecture states:

"Every even integer greater than 2 is the sum of two primes."

Let's write this more verbosely to make the use of quantification clearer:

For every even integer n greater than 2, there exist primes p and q such that $n = p + q$.

Let Evens be the set of even integers greater than 2, and let Primes be the set of primes. Then we can write Goldbach's Conjecture in logic notation as follows:

$$\underbrace{\forall n \in \text{Evens}}_{\substack{\text{for every even} \\ \text{integer } n \geq 2}} \underbrace{\exists p \in \text{Primes}\ \exists q \in \text{Primes}.}_{\substack{\text{there exist primes} \\ p \text{ and } q \text{ such that}}}\ n = p + q.$$

2.2.4 Order of Quantifiers

Swapping the order of different kinds of quantifiers (existential or universal) changes the meaning of a proposition. For another example, let's return to one of our initial, confusing statements:

"Every American has a dream."

This sentence is ambiguous because the order of quantifiers is unclear. Let A be the set of Americans, let D be the set of dreams, and define the predicate $H(a, d)$ to be "American a has dream d.". Now the sentence could mean there is a single dream that every American shares:

$$\exists d \in D\ \forall a \in A.\ H(a, d)$$

Or it could mean that every American has their personal dream:

$$\forall a \in A\ \exists d \in D.\ H(a, d)$$

Swapping quantifiers in Goldbach's Conjecture creates a patently false statement that every even number ≥ 2 is the sum of *the same* two primes:

$$\underbrace{\exists p \in \text{Primes}\ \exists q \in \text{Primes}}_{\substack{\text{there exist primes} \\ p \text{ and } q \text{ such that}}}\ \underbrace{\forall n \in \text{Evens}.}_{\substack{\text{for every even} \\ \text{integer } n \geq 2}}\ n = p + q.$$

Variables over One Domain

When all the variables in a formula are understood to take values from the same nonempty set, D, it's conventional to omit mention of D. For example, instead of $\forall x \in D\ \exists y \in D.\ Q(x, y)$ we'd write $\forall x \exists y.\ Q(x, y)$. The unnamed nonempty set that x and y range over is called the ***domain*** of the formula.

It's easy to arrange for all the variables to range over one domain. For example, Goldbach's Conjecture could be expressed with all variables ranging over the domain $\mathbb{N}$ as

$$\forall n.\ n \in \text{Evens} \longrightarrow (\exists p \exists q.\ p \in \text{Primes} \wedge q \in \text{Primes} \wedge n = p + q).$$

2.2.5 Negating Quantifiers

There is a simple relationship between the two kinds of quantifiers. The following two sentences mean the same thing:

> It is not the case that everyone likes to snowboard.
>
> There exists someone who does not like to snowboard.

In terms of logic notation, this follows from a general property of predicate formulas:

$$\neg \forall x.\ P(x) \qquad \text{is equivalent to} \qquad \exists x.\ \neg P(x).$$

Similarly, these sentences mean the same thing:

> There does not exist anyone who likes skiing over magma.
>
> Everyone dislikes skiing over magma.

We can express the equivalence in logic notation this way:

$$(\neg \exists x.\ Q(x)) \longleftrightarrow \forall x.\ \neg Q(x). \tag{2.1}$$

The general principle is that *moving a "not" across a quantifier changes the kind of quantifier.*

2.2.6 Validity

A propositional formula is called ***valid*** when it evaluates to T no matter what truth values are assigned to the individual propositional variables. For example, the propositional version of the Distributive Law is that $P \wedge (Q \vee R)$ is equivalent to $(P \wedge Q) \vee (P \wedge R)$. This is the same as saying that

$$[P \wedge (Q \vee R)] \longleftrightarrow [(P \wedge Q) \vee (P \wedge R)]$$

is valid.

The same idea extends to predicate formulas, but to be valid, a formula now must evaluate to true no matter what values its variables may take over any unspecified domain, and no matter what interpretation a predicate variable may be given. For example, we already observed that the rule for negating a quantifier is captured by the valid assertion (2.1).

Another useful example of a valid assertion is

$$\exists x \forall y.\ P(x, y) \longrightarrow \forall y \exists x.\ P(x, y).$$

We could prove this as follows:

Proof. Let D be the domain for the variables and P_0 be some binary predicate[4] on D. We need to show that if $\exists x \in D\, \forall y \in D.\ P_0(x, y)$ holds under this interpretation, then so does $\forall y \in D\, \exists x \in D.\ P_0(x, y)$.

So suppose $\exists x \in D\, \forall y \in D.\ P_0(x, y)$. So some element $x_0 \in D$ has the property that $P_0(x_0, y)$ is true for all $y \in D$. So for every $y \in D$, there is some $x \in D$, namely x_0, such that $P_0(x, y)$ is true. That is, $\forall y \in D \exists x \in D.\ P_0(x, y)$ holds, that is, $\forall y\, \exists x.\ P(x, y)$ holds under this interpretation, as required. □

On the other hand,

$$\forall y \exists x.\ P(x, y) \longrightarrow \exists x \forall y.\ P(x, y).$$

is *not* valid. We can prove this simply by describing an interpretation where the hypothesis, $\forall y \exists x.\ P(x, y)$, is true but the conclusion, $\exists x \forall y.\ P(x, y)$, is not true. For example, let the domain be the integers and $P(x, y)$ mean $x > y$. Then the hypothesis would be true because, given a value, n, for y we could choose the value of x to be $n + 1$, for example. But under this interpretation the conclusion asserts that there is an integer that is bigger than all integers, which is certainly false. An interpetation like this which falsifies an assertion is called a *counter model* to the assertion.

2.3 Mathematical Data Types

We've been assuming that the concepts of sets, sequences and functions are already familiar ones, and we've mentioned them repeatedly. Now we'll do a quick review of the definitions.

[4] That is, a predicate that depends on two variables.

Informally, a ***set*** is a bunch of objects, which are called the ***elements*** of the set. The elements of a set can be just about anything: numbers, points in space, or even other sets. The conventional way to write down a set is to list the elements inside curly-braces. For example, here are some sets:

$$\begin{array}{lcll} B & = & \{\text{Alex}, \text{Tippy}, \text{Shells}, \text{Shadow}\} & \text{dead pets} \\ C & = & \{\text{red}, \text{blue}, \text{yellow}\} & \text{primary colors} \\ D & = & \{\{a, b\}, \{a, c\}, \{b, c\}\} & \text{a set of sets} \end{array}$$

This works fine for small finite sets. Other sets might be defined by indicating how to generate a list of them:

$$A = \{1, 2, 4, 8, 16\} \qquad \text{the powers of 2}$$

The order of elements is not significant, so $\{x, y\}$ and $\{y, x\}$ are the same set written two different ways. Also, any object is, or is not, an element of a given set— there is no notion of an element appearing more than once in a set[5]. So writing $\{x, x\}$ is just indicating the same thing twice, namely, that x is in the set. In particular, $\{x, x\} = \{x\}$.

The expression $e \in S$ asserts that e is an element of set S. For example, $32 \in A$ and blue $\in C$, but Tailspin $\notin B$ —yet.

Sets are simple, flexible, and everywhere. You'll find at least one set mentioned on almost every page in these notes.

2.3.1 Some Popular Sets

Mathematicians have devised special symbols to represent some common sets.

symbol	**set**	**elements**
$\emptyset$	the empty set	none
$\mathbb{N}$	nonnegative integers	$\{0, 1, 2, 3, \ldots\}$
$\mathbb{Z}$	integers	$\{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$
$\mathbb{Q}$	rational numbers	$\frac{1}{2},\ -\frac{5}{3},\ 16,$ etc.
$\mathbb{R}$	real numbers	$\pi,\ e,\ -9,\ \sqrt{2},$ etc.
$\mathbb{C}$	complex numbers	$i,\ \frac{19}{2},\ \sqrt{2} - 2i,$ etc.

A superscript "$+$" restricts a set to its positive elements; for example, $\mathbb{R}^+$ denotes the set of positive real numbers. Similarly, $\mathbb{R}^-$ denotes the set of negative reals.

2.3.2 Comparing and Combining Sets

The expression $S \subseteq T$ indicates that set S is a ***subset*** of set T, which means that every element of S is also an element of T (it could be that $S = T$). For example, $\mathbb{N} \subseteq \mathbb{Z}$ and $\mathbb{Q} \subseteq \mathbb{R}$ (every rational number is a real number), but $\mathbb{C} \not\subseteq \mathbb{Z}$ (not every complex number is an integer).

[5] It's not hard to develop a notion of ***multisets*** in which elements can occur more than once, but multisets are not ordinary sets.

As a memory trick, notice that the $\subseteq$ points to the smaller set, just like a $\leq$ sign points to the smaller number. Actually, this connection goes a little further: there is a symbol $\subset$ analogous to $<$. Thus, $S \subset T$ means that S is a subset of T, but the two are *not* equal. So $A \subseteq A$, but $A \not\subset A$, for every set A.

There are several ways to combine sets. Let's define a couple of sets for use in examples:

$$X ::= \{1, 2, 3\}$$
$$Y ::= \{2, 3, 4\}$$

- The **union** of sets X and Y (denoted $X \cup Y$) contains all elements appearing in X or Y or both. Thus, $X \cup Y = \{1, 2, 3, 4\}$.
- The **intersection** of X and Y (denoted $X \cap Y$) consists of all elements that appear in *both* X and Y. So $X \cap Y = \{2, 3\}$.
- The **difference** of X and Y (denoted $X - Y$) consists of all elements that are in X, but not in Y. Therefore, $X - Y = \{1\}$ and $Y - X = \{4\}$.

Complement of a Set

Sometimes we are focussed on a particular domain, D. Then for any subset, A, of D, we define $\overline{A}$ to be the set of all elements of D *not* in A. That is, $\overline{A} ::= D - A$. The set $\overline{A}$ is called the ***complement*** of A.

For example, when the domain we're working with is the real numbers, the complement of the positive real numbers is the set of negative real numbers together with zero. That is,

$$\overline{\mathbb{R}^+} = \mathbb{R}^- \cup \{0\}\,.$$

Power Set

The collection of all the subsets of a set, A, is called the ***power set***, $\mathcal{P}(A)$, of A. So $B \in \mathcal{P}(A)$ iff $B \subseteq A$. For example, the elements of $\mathcal{P}(\{1, 2\})$ are $\emptyset, \{1\}, \{2\}$ and $\{1, 2\}$.

More generally, if A has n elements, then there are 2^n sets in $\mathcal{P}(A)$. For this reason, some authors use the notation 2^A instead of $\mathcal{P}(A)$.

2.3.3 Sequences

Sets provide one way to group a collection of objects. Another way is in a ***sequence***, which is a list of objects called ***terms*** or ***components***. Short sequences are commonly described by listing the elements between parentheses; for example, (a, b, c) is a sequence with three terms.

While both sets and sequences perform a gathering role, there are several differences.

- The elements of a set are required to be distinct, but terms in a sequence can be the same. Thus, (a, b, a) is a valid sequence of length three, but $\{a, b, a\}$ is a set with two elements —not three.

- The terms in a sequence have a specified order, but the elements of a set do not. For example, (a, b, c) and (a, c, b) are different sequences, but $\{a, b, c\}$ and $\{a, c, b\}$ are the same set.
- The empty set is usually denoted $\emptyset$, and the empty sequence is typically λ.

The product operation is one link between sets and sequences. A ***product*** of sets, $S_1 \times S_2 \times \cdots \times S_n$, is a new set consisting of all sequences where the first component is drawn from S_1, the second from S_2, and so forth. For example, $\mathbb{N} \times \{a, b\}$ is the set of all pairs whose first element is a natural number and whose second element is an a or a b:

$$\mathbb{N} \times \{a, b\} = \{(0, a), (0, b), (1, a), (1, b), (2, a), (2, b), \dots\}$$

A product of n copies of a set S is denoted S^n. For example, $\{0, 1\}^3$ is the set of all 3-bit sequences:

$$\{0, 1\}^3 = \{(0, 0, 0), (0, 0, 1), (0, 1, 0), (0, 1, 1), (1, 0, 0), (1, 0, 1), (1, 1, 0), (1, 1, 1)\}$$

2.3.4 Set Builder Notation

One specialized, but important use of predicates is in ***set builder notation***. We'll often want to talk about sets that cannot be described very well by listing the elements explicitly or by taking unions, intersections, etc. of easily-described sets. Set builder notation often comes to the rescue. The idea is to define a *set* using a *predicate*; in particular, the set consists of all values that make the predicate true. Here are some examples of set builder notation:

$$\begin{aligned} A &::= \{n \in \mathbb{N} \mid n \text{ is a prime and } n = 4k + 1 \text{ for some integer } k\} \\ B &::= \left\{x \in \mathbb{R} \mid x^3 - 3x + 1 > 0\right\} \\ C &::= \left\{a + bi \in \mathbb{C} \mid a^2 + 2b^2 \leq 1\right\} \end{aligned}$$

The set A consists of all nonnegative integers n for which the predicate

$$\text{“}n \text{ is a prime and } n = 4k + 1 \text{ for some integer } k\text{”}$$

is true. Thus, the smallest elements of A are:

$$5, 13, 17, 29, 37, 41, 53, 57, 61, 73, \dots.$$

Trying to indicate the set A by listing these first few elements wouldn't work very well; even after ten terms, the pattern is not obvious! Similarly, the set B consists of all real numbers x for which the predicate

$$x^3 - 3x + 1 > 0$$

is true. In this case, an explicit description of the set B in terms of intervals would require solving a cubic equation. Finally, set C consists of all complex numbers $a + bi$ such that:

$$a^2 + 2b^2 \leq 1$$

This is an oval-shaped region around the origin in the complex plane.

2.3.5 Functions

A ***function*** assigns an element of one set, called the ***domain***, to elements of another set, called the ***codomain***. The notation

$$f : A \to B$$

indicates that f is a function with domain, A, and codomain, B. The familiar notation "$f(a) = b$" indicates that f assigns the element $b \in B$ to a. Here b would be called the *value* of f at *argument* a.

Functions are often defined by formulas as in:

$$f_1(x) ::= \frac{1}{x^2}$$

where x is a real-valued variable, or

$$f_2(y, z) ::= y10yz$$

where y and z range over binary strings, or

$$f_3(x, n) ::= \text{ the pair } (n, x)$$

where n ranges over the nonnegative integers.

Finite functions can be specified by a table that shows the value of the function at each element of the domain, as in the function $f_4(P, Q)$ where P and Q are propositional variables:

P	Q	$f_4(P,Q)$
T	**T**	**T**
T	**F**	**F**
F	**T**	**T**
F	**F**	**T**

Notice that f_4 could also have been described by a formula: $f_4(P, Q) = [P \longrightarrow Q]$.

A function might also be defined by a procedure for computing its value at any element of its domain, or by some other kind of specification. For example, define $f_5(y)$ to be the length of a left to right search of the bits in the binary string y until a `1` appears, so

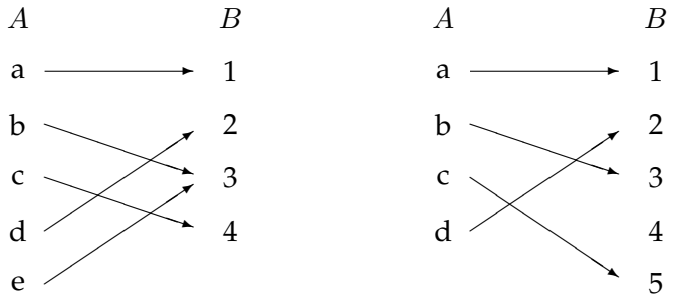
$$\begin{aligned} f_5(0010) &= 3, \\ f_5(100) &= 1, \\ f_5(0000) &\text{ is undefined.} \end{aligned}$$

There are a few properties of functions that will be useful when we take up the topic of counting because they imply certain relations between the sizes of domains and codomains. We say a function $f : A \to B$ is:

- ***total*** if every element of A is assigned to some element of B; otherwise, f is called a ***partial function***,

- ***surjective*** if every element of B is mapped to *at least once*[6],
- ***injective*** if every element of B is mapped to *at most once*, and
- ***bijective*** if f is total, surjective, and injective. In particular, each element of B is mapped to *exactly once*.

We can explain all these properties in terms of a diagram where all the elements of the domain, A, appear in one column (a very long one if A is infinite) and all the elements of the codomain, B, appear in another column, and we draw an arrow from a point a in the first column to a point b in the second column when $f(a) = b$. For example, here are diagrams for two functions:



Here is what the definitions say about such pictures:

- "f is a function" means that every point in the domain column, A, has at *most one arrow out of it*. (If more than one arrow came out of any point in the first column, then f would be a *relation*, but not a function. We'll take up the topic of relations in a couple of weeks.)
- "f is total" means that *every* point in the A column has *at least one arrow out of it*, which really means it has *exactly one arrow out of it* since f is a function.
- "f is surjective" means that *every* point in the codomain column, B, has *at least one arrow into it*.
- "f is injective" means that every point in the codomain column, B, has *at most one arrow into it*.
- "f is bijective" means that *every* point in the A column has exactly one arrow out of it, and *every* point in the B column has exactly one arrow into it.

So in the diagrams above, the function on the left is total and surjective (every element in the A column has an arrow out, and every element in the B column has at least one arrow in), but not injective (element 3 has two arrows going into it). The function on the right is total and injective (every element in the A column has an arrow out, and every element in the B column has at most one arrow in), but not surjective (element 4 has no arrow going into it).

[6] The names "surjective" and "injective" are unmemorable and nondescriptive. Some authors use the term ***onto*** for surjective and *one-to-one* for injective, which are shorter but arguably no more memorable.

Everything about a function is captured by three sets: it domain, its codomain, and the set

$$\{(a, b) \mid f(a) = b\}$$

which is called the *graph* of f. Notice that the graph of f simply describes where the arrows go in a diagram for f.

The graph of f does not determine by itself whether f is total or surjective; we also need to know what the domain is to determine if f is total, and we need to know the codomain to tell if it's surjective. For example, a function defined by the formula $1/x^2$, is total if its domain is $\mathbb{R}^+$ but partial if its domain is some set of real numbers including 0. It is bijective if its domain and codomain are both $\mathbb{R}^+$, but neither injective nor surjective if its domain and codomain are both $\mathbb{R}$.

Surjections and injections imply certain size relationships between domains and codomains. If A is a finite set, we let $|A|$ be its size, that is, the number of elements in A.

Lemma (Mapping Rule).

- *If $f : A \to B$ is surjective, then $|A| \geq |B|$.*
- *If $f : A \to B$ is total and injective, then $|A| \leq |B|$.*
- *If $f : A \to B$ is bijective, then $|A| = |B|$.*

It's often useful to find the set of values a function takes when applied to the elements in *a set* of arguments. So if $f : A \to B$, and $A' \subseteq A$, we define

$$\widehat{f}(A') ::= \left\{b \in B \mid f(a') = b \text{ for some } a' \in A'\right\}.$$

For example, if we let $[r, s]$ denote the interval from r to s on the real line, then $\widehat{f_1}([1, 2]) = [1/4, 1]$.

For another example, let's take the "search for a `1`" function, f_5. If we let X be the set of binary words which start with an even number of `0`'s followed by a `1`, then $\widehat{f_5}(X)$ would be the even nonnegative integers.

Applying $\widehat{f}$ to a set, A', of arguments is referred to as "applying f pointwise to A'." The distinction between f and $\widehat{f}$ is kind of picky, and it's common practice to omit the hat on $\widehat{f}$ and just write f. We'll do this too, since it's usually easy to figure out whether f is being applied to a single argument or pointwise to a set of them. But technically, f and $\widehat{f}$ are quite different functions: for example, the domain of $\widehat{f}$ is not A, but the set of subsets of A, that is, $\mathcal{P}(A)$, and the codomain of $\widehat{f}$ is $\mathcal{P}(B)$.

The set of values that arise from applying f to all possible arguments is called the *range* of f. That is

$$\text{range}\,(f) ::= \widehat{f}(\text{domain}\,(f)).$$

Some authors refer to the codomain as the range of a function, but they shouldn't: the distinction between the range and codomain is important. The range and codomain of f are the same only when f is surjective.

2.4 Does All This Really Work?

So this is where mainstream mathematics stands today: there is a handful of axioms from which everything else in mathematics can be logically derived. This sounds like a rosy situation, but there are several dark clouds, suggesting that the essence of truth in mathematics is not completely resolved.

- The ZFC axioms weren't etched in stone by God. Instead, they were mostly made up by some guy named Zermelo. Probably some days he forgot his house keys.
- No one knows whether the ZFC axioms are logically consistent; there is some possibility that one person might prove a proposition P and another might prove the proposition $\neg P$. Then Math would be broken. This sounds like a crazy situation, but it has happened before. At the beginning of the 20th century, the logician Gotlob Frege made an initial attempt to axiomatize set theory using a few very plausible axioms. Several mathematicians— most famously Bertrand Russell[7]— discovered that Frege's axioms actually *were* self-contradictory!
- While the ZFC axioms largely generate the mathematics everyone wants— where $3 + 3 = 6$ and other basic facts are true— they also imply some disturbing conclusions. For example, the Banach-Tarski Theorem says that a ball can be divided into six pieces and then the pieces can be rearranged to give *two* balls, each the same size as the original!
- In the 1930's, Gödel proved that, assuming the ZFC axioms *are* consistent, then they are not *complete*: that is, there exist propositions that are true, but do not logically follow from the axioms. As a matter of fact, the proposition that ZFC is consistent (which is not too hard to express as a formula about sets) is an example of a true proposition that cannot be proved.

  There seems to be no way out of this disturbing situation; simply adding more axioms does not eliminate the problem.

These problems will not trouble us in 6.042, but they are interesting to think about!

[7] Bertrand Russell was a Mathematician/Logician at Oxford University at the turn of the Twentieth Century. He reported that when he felt too old to do Mathematics, he began to study and write about Philosophy, and when he was no longer smart enough to do Philosophy, he began writing about Politics. He was jailed as a conscientious objector during World War I. He won two Nobel Prizes— one Literature Prize and one Peace Prize.

Russell's Paradox

Reasoning naively about sets quickly leads to the following contradiction— known as Russell's Paradox:

Let S be a variable ranging over all sets, and define

$$W ::= \{S \mid S \notin S\}.$$

So by definition,

$$S \in W \text{ iff } S \notin S,$$

for every set S. In particular, we can let S be W, and obtain the contradictory result that

$$W \in W \text{ iff } W \notin W.$$

This paradox revealed a fatal flaw in Frege's initial effort to axiomatize set theory. This was an astonishing blow to efforts to provide an axiomatic foundation for Mathematics.

But a way out was clear at the time: *we cannot assume that W is a set*. So the step in the proof where we let S be W is invalid, because S ranges over sets, and W is not a set.

But denying that W is a set means we must reject the axiom that every mathematically well-defined collection of elements is actually a set.

The problem faced by Logicians was how to axiomatize rules determining which well-defined collections are sets. Russell and his colleague Whitehead immediately went to work on this problem and spent a dozen years developing a huge new axiom system in an even huger monograph called *Principia Mathematica*.

The modern ZFC axioms for set theory are much simpler than the Russell/Whitehead system and are close to Frege's original axioms. They specify that sets must be built up from "simpler" sets in certain standard ways. In particular, no set is ever a member of itself. So the modern resolution of Russell's paradox goes as follows: since $S \notin S$ for all sets S, it follows that W, defined above, contains every set. So W can't be a set or it would be a member of itself.

These issues rarely come up in mainstream Mathematics. And they don't come up at all in Computer Science, where the focus is generally on "countable," and often just finite, sets. In practice, only Logicians and Set Theorists have to worry about collections that are too big to be sets.

2.5 In-Class Problems Week 2, Mon.

Problem 2.5.1. [8] A set of propositions is *consistent* if there is a possible situation in which they are all true. This problem examines whether the following specifications are consistent.

1. If the file system is not locked, then
   (a) new messages will be queued.
   (b) new messages will be sent to the messages buffer.
   (c) the system is functioning normally, and conversely, if the system is functioning normally, then the file system is not locked.
2. If new messages are not queued, then they will be sent to the messages buffer.
3. New messages will not be sent to the message buffer.

(a) Begin by translating the parts of the specification into propositional formulas using four propositional variables:

$$\begin{aligned} L &::= \text{file system locked,} \\ Q &::= \text{new messages are queued,} \\ B &::= \text{new messages are sent to the message buffer,} \\ N &::= \text{system functioning normally.} \end{aligned}$$

Solution. The translations of the specifications are:

$$\begin{aligned} \neg L &\longrightarrow Q && \text{(Spec. 1.(a))} \\ \neg L &\longrightarrow B && \text{(Spec. 1.(b))} \\ \neg L &\longleftrightarrow N && \text{(Spec. 1.(c))} \\ \neg Q &\longrightarrow B && \text{(Spec. 2.)} \\ \neg B & && \text{(Spec. 3.)} \end{aligned}$$

■

(b) To be precise, the specification is consistent if there is a way to assign a truth value to each of the variables L, Q, B, N, so that every one of the propositional formulas from the previous part are true. Explain how to use a *truth table* to determine whether the specification is consistent.

Solution. We can construct a truth table with sixteen lines—one for each way of assigning truth values to the four variables L, N, Q, and B. For each line, we could record the truth values of these five statements above.

If all five statements are true for some assignment of truth values to the variables, then the system is consistent. If for every one of the sixteen possible truth assignments, at least one of the five

[8] From Rosen, 5th edition, Exercise 1.1.36

statements is false, then the system is inconsistent. Carrying out the calculation shows that there is a unique assignment of True/False values to L, N, Q, and B (see below) that satisfies all the specifications. ■

(c) Use *simple reasoning by cases* to find a truth assignment that confirms that this system specification is consistent. Explain why there is only one such assignment.

Solution. We can avoid the full truthtable calculation if we reason by cases.

Case 1 (B is True): Then the last formula, (Spec. 3.), is false, and the whole specification is false.

Case 2 (B is False): Now (Spec. 2.) and (Spec. 1.(b)) can be true only if Q and L are true. Since L is true, (Spec. 1.(c)) can be true only if N is false. Thus, we have deduced that in order to be consistent, we must have

$$\begin{aligned} L &= \text{True} \\ N &= \text{False} \\ Q &= \text{True} \\ B &= \text{False.} \end{aligned}$$

From the way this assignment was constructed, we know it ensures that formulas from (Spec. 1.(b)) on are true. So all that remains is to check formula (Spec. 1.(a)), and indeed it is also true under this assignment.

So the system is consistent, and this is the only assignment that will satisfy it. ■

Problem 2.5.2. If we raise an irrational number to an irrational power, can the result be rational? Show that it can by considering $\sqrt{2}^{\sqrt{2}}$ and arguing by cases.

Solution. We want to find irrational numbers a, b such that a^b is rational. We argue by cases.

Case 1: [$\sqrt{2}^{\sqrt{2}}$ is rational]. Let $a = b = \sqrt{2}$. a and b are irrational since $\sqrt{2}$ is irrational as we know. Also, a^b is rational by case hypothesis. So we have found the required a and b in this case.

Case 2: [$\sqrt{2}^{\sqrt{2}}$ is irrational]. Let $a = \sqrt{2}^{\sqrt{2}}$ and $b = \sqrt{2}$. Then a is irrational by case hypothesis, we know b is irrational, and

$$a^b = \left(\sqrt{2}^{\sqrt{2}}\right)^{\sqrt{2}} = \sqrt{2}^{\sqrt{2}\cdot\sqrt{2}} = \sqrt{2}^2 = 2,$$

which is rational. So we have found the required a and b in this case also.

So in any case, there will be irrational a, b such that a^b is rational. ■

Problem 2.5.3. Generalize the proof from lecture (reproduced below) that $\sqrt{2}$ is irrational, for example, how about $\sqrt[3]{2}$? Remember that an irrational number is a number that cannot be expressed as a ratio of two integers.

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

Proof. The proof is by contradiction. Assume for purpose of contradiction that $\sqrt{2}$ is rational.

Then we can write $\sqrt{2} = m/n$ where m and n are integers and the fraction is in lowest terms. Squaring both sides gives $2 = m^2/n^2$, so $2n^2 = m^2$. This implies that m^2 is even, and hence that m is even; that is, m is a multiple of 2. But that means m^2 is actually a multiple of 4, say $m^2 = 4k$.

Now we have $2n^2 = m^2 = 4k$, so $n^2 = 2k$. So n^2 is even, and hence n is even. But since m and n are both even, the fraction m/n is not in lowest terms, a contradiction. □

Solution. We prove that for any $n > 1$, $\sqrt[n]{2}$ is irrational by contradiction.

Assume that $\sqrt[n]{2}$ is rational. Under this assumption, there exist integers a and b with $\sqrt[n]{2} = a/b$, where a and b have no common factors (so that the fraction a/b is in lowest terms). Now we prove that a and b are both even, which contradicts the existence of a, b.

$$\begin{aligned} \sqrt[n]{2} &= \frac{a}{b} \\ 2 &= \frac{a^n}{b^n} \\ 2b^n &= a^n. \end{aligned}$$

The lefthand side of the last equation is even, so a^n is even. This implies that a is even as well (see below for justification).

In particular, $a = 2c$ for some integer c. Thus,

$$\begin{aligned} 2b^n &= (2c)^n = 2^n c^n, \\ b^n &= 2^{n-1} c^n. \end{aligned}$$

Since $n - 1 > 0$, the righthand side of the last equation is an even number, so b^n is even. But this implies that b must be even as well, contradicting the fact that a/b is in lowest terms.

Now we justify the claim that if a^n is even, so is a.

There is a simple proof by contradiction: suppose a was not even, *i.e.*, a is odd. It's a familiar (and easily verified) fact that the product of *any* number of odd numbers is odd, so a^n would also be odd, contradicting the fact that a^n is even.

But more generally for *any* integers $m, k > 0$, if m^k is divisible by a prime number, p, then m must be divisible by p. This follows from the factorization of an integer into primes (which we'll discuss further in a coming lecture): the primes in the factorization of m^k are precisely the primes in the factorization of m repeated k times, so if there is a p in the factorization of m^k it must be one of k copies of a p in the factorization of m.

[Optional]

Here's a somewhat broader generalization of the proof that $\sqrt{2}$ is irrational.

Lemma. *Let the coefficients of the polynomial $a_0 + a_1 x + a_2 x^2 + \cdots + a_{n-1} x^{n-1} + x^n$ be integers. Then any real root of the polynomial is either integral or irrational.*

Notice that this Lemma directly implies that $\sqrt{2}$ is irrational: $x^2 - 2$ has no integer root because 2 is not a perfect square, so the real roots, namely, $\pm\sqrt{2}$, must be irrational. Similarly, if $m \in \mathbb{N}$ is not the kth power of an integer, then $x^k - m$ has no integer roots, so $\sqrt[k]{m}$ will be irrational.

Proof. Let $r \in \mathbb{R}$ be a root of the polynomial, that is,

$$a_0 + a_1 r + a_2 r^2 + \cdots + a_{n-1} r^{n-1} + r^n = 0.$$

There are three cases for r: it is either integral, irrational, or the ratio of two integers with no common divisors where the denominator is greater than one. We want to eliminate the last case.

So assume to the contrary that $r = s/t$ for integers s and t which have no common divisors and such that $t > 1$. Now t must have a prime divisor, p.

Substituting s/t for r and multiplying both sides of the above equation by t^n yields:

$$a_0 t^n + a_1 s t^{n-1} + a_2 s^2 t^{n-2} + \cdots + a_{n-1} s^{n-1} t + s^n = 0, \tag{2.2}$$

$$a_0 t^n + a_1 s t^{n-1} + a_2 s^2 t^{n-2} + \ldots + s^{n-1} t = -s^n. \tag{2.3}$$

Then p divides each term of the lefthand side of equation (2.3), and therefore also divides the righthand side, viz., $p \mid -s^n$. But this means that p must also divide s. So p is a common divisor of s and t, contradicting our choice of s and t.

□

■

2.6 In-Class Problems Week 2, Wed.

Problem 2.6.1. For each of the logical formulas, indicate whether or not it is true when the domain of discourse is $\mathbb{N}$ (the natural numbers 0, 1, 2, ...), $\mathbb{Z}$ (the integers), $\mathbb{Q}$ (the rationals), $\mathbb{R}$ (the real numbers), and $\mathbb{C}$ (the complex numbers).

$$\begin{array}{rrrcl} & \exists x & (x^2 & = & 2) \\ \forall x & \exists y & (x^2 & = & y) \\ \forall y & \exists x & (x^2 & = & y) \\ \forall x \neq 0 & \exists y & (xy & = & 1) \\ \exists x & \exists y & (x+2y & = & 2) \wedge (2x+4y=5) \end{array}$$

Solution.

Statement	$\mathbb{N}$	$\mathbb{Z}$	$\mathbb{Q}$	$\mathbb{R}$	$\mathbb{C}$
$\exists x\ (x^2 = 2)$	f	f	f	$t\ (x = \sqrt{2})$	t
$\forall x \exists y\ (x^2 = y)$	t	t	t	$t\ (y = x^2)$	t
$\forall y \exists x\ (x^2 = y)$	f	f	f	f (take $y < 0$)	t
$\forall x \neq 0 \exists y\ (xy = 1)$	f	f	t	$t\ (y = 1/x)$	t
$\exists x \exists y\ (x + 2y = 2) \wedge (2x + 4y = 5)$	f	f	f	f	f

■

Problem 2.6.2. The goal of this problem is to translate some assertions about binary strings into logic notation. The domain of discourse is the set of all finite-length binary strings: λ, 0, 1, 00, 01, 10, 11, 000, 001, ... (Here λ denotes the empty string.) In your translations, you may use all the ordinary logic symbols (including `=`), variables, and the binary symbols `0`, `1` denoting 0, 1.

A string like $\texttt{01}x\texttt{0}y$ of binary symbols and variables denotes the *concatenation* of the symbols and the binary strings represented by the variables. For example, if the value of x is `011` and the value of y is `1111`, then the value of $\texttt{01}x\texttt{0}y$ is the binary string `0101101111`.

Here are some examples of formulas and their English translations. Names for these predicates are listed in the third column so that you can reuse them in your solutions (as we do in the definition of the predicate NO-1S below).

Meaning	Formula	Name
x is a prefix of y	$\exists z\ (xz = y)$	PREFIX(x, y)
x is a substring of y	$\exists u \exists v\ (uxv = y)$	SUBSTRING(x, y)
x is empty or a string of 0's	$\neg$SUBSTRING$(\texttt{1}, x)$	NO-1S(x)

(a) x consists of three copies of some string.

Solution. $\exists y\ (x = yyy)$ ■

(b) x is an even-length string of 0's.

Solution. NO-1S$(x) \wedge \exists y\ (x = yy)$ ■

(c) x does not contain both a 0 and a 1.

Solution. $\neg[\text{SUBSTRING}(0, x) \wedge \text{SUBSTRING}(1, x)]$

■

(d) x is the binary representation of $2^k + 1$ for some integer $k \geq 0$.

Solution. $(x = 10) \vee (\exists y\ (x = 1y1 \wedge \text{NO-1S}(y)))$ ■

(e) An elegant, slightly trickier way to define NO-1S(x) is:

$$\text{PREFIX}(x, 0x). \qquad (*)$$

Explain why (*) is true only when x is a string of 0's.

Solution. Prefixing x with 0 rightshifts all the bits. So the nth symbol of x shifts into the $(n+1)$st symbol of $0x$. Now for x to be a prefix of $0x$, the $n+1$st symbol of $0x$ must match the $(n+1)$st symbol of x. So if x satisfies (*), the nth and $(n+1)$st symbols of x must match. This holds for all $n > 0$ up to the length of x, that is, *all* the symbols of x must be the same. In addition, if $x \neq \lambda$, it must start with 0. Therefore, if x satisfies (*), all its symbols must be 0's.

Note that it's easy to see, conversely, that if $x = \lambda$ or x is all 0's, then of course it satisfies (*). ■

Problem 2.6.3. A media tycoon has an idea for an all-news television network called LNN: The Logic News Network. Each segment will begin with a definition of the domain of discourse and a few predicates. The day's happenings can then be communicated concisely in logic notation. For example, a broadcast might begin as follows:

> "THIS IS LNN. The domain of discourse is $\{\text{Bill}, \text{Monica}, \text{Ken}, \text{Linda}, \text{Betty}\}$. Let $D(x)$ be a predicate that is true if x is deceitful. Let $L(x, y)$ be a predicate that is true if x likes y. Let $G(x, y)$ be a predicate that is true if x gave gifts to y."

Complete the broadcast by translating the following statements into logic notation.

(a) If neither Monica nor Linda is deceitful, then Bill and Monica like each other.

Solution.

$$(\neg(D(\text{Monica}) \vee D(\text{Linda}))) \longrightarrow (L(\text{Bill}, \text{Monica}) \wedge L(\text{Monica}, \text{Bill}))$$

■

(b) Everyone except for Ken likes Betty, and no one except Linda likes Ken.

Solution.

$$\forall x\ (x = \text{Ken} \wedge \neg L(x, \text{Betty})) \vee (x \neq \text{Ken} \wedge L(x, \text{Betty})) \wedge$$
$$\forall x\ (x = \text{Linda} \wedge L(x, \text{Ken})) \vee (x \neq \text{Linda} \wedge \neg L(x, \text{Ken}))$$

■

(c) If Ken is not deceitful, then Bill gave gifts to Monica, and Monica gave gifts to someone.

Solution.

$$\neg D(\text{Ken}) \longrightarrow (G(\text{Bill}, \text{Monica}) \wedge \exists\, x G(\text{Monica}, x))$$

■

(d) Everyone likes someone and dislikes someone else.

Solution.

$$\forall x \exists y \exists z\ (y \neq z) \wedge L(x, y) \wedge \neg L(x, z)$$

■

(e) How could you express "Everyone except for Ken likes Betty" using just propositional connectives *without* using any quantifiers ($\forall, \exists$)? Can you generalize to explain how *any* logical formula over this domain of discourse can be expressed without quantifiers? How big would the formula in the previous part be if it was expressed this way?

Solution.

$$L(\text{Bill}, \text{Betty}) \wedge L(\text{Monica}, \text{Betty}) \wedge L(\text{Linda}, \text{Betty}) \wedge L(\text{Betty}, \text{Betty}) \wedge \neg L(\text{Ken}, \text{Betty})$$

In general, quantifiers can be eliminated by treating $\forall x\ P(x)$ as an abbreviation for

$$P(\text{Bill}) \wedge P(\text{Monica}) \wedge P(\text{Ken}) \wedge P(\text{Linda}) \wedge P(\text{Betty}),$$

and $\exists x\ P(x)$ as an abbreviation for

$$P(\text{Bill}) \vee P(\text{Monica}) \vee P(\text{Ken}) \vee P(\text{Linda}) \vee P(\text{Betty}).$$

Expanded this way, the three-quantifier formula of the previous part would expand by a factor of $5 \times 5 \times 5 = 125$. So using quantifiers can pay off even when they are not strictly necessary. ■

Problem 2.6.4. **(a)** Describe a counter-model demonstrating that

$$(\forall x \exists y.\ P(x, y)) \longrightarrow \forall z.\ P(z, z)$$

is not valid.

Solution. Let $P(x, y)$ mean $x \neq y$. Then the conclusion $\forall z.\ z \neq z$ is always false, but in any domain with two or more elements, the hypothesis is true. ■

(b) Explain why

$$(\forall z.\ P(z,z)) \longrightarrow \forall x \exists y.\ P(x,y) \tag{2.4}$$

is valid.

Solution. *Proof.* Assume

$$\forall z.\ P(z,z) \tag{2.5}$$

is true for some domain and interpretation of the predicate P. We want to show that

$$\forall x \exists y.\ P(x,y) \tag{2.6}$$

also holds.

So let c be an element of the domain. Then $P(c,c)$ holds by assumption (2.5). So there is a y, namely $y = c$ such that $P(c,y)$ holds. That is, $\exists y.\ P(c,y)$ is true. But c could have been any element in the domain, so (by *Universal Generalization*), we conclude that (2.6) holds. □

■

Proof of a Validity about Quantifiers

Lemma. *The formula*

$$\forall x\ [P(x) \wedge Q(x)] \longrightarrow [(\forall y.\ P(y)) \wedge \forall z.\ Q(z)]$$

is valid.

Proof. Assume

$$\forall x\ [Q(x) \wedge P(x)] \tag{2.7}$$

holds (for some domain and interpretation of P and Q). We want to prove that

$$(\forall y.\ Q(y)) \wedge \forall z.\ P(z) \tag{2.8}$$

holds.

To do this, let c be an element of the domain. Then $Q(c) \wedge P(c)$ holds by hypothesis (2.7). In particular, $Q(c)$ holds. But since c could have been any element of the domain, we conclude (by Universal Generalization) that

$$\forall y.\ Q(y) \tag{2.9}$$

holds. We conclude similarly that

$$\forall z.\ P(z) \tag{2.10}$$

holds. Now (2.9) and (2.10) immediately yield (2.8), as required. □

2.7 In-Class Problems Week 2, Fri.

Problem 2.7.1. Subset take-away[9] is a two player game involving a fixed finite set, A. Players alternately choose nonempty subsets of A with the conditions that a player may not choose

- the whole set A, or
- any set containing a set that was named earlier.

The first player who is unable to move loses the game.

For example, if A is $\{1\}$, then there are no legal moves and the second player wins. If A is $\{1, 2\}$, then the only legal moves are $\{1\}$ and $\{2\}$. Each is a good reply to the other, and so once again the second player wins.

The first interesting case is when A has three elements. This time, if the first player picks a subset with one element, the second player picks the subset with the other two elements. If the first player picks a subset with two elements, the second player picks the subset whose sole member is the third element. Both cases produce positions equivalent to the starting position when A has two elements, and thus leads to a win for the second player.

Verify that when A has four elements, the second player still has a winning strategy.[10]

Solution. There are way too many cases to work out by hand if we tried to list all possible games. But the elements of A all behave the same, so we can cut to a small number of cases using the fact that permuting around the elements of A in any game yields another possible game. We can do this by not mentioning specific elements of A, but instead using the *variables* a, b, c, d whose values will be the four elements of A.

We consider two cases for the move of the Player 1 when the game starts:

1. Player 1 chooses a one element or a three element subset. Then Player 2 should choose the complement of Player one's choice. The game then becomes the same as playing the $n = 3$ game on the three element set chosen in this first round, where we know Player 2 has a winnging strategy.
2. Player 1 chooses a subset of 2 elements. Let a, b be these elements, that is, the first move is $\{a, b\}$. Player 2 should choose the complement, $\{c, d\}$, of Player 1's choice. We then have the following subcases:
   (a) Player 1's second move is a one element subset, $\{a\}$. Player 2 should choose $\{b\}$. The game is then reduced to the two element game on $\{c, d\}$ where Player 2 has a winnging strategy.
   (b) Player 1's second move is a two element subset, $\{a, c\}$. Player 2 should choose its complement, $\{b, d\}$. This leads to two subsubcases:

[9] From Christenson & Tilford, *David Gale's Subset Takeaway Game, American Mathematical Monthly, Oct. 1997*

[10] David Gale worked out some of the properties of this game and conjectured that the second player wins the game for any set A. This remains an open problem.

i. Player 1's third move is one the remaining sets of size two, $\{a, d\}$. Player 2 should choose its complement, $\{b, c\}$. The remaining possible moves are the four sets of size 1, where the Player 2 clearly wins after two more rounds.

ii. Player 1's third move is a one element set, $\{a\}$. Player 2 should choose $\{b\}$. The game is then reduced to the case two element game on $\{c, d\}$ where Player 2 has a winnging strategy.

So in all cases, Player 2 has a winning strategy in the Gale game for $n = 4$. ■

Problem 2.7.2. **(a)** Define a bijection between the positive integers and all integers.

Solution. One such bijection is defined by lining up the positive integers and all the integers as follows:

$$\begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & \dots \\ 0 & 1 & -1 & 2 & -2 & 3 & -3 & 4 & \dots \end{array}$$

We can define this bijection, $f : \mathbb{Z}^+ \to \mathbb{Z}$, by the rule

$$f(n) = \begin{cases} n/2, & \text{if } n \text{ is even;} \\ -(n-1)/2, & \text{if } n \text{ is odd.} \end{cases}$$

■

(b) Define a bijection between the positive integers and $\mathbb{Z} \times \mathbb{Z}$, the set of all the ordered pairs of integers:

$$\begin{array}{l} (0,0), (0,1), (0,-1), (0,2), (0,-2), (0,3), (0,-3), \dots \\ (1,0), (1,1), (1,-1), (1,2), (1,-2), (1,3), (1,-3), \dots \\ (-1,0), (-1,1), (-1,-1), (-1,2), (-1,-2), (-1,3), (-1,-3), \dots, \\ (2,0), (2,1), (2,-1), (2,2), (2,-2), (2,3), (2,-3), \dots \\ (-2,0), (-2,1), (-2,-1), (-2,2), (-2,-2), (-2,3), (-2,-3), \dots \\ \vdots \end{array}$$

Solution. There are $4k$ pairs of integers (i, j) such that $|i| + |j| = k \geq 0$. Let L_k be a list of these pairs in some simple order, for example,

$$\begin{array}{rcl} L_3 & ::= & (3,0), \\ & & (2,1), (2,-1), \\ & & (1,2), (1,-2), \\ & & (0,3), (0,-3), \\ & & (-1,2), (-1,-2), \\ & & (-2,1), (-2,-1), \\ & & (-3,0). \end{array}$$

Then form the infinite list consisting of the elements of L_0 then the elements of L_1, then $L_2, \ldots$:

$$\begin{array}{l}(0,0),\\ (1,0),(0,1),(0,-1),(-1,0),\\ (2,0),(1,1),(1,-1),(0,2),(0,-2),(-1,1),(-1,-1),(-2,0),\\ (3,0),(2,1),(2,-1),(1,2),(1,-2),\ldots,(-3,0),\\ (4.0),(3,1),\ldots\end{array}$$

Now the bijection will be the function, f, where $f(n)$ is defined to be the nth pair in the list, so $f(1) ::= (0,0), f(2) ::= (1,0),\ldots$. ■

(c) Conclude that the set of positive integers is the same size as the set of all rational numbers.

Solution. For infinite sets, "same size" means there is a bijection between them. One way to construct such a bijection is to take the list of all pairs, (i,j), of integers above, delete the pairs with $j = 0$,

$$\begin{array}{l}(0,1),(0,-1),\\ (1,1),(1,-1),(0,2),(0,-2),(-1,1),(-1,-1),\\ (2,1),(2,-1),(1,2),(1,-2),\ldots,\\ (3,1),\ldots,\end{array}$$

replace each remaining pair by the rational number i/j,

$$\begin{array}{l}0/1, 0/-1,\\ 1/1, 1/-1, 0/2, 0/-2, -1/1, -1/-1,\\ 2/1, 2/-1, 1/2, 1/-2, \ldots,\\ 3/1, \ldots,\end{array}$$

and, from left to right, delete all the occurrences of numbers that are already in the list.

$$0, 1, -1, 2, -2, 1/2, -1/2, 3, \ldots$$

Then define a bijection, r, where $r(n)$ is defined to be the nth pair in the remaining list, so

$$r(1) = 0, r(2) = 1, r(3) = -1, r(4) = 2, r(5) = -2, r(6) = 1/2, r(7) = -(1/2), r(8) = 3, \ldots.$$

■

Problem 2.7.3. Consider procedures in your favorite programming language — say, C++, Java, or Scheme — that can take a string of ASCII characters as an argument. Call a procedure of this kind a *string procedure*.

A set, $\mathcal{R}$, of ASCII strings is called *recognizable* (using your programming language) if there is a string procedure that returns the integer 1 when it is applied to a string that is in $\mathcal{R}$, and does not return 1 (it need not return any value at all) when it is applied to a string that is not in $\mathcal{R}$.

Now any program can be represented by a string of ASCII characters. So if, in particular, a string s is the definition of a string procedure, let's refer to that procedure as P_s. In case string s is *not* a definition of a string procedure, let's define P_s to be some fixed string procedure that never returns

a value no matter what string it's applied to. In this way, every string, s, gets associated with a (frequently useless) string procedure, P_s.

We're going to define a set, $\mathcal{V}$, of strings in a way similar to the set in Russell's Paradox. Namely, let $\mathcal{V}$ be the set of strings, s, such that P_s applied to argument s does *not* return the integer 1. So for every ASCII string, s, we have by definition

$$s \in \mathcal{V} \quad \text{iff} \quad P_s \text{ applied to } s \text{ does not return } 1. \tag{2.11}$$

Prove that $\mathcal{V}$ is not recognizable.

Solution. Assume to the contrary that $\mathcal{V}$ was recognizable by some string procedure. So letting r be the string defining this procedure, we have

$$s \in \mathcal{V} \quad \text{iff} \quad P_r \text{ applied to } s \text{ returns } 1. \tag{2.12}$$

Combining (2.11) and (2.12), we have that for every string, s,

$$P_s \text{ applied to } s \text{ does not return } 1 \quad \text{iff} \quad P_r \text{ applied to } s \text{ returns } 1. \tag{2.13}$$

Letting s equal r in (2.13), we reach the contradiction

$$P_r \text{ applied to } r \text{ does not return } 1 \quad \text{iff} \quad P_r \text{ applied to } r \text{ returns } 1.$$

So the assumption that $\mathcal{V}$ is recognizable must be false. ■

2.8 Problem Set 1

Problem 2.8.1. Show that $\log_7 n$ is either an integer or irrational, where n is a positive integer. Use whatever familiar facts about integers and primes you need, but explicitly state such facts. (This problem will be graded on the clarity and simplicity of your proof. If you can't figure out how to prove it, ask the staff for help and they'll tell you how.)

Solution. The statement to be proved is equivalent to the assertion that, for all positive integers, n, if $\log_7 n$ is rational, then it is an integer.

So we'll assume

$$\log_7 n = \frac{i}{j} \tag{2.14}$$

for some integers, i, j. Now if $i = 0$, then $\log_7 n$ is the integer 0, so we can asssume $i > 0$ and $n > 1$. Using the fact that $\log_7 x$ is positive for $x > 1$, we can conclude that $j > 0$, also.

Now, raising 7 to the power $\log_7 n$, we have from (2.14)

$$n = 7^{\log_7 n} = 7^{i/j}.$$

Then, taking jth powers,

$$n^j = (7^{i/j})^j = 7^i. \tag{2.15}$$

Since $i, j > 0$, both sides of equation (2.15) are integers. Also, since the only prime dividing the righthand of (2.15) is 7, *the fact that integers factor uniquely into primes* implies that the only prime factor of n^j, and hence the only prime factor of n, is 7. This means that n can only be a nonnegative power of 7, so $\log_7 n$ must be a nonnegative integer. ■

Problem 2.8.2. There are two kinds of people in the land of Paradox: liars and truth-tellers. As you might expect, liars claim a statement is true iff it is actually false, while truth-tellers do the opposite. There is only one direction to go to get out of the land of Paradox, either North or South. Bill is a resident of Paradox and you say to him "$L \oplus N$," where L means "You, Bill, are a liar" and N means "the direction out of Paradox is North." Draw a truth table showing that North is the direction out of Paradox iff Bill claims what you told him is true. (Remember, you start off with no idea whether Bill is a liar or truth-teller.)

Solution. Let R be Bill's **T**/**F** response to the proposition $L \oplus N$. There are four possible truth values for L and N, and these determine the values of $L \oplus N$ and R. We want to show that R is equivalent to N. The following truth table confirms this:

L	N	$L \oplus N$	R
T	**T**	**F**	**T**
T	**F**	**T**	**F**
F	**T**	**T**	**T**
F	**F**	**F**	**F**

Note that after Bill responds and the direction out of Paradox becomes known, you still have no idea whether Bill is a liar or a truth-teller. This make sense: Bill provides you a single **T**/**F** response, that is, one bit of information. Knowing the direction you should go *and* whether Bill is a liar would mean you learned two bits of information. Later in the course, we'll indicate how to make this reasoning about "bits" into a rigorous argument. ■

Problem 2.8.3. Describe a simple recursive procedure which, given a positive integer argument, n, produces a truth table whose rows are all the assignments of truth values to n propositional variables. For example, for $n = 2$, the table might look like:

T	**T**
T	**F**
F	**T**
F	**F**

Your description can be in English, or a simple program in some familiar language (say Scheme or Java), but if you do write a program, be sure to include some sample output.

Solution. Start with an $n = 1$ table, namely a one-column table whose first row consists of a **T** entry and second row an **F** entry. Build the $n + 1$ table recursively by taking an n table and attaching a **T** at the beginning of every row, then taking another n table and attaching a **F** at the beginning of every row, and finally placing the first table above the second table.

Here's a Scheme program that carries out this procedure:

```
(define (truth-values n)
   (if (= n 1) '((T) (F))
       (let ((table (truth-values (- n 1))))
            (append
               (map (lambda (row) (cons 'T row)) table)
               (map (lambda (row) (cons 'F row)) table)))))
(truth-values 3)
;Value 17:  ((t t t) (t t f) (t f t) (t f f)
             (f t t) (f t f) (f f t) (f f f))
```

■

Problem 2.8.4. Translate the following sentence into a predicate formula:

> There is a student who has emailed exactly two other people in the class, besides possibly herself.

The domain of discourse should be the set of students in the class; in addition, the only predicates that you may use are

- equality, and
- $E(x, y)$, meaning that "x has sent e-mail to y."

Solution. A good way to begin tackling this problem is by working "top-down" to translate the successive parts of the sentence. First of all, our formula must be of the form

$$\exists x.P(x)$$

where $P(x)$ should be a formula that says that "student x has e-mailed exactly two other people in the class, besides possibly herself".

One way to write $P(x)$ is to give names, say y and z, to the two students whom x has emailed. So we translate $P(x)$ as "besides x, there are two students, y and z, and ...":

$$\exists y, z.\ x \neq y \wedge x \neq z \wedge y \neq z \wedge \ldots$$

"x has emailed both y and z, and ...":

$$E(x, y) \wedge E(x, z) \wedge \ldots$$

"if x has emailed somebody, it's either x, y, or z.":

$$\forall s.\ E(x, s) \longrightarrow (s = x \vee s = y \vee s = z).$$

Putting these together, we get:

$$\begin{array}{lll} P(x)::= & \exists y, z. & x \neq y \wedge x \neq z \wedge y \neq z \wedge \\ & & E(x, y) \wedge E(x, z) \wedge \\ & & [\forall s.E(x, s) \longrightarrow (s = x \vee s = y \vee s = z)] \end{array}$$

■

Problem 2.8.5. Express each of the following predicates and propositions in formal logic notation. The domain of discourse is the nonnegative integers, $\mathbb{N}$. Moreover, in addition to the propositional operators, variables and quantifiers, you may define predicates using addition, multiplication, and equality symbols, but no *constants* (like 0, 1,...) and no *exponentiation* (like x^y). For example, the proposition "n is an even number" could be written

$$\exists m.\ (m + m = n).$$

(a) n is the sum of two fourth-powers (a fourth-power is k^4 for some integer k).

Solution.

$$\exists x \exists y.\ (x \cdot x \cdot x \cdot x + y \cdot y \cdot y \cdot y = n)$$

■

Since the constant 0 is not allowed to appear explicitly, the predicate "$x = 0$" can't be written directly, but note that it could be expressed in a simple way as:

$$x + x = x.$$

Then the predicate $x > y$ could be expressed

$$\exists w.\ (y + w = x) \wedge (w \neq 0).$$

Note that we've used "$w \neq 0$" in this formula, even though it's technically not allowed. But since "$w \neq 0$" is equivalent to the allowed formula "$\neg(w + w = w)$," we can use "$w \neq 0$" with the understanding that it abbreviates the real thing. And now that we've shown how to express "$x > y$," it's ok to use it too.

(b) $x = 1$.

Solution. One formula is $\forall y.\ xy = y$. Another is $(x \cdot x = x) \wedge (x \neq 0)$. ■

(c) m is a divisor of n (notation: $m \mid n$)

Solution.

$$m \mid n ::= \quad \exists k.\ k \cdot m = n$$

■

(d) n is a prime number (hint: use the predicates from the previous parts)

Solution.

$$\text{IS-PRIME}(n) ::= \quad (n \neq 1) \ \wedge \ \forall m.\ (m \mid n) \longrightarrow (m = 1 \vee m = n).$$

Note that the requirement $n \neq 1$ is necessary, or else our predicate would be satisfied by 1, which is not considered to be a prime number. Also note that $n \neq 1$ is given here as an abbreviation of the formula $\neg(n = 1)$; and thus of $\neg\forall y.yn = n$.

If we don't want to use the divisor predicate, we can write:

$$\text{IS-PRIME}(n) ::= \quad (n > 1) \ \wedge \ \neg\big(\exists x \exists y.\ (x > 1) \wedge (y > 1) \wedge (x \cdot y = n)\big).$$

Do you agree this formula is also correct? If so, note that the formulas $n > 1$, $x > 1$ and $y > 1$ are not allowed. Can you express them in terms of allowed formulas? ■

(e) n is a power of 3.

Solution. We can simply say that $n \neq 0$ and the only prime divisor of n is 3:

$$n \neq 0 \wedge \forall m.\ (\text{IS-PRIME}(m) \wedge m \mid n) \longrightarrow m = 3.$$

Still, $m = 3$ is not allowed, so we have to express it in terms of allowed formulas. Here is one way to do this:

$$m = 3 ::= \quad \exists z.z = 1 \ \wedge \ m = z + z + z$$

■

Problem 2.8.6. Prove that

$$[\forall x.\ \neg P(x)] \longrightarrow \neg\exists z.\ P(z).$$

is valid. (Use the validity proof from lecture 2W and from the subsection on Validity in Week 2 Notes as guides to writing your proof.)

Solution. *Proof.* Assume

$$\forall x.\ \neg P(x). \tag{2.16}$$

Now we prove by contradiction that $\neg\exists z.\ P(z)$ holds. Namely, we assume $\exists z.\ P(z)$ and reach a contradiction.

So suppose $\exists z.\ P(z)$ holds. So there is an element, c, such that $P(c)$. But from (2.16), we know that $\neg P(c)$ holds, contradicting the fact that $P(c)$ holds. Hence, the assumption $\exists z.\ P(z)$ must be false, that is, $\neg\exists z.\ P(z)$ is true. □

■

Problem 2.8.7. Let A, B, and C be sets. Prove that:

$$A \cup B \cup C = (A - B) \cup (B - C) \cup (C - A) \cup (A \cap B \cap C).$$

You are welcome to use a diagram to aid your own reasoning, but your proof must be text.

Solution. We prove that the left side is contained in the right side, and that the right side is contained in the left side.

First, we show that the left side is contained in the right side. Let x be any element of $A \cup B \cup C$. Then x belongs to at least one of A, B, and C. We distinguish two cases.

- x belongs to all three sets: Then x belongs to the intersection $A \cap B \cap C$.
- x does *not* belong to all three sets: Then at least one of A, B, C does not contain x. So overall, at least one set contains x and at least one set doesn't. We distinguish cases:
  - If A contains x, then one of B and C must not contain it.
    - If B does not contain it, then $x \in A - B$.
    - If B contains it, then C does not, therefore $x \in B - C$.
  - If A does *not* contain x, then one of B and C must contain it.
    - If C does, then $x \in C - A$.
    - If C does not contain it, then B does, therefore $x \in B - C$.

In all cases, we end up with x being a member of one of $A - B$, $B - C$, $C - A$, or $A \cap B \cap C$. Therefore, it belongs to the right side. Hence, the set on the left is contained in the set on the right.

Next, we show that the right side is contained in the left. This is easier. Let x belong to the right side. Then it belongs to one of $A - B$, $B - C$, $C - A$, or $A \cap B \cap C$. In the first case, we clearly know $x \in A$. In the second case, $x \in B$. In the third case, $x \in C$. In the last case, $x \in A$ again. So, in all cases, x belongs to one of A, B, or C. So x belongs to the left side. Therefore, the set on the right is contained in the set on the left.

Since each set is contained in the other, they are equal. ■

Problem 2.8.8. **(a)** Give an example where the following result fails:

False Theorem. *For sets A, B, C, and D, let*

$$L ::= (A \cup C) \times (B \cup D),$$
$$R ::= (A \times B) \cup (C \times D).$$

Then $L = R$.

Solution. If $A = D = \emptyset$ and B and C are both nonempty, then $L = C \times B \neq \emptyset$, but $R = \emptyset$. ■

(b) Identify the mistake in the following proof of the False Theorem.

Bogus proof. Since L and R are both sets of pairs, it's sufficient to prove that $(x, y) \in L \longleftrightarrow (x, y) \in R$ for all x, y.

The proof will be a chain of iff implications:

$(x, y) \in L$
iff $x \in A \cup C$ and $y \in B \cup D$
iff either $x \in A$ or $x \in C$, and either $y \in B$ or $y \in D$
iff ($x \in A$ and $y \in B$) or else ($x \in C$ and $y \in D$)
iff $(x, y) \in A \times B$, or $(x, y) \in C \times D$
iff $(x, y) \in (A \times B) \cup (C \times D)$
iff $(x, y) \in R$.

□

Solution. The mistake is in the third "iff." If [$x \in A$ or $x \in C$, and either $y \in B$ or $y \in D$], it does not necessarily follow that $(x, y) \in (A \times B) \cup (C \times D)$. It might be that (x, y) is in $A \times D$ instead. This happens, for example, if $A = \{1\}, B = \{2\}, C = \{3\}, D = \{4\}$, and $(x, y) = (1, 4)$. ■

(c) Fix the proof to show that $R \subseteq L$.

Solution. Replacing the third "iff" by "which will be true when," yields a correct proof that $(x, y) \in L$ will be true when $(x, y) \in R$, which implies that $R \subseteq L$. ■

Problem 2.8.9. If a set, A, is finite, then $|A| < 2^{|A|} = |\mathcal{P}(A)|$, and so there is no bijection from $\mathcal{P}(A)$ to A. Show that this is still true if A is infinite. *Hint:* Remember Russell's paradox and consider

$$\{f(B) \in A \mid B \subseteq A \text{ and } f(B) \notin B\}$$

where f is such a bijection.

Solution. We prove there is no bijection by contradiction: suppose there was a bijection $f : \mathcal{P}(A) \to A$ for some set A. Because f is a bijection, for any element $a \in A$ there is a unique subset $B_a \subseteq A$ such that $f(B_a) = a$. So consider the set

$$W ::= \{a \in A \mid a \notin B_a\}.$$

By the definition of this set, we know that for all $a \in A$:

$$(a \in W) \longleftrightarrow (a \notin B_a). \tag{2.17}$$

But $W \subseteq A$ (by the definition of W), and hence W is a member of $\mathcal{P}(A)$. So $f(W) \in A$ and $W = B_{f(W)}$. So we have from (2.17), that

$$(a \in B_{f(W)}) \longleftrightarrow (a \notin B_a) \tag{2.18}$$

for all $a \in A$. Substituting $f(W)$ for a in (2.18) yields a contradiction, proving that there cannot be such an f. ■

2.9 Miniquiz Feb. 21

Problem 2.9.1. Next to each of the following propositional formulas, indicate whether it is valid (V), satisfiable but not valid (S), or not satisfiable (N).

$$\begin{array}{rcll} (P \wedge Q) & \longrightarrow & (P \vee Q \vee R) & \_\_\_\_\_ \\ (P \vee Q \vee R) & \longrightarrow & (\overline{P} \wedge \overline{Q} \wedge \overline{R}) & \_\_\_\_\_ \\ (P \wedge Q \wedge R) & \longleftrightarrow & (P \vee Q \vee R) & \_\_\_\_\_ \end{array}$$

Solution. The first formula is **V**alid.

The second formula is **S**atisfiable but not valid — satisfied when P, Q, R all **F**, and false otherwise.

The third formula is also **S**atisfiable but not valid —satisfied when P, Q, R all have the same truth value, but false otherwise.

The second formula in this problem was actually a misprint. The "implies" was intended to be an "and," giving the unsatisfiable formula

$$(P \vee Q \vee R) \wedge (\overline{P} \wedge \overline{Q} \wedge \overline{R}). \qquad (*)$$

Regrettably, the grading for this formula was based on the intended version (*), not the one actually printed on the quiz, so everyone who got this part right was marked wrong, and many who got it wrong were marked right.

We're adjusting grades so everyone gets credit for this part, but people who were marked correct should be aware they got it wrong and be sure to see why. ■

Problem 2.9.2. Let A and B be finite sets.

(a) What relation among $\leq, <, =, \geq$, or $>$, best describes the relationship between

$$|A| + |B| \text{ and } |A \cup B|\,? \quad \_\_\_\_\_$$

Solution. $\geq$. ■

(b) One of the following conditions holds if and only if $|A| + |B| = |A \cup B|$. Circle this condition.

- $A \cap B = \emptyset$
- $A \cup B = \emptyset$
- $A - B = \emptyset$
- $A \subseteq B$

Solution. $A \cap B = \emptyset$. ■

Problem 2.9.3. Circle those logical formulas below, if any, that are implied by the formula $\forall x \exists y . P(x, y)$. (Heads up: the last formula ends with $P(y, x)$.)

$$\exists y \, \forall x . P(x, y)$$

$$\forall y \, \exists x . P(x, y)$$

$$\forall y \, \exists x . P(y, x)$$

Solution. Only the last one, which is the same as the original formula except that the names of x and y are exchanged. ■

Problem 2.9.4. Define a bijection between the nonnegative integers and all the integers.

Solution. One such bijection, f, is defined by lining up the nonnegative integers against all the integers as follows:

$$\begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & \ldots \\ 0 & -1 & 1 & -2 & 2 & -3 & 3 & -4 & \ldots \end{array}$$

We can also define $f : \mathbb{N} \to \mathbb{Z}$ by the formula,

$$f(2k) ::= k,$$
$$f(2k+1) ::= -(k+1),$$

for $k \in \mathbb{N}$. An equivalent formulation is:

$$f(n) ::= \begin{cases} n/2, & \text{if } n \text{ is even;} \\ -(n+1)/2, & \text{if } n \text{ is odd.} \end{cases}$$

■

Problem 2.9.5. The following proof ends with a contradiction, so unless Math is broken, this proof must be mistaken. In a sentence or two, explain what the mistake is.

Let x be a variable ranging over all sets, and define

$$W ::= \{x \mid x \notin x\}.$$

So by definition,

$$x \in W \quad \text{iff} \quad x \notin x,$$

for every set, x. Letting x be W, we get the contradiction

$$W \in W \quad \text{iff} \quad W \notin W.$$

Solution. Since $x \notin x$ is always true, the set W is the collection of *all* sets, which is not a set. So substituting W for x is not justified. ■

Problem 2.9.6. Prove that $\sqrt[3]{2}$ is irrational. You may assume the fact that if a positive integer power of an integer is even, then the integer is even. Your answer will be graded primarily on the clarity of your argument.

Solution. See Week 2 Monday, Class Problem 3. ■

Glossary

symbol	meaning
$\wedge$	AND
$\vee$	OR
$\longrightarrow$	IMPLIES
$\neg$	NOT
$\longleftrightarrow$	EQUIVALENT
$\overline{P}$	NOT P
$\cup$	SET UNION
$\cap$	SET INTERSECTION
$\lvert S\rvert$	size of finite set S
$\exists$	EXISTS
$\forall$	FOR ALL
$\in$	is a member of
$\subseteq$	is a subset of

Chapter 3

Relations; Induction

3.1 Binary Relations

Relations are another fundamental Mathematical data type. Equality and "less than" are the most familiar examples of Mathematical relations. These are called *binary* relations because they relate two objects – are the objects equal? is the first less than the second? In this section of Notes we'll define some basic properties of binary relations and then focus on *partial orders*, which are a class of binary relations of particular importance in Computer Science, with direct applications that include task scheduling, database concurrency control, and proving that computations terminate.

3.1.1 Binary Relations and Functions

Binary relations are far more general than equality or less-than. Here's the official definition:

Definition 3.1.1. A *binary relation*, R, consists of a set, A, called the *domain* of R, a set, B, called the *codomain* of R, and a subset of $A \times B$ called the *graph* of R.

For example, we can define an "is teaching relation" for Spring '07 at MIT to have domain equal to the names of all the teaching staff (faculty, T.A.'s, *etc*.) and codomain equal to all the subject numbers in the current catalogue. Its graph would contain pairs like

$$\begin{array}{l}(\texttt{Albert R. Meyer}, 6.042),\\ (\texttt{Tina Nolte}, 18.062),\\ (\texttt{Chiyoun Park}, 6.042),\\ (\texttt{Albert R. Meyer}, 18.062),\\ (\texttt{Srini Devadas}, 6.046),\\ (\texttt{Donald Sadoway}, 3.091),\\ \quad\vdots\end{array}$$

Notice that Definition 3.1.1 is exactly the same as the definition of a *function*, except that it doesn't require the functional condition that, for each domain element, a, there is *at most* one pair in the graph whose first coordinate is a. So a function is a special case of a binary relation.

A relation whose domain is A and codomain is B is said to be "between A and B", or "from A to B." When the domain and codomain are the same set, A, we simply say the relation is "on A." It's common to use infix notation "$a\,R\,b$" to mean that the pair (a, b) is in the graph of R.

3.1.2 Images and Inverse Images

Before we go any further, it's worth introducing some notation that we'll get a lot of mileage out of. If R is a binary relation from A to B, and C is any set, define

$$\begin{aligned} CR &::= \{b \in B \mid cRb \text{ for some } c \in C\}, \\ RC &::= \{a \in A \mid aRc \text{ for some } c \in C\}. \end{aligned}$$

The set CR is called the *image* of C under R. With this notation, we could have defined the *range* of R simply as AR.

The set RC is called the *inverse image* of C under R. Notice the clash with pointwise application notation from Notes 2 when R happens to be a function: $\widehat{R}(C) = CR$, not RC. Sorry about that.

3.1.3 Surjective and like that

A relation with the property that every codomain element is related to some domain element is called a *surjective* (or *onto*) relation —again, the same definition as for functions. More concisely, a relation, R, between A and B is surjective iff $AR = B$. Likewise, a relation with the property that every domain element is related to some codomain element is called a *total* relation; more concisely, R is total iff $A = RB$.

3.2 Partial Orders

The prerequisite structure among MIT subjects provides a nice illustration of partial orders. Here is a table indicating some of the prerequisites of subjects in the Course 6 program:

Direct Prerequisites	Subject
18.01	6.042
18.01	18.02
18.01	18.03
8.01	8.02
6.001	6.034
6.042	6.046
18.03, 8.02	6.002
6.001, 6.002	6.004
6.001, 6.002	6.003
6.004	6.033
6.033	6.857
6.046	6.840

Since 18.01 is a direct prerequisite for 6.042, a student must take 18.01 before 6.042. Also, 6.042 is a direct prerequisite for 6.046, so in fact, a student has to take *both* 18.01 and 6.042 before taking 6.046. So 18.01 is also really a prerequisite for 6.046, though an implicit or indirect one; we'll indicate this by writing

$$18.01 \to 6.046.$$

This prerequisite relation has a basic property known as *transitivity*: if subject a is an indirect prerequisite of subject b, and b is an indirect prerequisite of subject c, then a is also an indirect prerequisite of c.

In this table, a longest sequence of prerequisites is

$$18.01 \to 18.03 \to 6.002 \to 6.004 \to 6.033 \to 6.857$$

so a student would need at least six terms to work through this sequence of courses. But it would take a lot longer to complete a Course 6 major if the direct prerequisites led to a situation[1] where two subjects turned out to be prerequisites of *each other*! So another crucial property of the prerequisite relation is that if $a \to b$, then it is not the case that $b \to a$. This property is called *asymmetry*.

Another basic example of a partial order is the subset relation, $\subseteq$, on sets. In fact, we'll see that every partial order can be represented by the subset relation.

3.2.1 Axioms for Partial Orders

Definition 3.2.1. A binary relation, R, on a set A is:

- *transitive* iff
$$[a\,R\,b \text{ and } b\,R\,c] \quad \text{implies} \quad a\,R\,c,$$
for every $a, b, c \in A$,
- *asymmetric* iff
$$a\,R\,b \quad \text{implies} \quad \neg(b\,R\,a)$$
for all $a, b \in A$,
- a *strict partial order* iff it is transitive and asymmetric.

So the prerequisite relation, $\to$, on subjects in the MIT catalogue is a strict partial order. More familiar examples of strict partial orders are the relation, $<$, on real numbers, and the proper subset relation, $\subset$, on sets.

The subset relation, $\subseteq$, on sets and $\leq$ relation on numbers are examples of *reflexive* relations in which each element is related to itself. Reflexive partial orders are called *weak* partial orders:

Definition 3.2.2. A binary relation, R, on a set A, is

- *reflexive* iff $a\,R\,a$ for all $a \in A$,

[1] MIT's Committee on Curricula has the responsibility of watching out for such bugs that might creep into departmental requirements.

- *antisymmetric* if

$$a\,R\,b \quad \text{implies} \quad \neg(b\,R\,a)$$

for all $a \neq b \in A$,

- a *weak partial order* iff it is transitive, reflexive and antisymmetric.

Some authors define partial orders to be what we call weak partial orders, but we'll use the phrase "partial order" to mean either a weak or strict one.

For weak partial orders in general, we often write an ordering-style symbol like $\preceq$ or $\sqsubseteq$ instead of a letter symbol like R. (General relations are usually denoted by a letter like R instead of a cryptic squiggly symbol, so $\preceq$ is kind of like Prince.) Likewise, we generally use $\prec$ or $\sqsubset$ to indicate a strict partial order. We also write $b \succeq a$ to mean $a \preceq b$ and $b \succ a$ to mean $a \prec b$.

Two more examples of partial orders are worth mentioning:

Example 3.2.3. Let A be some family of sets and define aRb iff $a \supset b$. Then R is a strict partial order.

For integers, m, n we write $m \mid n$ to mean that m *divides* n, namely, there is an integer, k, such that $n = km$.

Example 3.2.4. The divides relation is a weak partial order on the nonnegative integers.

3.2.2 Representing Partial Orders by Set Containment

When a class of objects are defined by axioms, like the axioms for partial orders, it can help to have a way to "represent" them explicitly by known objects. Partial orders can be represented by the subset relation on a collection of sets. Namely, if R is a weak partial order on a set, A, we can let each element $a \in A$ correspond to the set $R\{a\}$. Since

$$a\,R\,b \quad \text{iff} \quad R\{a\} \subseteq R\{b\} \tag{3.1}$$

holds for all $a, b \in A$, we have completely captured the weak partial order R by the subset relation on the corresponding sets. A similar correspondence shows that strict partial orders can be represented by the proper subset relation, $\subset$.

Problem 3.2.1. Prove the iff assertion (3.1).

Problem 3.2.2. Verify that the relations in Examples 3.2.3 and 3.2.4 are partial orders.

Definition 3.2.5. A relation, R, on a set, A, is *irreflexive* iff for all $a \in A$, it is *not* true that $a\,R\,a$.

Problem 3.2.3. Prove that a binary relation is a strict partial order iff it is transitive and irreflexive.

3.2.3 Total Orders

The familiar order relations on numbers have an important additional property: given any two numbers, one will be bigger than the other. Partial orders with this property are said to be *total*[2] *orders*:

[2] "Total" is an overloaded term when talking about partial orders: being a total order is a much stronger condition than being a partial order that is a total relation. For example, any weak partial order such as $\subseteq$ is a total relation.

Definition 3.2.6. Let R be a binary relation on a set, A, and let a, b be elements of A. Then a and b are *comparable* with respect to R iff ($a\,R\,b$ or $b\,R\,a$). A partial order under which every two distinct elements are comparable is called a *total order*.

So $<$ and $\leq$ are total orders on $\mathbb{R}$. On the other hand, the subset relation is *not* total, since, for example, any two distinct finite sets of the same size will be incomparable under $\subseteq$. The prerequisite relation on Course 6 required subjects is also not total because, for example, neither 8.01 nor 6.001 is a prerequisite of the other.

3.2.4 Products of Relations

Taking the product of two relations is a useful way to construct new relations from old ones.

The product, $R_1 \times R_2$, of relations R_1 and R_2 is defined to be the relation with

$$\begin{aligned} \text{domain}\,(R_1 \times R_2) &::= \text{domain}\,(R_1) \times \text{domain}\,(R_2)\,, \\ \text{codomain}\,(R_1 \times R_2) &::= \text{codomain}\,(R_1) \times \text{codomain}\,(R_2)\,, \\ (a_1, a_2)\,(R_1 \times R_2)\,(b_1, b_2) &\quad \text{iff} \quad [a_1\,R_1\,b_1 \text{ and } a_2\,R_2\,b_2]. \end{aligned}$$

Example 3.2.7. Define a relation, Y, on age-height pairs of being younger *and* shorter. This is the relation on the set of pairs (y, h) where y is a natural number ≤ 2400 which we interpret as an age in months, and h is a natural number ≤ 120 describing height in inches. We define Y by the rule

$$(y_1, h_1)\,Y\,(y_2, h_2) \quad \text{iff} \quad y_1 \leq y_2 \wedge h_1 \leq h_2.$$

That is, Y is the product of the $\leq$-relation on ages and the $\leq$-relation on heights.

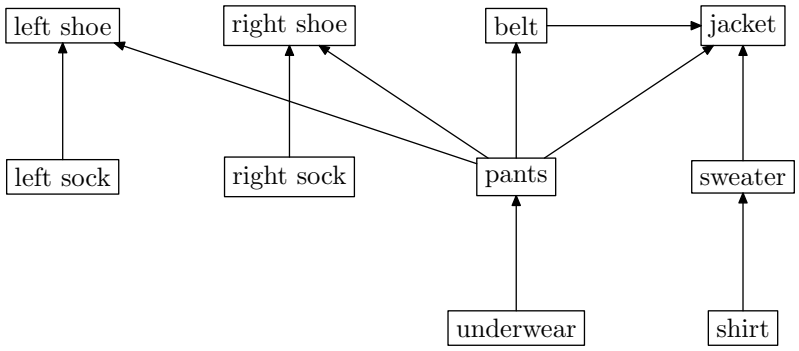
Products preserve several of the relational properties we have considered. Namely, it's not hard to verify that if R_1 and R_2 are both transitive, then so is $R_1 \times R_2$. The same holds for reflexivity, irreflexivity, and antisymmetry. This implies that if R_1 and R_2 are both partial orders, then so is $R_1 \times R_2$.

On the other hand, the property of being a total order is not preserved. For example, the age-height relation Y is the product of two total orders, but it is not total: the age 240 months, height 68 inches pair, (240,68), and the pair (228,72) are incomparable under Y.

3.2.5 Topological Sorting

Scheduling problems are a common source of partial orders: there is a set, A, of tasks and a set of constraints specifying that starting a certain task depends on other tasks being completed beforehand. We can picture the constraints by drawing labelled boxes corresponding to different tasks, with an arrow from one box to another if the first box corresponds to a task that must be completed before starting the second one.

Example 3.2.8. Here is a drawing describing the order in which you could put on clothes. The tasks are the clothes to be put on, and the arrows indicate what should be put on directly before what.



When we have a partial order of tasks to be performed, it can be useful to have an order in which to perform all the tasks, one at a time, while respecting the dependency constraints. This amounts to finding a total order that is consistent with the partial order. This task of finding a total ordering that is consistent with a partial order is known as *topological sorting*.

Definition 3.2.9. A *topological sort* of a partial order, $\prec$, on a set, A, is a total ordering, $\sqsubset$, on A such that

$$a \prec b \quad \text{implies} \quad a \sqsubset b.$$

For example,

$$\text{shirt} \sqsubset \text{sweater} \sqsubset \text{underwear} \sqsubset \text{leftsock} \sqsubset \text{rightsock} \sqsubset \text{pants} \sqsubset \text{leftshoe} \sqsubset \text{rightshoe} \sqsubset \text{belt} \sqsubset \text{jacket},$$

is one topological sort of the partial order of dressing tasks given by Example 3.2.8; there are several other possible sorts as well.

Topological sorts for partial orders on finite sets are easy to construct by starting from *minimal* elements:

Definition 3.2.10. Let $\preceq$ be a partial order on a set, A. An element $a \in A$ is *minimum* iff it is $\preceq$ every other element of A. The element a is minim*al* iff no other element is $\preceq a$.

In a total order, minimum and minimal elements are the same thing. But a partial order may no minimum element but lots of minimal elements. There are four minimal elements in the clothes example: leftsock, rightsock, underwear, and shirt.

To construct a total ordering for getting dressed, we pick one of these minimal elements, say shirt. Next we pick a minimal element among the remaining ones. For example, once we have removed shirt, sweater becomes minimal. We continue in this way removing successive minimal elements until all elements have been picked. The sequence of elements in the order they were picked will be a topological sort. This is how the topological sort above for getting dressed was constructed.

For this method of topological sorting to work, we need to be sure there is always a minimal element. This is sort of obvious, but noting that an infinite partially ordered set might have no minimal element —consider $<$ on the $\mathbb{Z}$ —it would be good to prove that minimal elements exist.

Lemma 3.2.11. *Every partial order on a nonempty finite set has a minimal element.*

Proof. Let R be a strict partial order on a set, A. Define the *weight* of an element $a \in A$ to be $|R\{a\}|$ —the number of elements in the set $R\{a\}$. Since A is finite, the weights of all elements in A are nonnegative integers, so there must be an $a_0 \in A$ with the smallest weight.

Now suppose $|R\{a_0\}| \neq 0$. Then there is an element $a_1 \in R\{a_0\}$, which implies (by transitivity of R) that $R\{a_1\} \subseteq R\{a_0\}$, and hence $|R\{a_1\}| \leq |R\{a_0\}|$. But since R is strict, $a_1 \in R\{a_0\} - R\{a_1\}$, so in fact $|R\{a_1\}| < |R\{a_0\}|$, contradicting the fact the a_0 has the smallest weight.

This contradiction implies that $|R\{a_0\}| = 0$, which means that no element is related by R to a_0, that is, a_0 is minimal.

A similar argument works in the case that R is a weak partial order.

□

So our construction shows:

Theorem 3.2.12. *Every partial order on a finite set has a topological sort.*

In fact, the domain of the partial order need not be finite: we won't prove it, but *all* partial orders, even infinite ones, have topological sorts.

There are many other ways of constructing topological sorts. For example, instead of starting "from the bottom" with minimal elements, we could start "from the top" by picking *maximal* elements:

Definition 3.2.13. Let $\preceq$ be a partial order on a set, A. An element $a \in A$ is *maximum* iff it is $\succeq$ every other element of A. The element a is maxim*al* iff no other element is $\succeq a$.

Problem 3.2.4. **(a)** Prove that there is at most one minim*um* element in any partial order.

(b) Give an example of a partial order with exactly one minimal element, but no minimum element. *Hint:* It will have be infinite.

3.2.6 Parallel Task Scheduling

For a partial order of task dependencies, topological sorting provides a way to execute tasks sequentially without violating the dependencies. But what if we have the ability to execute more than one task at the same time? For example, say tasks are programs, the partial order indicates data dependence, and we have a parallel machine with lots of processors instead of a sequential machine with only one. How should we schedule the tasks? Our goal should be to minimize the total *time* to complete all the tasks. For simplicity, let's say all the tasks take the same amount of time and all the processors are identical.

So, given a finite partially ordered set of tasks, how long does it take to do them all, in an optimal parallel schedule? We can also use partial order concepts to analyze this problem.

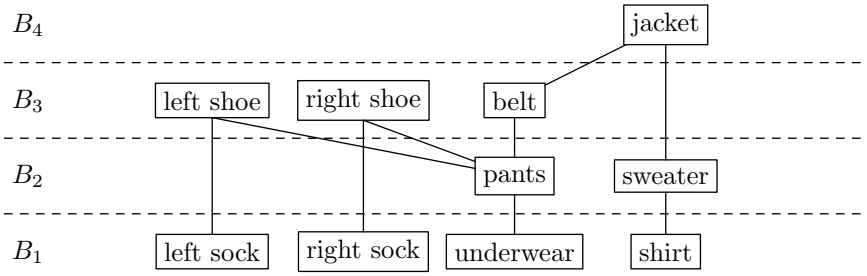
In the clothes example, we could do all the minimal elements first (leftsock, rightsock, underwear, shirt), remove them and repeat. We'd need lots of hands, or maybe dressing servants. We can do pants and sweater next, and then leftshoe, rightshoe, and belt, and finally jacket.

We can't do any better, because the sequence underwear, pants, belt, jacket must be done in that order. A set of tasks that must be done in sequence like this is called a *chain*.

Definition 3.2.14. A *chain* in a partial order is a set of elements such that any two elements in the set are comparable.

In other words, a chain is a totally ordered subset of the elements in a partial order. Clearly, the parallel time must be at least the size of any chain. For if we used less time, then two tasks in the chain would have to be done at the same time, violating the dependency constraints.

A largest chain is also known as a *critical path*. So we need at least t steps, where t is the size of a largest chain. Fortunately, it is always possible to use only t parallel steps. The idea is to let B_1 be all the minimal elements and schedule them first. Then remove all the elements in B_1, let B_2 be the elements that now become minimal, and schedule them next, and so on. For getting dressed, here is a picture of the schedule obtained in this way:



Theorem 3.2.15. *Let R be strict partial order on a set, A. If the longest chain in A is of size t, then there is a partition[3] of A into t blocks, $B_1, B_2, \ldots, B_t$, such that for each block, B_i, all tasks that have to precede tasks in B_i are in smaller-numbered groups. That is,*

$$RB_1 = \emptyset, \text{ and} \tag{3.2}$$

$$RB_i \subseteq B_1 \cup B_2 \cup \cdots \cup B_{i-1}, \tag{3.3}$$

for $1 < i \leq t$.

Corollary 3.2.16. *For R and t as above, it is possible to schedule all tasks in t steps.*

Proof. Schedule all the elements of B_i at time i. This satisfies the dependency requirements, because all the tasks that any task depends on are scheduled at preceding times. □

Corollary 3.2.17. *Parallel time = Size of largest chain.*

So it remains to prove Theorem 3.2.15:

Proof. A chain is said to *begin* with its smallest element and *end* with its largest element, if any.

Construct the sets B_i as follows:

$$B_i ::= \{a \in A \mid \text{the largest chain ending in } a \text{ is of size } i\}.$$

[3]Partitioning a set, A, means "cutting it up" into non-overlapping, nonempty pieces. The pieces are called the blocks of the partition. More precisely, a *partition* of A is a set $\mathcal{B}$ whose elements are nonempty subsets of A such that

- if $B, B' \in \mathcal{B}$ are distinct sets, then $B \cap B' = \emptyset$, and
- $\bigcup_{B \in \mathcal{B}} B = A$.

This gives just t sets, because the largest chain is of size t. Also, each $a \in A$ belongs to exactly one B_i. To complete the proof, notice that if $a \in B_1$, then a must be minimal, and since R is strict we have $RB_1 = \emptyset$ proving (3.2).

Now suppose $1 < i \leq t$, and assume for the sake of contradiction that (3.3) does not hold. That is, there is an $a \in B_i$ and $b \in A$ such that $b\,R\,a$, and $b \notin B_1 \cup B_2 \cup \cdots \cup B_{i-1}$. Then by definition of the B_j's, there is a chain of size $> i-1$ ending at b. Also, since R is strict, a is not in the chain ending at b. So we can add a to the end of the chain to obtain a chain of size $> i$ ending in a, contradicting the fact that that $a \in B_i$. □

So with an unlimited number of processors, the time to complete all the tasks is the size of the largest chain. It turns out that this theorem is good for more than parallel scheduling. It is usually stated as follows.

Definition 3.2.18. An *antichain* in a partial order is a set of elements such that any two elements in the set are incomparable.

Corollary 3.2.19. *If the largest chain in a partial order is of size, t, then the domain can be partitioned into t antichains.*

Proof. Let the antichains be the sets B_i defined as in the proof of Theorem 3.2.15.

We should verify that each B_i is an antichain, namely, if a, b are distinct elements of B_i, then they are incomparable. But suppose to the contrary that there exist two elements $a, b \in B_i$ such that a and b are comparable, say $a\,R\,b$. Then, as in the proof of Theorem 3.2.15, by adding b at the end of the chain of size i ending at a, we obtain a chain of size $i+1$ ending at b, contradicting the assumption that $b \in B_i$. □

3.2.7 Dilworth's Lemma

We can use the Corollary 3.2.19 to prove a famous result[4] about partially ordered sets:

Lemma 3.2.20 (Dilworth). *For all $t > 0$, every partially ordered set with n elements must have either a chain of size greater than t or an antichain of size at least n/t.*

Proof. Assume there is no chain of size greater than t, that is, the largest chain is of size $\leq t$. Then by Corollary 3.2.19, the n elements can be partitioned into at most t antichains. Let ℓ be the size of the largest antichain. Since every element belongs to exactly one antichain, and there are at most t antichains, there can't be more than ℓt elements, namely, $\ell t \geq n$. So there is an antichain with at least $\ell \geq n/t$ elements. □

Corollary 3.2.21. *Every partially ordered set with n elements has a chain of size greater than $\sqrt{n}$ or an antichain of size at least $\sqrt{n}$.*

Proof. Set $t = \sqrt{n}$ in Lemma 3.2.20. □

[4] Lemma 3.2.20 also follows from a more general result known as Dilworth's Theorem which we will not discuss.

Example 3.2.22. In the dressing partially ordered set, $n = 10$.

Try $t = 3$. There is a chain of size 4.

Try $t = 4$. There is no chain of size 5, but there is an antichain of size $4 \geq 10/4$.

Example 3.2.23. Suppose we have a class of 101 students. Then using the product partial order, Y, from Example 3.2.7, we can apply Dilworth's Lemma to conclude that there is a chain of 11 students who get taller as they get older, or an antichain of 11 students who get taller as they get younger, which makes for an amusing in-class demo.

Quick Exercise: What is the size of the longest chain that is guaranteed to exist in any partially ordered set of n elements? What about the largest antichain?

Solution. For $n > 0$, chain size is 1 in the "discrete" partial order in which every two distinct elements are incomparable. Antichain size is 1 if $n > 0$ and the partial order is total. ■

3.3 Induction

Induction is by far the most powerful and commonly-used proof technique in Discrete Mathematics and Computer Science. In fact, the use of induction is a defining characteristic of *discrete* —as opposed to *continuous* —Mathematics.

To understand how induction works, suppose there is a professor who brings to class a bottomless bag of assorted miniature candy bars. She offers to share in accordance with two rules. First, she numbers the students 0, 1, 2, 3, and so forth for convenient reference. Now here are the two rules:

1. Student 0 gets candy.
2. If student n gets candy, then student $n + 1$ also gets candy, for every $n \in \mathbb{N}$.

You can think of the second rule as a compact way of writing a whole sequence of statements, one for each natural value of n:

- If student 0 gets candy, then student 1 also gets candy.
- If student 1 gets candy, then student 2 also gets candy.
- If student 2 gets candy, then student 3 also gets candy. ⋮

Now suppose you are student 17. By these rules, are you entitled to a miniature candy bar? Well, student 0 gets candy by the first rule. Therefore, by the second rule, student 1 also gets candy, which means student 2 gets candy as well, which means student 3 get candy, and so on. So the professor's two rules actually guarantee candy for *every* student, no matter how large the class. You win!

This kind of reasoning is an instance of

The Principle of Induction. Let $P(n)$ be a predicate. If

- $P(0)$ is true, and
- for all $n \in \mathbb{N}$, $P(n)$ implies $P(n+1)$,

then $P(n)$ is true for all $n \in \mathbb{N}$.

Here's the correspondence between the induction principle and sharing candy bars. Suppose that $P(n)$ is the predicate, "student n gets candy". Then the professor's first rule asserts that $P(0)$ is true, and her second rule is that for all $n \in \mathbb{N}$, $P(n)$ implies $P(n+1)$. Given these facts, the induction principle says that $P(n)$ is true for all $n \in \mathbb{N}$. In other words, everyone gets candy.

The intuitive justification for the general induction principle is the same as for everyone getting a candy bar under the professor's two rules. So the Principle of Induction is universally accepted as an obvious, sound proof method. What's not so obvious is how much mileage we get by using it.

3.4 Using Induction

Induction often works directly in proving that some statement about natural numbers holds for all of them. For example, here is a classic formula:

Theorem 3.4.1. *For all $n \in \mathbb{N}$,*

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2} \tag{3.4}$$

The left side of equation (3.4) represents the sum of all the numbers from 1 to n. Here the dots ($\cdots$) indicate a pattern you're supposed to be able to guess so you can mentally fill in the remaining terms.

The meaning of this sum is not so obvious in a couple of special cases:

- If $n = 1$, then there is only one term in the summation, and so $1 + 2 + 3 + \cdots + n = 1$. Don't be misled by the appearance of 2 and 3 and the suggestion that 1 and n are distinct terms!
- If $n \leq 0$, then there are no terms at all in the summation. By convention, the sum in this case is 0.

So while the dots notation is convenient, you have to watch out for these special cases where the notation is misleading! (In fact, whenever you see the dots, you should be on the lookout to be sure you understand the pattern.)

We could eliminate the need for guessing by rewriting the left side of (3.4) with ***summation notation***:

$$\sum_{i=1}^{n} i \qquad \text{or} \qquad \sum_{1 \leq i \leq n} i.$$

Both of these expressions denote the sum of all values taken on by the expression to the right of the sigma as the variable, i, ranges from 1 to n. Both these summation expressions make it clear what (3.4) means when $n = 1$. The second expression makes it clear that when $n = 0$, there are no terms in the sum, though you still have to know the convention that a sum of no numbers equals 0 (the *product* of no numbers is 1, by the way).

Now let's use the induction principle to prove Theorem 3.4.1. Suppose that we define predicate $P(n)$ to be "$1+2+3+\cdots+n = n(n+1)/2$". Recast in terms of this predicate, the theorem claims that $P(n)$ is true for all $n \in \mathbb{N}$. This is great, because the induction principle lets us reach precisely that conclusion, provided we establish two simpler facts:

- $P(0)$ is true.
- For all $n \in \mathbb{N}$, $P(n)$ implies $P(n+1)$.

So now our job is reduced to proving these two statements. The first is true because $P(0)$ asserts that a sum of zero terms is equal to $0(0+1)/2 = 0$.

The second statement is more complicated. But remember the basic plan for proving the validity of any implication: *assume* the statement on the left and then *prove* the statement on the right. In this case, we assume $P(n)$:

$$1+2+3+\cdots+n = \frac{n(n+1)}{2} \tag{3.5}$$

in order to prove $P(n+1)$:

$$1+2+3+\cdots+n+(n+1) = \frac{(n+1)(n+2)}{2} \tag{3.6}$$

These two equations are quite similar; in fact, adding $(n+1)$ to both sides of equation (3.5) and simplifying the right side gives the equation (3.6):

$$\begin{aligned} 1+2+3+\cdots+n+(n+1) &= \frac{n(n+1)}{2} + (n+1) \\ &= \frac{(n+2)(n+1)}{2} \end{aligned}$$

Thus, if $P(n)$ is true, then so is $P(n+1)$. This argument is valid for every natural number n, so this establishes the second fact required by the induction principle. In effect, we've just proved that $P(0)$ implies $P(1)$, $P(1)$ implies $P(2)$, $P(2)$ implies $P(3)$, etc., all in one fell swoop.

With these two facts in hand, the induction principle says that the predicate $P(n)$ is true for all natural n, so the theorem is proved.

3.4.1 A Template for Induction Proofs

The proof of Theorem 3.4.1 was relatively simple, but even the most complicated induction proof follows exactly the same template. There are five components:

1. **State that the proof uses induction.** This immediately conveys the overall structure of the proof, which helps the reader understand your argument.

2. **Define an appropriate predicate $P(n)$.** The eventual conclusion of the induction argument will be that $P(n)$ is true for all natural n. Thus, you should define the predicate $P(n)$ so that your theorem is equivalent to (or follows from) this conclusion. Often the predicate can be lifted straight from the claim, as in the example above. The predicate $P(n)$ is called the "induction hypothesis". Sometimes the induction hypothesis will involve several variables, in which case you should indicate which variable serves as n.

3. **Prove that $P(0)$ is true.** This is usually easy, as in the example above. This part of the proof is called the "base case" or "basis step". (Sometimes the base case will be $n = 1$ or even some larger number, in which case the starting value of n also should be stated.)

4. **Prove that $P(n)$ implies $P(n+1)$ for every natural number n.** This is called the "inductive step" or "induction step". The basic plan is always the same: assume that $P(n)$ is true and then use this assumption to prove that $P(n+1)$ is true. These two statements should be fairly similar, but bridging the gap may require some ingenuity. Whatever argument you give must be valid for every natural number n, since the goal is to prove the implications $P(0) \to P(1)$, $P(1) \to P(2)$, $P(2) \to P(3)$, etc. all at once.

5. **Invoke induction.** Given these facts, the induction principle allows you to conclude that $P(n)$ is true for all natural n. This is the logical capstone to the whole argument, but many writers leave this step implicit.

Explicitly labeling the *base case* and *inductive step* may make your proofs clearer.

3.4.2 A Clean Writeup

The proof of Theorem 3.4.1 given above is perfectly valid; however, it contains a lot of extraneous explanation that you won't usually see in induction proofs. The writeup below is closer to what you might see in print and should be prepared to produce yourself.

Proof. We use induction. The induction hypothesis, $P(n)$, will be equation (3.4).

Base case: $P(0)$ is true, because both sides of equation (3.4) equal zero when $n = 0$.

Inductive step: Assume that $P(n)$ is true, where n is any natural number. Then

$$\begin{aligned} 1+2+3+\cdots+n+(n+1) &= \frac{n(n+1)}{2}+(n+1) && \text{by induction hypothesis} \\ &= \frac{(n+1)(n+2)}{2} && \text{by simple algebra} \end{aligned}$$

which proves $P(n+1)$.

So it follows by induction that $P(n)$ is true for all natural n. □

Induction was helpful for *proving the correctness* of this summation formula, but not helpful for *discovering* it in the first place. We'll show you some tricks for finding such formulas in a few weeks.

3.4.3 Powers of Odd Numbers

A proof in class that $\sqrt[n]{2}$ is irrational used the "obvious":

Fact. The nth power of an odd number is odd, for all nonnegative integers, n.

Instead of taking this fact for granted, we can prove it by induction. The proof will require a simple Lemma.

Lemma. *The product of two odd numbers is odd.*

To prove the Lemma, note that the odd numbers are, by definition, the numbers of the form $2k+1$ where k is an integer. But

$$(2k+1)(2k'+1) = 2(2kk' + k + k') + 1,$$

so the product of two odd numbers also has the form of an odd number, which proves the Lemma.

Now we will prove the Fact using the induction hypothesis

$$P(n) ::= \text{if } a \text{ is an odd integer, then so is } a^n.$$

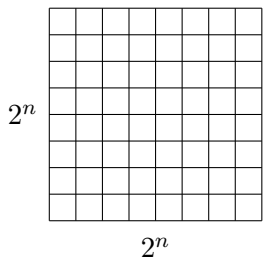
The base case $P(0)$ holds because $a^0 = 1$, and 1 is odd.

For the inductive step, suppose $n \geq 0$, a is an odd number and $P(n)$ holds. So a^n is an odd number. Therefore, $a^{n+1} = a^n a$ is a product of odd numbers, and by the Lemma a^{n+1} is also odd. This proves $P(n+1)$, and we conclude by induction that $P(n)$ holds for nonnegative integers n.

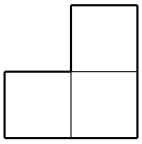
3.4.4 Courtyard Tiling

Induction served purely as a proof technique in the preceding examples. But induction sometimes can serve as a more general reasoning tool.

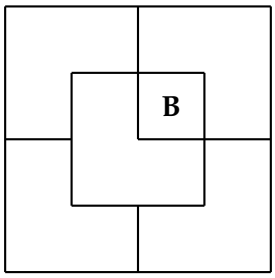
MIT recently constructed the Stata Center which houses the Computer Science and AI Laboratory. During development, the project went further and further over budget, and there were some radical fundraising ideas. One rumored plan was to install a big courtyard with dimensions $2^n \times 2^n$:



One of the central squares would be occupied by a statue of a wealthy potential donor. Let's call him "Bill". (In the special case $n = 0$, the whole courtyard consists of a single central square; otherwise, there are four central squares.) A complication was that the building's unconventional architect, Frank Gehry, supposedly insisted that only special L-shaped tiles be used:



A courtyard meeting these constraints exists, at least for $n = 2$:



For larger values of n, is there a way to tile a $2^n \times 2^n$ courtyard with L-shaped tiles and a statue in the center? Let's try to prove that this is so.

Theorem 3.4.2. *For all $n \geq 0$ there exists a tiling of a $2^n \times 2^n$ courtyard with Bill in a central square.*

Proof. (doomed attempt) The proof is by induction. Let $P(n)$ be the proposition that there exists a tiling of a $2^n \times 2^n$ courtyard with Bill in the center.

Base case: $P(0)$ is true because Bill fills the whole courtyard.

Inductive step: Assume that there is a tiling of a $2^n \times 2^n$ courtyard with Bill in the center for some $n \geq 0$. We must prove that there is a way to tile a $2^{n+1} \times 2^{n+1}$ courtyard with Bill in the center ....

□

Now we're in trouble! The ability to tile a smaller courtyard with Bill in the center isn't much help in tiling a larger courtyard with Bill in the center. We haven't figured out how to bridge the gap between $P(n)$ and $P(n+1)$.

So if we're going to prove Theorem 3.4.2 by induction, we're going to need some *other* induction hypothesis than simply the statement about n that we're trying to prove.

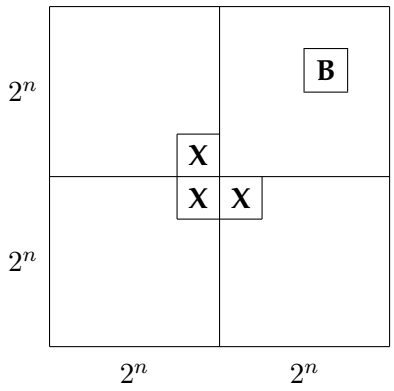
When this happens, your first fallback should be to look for a *stronger* induction hypothesis; that is, one which implies your previous hypothesis. For example, we could make $P(n)$ the proposition that for *every* location of Bill in a $2^n \times 2^n$ courtyard, there exists a tiling of the remainder.

This advice may sound bizarre: "If you can't prove something, try to prove something grander!" But for induction arguments, this makes sense. In the inductive step, where you have to prove $P(n) \longrightarrow P(n+1)$, you're in better shape because you can *assume* $P(n)$, which is now a more powerful statement. Let's see how this plays out in the case of courtyard tiling.

Proof. (successful attempt) The proof is by induction. Let $P(n)$ be the proposition that for every location of Bill in a $2^n \times 2^n$ courtyard, there exists a tiling of the remainder.

Base case: $P(0)$ is true because Bill fills the whole courtyard.

Inductive step: Assume that $P(n)$ is true for some $n \geq 0$; that is, for every location of Bill in a $2^n \times 2^n$ courtyard, there exists a tiling of the remainder. Divide the $2^{n+1} \times 2^{n+1}$ courtyard into four quadrants, each $2^n \times 2^n$. One quadrant contains Bill (**B** in the diagram below). Place a temporary Bill (**X** in the diagram) in each of the three central squares lying outside this quadrant:



Now we can tile each of the four quadrants by the induction assumption. Replacing the three temporary Bills with a single L-shaped tile completes the job. This proves that $P(n)$ implies $P(n+1)$ for all $n \geq 0$. The theorem follows as a special case. □

This proof has two nice properties. First, not only does the argument guarantee that a tiling exists, but also it gives an algorithm for finding such a tiling. Second, we have a stronger result: if Bill wanted a statue on the edge of the courtyard, away from the pigeons, we could accommodate him!

Strengthening the induction hypothesis is often a good move when an induction proof won't go through. But keep in mind that the stronger assertion must actually be *true*; otherwise, there isn't much hope of constructing a valid proof! Sometimes finding just the right induction hypothesis requires trial, error, and insight. For example, mathematicians spent almost twenty years trying to prove or disprove the conjecture that "Every planar graph is 5-choosable"[5]. Then, in 1994, Carsten Thomassen gave an induction proof simple enough to explain on a napkin. The key turned out to be finding an extremely clever induction hypothesis; with that in hand, completing the argument is easy!

3.4.5 A Faulty Induction Proof

False Theorem. *All horses are the same color.*

Notice that no n is mentioned in this assertion, so we're going to have to reformulate it in a way that makes an n explicit. In particular, we'll (falsely) prove that

False Theorem 3.4.3. *In every set of $n \geq 1$ horses, all are the same color.*

[5] 5-choosability is a slight generalization of 5-colorability. Although every planar graph is 4-colorable and therefore 5-colorable, not every planar graph is 4-choosable. If this all sounds like nonsense, don't panic. We'll discuss graphs, planarity, and coloring in two weeks.

This a statement about all integers $n \geq 1$ rather ≥ 0, so it's natural to use a slight variation on induction: prove $P(1)$ in the base case and then prove that $P(n)$ implies $P(n+1)$ for all $n \geq 1$ in the inductive step. This is a perfectly valid variant of induction and is *not* the problem with the proof below.

Proof. The proof is by induction on n. The induction hypothesis, $P(n)$, will be

$$\text{In every set of } n \text{ horses, all are the same color.} \tag{3.7}$$

Base case: ($n = 1$). $P(1)$ is true, because in a set of horses of size 1, there's only one horse, and this horse is definitely the same color as itself.

Inductive step: Assume that $P(n)$ is true for some $n \geq 1$. that is, assume that in every set of n horses, all are the same color. Now consider a set of $n+1$ horses:

$$h_1,\ h_2,\ \ldots,\ h_n,\ h_{n+1}$$

By our assumption, the first n horses are the same color:

$$\underbrace{h_1,\ h_2,\ \ldots,\ h_n,}_{\text{same color}}\ h_{n+1}$$

Also by our assumption, the last n horses are the same color:

$$h_1,\ \underbrace{h_2,\ \ldots,\ h_n,\ h_{n+1}}_{\text{same color}}$$

So h_1 is the same color as the remaining horses besides h_{n+1}, and likewise h_{n+1} is the same color as the remaining horses besides h_1. So h_1 and h_{n+1} are the same color. That is, horses $h_1, h_2, \ldots, h_{n+1}$ must all be the same color, and so $P(n+1)$ is true. Thus, $P(n)$ implies $P(n+1)$.

By the principle of induction, $P(n)$ is true for all $n \geq 1$. □

We've proved something false! Is Math broken? Should we all become poets?

The error in this argument is in the sentence that begins, "So h_1 and h_{n+1} are the same color." The "..." notation creates the impression that there are some remaining horses besides h_1 and h_{n+1}. However, this is not true when $n = 1$. In that case, the first set is just h_1 and the second is h_2, and there are no remaining horses besides them. So h_1 and h_2 need not be the same color!

This mistake knocks a critical link out of our induction argument. We proved $P(1)$ and we *correctly* proved $P(2) \longrightarrow P(3)$, $P(3) \longrightarrow P(4)$, etc. But we failed to prove $P(1) \longrightarrow P(2)$, and so everything falls apart: we can not conclude that $P(2)$, $P(3)$, etc., are true. And, of course, these propositions are all false; there are horses of a different color.

Students sometimes claim that the mistake in the proof is because $P(n)$ is false for $n \geq 2$, and the proof assumes something false, namely, $P(n)$, in order to prove $P(n+1)$. You should think about how to explain to such a student why this claim would get no credit on a 6.042 exam.

3.5 Strong Induction

3.5.1 The Strong Induction Principle

A useful variant of induction is called *strong induction*. Strong induction and ordinary induction are used for exactly the same thing: proving that a predicate $P(n)$ is true for all $n \in \mathbb{N}$.

Principle of Strong Induction. Let $P(n)$ be a predicate. If

- $P(0)$ is true, and
- for all $n \in \mathbb{N}$, $P(0)$, $P(1)$, ..., $P(n)$ *together* imply $P(n+1)$,

then $P(n)$ is true for all $n \in \mathbb{N}$.

The only change from the ordinary induction principle is that strong induction allows you to assume more stuff in the inductive step of your proof! In an ordinary induction argument, you assume that $P(n)$ is true and try to prove that $P(n+1)$ is also true. In a strong induction argument, you may assume that $P(0)$, $P(1)$, ..., and $P(n)$ are *all* true when you go to prove $P(n+1)$. These extra assumptions can only make your job easier.

3.5.2 Products of Primes

As a first example, we'll use strong induction to prove one of those familiar facts that is almost, but maybe not entirely, obvious:

Lemma 3.5.1. *Every integer greater than 1 is a product of primes.*

Note that, by convention, any number is considered to be a product consisting of one term, namely itself. In particular, every prime is considered to be a product whose terms are all primes.

Proof. We will prove Lemma 3.5.1 by strong induction, letting the induction hypothesis, $P(n)$, be

$$n+2 \text{ is a product of primes.}$$

So Lemma 3.5.1 will follow if we prove that $P(n)$ holds for all $n \geq 0$.

Base Case: $P(0)$ is true because $0+2$ is prime, and so is a product of primes by convention.

Inductive step: Suppose that $n \geq 0$ and that $i+2$ is a product of primes for every natural number $i < n+1$. We must show that $P(n+1)$ holds, namely, that $n+3$ is also a product of primes. We argue by cases:

If $n+3$ is itself prime, then it is a product of primes by convention, so $P(n+1)$ holds in this case.

Otherwise, $n+3$ is not prime, which by definition means $n+3 = km$ for some natural numbers k, m such that $2 \leq k, m < n+3$. So $k-2$ is a natural number less than $n+1$, which means

that $(k-2)+2$ is a product of primes by induction hypothesis. That is, k is a product of primes. Likewise, m is a product of primes. So $km = n+3$ is also a product of primes. Therefore, $P(n+1)$ holds in this case as well.

So $P(n+1)$ holds in any case, which completes the proof by strong induction that $P(n)$ holds for all natural numbers, n.

□

Despite the name, strong induction is actually no more powerful than ordinary induction. In other words, any theorem that can be proved with strong induction could also be proved with ordinary induction (using a slightly more complicated induction hypothesis). But strong induction can make some proofs a bit easier. On the other hand, if $P(n)$ is easily sufficient to prove $P(n+1)$, then it's better to use ordinary induction for simplicity.

3.5.3 Making Change

The country Inductia, whose unit of currency is the Strong, has coins worth 6S (6 Strongs), 10S and 15S. Although the Inductians have some trouble making small change like 11S or 29S, it turns out that they can collect coins to make change for any number of Strongs greater than 29S.

Strong induction makes this easy to prove for $n+1 > 35$, because then $(n+1)-6 > 29$, so by strong induction the Inductians can make change for exactly $((n+1)-6)$S, and then they can add a 6S coin to get $(n+1)$S. So the only thing to do is check that they can make change for all the amounts from 30 to 35, which is not too hard to do.

Here's a detailed writeup using the official format:

Proof. We prove the Inductians can make change for any amount greater than 29S by strong induction. The induction hypothesis, $P(n)$ will be:

> If $n > 29$, then there is a collection of coins whose value is n Strongs.

Notice that $P(n)$ is an implication. When the hypothesis of an implication is false, we know the whole implication is true. In this situation, the implication is said to be *vacuously* true. So $P(n)$ will be vacuously true whenever $n \leq 29$.[6]

We now proceed with the induction proof:

Base case: $P(0)$ is vacuously true.

Inductive step: We assume $P(i)$ holds for all $i \leq n$, and prove that $P(n+1)$ holds. We argue by cases:

Case ($n+1 \leq 29$): $P(n+1)$ is vacuously true in this case.

Case ($n+1 = 30$): $P(30)$ holds because the Inductians can use five 6S coins.

[6] A more elegant approach that avoids these vacuous cases is to define

$$P'(n) ::= \text{there is a collection of coins whose value is } n+30 \text{ Strongs}$$

and prove that $P'(n)$ holds for all $n \geq 0$.

Case ($n + 1 = 31$): Use a 6S coin, a 10S coin and a 15S coin.

Case ($n + 1 = 32$): Use two 6S coins, and two 10S coins.

Case ($n + 1 = 33$): Use three 6S coins, and a 15S coin.

Case ($n + 1 = 34$): Use a four 6S coins, and a 10S coin.

Case ($n + 1 = 35$): Use a two 10S coins and a 15S coin.

Case ($n+1 > 35$): Then $n \geq (n+1)-6 > 29$, so by the strong induction hypothesis, the Inductians can make change for $((n + 1) - 6)$S. Now by adding a 6S coin, they can make change for $(n + 1)$S.

So in any case, $P(n + 1)$ is true, and we conclude by strong induction that for all $n > 29$, the Inductians can make change for nS.

□

3.5.4 Unstacking

Here is another exciting 6.042 game that's surely about to sweep the nation!

You begin with a stack of n boxes. Then you make a sequence of moves. In each move, you divide one stack of boxes into two nonempty stacks. The game ends when you have n stacks, each containing a single box. You earn points for each move; in particular, if you divide one stack of height $a + b$ into two stacks with heights a and b, then you score ab points for that move. Your overall score is the sum of the points that you earn for each move. What strategy should you use to maximize your total score?

As an example, suppose that we begin with a stack of $n = 10$ boxes. Then the game might proceed as follows:

Stack Heights										**Score**
$\underline{10}$										
5	$\underline{5}$									25 points
$\underline{5}$	3	2								6
$\underline{4}$	3	2	1							4
2	$\underline{3}$	2	1	2						4
$\underline{2}$	2	2	1	2	1					2
1	$\underline{2}$	2	1	2	1	1				1
1	1	$\underline{2}$	1	2	1	1	1			1
1	1	1	1	$\underline{2}$	1	1	1	1		1
1	1	1	1	1	1	1	1	1	1	1
								Total Score	$=$	45 points

On each line, the underlined stack is divided in the next step. Can you find a better strategy?

Analyzing the Game

Let's use strong induction to analyze the unstacking game. We'll prove that your score is determined entirely by the number of boxes —your strategy is irrelevant!

Theorem 3.5.2. *Every way of unstacking n blocks gives a score of $n(n-1)/2$ points.*

There are a couple technical points to notice in the proof:

- The template for a strong induction proof is exactly the same as for ordinary induction.
- As with ordinary induction, we have some freedom to adjust indices. In this case, we prove $P(1)$ in the base case and prove that $P(1), \ldots, P(n)$ imply $P(n+1)$ for all $n \geq 1$ in the inductive step.

Proof. The proof is by strong induction. Let $P(n)$ be the proposition that every way of unstacking n blocks gives a score of $n(n-1)/2$.

Base case: If $n = 1$, then there is only one block. No moves are possible, and so the total score for the game is $1(1-1)/2 = 0$. Therefore, $P(1)$ is true.

Inductive step: Now we must show that $P(1), \ldots, P(n)$ imply $P(n+1)$ for all $n \geq 1$. So assume that $P(1), \ldots, P(n)$ are all true and that we have a stack of $n+1$ blocks. The first move must split this stack into substacks with positive sizes a and b where $a+b = n+1$ and $0 < a, b \leq n$. Now the total score for the game is the sum of points for this first move plus points obtained by unstacking the two resulting substacks:

$$\begin{aligned}
\text{total score} &= (\text{score for 1st move}) \\
&\quad + (\text{score for unstacking } a \text{ blocks}) \\
&\quad + (\text{score for unstacking } b \text{ blocks}) \\
&= ab + \frac{a(a-1)}{2} + \frac{b(b-1)}{2} && \text{by } P(a) \text{ and } P(b) \\
&= \frac{(a+b)^2 - (a+b)}{2} = \frac{(a+b)((a+b)-1)}{2} \\
&= \frac{(n+1)n}{2}
\end{aligned}$$

This shows that $P(1), P(2), \ldots, P(n)$ imply $P(n+1)$.

Therefore, the claim is true by strong induction. □

Problem 3.5.1. Define the ***potential***, $p(S)$, of a stack, S, of blocks to be $k(k+1)/2$ where k is the number of blocks in S. Define the potential, $p(A)$, of a set, A, of stacks to be the sum of the potentials of the stacks in A.

Generalize Theorem 3.5.2 to show that for any set, A, of stacks, if a sequence of moves starting with A leads to another set, B, of stacks, then the score for this sequence of moves is $p(A) - p(B)$.

3.6 The Well Ordering Principle

Another proof method closely related to induction depends on the

Well Ordering Principle. Every *nonempty* set of *nonnegative integers* has a *smallest* element.

Do you believe this statement? Seems sort of obvious, right? But notice how tight it is: it requires a *nonempty* set —it's false for the empty set which has *no* smallest element because it has no elements at all! And it requires a set of *nonnegative* integers —it's false for the set of *negative* integers and also false for some sets of nonnegative *rationals* —for example, the set of positive rationals. So, the Well Ordering Principle captures something special about the natural numbers.

While the Well Ordering Principle may seem obvious, it looks nothing like the induction axiom, and it's harder to see offhand why it is useful. But in fact, it's as powerful as strong induction. We'll explain this after we introduce a template for well ordering principle proofs resembling the template in Section 3.4.1 for a proof by strong induction.

In fact, looking back, we took the Well Ordering Principle for granted in proving Lemma 3.2.11, that every finite partial order has a minimal element. We even implicitly relied on the Well Ordering Principle in the proof in Week 2 Notes that $\sqrt{2}$ is irrational. That proof assumed that any nonzero fraction can be written in *lowest terms*, that is, in the form m/n where m and n are integers with no common factors. How do we know this is always possible?

Suppose to the contrary that there is a nonzero fraction that cannot be written in lowest terms. Now let C be the set of positive integers that are numerators of such fractions. Then C is nonempty. To prove this, suppose m/n is a nonzero fraction that cannot be written in lowest terms. Then neither can $-m/(-n)$, so one of m or $-m$ must be in C.

Therefore, by Well Ordering, there must be a smallest integer, $m_0 \in C$. So by definition of C, there must be some integer, n_0, such that the fraction m_0/n_0 cannot be in written in lowest terms. This means that m_0 and n_0 must have a common factor, $p > 1$. But then $(m_0/p)/(n_0/p)$ cannot be in lowest terms either, since it equals m_0/n_0. So $m_0/p \in C$ by definition of C. But $m_0/p < m_0$, which contradicts the fact that m_0 is the smallest element of C.

Since the assumption that C is nonempty leads to a contradiction, it follows that C must be empty. That is, that there are no numerators of fractions that can't be written in lowest terms, and hence there are no such fractions at all.

We've been using the Well Ordering Principle on the sly from early on!

Here is a standard way to organize a well ordering proof.

To prove that "$P(n)$ is true for all $n \in \mathbb{N}$" using the Well Ordering Principle:

- Define the set, C, of *counterexamples* to P being true. Namely, define

$$C ::= \{n \in \mathbb{N} \mid \neg P(n)\}.$$

- Assume for proof by contradiction that C is nonempty.
- By the Well Ordering Principle, there will be a smallest element, s, in C.
- Reach a contradiction (somehow) —often by showing how to use s to find another member of C that is smaller than s. (This is the open-ended part of the proof task.)
- Conclude that C must be empty, that is, no counterexamples exist. QED

Now we can explain why the Well Ordering Principle is as powerful a proof method as Strong Induction. In fact, we will explain hwo to take any proof by Strong Induction and reformat it into a Well Ordering proof.

Here's how: suppose that we have a proof Strong Induction with induction hypothesis $P(n)$. Then we start a Well Ordering proof by defining the set of counterexamples to P, and then assuming there is a smallest counterexample, s. This means that $P(s)$ is false, but also $P(0), P(1), \ldots, P(s-1)$ are all true. At this point we reuse the proof of the inductive step in the Strong Induction proof, which shows that since $P(0), P(1), \ldots, P(s-1)$ are all true, then $P(s)$ is also true. This contradicts the assumption that $P(s)$ is false, so we have the contradiction needed to complete the Well Ordering Proof that $\forall n.\, P(n)$.

Problem 3.6.1. Use strong induction to prove the Well Ordering Principle. *Hint:* Prove that if a set of nonnegative integers contains an integer, n, then it has a smallest element.

Mathematicians commonly use the Well Ordering Principle because it can lead to shorter proofs than induction. On the other hand, well ordering proofs typically involve proof by contradiction, so using it is not always the best approach. The choice of method is really a matter of style—but style does matter.

3.7 In-Class Problems Week 3, Tue.

Problem 3.7.1. A pair of 6.042 TAs, Chiyoun and Tina, have decided to devote some of their spare time this term to establishing dominion over the entire galaxy. Recognizing this as an ambitious project, they worked out the following table of tasks on the back of Tina's copy of the lecture notes.

1. **Devise a logo** and cool imperial theme music - 8 days.
2. **Build a fleet** of Hyperwarp Stardestroyers out of eating paraphernalia swiped from Lobdell - 18 days.
3. **Seize control** of the United Nations - 9 days, after task #1.
4. **Get shots** for Chiyoun's cat, Tailspin - 11 days, after task #1.
5. **Open a Starbucks chain** for the army to get their caffeine - 10 days, after task #3.
6. **Train an army** of elite interstellar warriors by dragging people to see *The Phantom Menace* dozens of times - 4 days, after tasks #3, #4, and #5.
7. **Launch the fleet** of Stardestroyers, crush all sentient alien species, and establish a Galactic Empire - 6 days, after tasks #2 and #6.
8. **Defeat Microsoft** - 8 days, after tasks #2 and #6.

We picture this information in Figure 1 below by drawing a point for each task, and labelling it with the name and weight of the task. An edge between two points indicates that the task for the higher point must be completed before beginning the task for the lower one.

(a) Give some valid order in which the tasks might be completed.

Solution. We can easily find several of them. The most natural one is valid, too: #1, #2, #3, #4, #5, #6, #7, #8. ■

Chiyoun and Tina want to complete all these tasks in the shortest possible time. However, they have agreed on some constraining work rules.

- Only one person can be assigned to a particular task; they can not work together on a single task.
- Once a person is assigned to a task, that person must work exclusively on the assignment until it is completed. So, for example, Chiyoun cannot work on building a fleet for a few days, run to get shots for Tailspin, and then return to building the fleet.

(b) Chiyoun and Tina want to know how long conquering the galaxy will take. Tina suggests dividing the total number of days of work by the number of workers, which is two. What lower bound on the time to conquer the galaxy does this give, and why might the actual time required be greater?

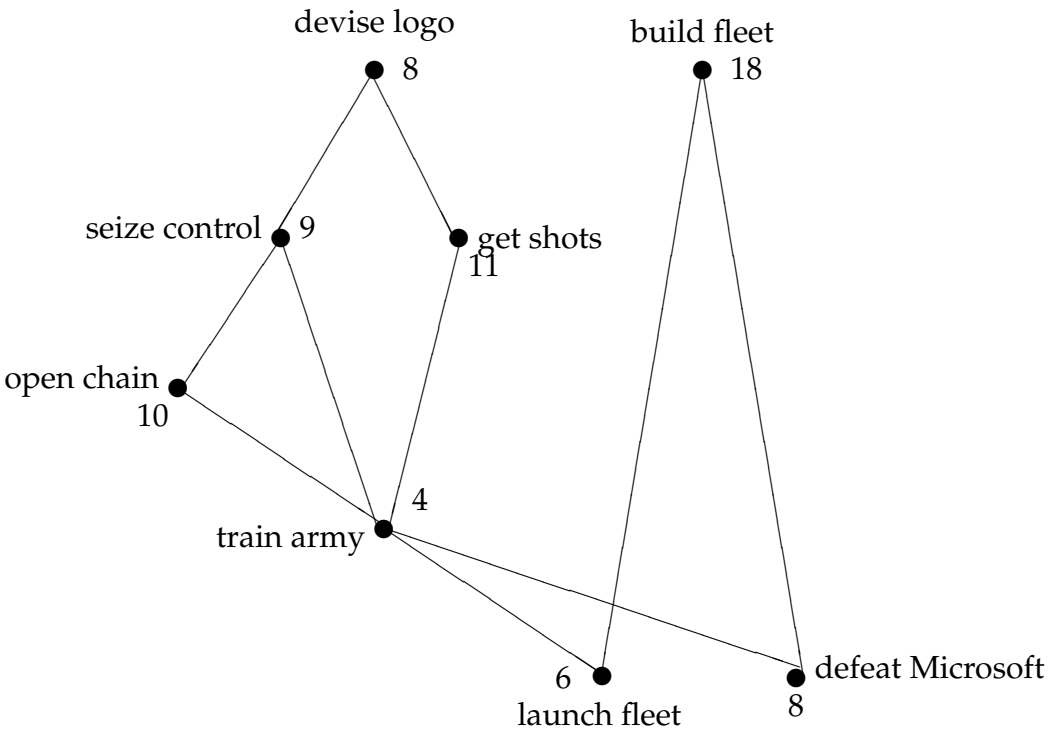


Figure 3.1: Graph representing the task precedence constraints.

Solution.

$$\frac{8+18+9+11+10+4+6+8}{2} = 37 \text{ days}$$

If working together and interrupting work on a task were permitted, then this answer would be correct. However, the rules may prevent Chiyoun and Tina from both working all the time. For example, suppose the only task was building the fleet. It will take 18 days, not 18/2 days, to complete, because only one person can work on it and the other must sit idle. ■

(c) Chiyoun proposes a different method for determining the duration of their project. He suggests looking at the duration of the "critical path", the most time-consuming sequence of tasks such that each depends on the one before. What lower bound does this give, and why might it also be too low?

Solution. The longest sequence of tasks is devising a logo (8 days), seizing the U. N. (9 days), opening a Starbucks (10 days), training the army (4 days), and then defeating Microsoft (8 days). Since these tasks must be done sequentially, galactic conquest will require at least 39 days.

If there were enough workers, this answer would be correct; however, with only two workers, Chiyoun and Tina may be unable to make progress on the critical path every day. For example, suppose there were only four tasks: devise logo, build fleet, seize control, get shots. Now the critical path consists of two critical tasks: devise logo, get shots, which take 19 days. But to get through this path in 19 days, some worker must be working on a critical task at all times for the 19 days. This leaves only one worker free to complete building the fleet and seizing control, which

will take at least 27 days. So in fact, 27 days is the minimum time for two workers to complete these four tasks. ■

(d) What is the minimum number of days that Chiyoun and Tina need to conquer the galaxy? No proof is required.

Solution. 40 days. Tasks could be divided as follows:

Tina: #1 (days 1-8), #3 (days 9-17), #4 (days 18-28), #8 (days 33-40).

Chiyoun: #2 (days 1-18), #5 (days 19-28), #6 (days 29-32), #7 (days 33-38). ■

Problem 3.7.2. Verify that each of the following relations is a partial order. For each, indicate if it is strict, weak, and/or a total order. (Definitions are in the Appendix.)

(a) The superset relation, $\supseteq$, on some family of sets.

Solution. To prove $\supseteq$ is transitive, suppose $A \supseteq B$ and $B \supseteq C$. Then by definition of $\supseteq$, every $b \in B$ is also in A, and every $c \in C$ is also in B. Hence, every $c \in C$ is also in A, which proves that $A \supseteq C$. This proves that $\supseteq$ is transitive. It is also antisymmetric, because $A \supseteq B$ and $B \supseteq A$ implies $A = B$. Finally, it is reflexive, since $A \supseteq A$. So $\supseteq$ is a weak partial order.

There are lots of pairs of sets neither of which contains the other, so $\supseteq$ will not, in general, be a total order. ■

(b) The "divides" relation on natural numbers.

Solution. Suppose $k \mid m$ and $m \mid n$. So there are integers, k_1, k_2 such that $m = k_1 k$ and $n = k_2 m$. Hence, $n = k_2(k_1 k) = (k_2 k_1)k$, which means that $k \mid n$. So "divides" is transitive. Similarly, if $m \mid n$ and $n \mid m$, then $m = n$, so divides is antisymmetric, and so is a partial order. It is weak, because every integer divides itself. It is not total, since any two primes, for example, are incomparable. ■

Problem 3.7.3. **(a)** What are the maxim*al* and minim*al* elements, if any, of the set, $\mathbb{N}$, of all nonnegative integers under divisibility? Is there a minim*um* or maxim*um* element?

Solution. The minimum (and therefore unique minimal) element is 1 since 1 divides all natural numbers. The maximum (and therefore unique maximal) element is 0 since all numbers divide 0. ■

(b) What are the minimal and maximal elements, if any, of the set of integers ≥ 2 under divisibility?

Solution. All prime numbers are minimal elements, since no numbers divide them.

There is no maximal element, because for any $n \geq 2$, there is a "larger" number under the divisibility partial order, namely, mn, for any $m > 1$. ■

(c) What is the size of the longest chain that is guaranteed to exist in any partially ordered set of $n > 0$ elements? What about the largest antichain?

Solution. Chain size is 1 in the "discrete" partial order in which every two distinct elements are incomparable. Antichain size is 1 if the partial order is total. ■

(d) Describe a partially ordered set that has no minimal or maximal elements.

Solution. $\mathbb{Z}$, $\mathbb{R}$, etc. ■

(e) Describe a partially ordered set that has a *unique minimal* element, but no minimum element. *Hint:* It will have to be infinite.

Solution. $\mathbb{Z} \cup \{i\}$ where i is a root of -1, under the usual order $\mathbb{Z}$. So i is incomparable to everything but itself, and is therefore minimal. ■

Problem 3.7.4. Prove that a binary relation, R, on a set, A, is a strict partial order iff it is transitive and irreflexive.

Solution. For the left-to-right direction of the "iff": Suppose R is a strict partial order; namely, R is transitive and asymmetric. We need to prove that it is transitive and irreflexive. That R is transitive we know already. To prove that it is irreflexive, we need to prove that

$$\text{for every } a \in A: \quad aRa \text{ is false.}$$

So, pick any $a \in A$. Towards a contradiction, assume that aRa is true. Then, by the asymmetry of R, we know that $\neg(aRa)$ is also true. So, we have both aRa and $\neg(aRa)$. This is a contradiction. Hence, R must be irreflexive.

For the right-to-left direction of the "iff": Suppose R is transitive and irreflexive. To show that R is a strict partial order, we need to show that it is transitive and asymmetric. That R is transitive we know already. To prove that it is asymmetric, we need to prove that

$$\text{for every } a, b \in A: \quad aRb \longrightarrow \neg(bRa).$$

So, pick any $a, b \in A$. We need to prove that the implication $aRb \longrightarrow \neg(bRa)$ holds. Towards a contradiction, assume the implication is false. In other words, assume that both aRb and bRa are true. Then we would also have aRa (by transitivity). But this would contradict the fact that R is irreflexive. Therefore, R must be asymmetric. ■

Appendix

Relational Properties

A binary relation, R, on a set, A, is

- *transitive* if for every $a, b, c \in A$, aRb and bRc implies aRc.

- *asymmetric* if for every $a, b \in A$, aRb implies $\neg(bRa)$,
- *reflexive* if aRa for every $a \in A$,
- *antisymmetric* if for every $a \neq b \in A$, aRb implies $\neg(bRa)$,
- *irreflexive* if aRa holds for no $a \in A$.

Partial Order

A binary relation is a *strict partial order* iff it is transitive and asymmetric. It is a *weak partial order* iff it is transitive, reflexive, and antisymmetric.

Let $\preceq$ be a weak (reflexive) partial order on a set, A.

- An element $a \in A$ is *minimal* iff there is no element in A that is $\preceq a$ except possibly a itself. Similarly, an element $a \in A$ is *maximal* iff there is no element in A that is $\succeq a$ except possibly a itself.
- An element $a \in A$ is a *lower bound* for a subset, $S \subseteq A$ iff $a \preceq s$ for every $s \in S$. Similarly, an element $a \in A$ is an *upper bound* for a subset, $S \subseteq A$ iff $s \preceq a$ for every $s \in S$.
- An element $a \in A$ is the *minimum* element iff a is a lower bound on A. Similarly, an element $a \in A$ is the *maximum* element iff a is an upper bound on A.
- Elements $a, b \in A$ are *comparable* iff either $a \preceq b$ or $b \preceq a$. Two elements are *incomparable* iff they are not comparable.
- A subset, $S \subseteq A$ is *totally ordered* iff every two distinct elements in S are comparable.
- A *chain* is a totally ordered subset of A.
- An *antichain* is a subset of A, such that no two elements in it are comparable.

3.8 In-Class Problems Week 3, Wed.

Problem 3.8.1. Use induction to prove that

$$1^3 + 2^3 + \cdots + n^3 = \left(\frac{n(n+1)}{2}\right)^2. \tag{3.8}$$

for all $n \geq 1$.

Remember to formally

1. Declare proof by induction.
2. Identify the induction hypothesis $P(n)$.
3. Establish the base case.
4. Prove that $P(n) \Rightarrow P(n+1)$.
5. Conclude that $P(n)$ holds for all $n \geq 1$.

as in the five part template.

Solution. We proceed by induction. The induction hypothesis, $P(n)$, will be the equation (3.8).

Base case: First, we must show that $P(1)$ is true. This is immediate, since:

$$1^3 = \left(\frac{1(1+1)}{2}\right)^2$$

Inductive step: Next, we must show that $P(n)$ implies $P(n+1)$ for all $n \geq 1$. Assuming that $P(n)$ is true, we can reason as follows:

$$\begin{aligned} 1^3 + 2^3 + \cdots + n^3 + (n+1)^3 &= \left(\frac{n(n+1)}{2}\right)^2 + (n+1)^3 \\ &= \left(\frac{(n+1)(n+2)}{2}\right)^2 \end{aligned}$$

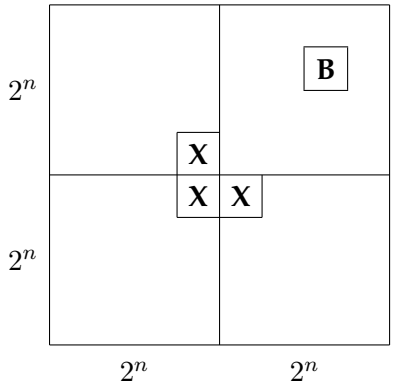
The first step uses the assumption $P(n)$, and the second uses only algebra. This shows that $P(n+1)$ is true. Therefore, $P(n)$ is true for all $n \geq 1$ by induction. ■

Problem 3.8.2. **(a)** Prove by induction that a $2^n \times 2^n$ courtyard with a 1×1 statue of Bill in *any position* can be covered with L-shaped tiles.

Solution. Let $P(n)$ be the proposition that for every location of Bill in a $2^n \times 2^n$ courtyard, there exists a tiling of the remainder.

Base case: $P(0)$ is true because Bill fills the whole courtyard.

Inductive step: Asume that $P(n)$ is true for some $n \geq 0$; that is, for every location of Bill in a $2^n \times 2^n$ courtyard, there exists a tiling of the remainder. Divide the $2^{n+1} \times 2^{n+1}$ courtyard into four quadrants, each $2^n \times 2^n$. One quadrant contains Bill (**B** in the diagram below). Place a temporary Bill (**X** in the diagram) in each of the three central squares lying outside this quadrant:



Now we can tile each of the four quadrants by the induction assumption. Replacing the three temporary Bills with a single L-shaped tile completes the job. This proves that $P(n)$ implies $P(n+1)$ for all $n \geq 0$. The theorem follows as a special case.

This proof has two nice properties. First, not only does the argument guarantee that a tiling exists, but also it gives a recursive procedure for finding such a tiling. Second, we have a stronger result: if Bill wanted a statue on the edge of the courtyard, away from the pigeons, we could accommodate him! ■

(b) *(Discussion Question)* In part (a) we saw that it can be easier to prove a stronger theorem. Does this surprise you? How would you explain this phenomenon?

Solution. It might seem that it ought to be harder to prove a more general theorem than a less general one, but sometimes not. For example, the more general result might actually be easier because it involves fewer assumptions, and this can help in avoiding the complications of unnecessary hypotheses.

But for an induction proof in particular, using a more general induction hypothesis means we can make a stronger *assumption* in the induction step (namely, we can assume a stronger $P(n)$), which can make it easier to prove the conclusion of the induction step (namely, $P(n+1)$). ■

Problem 3.8.3. Notes 2 contains a proof by induction that:

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}$$

But now we're going to prove a *contradictory* theorem!

False Theorem. *For all $n \geq 0$,*

$$2 + 3 + 4 + \cdots + n = \frac{n(n+1)}{2}$$

Proof. We use induction. Let $P(n)$ be the proposition that $2 + 3 + 4 + \cdots + n = n(n+1)/2$.

Base case: $P(0)$ is true, since both sides of the equation are equal to zero. (Recall that a sum with no terms is zero.)

Inductive step: Now we must show that $P(n)$ implies $P(n+1)$ for all $n \geq 0$. So suppose that $P(n)$ is true; that is, $2 + 3 + 4 + \cdots + n = n(n+1)/2$. Then we can reason as follows:

$$\begin{aligned} 2 + 3 + 4 + \cdots + n + (n+1) &= \left[2 + 3 + 4 + \cdots + n\right] + (n+1) \\ &= \frac{n(n+1)}{2} + (n+1) \\ &= \frac{(n+1)(n+2)}{2} \end{aligned}$$

Above, we group some terms, use the assumption $P(n)$, and then simplify. This shows that $P(n)$ implies $P(n+1)$. By the principle of induction, $P(n)$ is true for all $n \in \mathbb{N}$. □

Where exactly is the error in this proof?

Solution. The short answer is that we failed to prove $P(0) \longrightarrow P(1)$, just as in the colored horses problem in lecture. In fact, once again, the error is rooted in the misleading nature of the "$\cdots$" notation.

More precisely, in the inductive step we are required to prove that $P(n)$ implies $P(n+1)$ for all $n \geq 0$. However, the argument given above breaks down when $n = 0$. Let's look more closely at the first equation in the indutive step to see why:

$$2 + 3 + 4 + \cdots + n + (n+1) = \left[2 + 3 + 4 + \cdots + n\right] + (n+1)$$

This seems completely innucuous; after all, we've only grouped terms! However, the left side contains *no terms* when $n = 0$. The "$\cdots$" is completely misleading in this case; 2, 3, 4, and $n+1$ are actually *not* in the sum. This misimpression becomes an error when we "pull out" the $(n+1)$ term on the right side, disregarding the fact that no such term actually existed on the left. Thus, for $n = 0$, the equation we've just written down says:

$$\underbrace{2 + 3 + 4 + \cdots + n + (n+1)}_{=0} = \left[\underbrace{2 + 3 + 4 + \cdots + n}_{=0}\right] + \underbrace{(n+1)}_{=1}$$

The assertion $0 = 0 + 1$ is false, and so we have not shown that $P(0)$ implies $P(1)$. There is no way to fix this problem and correctly prove that $P(0)$ implies $P(1)$, because actually $P(0)$ is true and $P(1)$ is false.

Thus, we've only established $P(0)$, $P(1) \longrightarrow P(2)$, $P(2) \longrightarrow P(3)$, and so forth. The induction argument falls apart because of the missing link $P(0) \not\longrightarrow P(1)$. ■

3.9 In-Class Problems Week 3, Fri.

Problem 3.9.1. Given an unlimited supply of 5 cent and 7 cent stamps, what postages are possible? Prove it using Strong Induction. *Hint:* Try some examples! Which postage values between 1 and 25 cents can you construct from 5 cent and 7 cent stamps?

Solution. Let's use our examples to first try to guess the answer and then try to prove it. Let's begin filling in a table that shows the values of all possible combinations of 5 and 7 cent stamps. The column heading is the number of 5 cent stamps and the row heading is the number of 7 cent stamps.

	0	1	2	3	4	5	...
0	0	5	10	15	20	25	...
1	7	12	17	22	27	...	
2	14	19	24	29	...		
3	21	26	31	36	...		
4	28	33	38	...			
5	35	40	...				
⋮	⋮	⋮					

Looking at the table, a guess is that the possible postages are those shown and every value of 24 or more cents. Let's try to prove this last part using strong induction.

Claim 3.9.1. *For all $n \geq 24$, it is possible to produce n cents of postage from 5¢ and 7¢ stamps.*

Now let's preview the proof. The induction hypothesis will be

$$P(n) \quad ::= \quad n + 24¢ \text{ postage can be produced using } 5¢ \text{ and } 7¢ \text{ stamps.} \tag{3.9}$$

A proof by strong induction will have the same five-part structure as an ordinary induction proof. The base case, $P(0)$, is that 24¢ postage can be made, and from the table, it can, using two 5¢ and two 7¢ stamps.

In the inductive step we have to show how to produce $(n+1)+24$ cents of postage, assuming the strong induction hypothesis that we know how to produce $k + 24¢$ of postage for all values of k between 0 and n. A simple way to do this is to let $k = (n-4)+24$, produce $k¢$ of postage, then add a 5¢ stamp to get $(n+1)+24$ cents.

But we have to be careful. If $n < 4$, this method may not work, because we would have to produce postage of less than 24¢, which may not be possible. So in this case we cannot use the trick of creating $(n+1)+24$ cents of postage from $n-4+24$ cents and a 5 cent stamp. Fortunately, making $(n+1)+24$ cents of postage for $n < 4$ can easily be done directly.

Proof. The proof is by strong induction. The induction hypothesis, $P(n)$, is given by (3.9).

Base case: $n = 0$: $P(0)$ holds because we can make 24¢ postage using two 5¢ and two 7¢ stamps.

Inductive step: In the inductive step, we assume that it is possible to produce postage worth $24, 25, \ldots, n + 24$ cents in order to prove that it is possible to produce postage worth $(n + 1) + 24$ cents.

There are five cases:

1. $n + 1 = 1$: To make $(n + 1) + 24 = 25¢$ postage, use five $5¢$ stamps
2. $n + 1 = 2$: To make $26¢$ postage, use one $5¢$ stamp and three $7¢$ stamps.
3. $n + 1 = 3$: To make $27¢$ postage, use four $5¢$ stamps and one $7¢$ stamp.
4. $n + 1 = 4$: To make $28¢$ postage, use four $7¢$ stamps.
5. $n + 1 > 4$: We have $n \geq 4$, so $n - 4 \geq 0$ and by strong induction we may assume we can produce exactly $(n - 4) + 24$ cents of postage. With an additional $5¢$ stamp we can therefore produce $(n + 1) + 24$ cents of postage.

So in every case, $P(0) \wedge P(1) \wedge \ldots P(n) \longrightarrow P(n + 1)$. By strong induction, we have concluded that $P(n)$ is true for all $n \in \mathbb{N}$. □

■

Problem 3.9.2. We consider integer solutions to the equation

$$4a^3 + 2b^3 = c^3 \tag{3.10}$$

(a) Use the Well-ordering Principle to prove that there is no integer solution to equation (3.10) with $a > 0$.

Solution. The proof is by contradiction.

Let S be the set of all positive integers, a, such that there exist integers, b, and, c, that satisfy equation (3.10).

Assume for the purpose of obtaining a contradiction that S is nonempty. Then S contains a smallest element, $a_0 > 0$, by the Well-ordering Principle. By the definition of S, there exist corresponding integers, b_0, and, c_0, such that:

$$4a_0^3 + 2b_0^3 = c_0^3.$$

The left side of this equation is even, so c_0^3 is even, and therefore c_0 is also even. Thus, there exists an integer, c_1, such that $c_0 = 2c_1$. Now substituting $2c_1$ for c_0 in this equation and then dividing both sides by 2 gives:

$$2a_0^3 + b_0^3 = 4c_1^3, \tag{3.11}$$

so

$$b_0^3 = 2(2c_1^3 - a_0^3).$$

This impies that b_0^3 is even, so b_0 is even. Thus, there exists an integer, b_1, such that $b_0 = 2b_1$. Substituting $2b_1$ for b_0 in (3.11) and dividing both sides by 2 gives:

$$a_0^3 + 4b_1^3 = 2c_1^3 \tag{3.12}$$

From (3.12), we conclude that a_0^3 is even, so a_0 is also even. Thus, there exists an integer, a_1 such that $a_0 = 2a_1$, where $a_1 > 0$ since $a_0 > 0$. Substituting $2a_1$ for a_0 in (3.12) and dividing by 2 one final time gives:

$$4a_1^3 + 2b_1^3 = c_1^3$$

So evidently, $a = a_1$, $b = b_1$, and $c = c_1$ is another solution to the original equation (3.10), and so a_1 is an element of S. But this is a contradiction, because $a_1 < a_0$ and a_0 was defined to be the smallest element of S. Therefore, our assumption was wrong, and the original equation has no integer solutions with $a > 0$. ■

(b) Show that the only integer solution to equation (3.10) is $a = b = c = 0$.

Solution. There are two cases left to consider: when $a < 0$ or when $a = 0$.

So let S' be the set of all integers, $n > 0$, such that there is an integer solution to equation (3.10) with $a = -n$. The previous argument now goes through with S' in place of S, proving that S' is empty. So there is no solution with to equation (3.10) with $a < 0$.

Finally, with $a = 0$, to solve (3.10), we need integers b, c such that

$$2b^3 = c^3. \tag{3.13}$$

Now if $b = 0$ in (3.13), then obviously $c = 0$, giving us the all-zero solution. On the other hand, if $b \neq 0$ in (3.13), then c/b would be a rational cube root of 2, and we know there is none. So the only possibility is the all-zero solution. ■

3.10 Problem Set 2

Problem 3.10.1. Consider the proper subset partial order, $\subset$, on the power set $\mathcal{P}\{1, 2, \ldots 5\}$.

(a) What is the size of a maximal chain in this partial order? Describe one.

Solution. Size 6, for example,

$$\{\emptyset, \{1\}, \{1, 2\}, \{1, 2, 3\}, \{1, 2, 3, 4\}, \{1, 2, 3, 4, 5\}\}.$$

■

(b) Describe the largest antichain you can find in this partial order.

Solution. All the size 3 subsets of $\{1, 2, \ldots 5\}$ form an antichain of size 10; likewise the 2-element subsets. These are actually the largest, though proving this can be a challenge, especially trying to generalize to the power set of an n element set. ■

(c) What are the maximal and minimal elements? Are they maximum and minimum?

Solution. $\emptyset$ is minimum and $\{1, 2, \ldots 5\}$ is maximum. ■

(d) Answer the previous part for the $\subset$ partial order on the set $\mathcal{P}\{1, 2, \ldots 5\} - \emptyset$.

Solution. Now the five size 1 subsets are minimal and there is no minimum. $\{1, 2, \ldots 5\}$ is still maximum. ■

Problem 3.10.2. Let S be a sequence of n different numbers. A *subsequence* of S is a sequence that can be obtained by deleting elements of S.

For example, if

$$S = (6, 4, 7, 9, 1, 2, 5, 3, 8)$$

Then 647 and 7253 are both subsequences of S (for readability, we have dropped the parentheses and commas in sequences, so 647 abbreviates $(6, 4, 7)$, for example).

An *increasing subsequence* of S is a subsequence of whose successive elements get larger. For example, 1238 is an increasing subsequence of S. Decreasing subsequences are defined similarly; 641 is a decreasing subsequence of S.

(a) List all the maximum length increasing subsequences of S, and all the maximum length decreasing subsequences.

Solution. The maximum length increasing subsequences are 1238 and 1258. The maximum length decreasing subsequences are

$$641, 642, 643, 653, 753, 953$$

■

Now let A be the *set* of numbers in S. (So $A = \{1, 2, 3, \ldots, 9\}$ for the example above.) There are two straightforward ways to totally order A. The first is to order its elements numerically, that is, to order A with the $<$ relation. The second is to order the elements by which comes first in S; call this order $<_S$. So for the example above, we would have

$$6 <_S 4 <_S 7 <_S 9 <_S 1 <_S 2 <_S 5 <_S 3 <_S 8$$

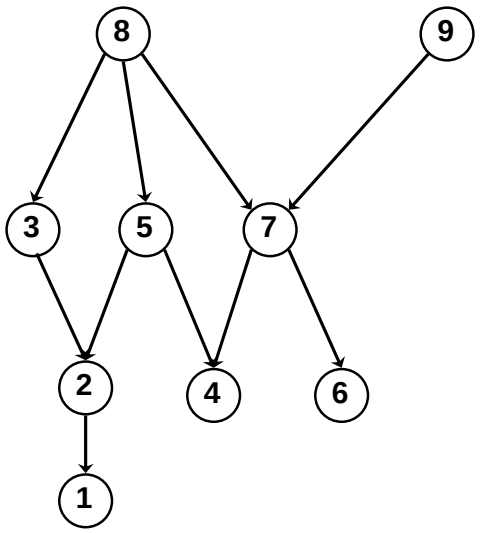
Next, define the partial order $\prec$ on A defined by the rule

$$a \prec a' \quad ::= \quad a < a' \text{ and } a <_S a'.$$

(It's not hard to prove that $\prec$ is strict partial order, but you may assume it.)

(b) Draw a diagram of the partial order, $\prec$, on A. What are the maximal elements,... the minimal elements?

Solution. The maximal elements are 8 and 9; the minimal are 1, 4,and 6:



■

(c) Explain the connection between increasing and decreasing subsequences of S, and chains and anti-chains under $\prec$.

Solution. A *chain*, with its elements listed in numerically increasing order, is an *increasing* subsequence and an *antichain*, with its elements listed in numerically decreasing order, is a *decreasing* subsequence. ■

(d) Prove that every sequence, S, of length n has an increasing subsequence of length greater than $\sqrt{n}$ or a decreasing subsequence of length at least $\sqrt{n}$.

Solution. By Dilworth's Lemma, either a chain or an antichain must have size at least $\sqrt{n}$, which, by the previous problem part, means there is either an increasing or a decreasing subsequence of this size. ■

(e) (Optional, and tricky) Describe a procedure that scans a length n sequence of numbers once from left to right and returns a subsequence of length at least $\sqrt{n}$ that is either increasing or decreasing.

Solution. REF TO FLOYD'S ALGORITHM NEEDED ■

Problem 3.10.3. Use induction to prove that the following equation holds for all $n \geq 2$:

$$\left(1 - \frac{1}{2}\right)\left(1 - \frac{1}{3}\right)\cdots\left(1 - \frac{1}{n}\right) = \frac{1}{n}$$

Solution. *Proof.* The proof is by induction on n. Let $P(n)$ be the proposition that the above equation holds.

base case ($n = 2$): $P(2)$ is true because $(1 - \frac{1}{2}) = \frac{1}{2}$.

inductive step: The induction hypothesis is $P(n)$. Assuming the hypothesis holds for some $n \geq 2$, we prove $P(n+1)$:

$$\begin{aligned}
&\left(1 - \frac{1}{2}\right)\left(1 - \frac{1}{3}\right)\cdots\left(1 - \frac{1}{n}\right)\left(1 - \frac{1}{n+1}\right) \\
&= \frac{1}{n}\left(1 - \frac{1}{n+1}\right) && \text{(induction hypothesis)} \\
&= \frac{1}{n} \cdot \frac{n}{n+1} && \text{(algebra)} \\
&= \frac{1}{n+1} && \text{(algebra)}
\end{aligned}$$

This shows that $P(n)$ implies $P(n+1)$, and the claim is proved by induction. □

■

Problem 3.10.4. A chocolate bar is molded into m rows, with n squares of chocolate in each row. The bar can be split by cutting between rows. For example, if a 8×10 bar is cut between its third and fourth rows, the result would be a two bars, one 3×10 and one 5×10. Similarly, a bar can also be split by cutting between columns. We want to keep making cuts until the bar is completely split into separate squares of chocolate.

Use strong induction to prove that exactly $mn - 1$ cuts are required to split the bar into individual squares.

Solution. This result does not immediately lend itself to an induction proof, since there are two variables, and it isn't clear whether to do induction on one or the other. In this case, a better approach is to do induction on another variable. Namely, the first cut in any process for completely splitting the bar leaves two *smaller* bars, and then we can break up the two bars recursively. This suggests a proof using strong induction on the number of individual squares of the chocolate bar. That is, the induction will be on the *size* of the bar, where the size of an $m \times n$ bar is mn. Now instead of a problem involving two variables (the two dimensions), we have a problem in one variable— the size. With this simplification, we can prove the claim using strong induction.

Proof. The proof is by strong induction on the size of the chocolate bar. Let $P(k)$ be the proposition that a chocolate bar of size k requires exactly $k-1$ cuts to be completely split into squares.

Base case ($k = 1$): $P(1)$ is true because there is only a single square of chocolate, and $1 - 1 = 0$ splits are required.

Induction step: We suppose $k \geq 1$ and assume the strong induction hypothesis that any chocolate bar of size s, where $1 \leq s \leq k$, requires exactly $s-1$ splits. We must now show that to split a chocolate bar of size $k+1$ requires exactly $(k+1)-1$ splits.

To do this, first break the chocolate bar of size $k+1$ into two smaller pieces of size p and q where $p+q = k+1$. This is certainly possible because the size of the bar is at least two. Now the pieces of sizes p and q are between one and k, so by strong induction, breaking these two pieces into single squares requires exactly $p-1$ and $q-1$ splits, respectively. The total number of splits required to break the bar of size $k+1$ into single squares is therefore exactly $1+(p-1)+(q-1) = p+q-1 = (k+1)-1$.

This shows that $P(1), \cdots, P(k)$ implies $P(k+1)$, and the claim is proved by strong induction. □

■

Problem 3.10.5. The following Lemma is true, but the *proof* given for it below is defective. Pinpoint, and illustrate with a counterexample, *exactly* where the proof goes wrong.

Lemma 3.10.1. *For any prime p and positive integers $n, x_1, x_2, \ldots, x_n$, if $p \mid x_1 x_2 \ldots x_n$, then $p \mid x_i$ for some $1 \leq i \leq n$.*

False proof. Proof by strong induction on n.

Base case $n = 1$: When $n = 1$, we have $p \mid x_1$, therefore we can let $i = 1$ and conclude $p \mid x_i$.

Induction step: Now assuming the claim holds for all $k \leq n$, we must prove it for $n+1$.

So suppose $p \mid x_1 x_2 \ldots x_{n+1}$. Let $y_n = x_n x_{n+1}$, so $x_1 x_2 \ldots x_{n+1} = x_1 x_2 \ldots x_{n-1} y_n$. Since the righthand side of this equality is a product of n terms, we have by induction that p divides one of them. If $p \mid x_i$ for some $i < n$, then we have the desired i. Otherwise $p \mid y_n$. But since y_n is a product of the two terms x_n, x_{n+1}, we have by strong induction that p divides one of them. So in this case $p \mid x_i$ for $i = n$ or $i = n+1$. □

Solution. Notice that nowhere in the proof is the fact that p is prime used. So if this proof were correct, the Lemma would hold not just for prime p, but for any positive integer p. But of course, the Lemma is false when p is not prime, for example if $p = 6$, $x_1 = 3$ and $x_2 = 4$, we have $p \mid x_1 x_2$ but $\neg(p \mid x_1)$ and $\neg(p \mid x_2)$. So there has to be something wrong somewhere.

The statement "we have by strong induction that p divides one of them" is the place where the proof breaks down: it appeals to strong induction to justify applying the induction hypothesis for $2 = k \leq n$. But the base case was $n = 1$, so we can't assume $2 \leq n$. Note that the reasoning above is fine for every $n \geq 2$, so the whole proof would be fine if we had an argument to prove the claim for $n + 1 = 2$.

Now in fact, if a prime, p divides $x_1 x_2$, it must divide x_1 or x_2; this follows by prime factorization of integers (and we'll show you another proof later in the term). But the proof here never made use of this fact. ■

Problem 3.10.6. While 6.042 was running in the TEAL room, the staff considered having students work in teams of exactly 4 or exactly 7 students. But TEAL tables hold up to nine students, so this plan was abandoned when it was realized then that no full table of nine could be divided into teams.

(a) What is the smallest number, k, such that a group of k or more students can be split into teams of 4 and 7? Prove that $k - 1$ students *cannot* be split in this way.

Solution. The smallest number k such that all tutorials of k or more students *can* be split into teams of 4 and 7 is 18.

A tutorial of size 17 *cannot* be split into teams of 4 and 7. In particular, a tutorial of size 17 can contain at most 2 groups of 7. Consider the various possible splits of the tutorial:

1. no groups of 7: the only other group size is 4. We cannot split the tutorial of 17, since it is not divisible by 4;
2. one group of 7: again, the remaining number, 10, of students is not divisible by 4;
3. two groups of 7: the remaining 3 students are not a group of 4 or 7;
4. three or more groups of 7: clearly, this accounts for $21 > 17$ students and is inadmissible.

■

(b) Now use strong induction to complete the proof of your answer to part (a). Namely, prove by strong induction that all tutorials of k or more students *can* be split into teams of size 4 and 7.

Solution. *Proof.* (strong induction). Define $P(n)$ to be the predicate that a tutorial of size n can be split into groups of size 4 and 7. Being "split" really means that $n = 4x + 7y$ for some nonnegative numbers x, y. So we can give a completely precise definition:

$$P(n) ::= \quad n \geq 18 \longrightarrow \exists x, y \in \mathbb{N}\, [n = 4x + 7y].$$

We must show that $P(n)$ holds for all $n \in \mathbb{N}$. By strong induction, we consider $n \geq 18$ and assume the strong induction hypothesis

$$\forall m\, [18 \leq m < n \longrightarrow \exists x_m, y_m \in \mathbb{N}\, (m = 4x_m + 7y_m)]. \tag{3.14}$$

in order to prove that $P(n)$ holds.

The proof divides into cases

Case 0: ($n < 18$) $P(n)$ holds trivially since the hypothesis $n \geq 18$ is false.

Case 1: ($n = 18$). $18 = 4 \cdot 1 + 7 \cdot 2$, so $P(18)$ is proved.

Case 2: ($n = 19$). $19 = 4 \cdot 3 + 7 \cdot 1$, so $P(19)$ is proved.

Case 3: ($n = 20$). $20 = 4 \cdot 5 + 7 \cdot 0$, so $P(20)$ is proved.

Case 4: ($n = 21$). $21 = 4 \cdot 0 + 7 \cdot 3$, so $P(21)$ is proved.

Notice that we proved Cases 0–4, without needing to use the induction hypothesis (3.14).

Case 5: ($n \geq 22$). In this case, letting $m ::= n - 4$, we have $18 \leq m < n$, so by induction hypothesis (3.14), we conclude that there are nonnegative integers x_m, y_m such that

$$m = 4x_m + 7y_m.$$

So

$$n = m + 4 = 4(x_m + 1) + 7y_m. \tag{3.15}$$

Letting $x ::= x_m + 1$ and $y ::= y_m$ in (3.15) implies

$$n = 4x + 7y,$$

which proves $P(n)$. □

■

Problem 3.10.7. Notes 3 proved that the stacking game with n blocks always ended with the same score.

Define the ***potential***, $p(S)$, of a stack, S, of blocks to be $k(k+1)/2$ where k is the number of blocks in S. Define the potential, $p(A)$, of a set, A, of stacks to be the sum of the potentials of the stacks in A.

Generalize the result in Notes 3 by showing that for any set, A, of stacks, if a sequence of moves starting with A leads to another set, B, of stacks, then the score for this sequence of moves is $p(A) - p(B)$.

Hint: Prove the

Lemma. *If B is the result of* one *move (stack split) from A, then the score for making this move is $p(A) - p(B)$.*

Solution. We first prove the Lemma.

Proof. Let A be a set of n stacks and B be another set of stacks made by one move from A, namely splitting some stack, $S \in A$ of size $s > 1$ into two stacks S_1 and S_2 of sizes of sizes s_1, s_2 where $s_1 + s_2 = s$; so the score of this move is $s_1 s_2$.

Since the other stacks in A do not change, they also appear in B. So by definition of potential, the difference between $p(A)$ and $p(B)$ is simply the difference between the potential of S and the potential of $\{S_1, S_2\}$, namely,

$$\begin{aligned} p(A) - p(B) &= p(S) - (p(S_1) + p(S_2)) \\ &= \frac{s(s+1)}{2} - \left(\frac{s_1(s_1+1)}{2} + \frac{s_2(s_2+1)}{2}\right) \\ &= \frac{(s_1+s_2)(s_1+s_2+1) - s_1(s_1+1) - s_2(s_2+1)}{2} \\ &= s_1 s_2, \end{aligned}$$

which is exactly the score of this move, as claimed. □

The general result now follows directly by induction on n with hypothesis:

$$\begin{aligned} P(n) ::= \quad & \forall A, B. \text{ if a sequence of } n \text{ moves goes from } A \text{ to } B, \\ & \text{then the score for these moves is } p(A) - p(B). \end{aligned}$$

Proof. **Base case** $n = 0$**:** 0 moves go from A to A with score 0, which equals $p(A) - p(A)$, proving $P(0)$.

Inductive step: Assume $P(n)$ is true.

Suppose $n+1$ moves take A to B, and let C be the set of stacks after the first move. Then the score for these moves equals the score for the move from A to C plus the score for the subsequent n moves from C to B.

By the Lemma, the score from A to C with the first move is $p(A) - p(C)$. By the induction hypothesis, the score for the n moves from C to B is $p(C) - p(B)$. Therefore the total score from A to B is

$$(p(A) - p(C)) + (p(C) - p(B)) = p(A) - p(B).$$

Since the choice of A and the sequence of $n+1$ moves to B were arbitrary, we conclude (by UG) that the score is $p(A) - p(B)$ for all A, B. This proves $P(n+1)$, completing the proof of the inductive step.

Therefore, $P(n)$ is true for all n by induction.

□

■

Problem 3.10.8. Consider the following equivalent way of viewing the subset take-away game from the Week 2, Friday class problem: for a fixed, finite set, A, let $\mathcal{S}$ initially be all the nonempty proper subsets of A. Players alternately choose a set $B \in \mathcal{S}$ and remove B and all sets that contain B from $\mathcal{S}$; they then continue playing on the updated $\mathcal{S}$. The player who chooses the last set in $\mathcal{S}$ wins.

Use the Well Ordering Property to show that, in any game, one of the players must have a winning strategy. *Hint:* Consider games whose initial set, $\mathcal{S}$, is an arbitrary collection of subsets of A, not necessarily all the proper subsets of A. Reach a contradiction by considering a minimum size game with no winning strategy for either player. What is a useful measure of size of a game?

Solution. Towards a contradiction, suppose that there is a game (i.e., a collection of subsets of A) where neither player has a winning strategy. Consider the set $\mathbb{S}$ of all such games. By our assumption, $\mathbb{S}$ is non-empty. By the Well Ordering Principle (applied on the set of sizes of games in $\mathbb{S}$), we know there exists a game $\mathcal{S}^* \in \mathbb{S}$ whose size is smallest. Consider all possile moves in this game $\mathcal{S}^*$ for the player whose turn it is —call him Bob, and let Alice be the other player. Say there are k of them.

We know $k \neq 0$. To see why, suppose $k = 0$. Then Bob can make no move. That is, there are no subsets for him to select from. So, Alice has already chosen the last subset, and won. Therefore Alice has a winning strategy in $\mathcal{S}^*$ —the strategy being to just sit there and do nothing, since the game is over. This contradicts the selection of $\mathcal{S}^*$ as a game where no player has a winning strategy.

So, $k \geq 1$. For each of Bob's possible moves, consider the resulting game if he makes that move. Let $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k$ be these new games. Clearly, each $\mathcal{S}_i$ is strictly smaller than $\mathcal{S}^*$ (contains fewer subsets of A). Therefore, in each $\mathcal{S}_i$ at least one of the players has a winning strategy —by the selection of $\mathcal{S}^*$ as a minimum-size game where this is not true. We now distinguish two cases.

Case 1: Bob has a winning strategy in at least one $\mathcal{S}_i$. Then Bob has the following winning strategy in $\mathcal{S}^*$: first, make the i-th move, to get yourself into $\mathcal{S}_i$; from then on, follow the strategy that you have in $\mathcal{S}_i$. Therefore, there *is* a player with a winning strategy in $\mathcal{S}^*$ (Bob), which contradicts the selection of $\mathcal{S}^*$.

Case 2: Bob has a winning strategy in none of the $\mathcal{S}_i$. We already know that, in each of the $\mathcal{S}_i$, at least one of the players has a winning strategy. Since this player is never Bob, we conclude that Alice has a winning strategy in each one of the $\mathcal{S}_i$. But then Alice has a winning strategy in $\mathcal{S}^*$, as well: just sit there and wait for Bob to move; then follow the strategy that you have in the resulting $\mathcal{S}_i$. Therefore, there *is* a player with a winning strategy in $\mathcal{S}^*$ (Alice), which again contradicts the selection of $\mathcal{S}^*$.

In either case, we reached a contradiction. Therefore, the original assumption is false: the game does admit a winning strategy for one of the players. Of course, this proof gives us no clue as to *who* this player is, let alone *what* his strategy is... ■

3.11 Miniquiz Feb. 28

Problem 3.11.1. For each of the following relations indicated below —the first two are on the set $\{1, 2, 3, 4\}$ —indicate whether it is **R**eflexive, **A**ntisymmetric, **TRAN**sitive, **TOT**al, or **N**one of the above. (More than one property may hold for some relations.)

$\{(1,1),(1,3),(3,1)\}$_____

$\{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}$_____

$\{(p,q) \mid p \text{ and } q \text{ are born in the same town}\}$_____

Note that every person is born in the same town as themselves.

Solution. **N** for the first: Not reflexive since (2,2), (3,3), and (4,4) are not in the relation. Not antisymmetric since (1,3) and (3,1) are both in the relation. Not transitive since (3,1) and (1,3) are in the relation, but (3,3) is not. Not total since 2 and 4 do not relate to any member.

A, TRAN for the second: Not reflexive since $(1,1), (2,2), (3,3)$, and $(4,4)$ are not in the relation. Antisymmetric since for each pair, the reverse pair is not in the relation; for example, $(1,2)$ is in the relation, but $(2,1)$ is not. Transitive since if (a,b) and (b,c) are in this relation, then (a,c) is also in this relation. Not total since 4 does not relate to any member.

R, TRAN, TOT for the third: Reflexive since every person is born in the same town as themselves. Not antisymmetric since if A is born in the same town as B, then B must be born in the same town as A. Transitive since if A is born in the same town as B, and B is born in the same town as C, then A must be born in the same town as C. Total since every person relates to at least themselves. ■

Problem 3.11.2. We use the notation $m \mid n$ to indicate that m divides n, and, for purposes of defining minimal elements in the divides partial order, when $m \mid n$, we'll consider m to be "smaller than or equal to" n.

Let $S ::= \{2, 6, 9, 12, 18, 27, 48, 72\}$ be partially ordered by the divides relation.

(a) Which are the minimal elements of S?_____

Solution. 2,9 ■

(b) Which are the maximal elements?_____

Solution. 27,48,72 ■

(c) Give an example of a maximum-size chain in S._____

Solution. There are three possible solutions: $\{2, 6, 12, 48\}$, $\{2, 6, 12, 72\}$, $\{2, 6, 18, 72\}$. ■

(d) Give an example of a maximum-size antichain in S._____

Solution. There are three possible solutions: $\{12, 18, 27\}, \{18, 27, 48\}, \{27, 48, 72\}$. ■

Problem 3.11.3. **(a)** Express the sum of the first n odd numbers, $1 + 3 + 5 + \ldots$, without the dots by filling in the two missing parts in the following $\sum$-expression:

$$\underbrace{1 + 3 + 5 + \ldots}_{n \text{ terms}} = \sum_{i=0}^{(\quad)} (\qquad)$$

Solution.

$$\sum_{i=0}^{(n-1)} (2i + 1)$$

■

(b) Prove by induction that this sum is n^2.

Solution. We will prove this by induction on n with **Induction hypothesis:**

$$P(n) ::= \sum_{i=0}^{n-1} (2i + 1) = n^2.$$

Base case $n = 0$**:** The left hand sum is 0 in this case, since it has no terms in it. Likewise, the right hand side is a sum over an empty set ($\{i \mid i \geq 0 \text{ and } i \leq -1\}$), and so is also 0 by convention.

Induction step: Assume that $P(n)$ is true for some $n \geq 0$, which means

$$1 + 3 + \cdots + (2(n-1) + 1) = n^2.$$

Now,

$$\begin{aligned} \sum_{i=0}^{(n+1)-1} (2i + 1) &= \left(\sum_{i=0}^{n-1} (2i + 1) \right) + 2((n + 1) - 1) + 1 && \text{(def. of } \Sigma) \\ &= n^2 + 2((n + 1) - 1) + 1 && \text{(ind. hyp)} \\ &= n^2 + 2n + 1 = (n + 1)^2, \end{aligned}$$

which proves $P(n + 1)$.

Therefore, by the principle of induction, $P(n)$ holds for all $n \geq 0$. ■

Problem 3.11.4. The faulty horses-of-the-same color proof from Notes 3 appears in the box below. Among the following statements, circle the one that *best* explains the mistake in the proof. Note that most of the statements are correct, but one of them is clearly the best explanation.

1. The inductive step assumes that for any two sets of n out of $n+1$ horses, there is a horse common to both sets, but this is false for $n = 1$.

2. The proof of the inductive step is false when $n = 2$.

3. It's silly to say that a horse is the same color as itself, so the claim that the base case is "certainly" true is wrong.

4. The base case should be for $n = 0$.

5. The inductive step assumes that "being the same color" is a transitive relation on a set of 2 horses.

Solution. The first statement

> 1. The inductive step assumes that for any two sets of n out of $n+1$ horses, there is a horse common to both sets, but this is false for $n = 1$.

is the best explanation.

Let's look at the other statements: 2. The inductive step is not false when $n = 2$. When $n = 2$, the inductive step considers a set of $n + 1$ horses, or a set of 3 horses. If the first 2 horses are the same color and the last 2 horses are the same color, then all 3 horses must be the same color. The inductive step is false when $n = 1$.

3. A horse is the same color as itself. It might be silly to say that, but it doesn't explain the mistake in the proof.

4. We are trying to prove that "In every set of $n \geq 1$ horses, all are the same color." Therefore, it makes sense to start with a base case of 1. Base cases are not restricted to being for $n = 0$.

5. "Being the same color" is a transitive relation on two horses. For example, if horse 1 is the same color as horse 2, and horse 2 is the same color as horse 2, then horse 1 must be the same color as horse 2. ■

False Theorem 3.11.1. *In every set of $n \geq 1$ horses, all are the same color.*

Proof. The proof is by induction on n. The induction hypothesis, $P(n)$, will be

$$\text{In every set of } n \text{ horses, all are the same color.} \tag{3.16}$$

Base case: ($n = 1$). $P(1)$ is true, because in a set of horses of size 1, there's only one horse, and this horse is definitely the same color as itself.

Inductive step: Assume that $P(n)$ is true for some $n \geq 1$. that is, assume that in every set of n horses, all are the same color. Now consider a set of $n + 1$ horses:

$$h_1,\ h_2,\ \ldots,\ h_n,\ h_{n+1}$$

By our assumption, the first n horses are the same color:

$$\underbrace{h_1,\ h_2,\ \ldots,\ h_n,}_{\text{same color}}\ h_{n+1}$$

Also by our assumption, the last n horses are the same color:

$$h_1,\ \underbrace{h_2,\ \ldots,\ h_n,\ h_{n+1}}_{\text{same color}}$$

So h_1 is the same color as the remaining horses besides h_{n+1}, and likewise h_{n+1} is the same color as the remaining horses besides h_1. So h_1 and h_{n+1} are the same color. That is, horses $h_1, h_2, \ldots, h_{n+1}$ must all be the same color, and so $P(n + 1)$ is true. Thus, $P(n)$ implies $P(n + 1)$.

By the principle of induction, $P(n)$ is true for all $n \geq 1$. □

Problem 3.11.5. The next page contains a proof using the Well Ordering Principle that every amount of postage that can be paid exactly using only 6 cent and 15 cent stamps, is divisible by 3. That is, letting $S(n)$ mean that exactly n cents postage can be paid using only 6 and 15 cent stamps, the proof shows that

$$\forall n \in \mathbb{N}.\ S(n) \text{ implies } 3 \mid n. \tag{*}$$

Fill in the missing portions (indicated by "…") of the following proof of (*).

So let C be the set of *counterexamples* to (*), namely

$$C ::= \{n \mid \textbf{. . . circle the correct choice below . . .}\}$$

- $\neg[S(n)$ equiv $3 \mid n]$
- $S(n)$ and $\neg(3 \mid n)$
- $\neg S(n)$ and $\neg(3 \mid n)$
- $\neg S(n)$ or $3 \mid n$

Solution. n is a counterexample to (*) if n cents postage can be made and n is not divisible by 3, so the second statement

$$S(n) \text{ and } \neg(3 \mid n)$$

defines the set, C, of counterexamples. ■

If C is nonempty, then by the WOP, there is a smallest number, $m \in C$. This m must be positive because. . .

Solution. $3 \mid 0$, so 0 is not a counterexample. ■

But if $S(m)$ holds and m is positive, then $S(m-6)$ or $S(m-15)$ must hold, because. . .

Solution. if $m > 0$ cents postage is made from 6 and 15 cent stamps, at least one stamp must have been used, so removing this stamp will leave another amount of postage that can be made. ■

So suppose $S(m-6)$ holds. Then $3 \mid (m-6)$, because. . .

Solution. if $\neg(3 \mid (m-6))$, then $m-6$ would be a counterexample smaller than m, contradicting the minimality of m. ■

But if $3 \mid (m-6)$, then obviously $3 \mid m$, contradicting the fact that m is a counterexample.

Next suppose $S(m-15)$ holds. Then the proof for $m-6$ carries over directly for $m-15$ to yield a contradiction in this case as well. Since we get a contradiction in both cases, we conclude that. . .

Solution. C must be empty. That is, there are no counterexamples to (*), ■

which proves that (*) holds.

Appendix

Relational Properties

A binary relation, R, on a set, A, is

- total (as a relation; *not* the same as total order) if every element of the domain, A is related to some element of A:
$$\forall a_1 \in A\ \exists a_2 \in A.\ a_1 R a_2.$$
- *transitive* if for every $a, b, c \in A$, aRb and bRc implies aRc.
- *asymmetric* if for every $a, b \in A$, aRb implies $\neg(bRa)$,
- *reflexive* if aRa for every $a \in A$,
- *antisymmetric* if for every $a \neq b \in A$, aRb implies $\neg(bRa)$,

Partial Order

A binary relation is a *strict partial order* iff it is transitive and asymmetric. It is a *weak partial order* iff it is transitive, reflexive, and antisymmetric.

Let $\preceq$ be a weak (reflexive) partial order on a set, A.

- An element $a \in A$ is *minimal* iff there is no element in A that is $\preceq a$ except possibly a itself. Similarly, an element $a \in A$ is *maximal* iff there is no element in A that is $\succeq a$ except possibly a itself.
- An element $a \in A$ is a *lower bound* for a subset, $S \subseteq A$ iff $a \preceq s$ for every $s \in S$. Similarly, an element $a \in A$ is an *upper bound* for a subset, $S \subseteq A$ iff $s \preceq a$ for every $s \in S$.
- An element $a \in A$ is the *minimum* element iff a is a lower bound on A. Similarly, an element $a \in A$ is the *maximum* element iff a is an upper bound on A.
- Elements $a, b \in A$ are *comparable* iff either $a \preceq b$ or $b \preceq a$. Two elements are *incomparable* iff they are not comparable.
- A subset, $S \subseteq A$ is *totally ordered* iff every two distinct elements in S are comparable.
- A *chain* is a totally ordered subset of A.
- An *antichain* is a subset of A, such that no two elements in it are comparable.

Chapter 4

Structural Induction; State Machines

4.1 Recursive Data Types

Recursive data types play a central role in modern programming languages. From a Mathematical point of view, recursive data types are what induction is about. Recursive data types are specified by *recursive definitions* that say how to build something from its parts. These definitions have two parts:

- **Base case(s)** that don't depend on anything else.
- **Constructor case(s)** that depend on previous cases.

Example 4.1.1. Define a set, E, recursively as follows:

- **Base case:** $0 \in E$,
- **Constructor cases:** if $n \in E$, then
  1. $n + 2 \in E$, when $n \geq 0$;
  2. $-n \in E$, when $n > 0$.

Using this definition, we can see that $0 \in E$ by the Base case, so $0 + 2 = 2 \in E$ by Constructor case 1., and so $2 + 2 = 4 \in E$, $4 + 2 = 6 \in E$, ..., and in fact any nonnegative even number is in E by successive application of case 1. Also, by case 2., $-2, -4, -6, \cdots \in E$. So clearly all the even integers are in E.

Is anything else in E? The definition doesn't say so explicitly, but an implicit condition on a recursive definition is that the only way things get into E is as a consequence of the Base and Constructor cases. In other words, E will be exactly the set of even integers.

Another example is the set, M, of strings of *matched* right and left parentheses. These are the strings that would be obtained if we took a sequence of fully parenthesized arithmetic (or Scheme) expressions and erased all the characters except the parentheses. Here's a recursive definition:

Example 4.1.2. Define the set, M, of strings of matched right and left parentheses recursively as follows:

- **Base case:** $\lambda \in M$, where λ is the *empty* string,
- **Constructor case:** if $s, t \in M$, then $(\,s\,)t \in M$.

Here we're writing $(\,s\,)t$ to indicate the string that starts with a left parenthesis, followed by the sequence of parentheses (if any) in the string s, followed by a right parenthesis, and ending with the sequence of parentheses in the string t.

Using this definition, we can see that $\lambda \in S$ by the Base case, so

$$(\,\lambda\,)\lambda = (\,) \in M$$

by the Constructor case, and so

$$\begin{aligned}(\,\lambda\,)(\,) &= (\,)(\,),\\ (\,(\,)\,)\lambda &= (\,(\,)\,),\\ (\,(\,)\,)(\,) &\end{aligned}$$

are further strings in M by repeated applications of the Constructor case.

4.1.1 Tagged data

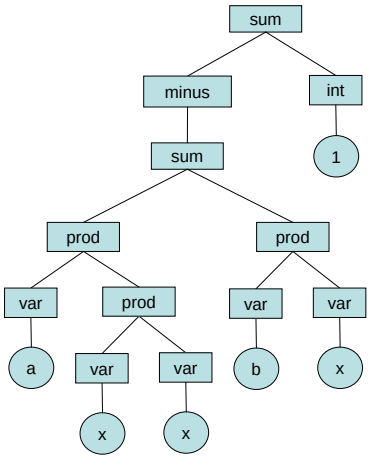


Figure 4.1: Parse tree for $-(a(x \cdot x) + bx) + 1$.

Arithmetic expressions like

$$-(a(x \cdot x) + bx) + 1 \tag{4.1}$$

are another important example of a recursive data type. We could define them as parenthesized strings with symbols for arithmetic operators, but a more useful representation uses the *parse trees* of the expressions, rather than strings. Figure 4.1 shows a parse tree for expression (4.1).

Such a tree would be represented by pairs or triples that begin with a *tag* equal to the label of the top node of the parse tree. We'll call these tagged data items Aexp's. They are defined recursively as follows:

Example 4.1.3. The set, Aexp, of *Arithmetic expressions* over a set of *variables*, V, is defined recursively as follows:

- **Base cases:**
  1. If $n \in \mathbb{Z}$, then $\langle \texttt{int}, n\rangle \in$ Aexp.
  2. If $v \in V$, then $\langle \texttt{var}, v\rangle \in$ Aexp.
- **Constructor cases:** if $e, e' \in$ Aexp, then
  1. $\langle \texttt{sum}, e, e'\rangle \in$ Aexp,
  2. $\langle \texttt{prod}, e, e'\rangle \in$ Aexp, and
  3. $\langle \texttt{minus}, e\rangle \in$ Aexp.

So the Aexp corresponding to the parse tree of Figure 4.1 is

$$\langle\, \texttt{sum},\ \langle\, \texttt{minus},\ \langle\, \texttt{sum},\ \langle \texttt{prod},\ \langle \texttt{var},\ a\rangle, \langle \texttt{prod},\ \langle \texttt{var},\ x\rangle,\ \langle \texttt{var},\ x\rangle\rangle\rangle,\ \langle \texttt{prod},\ \langle \texttt{var},\ b\rangle,\ \langle \texttt{var},\ x\rangle\rangle\rangle\rangle,\ \langle \texttt{int},\ 1\rangle\rangle\rangle$$

4.2 Structural Induction on Recursive Data Types

Structural induction is a method for proving some property P of all the elements of a recursively-defined data type. The proof consists of two steps:

- Prove P for the base cases of the definition.
- Prove P for the constructor cases of the definition, assuming that it is true for the component data items.

To illustrate this, we'll prove that strings of matched parentheses always have an equal number of left and right parentheses. To do this, define a predicate on strings

$$P(s) ::= \quad s \text{ has an equal number of left and right parentheses.}$$

Proof. We'll prove that $\forall s \in M.\, P(s)$ by structural induction on the definition of M, using $P(s)$ as the induction hypothesis.

Base case: $P(\lambda)$ holds because the empty string has zero left and zero right parentheses.

Constructor case: For $r = \texttt{(}s\texttt{)}t$, we must show that $P(r)$ holds, given that $P(s)$ and $P(t)$ holds. So let n_s, n_t be, respectively, the number of left parentheses in s and t. So the number of left parentheses in r is $1 + n_s + n_t$.

Now from the respective hypotheses $P(s)$ and $P(t)$, we conclude that the number of right parentheses in s and t, respectively, is also n_s and n_t. So the number of right parentheses in r is $1+n_s+n_t$, which is the same as the number of left parentheses. This proves $P(r)$. We conclude by structural induction that $\forall s \in M.\, P(s)$. □

Returning to the recursive definition of the set E in Example 4.1.1, the observation that all even integers are in E was easy to see, and is easy to prove by induction (with induction hypothesis $P(n) ::= 2n \in E \wedge -2n \in E$). To verify the observation E actually equals the even numbers, we need only show that every integer in E is even. This follows trivially by structural induction with hypothesis:

$$Q(n) ::= n \text{ is even.}$$

Base case: $Q(0)$ holds since 0 is even.

Constructor cases: assuming $n \in E$ and $Q(n)$ holds, prove that

1. $Q(n+2)$ holds. This is immediate, since adding 2 to an even number gives an even number.
2. $Q(-n)$ holds. This is also immediate, since n is even iff $-n$ is even.

We could understand induction on the length of strings as structural induction if we think of the strings as being represented as tagged data:

Definition 4.2.1. The set, A^*, of strings over a set, A, called the *alphabet*, is defined recursively as follows:

- **Base case:** $\langle \texttt{emptystring} \rangle \in A^*$.
- **Constructor case:** if $s \in A^*$ and $a \in A$, then $\langle \texttt{successor-string}, s, a \rangle \in A^*$.

Here, of course, $\langle \texttt{emptystring} \rangle$ is a tagged representation of the emptystring, λ, and

$$\langle \texttt{successor-string}, s, a \rangle$$

is a tagged representation of the string, sa, equal to the string s, followed by the character, a.

In fact, ordinary induction can be understood as an instance of structural induction if we think of the nonnegative integers as being represented as tagged data. To start, we might represent 0 as a length one sequence consisting of the tag `zero`:

Definition 4.2.2. The nonnegative integers can be defined recursively as follows:

- **Base case** $\langle \texttt{zero} \rangle \in \mathbb{N}$.
- **Constructor case** if $n \in \mathbb{N}$, then $\langle \texttt{successor}, n \rangle \in \mathbb{N}$.

4.2.1 Functions on Recursively-defined Data Types

Functions on recursively-defined data types can be defined recursively using the same cases as the data type definition. Namely, to define a function, F, on a recursive data type, define the value of F for the base cases of the data type definition, and then define the value of F in each constructor case in terms of the values of F on the component data items.

For example, from the recursive Definition 4.2.1 of strings, we can define:

Definition 4.2.3. The *length*, $|s|$, of a string, s, is defined recursively by the rules:

- $|\lambda| ::= 0$
- $|sa| ::= 1 + |s|$.

Definition 4.2.4. The *concatenation*, st, of strings s and t over an alphabet, A, is defined recursively on t by the rules:

- $s\lambda ::= s$.
- $s(ta) ::= (st)a$ for $a \in A$.

For the set, M, of strings of matched parentheses, we define:

Definition 4.2.5. The *depth*, $d(s)$, of a string, $s \in M$, is defined recursively by the rules:

- $d(\lambda) ::= 0$.
- $d(\texttt{(}\, s\, \texttt{)}t) ::= \max\{d(s) + 1, d(t)\}$

> **Warning:** When a recursive definition of a data type allows the same element to be constructed in more than one way, the definition is said to be *ambiguous*. A function defined recursively from an ambiguous definition of a data type will not be well-defined unless the values specified for the different ways of constructing the element agree.

We were careful to choose *un*ambiguous definitions of the sets M and E to ensure that functions defined recursively on the definitions of these data types would always be well-defined. Recursive definitions of tagged data types, where the tag uniquely determines the rule used to construct an element, are guaranteed to be unambiguous.

4.2.2 Evaluation and Substitution

We'll define some recursive functions on arithmetic expressions that suggest the role of recursive definitions in programming. For simplicity, we'll work with arithmetic expressions with only one variable —call it x. Now given such an expression, $e \in$ Aexp, and an integer value, n, for the variable, x, we can evaluate e in the usual way to arrive at an integer value, $\text{eval}(e, n)$. The eval function has a familiar recursive definition:

Definition 4.2.6. The function eval is defined recursively on an expression, e, and integer, n, as follows:

- $\text{eval}(\langle \texttt{integer}, k\rangle, n) ::= k$, (the value of the constant, k, is k, no matter what x equals),
- $\text{eval}(\langle \texttt{variable}, x\rangle, n) ::= n$, (the value of the variable, x, is given to be n),
- $\text{eval}(\langle \texttt{sum}, e, e'\rangle, n) = \text{eval}(e, n) + \text{eval}(e', n)$,
- $\text{eval}(\langle \texttt{product}, e, e'\rangle, n) = \text{eval}(e, n) \cdot \text{eval}(e', n)$,

- $\text{eval}(\langle \texttt{minus}, e\rangle, n) = -\text{eval}(e, n)$.

Another useful operation on arithmetic expressions is substituting one into another. Let $\text{subst}(e, f)$ be the result of substituting expression f for all occurrences of the variable x in e.

Definition 4.2.7. The function subst is defined recursively on expressions e and f as follows:

- $\text{subst}(\langle \texttt{integer}, k\rangle, f) ::= k$, (the constant, k, has no x's in it to substitute for),
- $\text{subst}(\langle \texttt{variable}, x\rangle, f) ::= f$,
- $\text{subst}(\langle \texttt{sum}, e, e'\rangle, f) = \langle \texttt{sum}, \text{subst}(e, f), \text{subst}(e', f)\rangle$,
- $\text{subst}(\langle \texttt{product}, e, e'\rangle, f) = \langle \texttt{product}, \text{subst}(e, f), \text{subst}(e', f)\rangle$,
- $\text{subst}(\langle \texttt{minus}, e\rangle, f) = \langle \texttt{minus}, \text{subst}(e, f)\rangle$.

Now suppose we substitute another expression, f, for all the x's in e to obtain a new expression, $e' ::= \text{subst}(e, f)$. Now we could evaluate e' when x has the value n by directly applying eval. But another, usually more efficient, approach would be to find the value of f when x has the value n, and then evaluate e with x given the value obtained from f.[1]

For example, suppose (using ordinary formula notation for readability)

$$\begin{array}{ll} e \text{ is } x + x & \text{and} \\ f \text{ is } 3x & \text{so} \\ g ::= \text{subst}(e, f) = 3x + 3x. & \end{array}$$

Then in the evaluation of g at $x = 1$, two multiplications by 3 would be performed. But evaluating f at $x = 1$ involves only one multiplication by 3, and then evaluating e with $x = 3 \cdot 1 = 3$ involves no further multiplications.

We will prove that both approaches yield the same answer. More precisely, what we want to prove is

Theorem 4.2.8. *For all expressions $e, f \in$ Aexp and $n \in \mathbb{Z}$,*

$$\textit{eval}(\textit{subst}(e, f), n) = \textit{eval}(e, \textit{eval}(f, n)). \tag{4.2}$$

Proof. The proof is by structural induction on e.[2]

Base cases:

- $e = \langle \texttt{integer}, k\rangle$. Then the lefthand side of equation (4.2) equals k by this base case in the definition of subst, and the righthand side equals k by this base case of the definition of eval.
- $e = \langle \texttt{variable}, x\rangle$. Then the lefthand side of equation (4.2) equals $\text{eval}(f, n)$ by this base case in the definition of subst, and the righthand side also equals $\text{eval}(f, n)$ by this base case in the definition of eval.

[1] In Lisp programming terminology, the first approach corresponds to evaluation using a "substitution model," and the second approach corresponds to evaluation using an "environment model."

[2] This is an example of why it's useful to notify the reader what the induction variable is —in this case it isn't n.

Constructor cases:

- $e = \langle \texttt{sum}, e_1, e_2 \rangle$. By the structural induction hypothesis (4.2), we may assume that for all $f \in \text{Aexp}$ and $n \in \mathbb{N}$,

$$\text{eval}(\text{subst}(e_i, f), n) = \text{eval}(e_i, \text{eval}(f, n)) \tag{4.3}$$

for $i = 1, 2$. We wish to prove that

$$\text{eval}(\text{subst}(\langle \texttt{sum}, e_1, e_2 \rangle, f), n) = \text{eval}(\langle \texttt{sum}, e_1, e_2 \rangle, \text{eval}(f, n)). \tag{4.4}$$

But the lefthand side of (4.4) equals

$$\text{eval}(\langle \texttt{sum}, \text{subst}(e_1, f), \text{subst}(e_2, f) \rangle, n)$$

by definition of subst for a sum expression, which equals

$$\text{eval}(\text{subst}(e_1, f), n) + \text{eval}(\text{subst}(e_2, f), n)$$

by definition of eval for a sum expression. By induction hypothesis (4.3), this equals

$$\text{eval}(e_1, \text{eval}(f, n)) + \text{eval}(e_2, \text{eval}(f, n)),$$

which equals the righthand side of (4.4) by definition of eval for a sum expression. This proves (4.4) in this case.

- $e = \langle \texttt{product}, e_1, e_2 \rangle$ or $e = \langle \texttt{minus}, e_1 \rangle$. Similar.

□

4.2.3 Recursive Functions on Nonnegative Integers

Definition 4.2.2 of the nonnegative integers as a recursive tagged data type justifies the familiar recursive definitions of functions on the nonnegative integers. Here are some examples.

The Factorial function. This function is often written "$n!$." You will see a lot of it later in the term. Here we'll use the notation $\text{fac}(n)$:

- $\text{fac}(0) ::= 1$.
- $\text{fac}(n+1) ::= (n+1) \cdot \text{fac}(n)$ for $n \geq 0$.

The Fibonacci numbers. These form interesting sequence of numbers that arise, for example, in modeling growth processes of plants, cells, and animal populations. Letting $\text{fib}(n)$ be the nth Fibonacci number, fib can be defined recursively by:

$$\begin{aligned} \text{fib}(0) &::= 0, \\ \text{fib}(1) &::= 1, \\ \text{fib}(n) &::= \text{fib}(n-1) + \text{fib}(n-2) \qquad \text{for } n \geq 2. \end{aligned}$$

Here the recursive step starts at $n = 2$ with base cases for 0 and 1. This is needed since the constructor case relies on two previous values.

What is $\text{fib}(4)$? Well, $\text{fib}(2) = \text{fib}(1) + \text{fib}(0) = 1$, $\text{fib}(3) = \text{fib}(2) + \text{fib}(1) = 2$, so $\text{fib}(4) = 3$. The sequence starts out $0, 1, 1, 2, 3, 5, 8, 13, 21, \ldots$.

Sum-notation. Let "$S(n)$" abbreviate the expression "$\sum_{i=1}^{n} f(i)$." We can recursively define the meaning of $S(n)$ with the rules

- $S(0) ::= 0$.
- $S(n+1) ::= f(n+1) + S(n)$ for $n \geq 0$.

Ill-formed Function Definitions

There are some blunders to watch out for when defining functions recursively. Below are some function specifications that resemble good definitions of functions on the nonnegative integers, but they aren't.

$$f_1(n) ::= 2 + f_1(n-1). \tag{4.5}$$

This "definition" has no base case. If some function, f_1, satisfied (4.5), so would a function obtained by adding a constant, k, to the value of f_1. So equation (4.5) does not uniquely define f_1.

$$f_2(n) ::= \begin{cases} 0, & \text{if } n \text{ is divisible by 2,} \\ 1, & \text{if } n \text{ is divisible by 3,} \\ 2, & \text{otherwise.} \end{cases} \tag{4.6}$$

This "definition" is inconsistent: it requires $f_2(6) = 0$ and $f_2(6) = 1$, so (4.6) doesn't define anything.

$$f_3(n) ::= \begin{cases} 0, & \text{if } n = 0, \\ f_3(n+1) + 1, & \text{otherwise.} \end{cases} \tag{4.7}$$

"Definition" (4.7) implies that $f_3(1) > f_3(2) > f_3(3) > \cdots$, so $f_3(1)$ cannot equal any integer. No total function on the nonnegative integers can satisfy this definition.[3]

$$f_4(n) ::= \begin{cases} 0, & \text{if } n = 0, \\ f_4(n+1) & \text{otherwise.} \end{cases} \tag{4.8}$$

Any function that is 0 at 0 and constant everywhere else satisfies (4.8), so it does not uniquely define anything.

[3]It does uniquely determine a *partial* function on the nonnegative integers, namely, $f_3(0) = 0$ and f_3 is undefined everywhere else, if we accept the convention that $f(a) = f(b)$ holds when f is not defined for both a and b.

A Mysterious Function

Mathematicians have been wondering about this function specification for a while:

$$f_5(n) ::= \begin{cases} 1, & \text{if } n \leq 1, \\ f_5(n/2) & \text{if } n > 1 \text{ is even}, \\ f_5(3n+1) & \text{if } n > 1 \text{ is odd}. \end{cases} \tag{4.9}$$

For example, $f_5(3) = 1$ because

$$f_5(3) ::= f_5(10) ::= f_5(5) ::= f_5(16) ::= f_5(8) ::= f_5(4) ::= f_5(2) ::= f_5(1) ::= 1.$$

The constant function equal to 1 will satisfy (4.9), but it's not known if another function does too. The problem is that the third case specifies $f_5(n)$ in terms of f_5 at arguments larger than n, and so cannot be justified by induction on $\mathbb{N}$. It's known that any f_5 satisfying (4.9) equals 1 for all n up to over a billion.

Quick exercise: Why does the constant function 1 satisfy (4.9)?

4.3 Games as a Recursive Data Type

Chess, Checkers, and Tic-Tac-Toe are examples of *two-person terminating games of perfect information*, —2PTG's for short. These are games in which two players alternate moves that depend only on the visible board position or state of the game. "Perfect information" means that the players know the complete state of the game at each move. (Most card games are *not* games of perfect information because neither player can see the other's hand.) "Terminating" means that play cannot go on forever —it must end after a finite number of moves.[4]

We will define 2PTG's in a straightforward way as a tagged recursive data type. To see how this will work, let's use the game of Tic-Tac-Toe as an example.

4.3.1 Tic-Tac-Toe

Tic-Tac-Toe is played on a 3×3 grid whose nine cells start off empty. Two players alternate moves, where a move for the first player is to write an "X" in an empty cell, and likewise the second player writes an "O". Three copies of the same letter filling a row, column, or diagonal of the grid is called a *tic-tac-toe*, and the first player who gets a tic-tac-toe wins the game.

At any point in the game, the "board position" is the pattern of X's and O's on the grid. From any such Tic-Tac-Toe pattern, there are a number of next patterns that might result from a move. For example, from the initial empty grid, there are nine possible next patterns, each with a single X in some grid cell and the other eight cells empty. From any of these patterns, there are eight possible next patterns gotten by placing an O in an empty cell. These move possibilities are given by the *game tree* for Tic-Tac-Toe outlined in Figure 4.2.

[4]Since board positions can repeat in chess and checkers, termination is enforced by rules that prevent any position from being repeated more than a fixed number of times. So the "state" of these games is the board position *plus* a record of how many times positions have been reached.

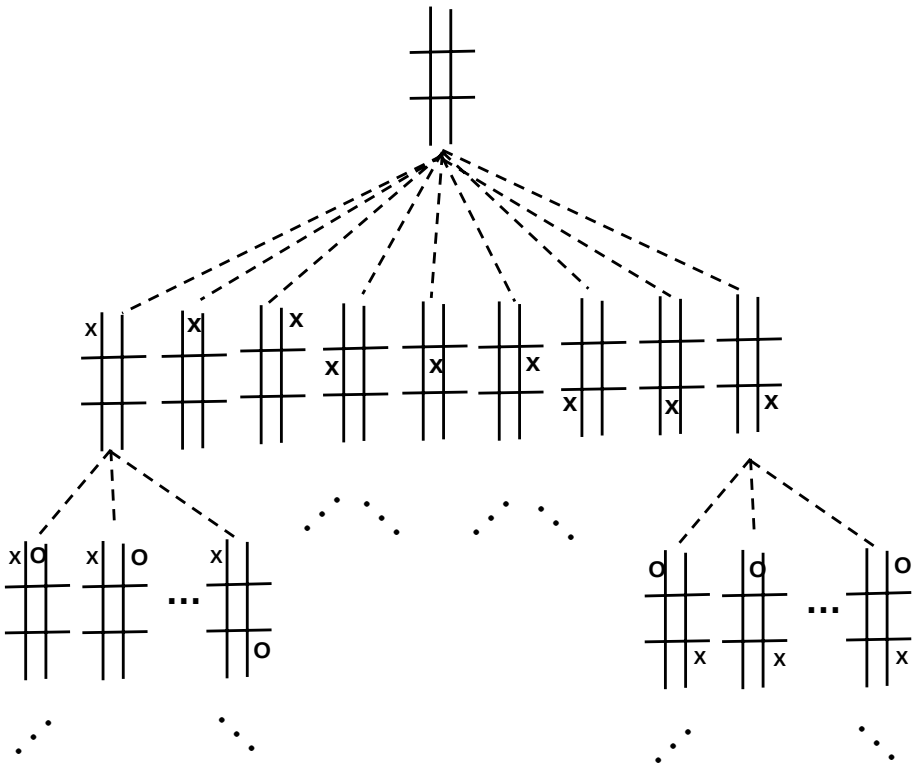


Figure 4.2: The Top of the Game Tree for Tic-Tac-Toe.

Definition 4.3.1. A Tic-Tac-Toe *pattern* is a 3×3 grid each of whose 9 cells contains either the single letter, X, the single letter, O, or is empty. Moreover, there must be either

- one more X than O's, with at most two tic-tac-toes of X's, and no tic-tac-toe of O's or
- an equal number of X's and O's, with at most one tic-tac-toes of O's, and no tic-tac-toe of X's.

If P is a Tic-Tac-Toe pattern, then the following are *Tic-Tac-Toe games*:

Base Cases:

- if P has a tic-tac-toe of X's:
$$\langle P, \langle \texttt{win} \rangle \rangle ,$$
- if P has a tic-tac-toe of O's:
$$\langle P, \langle \texttt{lose} \rangle \rangle ,$$
- if all nine cells of P are filled with letters and there are no tic-tac-toes:
$$\langle P, \langle \texttt{tie} \rangle \rangle .$$

These three kinds of patterns are called the *terminated* patterns.

A pattern, Q, is a *next pattern* after pattern, P, providing P is not terminated, and

- if P has an equal number of X's and O's, and Q is the same as P except that a cell that was empty in P has an X in Q, or
- if P has one more X than O's, and Q is the same as P except that a cell that was empty in P has an O in Q.

A Tic-Tac-Toe *game* with a tag that is a next pattern after P is called a *next move* from P. Let $\mathcal{G}_P$ be the set of next moves from P. Notice that $\mathcal{G}_P = \emptyset$ iff P is a terminated.

Constructor case: If P is a non-terminated Tic-Tac-Toe pattern, then
$$\langle P, \mathcal{G}_P \rangle$$
is a Tic-Tac-Toe game.

For example, if

$$P = \begin{array}{|c|c|c|}\hline X & O & X \\ \hline O & X & O \\ \hline O & & \\ \hline \end{array}$$

$$Q_1 = \begin{array}{|c|c|c|}\hline X & O & X \\ \hline O & X & O \\ \hline O & & X \\ \hline \end{array}$$

$$Q_2 = \begin{array}{|c|c|c|}\hline X & O & X \\ \hline O & X & O \\ \hline O & X & \\ \hline \end{array}$$

$$R = \begin{array}{|c|c|c|}\hline X & O & X \\ \hline O & X & O \\ \hline O & X & O \\ \hline \end{array}$$

Then,

$$\langle P, \{\langle Q_1, \langle \texttt{win} \rangle\rangle, \langle Q_2, \{\langle R, \langle \texttt{tie} \rangle\rangle\}\rangle\}\rangle \tag{4.10}$$

is the tagged recursive datum that corresponds to a Tic-Tac-Toe "end game" that starts with P. This game is easier to understand by looking at its game tree in Figure 4.3. Notice that the game tree —which so far we haven't actually defined —is simply the parse tree of the tagged datum.

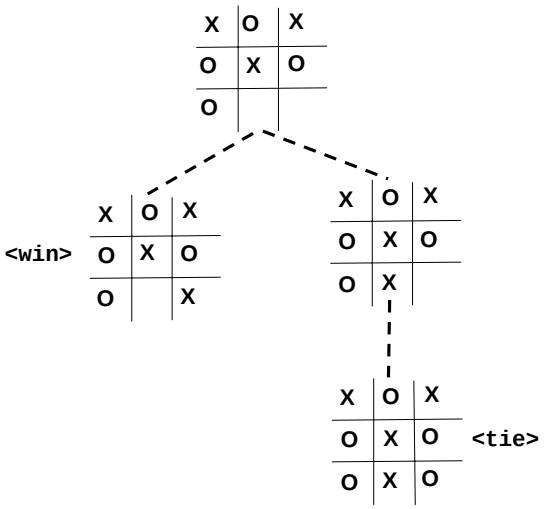


Figure 4.3: Game Tree for Tagged Datum (4.10), the Tic-Tac-Toe "End Game."

So the leaves at the bottom of the tree correspond to terminated games, and a path from the root (top node) to a leaf describes a complete *play* of the game. (In English, "game" can be used in two senses: first we can say that Chess is a game, and second we can play a game of Chess. The first usage refers to the data type of Chess games, and the second usage refers to a "play.")

Hmmm. Tic-Tac-Toe is pretty simple to understand —simpler than understanding this picky, precise definition. So why bother with this definition? Well, if all you were doing was explaining the game to a child, it would be nuts to use this definition. But not if you had to write a Tic-Tac-Toe-playing *computer program*. For the program to play right, you'd need this kind of picky precision.

4.3.2 Infinite Tic-Tac-Toe Games

At any point in a Tic-Tac-Toe game, there are at most nine possible next moves, and no play can continue for more than nine moves. But there are several ways to extend Tic-Tac-Toe into an amusing game on an $n \times n$ grid, for any $n \geq 3$. For these extensions, the number of possible next

moves from any point can be bounded by n^2, as can the length of any possible play. So such an $n \times n$ Tic-Tac-Toe still has a finite game tree.

But there's a natural way to combine all these games into a "meta-Tic-Tac-Toe" game. Namely, let the first player in meta-Tic-Tac-Toe choose any $n \geq 3$, after which the game continues an $n \times n$ grid. Now there are infinitely many possible first moves; but still, it's obvious that every possible play of the game is finite. It's not possible to keep moving forever —even though the game tree is infinite.

Meta-Tic-Tac-Toe isn't very hard to understand, but there is an important difference between it and the $n \times n$ games: even though every play must come to an end, there is no longer any finite bound on how many moves might be made before the end —a play might end after 100 moves, or 1000 moves, or 10^{10} moves; it just can't continue for an infinite number of moves.

Now that we understand meta-Tic-Tac-Toe, we can consider meta-meta-Tic-Tac-Toe —where the first player can choose either meta-Tic-Tac-Toe or an $n \times n$ Tic-Tac-Toe, after which the play continues in whatever game he chose. Then, of course, there's meta-meta-meta Tic-Tac-Toe....

4.3.3 Two Person Terminating Games

Familiar games like Tic-Tac-Toe, Checkers, and Chess can all end in ties, but for simplicity we'll only consider win/lose games —no "everybody wins"-type games at MIT. `:-)`. But everything we show about win/lose games will extend easily to games with ties.

Like Tic-Tac-Toe, the idea behind the definition of 2PTG's as a tagged recursive data type is that making a move in a 2PTG leads to a new position that defines the start of a new game. For Tic-Tac-Toe, the data tags were Tic-Tac-Toe patterns, but in general we use tags from an arbitrary set, Tags. This leads to the following very simple —perhaps deceptively simple —general definition.

Definition 4.3.2. The set, 2PTG, of *two-person terminating games of perfect information* is defined recursively as follows:

- **Base cases:**

$$\begin{array}{ll} \langle t, \texttt{win} \rangle & \in \text{2PTG, and} \\ \langle t, \texttt{lose} \rangle & \in \text{2PTG,} \end{array}$$

  where $t \in$ Tags.

- **Constructor case:** if $\mathcal{G}$ is a nonempty set of 2PTG's and $t \in$ Tags, then

$$G ::= \langle t, \mathcal{G} \rangle \in \text{2PTG.}$$

  The games in $\mathcal{G}$ are called the possible *next moves* of G.

These games are called "terminating" because, even though a 2PTG may be (very) infinite datum like meta-Tic-Tac-Toe, every play of a 2PTG must terminate. This is something we can now prove, after we give a precise definition of "play":

Definition 4.3.3. A *play* of a 2PTG, G, is a (potentially infinite) sequence of 2PTG's starting with G and such that if G_1 and G_2 are consecutive 2PTG's in the play, then G_2 is a possible next move of G_1.

If a 2PTG has no infinite play, it is called a *terminating* game.

Theorem 4.3.4. *Every 2PTG is terminating.*

Proof. By induction on the definition of a 2PTG, G, with induction hypothesis

$$G \text{ is terminating.}$$

Base case: If $G = \langle t, \texttt{win} \rangle$ or $G = \langle t, \texttt{lose} \rangle$ then the only possible play of G is the length one sequence consisting of G. Hence G terminates.

Constructor case: If $G = \langle t, \mathcal{H} \rangle$. Then any play of G is, by definition, a sequence starting with G and followed by a play of some $H \in \mathcal{H}$.

Now suppose G had an infinite play. Then this play starts with G and continues with an infinite play of some $H_0 \in \mathcal{H}$. Because H_0 has an infinite play, it is, by definition, not terminating. But by induction hypothesis, *every* $H \in \mathcal{H}$ is terminating, a contradiction. Hence G cannot have an infinite play, that is, G terminates.

This completes the structural induction, proving that

$$\forall \text{2PTG's } G.\ G \text{ is terminating.}$$

□

4.3.4 Game Strategies

A key question about a game is whether a player has a winning strategy. A *strategy* for a player in a game specifies which move the player should make at any point in the game when it is that player's turn. A *winning* strategy ensures that the player will win no matter what moves the other player makes.

In Tic-Tac-Toe for example, most elementary school children figure out strategies for both players that each ensure that the game ends with no tic-tac-toes, that is, it ends in a tie. Of course the first player can win if his opponent plays childishly, but not if the second player follows the winning strategy. In more complicated games like Checkers or Chess, it's not who clear that anyone has a winning strategy, even if we agreed to count ties as wins for the second player.

But structural induction makes it easy to prove that in any 2PTG, *somebody* has the winning strategy!

Theorem 4.3.5. ***Fundamental Theorem for Two-Person Games:*** *For every two-person terminating game of perfect information, there is a winning strategy for one of the players.*

Proof. The proof is by structural induction on the definition of a 2PTG, G. The induction hypothesis, is: there is a winning strategy for game G.

Base cases:

1. $G = \langle t, \texttt{win} \rangle$. Then the first player has the winning strategy: "make the winning move."
2. $G = \langle t, \texttt{lose} \rangle$. Then the second player has a winning strategy: "Let the first player make the losing move."

Constructor case: Suppose $G = \langle t, \mathcal{H} \rangle$. By structural induction, we may assume that some player has a winning strategy for each $H \in \mathcal{H}$. There are two cases to consider:

- some $H_0 \in \mathcal{H}$ has a winning strategy for its second player. Then the first player in G has a winning strategy: make the move to H_0 and then follow the second player's winning strategy in H_0.
- every $H \in \mathcal{G}$ has a winning strategy for its first player. Then the second player in G has a winning strategy: if the first player's move in G is to $H \in \mathcal{H}$, then follow the winning strategy for the first player in H.

So in any case, one of the players has a winning strategy for G, which completes the proof of the constructor case.

It follows by structural induction that there is a winning strategy for every 2PTG, G. □

Notice that although Theorem 4.3.5 guarantees a winning strategy, its proof gives no clue which player has it. For the Subset Takeaway Game (Team Problem, Friday, Week 2), and most other familiar 2PTG's like Checkers, Chess, Go, ..., no one knows which player has a winning strategy.

4.3.5 Structural Induction versus Ordinary Induction

In Computer Science, structural induction is the natural, preferred approach to proving properties of recursive data types. So you really should learn it.

Students will sometimes try to avoid structural induction in favor of ordinary induction by assigning nonnegative integer sizes to data items and then using ordinary induction on size. This works pretty generally, since each recursive datum can be assigned a size equal to the *smallest number of constructor steps* needed to build it. In this way, ordinary induction remains a viable, though more cumbersome, alternative approach to proofs about nearly all recursive data types that come up in practice, including all the examples we considered before we got to infinite games.

But infinite games are different. Not only are such games infinite, but the number of constructor steps to build them is infinite, so there's no apparent integer measure of game size that allows structural induction to be replaced by ordinary induction. In fact, it can't be done: it's a *meta*mathematical fact (that is, a mathematical fact *about* properties of Mathematics) that structural induction is more powerful than ordinary induction when it comes to reasoning about such infinite data items.[5]

So let the truth be known: while it's kind of fun to think about infinite games, the real point of examining them in these Notes is to set up an example where there is no alternative to structural induction proofs.

[5]There is a generalization of induction on nonnegative integers to induction on things called *ordinals* which is as powerful as structural induction, but the theory of ordinals is not a topic that belongs in an introduction to discrete Math.

4.4 State machines

State machines are an abstract model of step-by-step processes, and accordingly, they come up in many areas of Computer Science. You may already have seen them in a digital logic course, a compiler course, or a probability course.

4.4.1 Basic definitions

A state machine is really nothing more than a binary relation on a set, except that the elements of the set are called "states" and a pair (p, q) in the graph of the relation is called a "transition." The transition from state p to state q will be written $p \to q$. A state machine also comes equipped with a designated *start state*.

State machines used in digital logic and compilers usually have only a finite number of states, but machines that model continuing computations typically have an infinite number of states. In many applications, the states, and/or the transitions have labels indicating input or output values, costs, capacities, or probabilities, but for our purposes, unlabelled states and transitions are all we need.[6]

Example 4.4.1. A bounded counter, which counts from 0 to 99 and overflows at 100. The transitions are pictured in Figure 4.4, with start state zero.

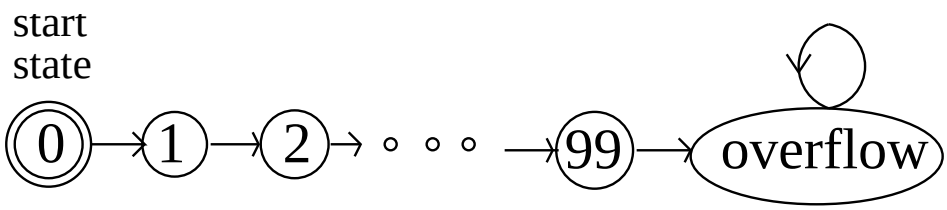


Figure 4.4: *State transitions for the 99-bounded counter.*

This machine isn't much use once it overflows, since it has no way to get out of its overflow state.

Example 4.4.2. An unbounded counter is similar, but has an infinite state set. This is harder to draw :-)

Example 4.4.3. In the movie *Die Hard 3: With a Vengeance*, the characters played by Samuel L. Jackson and Bruce Willis have to disarm a bomb planted by the diabolical Simon Gruber:

[6] We do name states, as in Figure 4.4, so we can talk about them, but the names aren't part of the state machine.

Simon: On the fountain, there should be 2 jugs, do you see them? A 5-gallon and a 3-gallon. Fill one of the jugs with exactly 4 gallons of water and place it on the scale and the timer will stop. You must be precise; one ounce more or less will result in detonation. If you're still alive in 5 minutes, we'll speak.

Bruce: Wait, wait a second. I don't get it. Do you get it?

Samuel: No.

Bruce: Get the jugs. Obviously, we can't fill the 3-gallon jug with 4 gallons of water.

Samuel: Obviously.

Bruce: All right. I know, here we go. We fill the 3-gallon jug exactly to the top, right?

Samuel: Uh-huh.

Bruce: Okay, now we pour this 3 gallons into the 5-gallon jug, giving us exactly 3 gallons in the 5-gallon jug, right?

Samuel: Right, then what?

Bruce: All right. We take the 3-gallon jug and fill it a third of the way...

Samuel: No! He said, "Be precise." Exactly 4 gallons.

Bruce: Sh - -. Every cop within 50 miles is running his a - - off and I'm out here playing kids games in the park.

Samuel: Hey, you want to focus on the problem at hand?

Fortunately, they find a solution in the nick of time. We'll let the reader work out how.

The *Die Hard* series is getting tired, so we propose a final *Die Hard Once and For All*. Here Simon's brother returns to avenge him, and he poses the same challenge, but with the 5 gallon jug replaced by a 9 gallon one.

We can model jug-filling scenarios with a state machine. In the scenario with a 3 and a 5 gallon water jug, the states will be pairs, (b, l) of real numbers such that $0 \leq b \leq 5, 0 \leq l \leq 3$. We let b and l be arbitrary real numbers. (We can prove that the values of b and l will only be nonnegative integers, but we won't assume this.) The start state is $(0, 0)$, since both jugs start empty.

Since the amount of water in the jug must be known exactly, we will only consider moves in which a jug gets completely filled or completely emptied. There are several kinds of transitions:

1. Fill the little jug: $(b, l) \to (b, 3)$ for $l < 3$.
2. Fill the big jug: $(b, l) \to (5, l)$ for $b < 5$.
3. Empty the little jug: $(b, l) \to (b, 0)$ for $l > 0$.
4. Empty the big jug: $(b, l) \to (0, l)$ for $b > 0$.

5. Pour from the little jug into the big jug: for $l > 0$,

$$(b, l) \to \begin{cases} (b + l, 0) & \text{if } b + l \leq 5, \\ (5, l - (5 - b)) & \text{otherwise.} \end{cases}$$

6. Pour from big jug into little jug: for $b > 0$,

$$(b, l) \to \begin{cases} (0, b + l) & \text{if } b + l \leq 3, \\ (b - (3 - l), 3) & \text{otherwise.} \end{cases}$$

Note that in contrast to the 99-counter state machine, there is more than one possible transition out of states in the Die Hard machine.

Quickie exercise: Which states of the Die Hard 3 machine have direct transitions to exactly two states?

4.4.2 Reachability and Invariants

The Die Hard 3 machine models every possible way of pouring water among the jugs according to the rules. Die Hard properties that we want to verify can now be expressed and proved using the state machine model. For example, Bruce's character will disarm the bomb if he can get to some state of the form $(4, l)$.

A (possibly infinite) sequence of transitions through successive states beginning at the start state corresponds to a possible system behavior; such a sequence is called an *execution* of the state machine. A state is called *reachable* if it appears in some execution. The bomb in Die Hard 3 gets disarmed successfully because the state (4,3) is reachable.

A useful approach in analyzing state machine is to identify *invariant* properties of states.

Definition 4.4.4. An *invariant* for a state machine is a predicate, P, on states, such that whenever $P(q)$ is true of a state, q, and $q \to r$ for some state, r, then $P(r)$ holds.

Now we can reformulate Induction in a convenient form for state machines:

The Invariant Principle

If a predicate is an invariant of a state machine, and the predicate holds for the start state, then it holds for all reachable states.

Die Hard Once and For All

Now back to Die Hard Once and For All. This time there is a 9 gallon jug instead of the 5 gallon jug. We can model this with a state machine whose states and transitions are specified the same way as for the Die Hard 3 machine, with all occurrences of "5" replaced by "9."

Now reaching any state of the form $(4, l)$ is impossible. We prove this using the Invariant Principle. Namely, we define the invariant predicate, $P(b, l)$, to be that b and l are nonnegative integer mulitples of 3. So P obviously holds for the state state $(0, 0)$.

To prove that P is an invariant, we assume $P(b, l)$ holds for some state (b, l) and show that if $(b, l) \to (b', l')$, then $P(b', l')$. The proof divides into cases, according to which transition rule is used. For example, suppose the transition followed from the "fill the little jug" rule. This means $(b, l) \to (b, 3)$. But $P(b, l)$ implies that b is an integer multiple of 3, and of course 3 is an integer multiple of 3, so P still holds for the new state $(b, 3)$. Another example is when the transition rule used is "pour from big jug into little jug" for the subcase that $b + l > 3$. Then state is $(b, l) \to (b - (3 - l), 3)$. But since b and l are integer multiples of 3, so is $b - (3 - l)$. So in this case too, P holds after the transition.

We won't bother to crank out the remaining cases, which can all be checked just as easily. Now by the Invariant Principle, we conclude that every reachable state satisifies P. But since no state of the form $(4, l)$ satisifies P, we have proved rigorously that Bruce dies once and for all!

A Robot on a Grid

There is a robot. It walks around on a grid, and at every step it moves diagonally in a way that changes its position by one unit up or down *and* one unit left or right. The robot starts at position $(0, 0)$. Can the robot reach position $(1, 0)$?

To get some intuition, we can simulate some robot moves. For example, starting at (0,0) the robot could move northeast to (1,1), then southeast to (2,0), then southwest to (1,-1), then southwest again to (0,-2).

Let's model the problem as a state machine and then prove a suitable invariant. A state will be a pair of integers corresponding to the coordinates of the robot's position. State (i, j) has transitions to four different states: $(i \pm 1, j \pm 1)$.

The problem is now to choose an appropriate invariant predicate, P, that is true for the start state $(0, 0)$ and false for $(1, 0)$. The Invariant Theorem then will imply that the robot can never reach $(1, 0)$. A direct attempt at an invariant is to let $P(q)$ be the predicate that $q \neq (1, 0)$.

Unfortunately, this is not going to work. Consider the state $(2, 1)$. Clearly $P(2, 1)$ holds because $(2, 1) \neq (1, 0)$. And of course $P(1, 0)$ does not hold. But $(2, 1) \to (1, 0)$, so this choice of P will not yield an invariant.

We need a stronger predicate. Looking at our example execution you might be able to guess a proper one, namely, that the sum of the coordinates is even! If we can prove that this is an invariant, then we have proven that the robot never reaches $(1, 0)$ because the sum $1 + 0$ of its coordinates is not an even number, but the sum $0 + 0$ of the coordinates of the start state is an even number.

Theorem 4.4.5. *The sum of the robot's coordinates is always even.*

Proof. The proof uses the Invariant Principle.

Let $P(i, j)$ be the predicate that $i + j$ is even.

First, we must show that the predicate holds for the start state $(0, 0)$. Clearly, $P(0, 0)$ is true because $0 + 0$ is even.

Next, we must show that P is an invariant. That is, we must show that for each transition $(i, j) \to (i', j')$, if $i+j$ is even, then $i'+j'$ is even. But $i' = i \pm 1$ and $j' = j \pm 1$ by definition of the transitions. Therefore, $i' + j'$ is equal to $i + j$ or $i + j \pm 2$, all of which are even. □

Corollary 4.4.6. *The robot cannot reach* $(1, 0)$.

Problem 4.4.1. A robot moves on the two-dimensional integer grid. It starts out at $(0, 0)$, and is allowed to move in any of these four ways:

1. (+2,-1) Right 2, down 1
2. (-2,+1) Left 2, up 1
3. (+1,+3)
4. (-1,-3)

Prove that this robot can never reach (1,1).

Robert W. Floyd

The Invariant Principle was formulated by Robert Floyd at Carnegie Tech in 1967[a]. Floyd was already famous for work on formal grammars which transformed the field of programming language parsing; that was how he got to be a professor even though he never got a Ph.D. (He was admitted to a PhD program as a teenage prodigy, but flunked out and never went back.)

In that same year, Albert R. Meyer was appointed Assistant Professor in the Carnegie Tech Computer Science Department where he first met Floyd. Floyd and Meyer were the only theoreticians in the department, and they were both delighted to talk about their shared interests. After just a few conversations, Floyd's new junior colleague decided that Floyd was the smartest person he had ever met.

Naturally, one of the first things Floyd wanted to tell Meyer about was his new, as yet unpublished, Invariant Principle. Floyd explained the result to Meyer, and Meyer wondered (privately) how someone as brilliant as Floyd could be excited by such a trivial observation. Floyd had to show Meyer a bunch of examples before Meyer understood Floyd's excitement —not at the truth of the utterly obvious Invariant Principle, but rather at the insight that such a simple theorem could be so widely and easily applied in verifying programs.

Floyd left for Stanford the following year. He won the Turing award —the "Nobel prize" of Computer Science— in the late 1970's, in recognition both of his work on grammars and on the foundations of program verification. He remained at Stanford from 1968 until his death in September, 2001. A eulogy describing Floyd's life and work by his closest colleague, Don Knuth, can be found at `http://www.acm.org/pubs/membernet/stories/floyd.pdf`.

[a] The following year, Carnegie Tech was renamed Carnegie-Mellon Univ.

4.4.3 Sequential algorithm examples

Proving Correctness

Robert Floyd, who pioneered modern approaches program verification, distinguished two aspects of state machine or process correctness:

1. The property that the final results, if any, of the process satisfy system requirements. This is called *partial correctness*.

   You might suppose that if a result was only partially correct, then it might also be partially incorrect, but that's not what he meant. The word "partial" comes from viewing a process that might not terminate as computing a *partial function*. So partial correctness means that when there is a result, it is correct, but the process might not always produce a result, perhaps because it gets stuck in a loop.

2. The property that the process always finishes, or is guaranteed to produce some legitimate final output. This is called *termination*.

Partial correctness can commonly be proved using the Invariant Principle. Termination can commonly be proved using the Well Ordering Principle. We'll illustrate Floyd's ideas by verifying the Euclidean Greatest Common Divisor (GCD) Algorithm.

The Euclidean Algorithm

The Euclidean algorithm is a three-thousand-year-old procedure to compute the greatest common divisor, $\gcd(a, b)$ of integers a and b. We can represent this algorithm as a state machine. A state will be a pair of integers (x, y) which we can think of as integer registers in a register program. The state transitions are defined by the rule

$$(x, y) \to (y, \text{remainder}(x, y))$$

for $y \neq 0$. The algorithm terminates when no further transition is possible, namely when $y = 0$. The final answer is in x.

We want to prove:

1. starting from the state with $x = a$ and $y = b > 0$, if we ever finish, then we have the right answer. That is, at termination, $x = \gcd(a, b)$. This is a *partial correctness* claim.

2. we do actually finish. This is a process *termination* claim.

Partial Correctness of GCD First let's prove that if GCD gives an answer, it is a correct answer. Specifically, let $d ::= \gcd(a, b)$. We want to prove that *if* the procedure finishes in a state (x, y), then $x = d$.

Proof. Define the state predicate

$$P(x,y) ::= \quad [\gcd(x,y) = d \text{ and } (x > 0 \text{ or } y > 0)].$$

P holds for the start state (a,b), by definition of d and the requirement that b is positive. Also, the invariance of P follows immediately from

Lemma 4.4.7. *For all $m, n \in \mathbb{N}$ such that $n \neq 0$,*

$$\gcd(m,n) = \gcd(n, \text{remainder}(m,n)). \tag{4.11}$$

Lemma 4.4.7 is easy to prove, and we'll leave it to the reader (a proof will appear in later Notes on elementary Number Theory). So by the Invariant Principle, P holds for all reachable states.

Since the only rule for termination is that $y = 0$, it follows that if (x,y) is a terminal state, then $y = 0$. If this terminal state is reachable, then the invariant holds for (x,y). This implies that $\gcd(x,0) = d$ and that $x > 0$. We conclude that $x = \gcd(x,0) = d$. □

Termination of GCD Now we turn to the second property, that the procedure must terminate. To prove this, notice that y gets strictly smaller after any one transition. That's because the value of y after the transition is the remainder of x divided by y, and this remainder is smaller than y by definition. But the value of y is always a natural number, so by the Well Ordering Principle, it reaches a minimum value among all its values at reachable states. But there can't be a transition from a state where y has its minimum value, because the transition would decrease y still further. So the reachable state where y has its minimum value is a state at which no further step is possible, that is, at which the procedure terminates.

Note that this argument does not prove that the minimum value of y is zero, only that the minimum value occurs at termination. But we already noted that the only rule for termination is that $y = 0$, so it follows that the minimum value of y must indeed be zero.

The Extended Euclidean Algorithm

An important fact about the $\gcd(a,b)$ is that it equals an integer linear combination of a and b, that is,

$$\gcd(a,b) = sa + tb \tag{4.12}$$

for some $s, t \in \mathbb{N}$. We'll see some nice proofs of (4.12) in later Notes, but there is also an extension of the Euclidean Algorithm that efficiently, if obscurely, produces the desired s and t. In particular, given nonnegative integers x,y, with $y > 0$, we claim the following procedure[7] halts with integers s, t in registers `S` and `T` satisfying (4.12).

Inputs: $a, b \in \mathbb{N}, b > 0$.

Registers: `X,Y,S,T,U,V,Q`.

Extended Euclidean Algorithm:

[7] This procedure is adapted from Aho, Hopcroft, and Ullman's text on algorithms.

```
X := a; Y := b; S := 0; T := 1; U := 1; V := 0;
loop:
if Y divides X, then halt
else
  Q := quotient(X,Y);
         ;;the following assignments in braces are SIMULTANEOUS
 {X := Y,
  Y := remainder(X,Y);
  U := S,
  V := T,
  S := U - Q * S,
  T := V - Q * T};
goto loop;
```

Note that X,Y behave exactly as in the Euclidean GCD algorithm in Section 4.4.3, except that this extended procedure stops one step sooner, ensuring that $\gcd(x, y)$ is in Y at the end. So for all inputs x, y, this procedure terminates for the same reason as the Euclidean algorithm: the contents, y, of register Y is a nonnegative integer-valued variable that strictly decreases each time around the loop.

We claim that invariant properties that can be used to prove partial correctness are:

$$\gcd(X, Y) = \gcd(a, b), \tag{4.13}$$

$$Sa + Tb = Y, \text{ and} \tag{4.14}$$

$$Ua + Vb = X. \tag{4.15}$$

To verify these invariants, note that invariant (4.13) is the same one we observed for the Euclidean algorithm. To check the other two invariants, let x, y, s, t, u, v be the contents of registers X,Y,S,T,U,V at the start of the loop and assume that all the invariants hold for these values. We must prove that (4.14) and (4.15) hold (we already know (4.13) does) for the new contents x', y', s', t', u', v' of these registers at the next time the loop is started.

Now according to the procedure, $u' = s, v' = t, x' = y$, so invariant (4.15) holds for u', v', x' because of invariant (4.14) for s, t, y. Also, $s' = u - qs, t' = v - qt, y' = x - qy$ where $q = \text{quotient}(x, y)$, so

$$s'a + t'b = (u - qs)a + (v - qt)b = ua + vb - q(sa + tb) = x - qy = y',$$

and therefore invariant (4.14) holds for s', t', y'.

Also, it's easy to check that all three invariants are true just before the first time around the loop. Namely, at the start $X = a, Y = b, S = 0, T = 1$ so $Sa + Tb = 0a + 1b = b = Y$ so (b) holds; also $U = 1, V = 0$ and $Ua + Vb = 1a + 0b = a = X$ so (4.15) holds. So by the Invariant Principle, they are true at termination. But at termination, the contents, Y, of register Y divides the contents, X, of register X, so invariants (4.13) and (4.14) imply

$$\gcd(a, b) = \gcd(X, Y) = Y = Sa + Tb.$$

So we have the gcd in register Y and the desired coefficients in S, T.

4.4.4 Derived Variables

The preceding termination proofs involved finding a natural-number-valued measure to assign to states. We might call this measure the "size" of the state. We then showed that the size of a state decreased with every state transition. By the Well Ordering Principle, the size can't decrease indefinitely, so when a minimum size state is reached, there can't be any transitions possible: the process has terminated.

More generally, the technique of assigning values to states —not necessarily nonnegative integers and not necessarily decreasing under transitions— is often useful in the analysis of algorithms. *Potential functions* play a similar role in physics. In the context of computational processes, such value assignments for states are called *derived variables*.

For example, for the Die Hard machines we could have introduced a derived variable, $f : \text{states} \to \mathbb{R}$, for the amount of water in both buckets, by setting $f((a,b)) ::= a + b$. Similarly, in the robot problem, the position of the robot along the x-axis would be given by the derived variable x-coord, where $x\text{-coord}((i,j)) ::= i$.

We can formulate our general termination method as follows:

Definition 4.4.8. A derived variable $f : \text{states} \to \mathbb{R}$ is *strictly decreasing* iff

$$q \to q' \text{ implies } f(q') < f(q).$$

Theorem 4.4.9. *If f is a strictly decreasing derived variable of a state machine that takes only nonnegative integer values, then the length of any execution starting at state q is at most $f(q)$.*

Of course we could prove Theorem 4.4.9 by induction on the value of $f(q)$. But think about what it says: "If you start counting down at some natural number $f(q)$, then you can't count down more than $f(q)$ times." Put this way, it's obvious.

Corollary 4.4.10. *If there exists a strictly decreasing natural-number-valued derived variable for some state machine, then every execution of that machine terminates.*

We now define some other useful flavors of derived variables taking values over partial ordered sets. We'll use the notational convention that when $\prec$ denotes a strict partial order on some set, then $\preceq$ is the corresponding *weak* partial order

$$a \preceq a' ::= \quad a \prec a' \vee a = a'.$$

Definition 4.4.11. Let $\prec$ be a strict partial order on a set, A. A derived variable $f : Q \to A$ is *strictly decreasing* with respect to $\prec$ iff

$$q \to q' \text{ implies } f(q') \prec f(q).$$

It is *weakly decreasing* iff

$$q \to q' \text{ implies } f(q') \preceq f(q).$$

Strictly increasing and *weakly increasing* derived variables are defined similarly.[8]

The existence of a natural-number-valued *weakly* decreasing derived variable does not guarantee that every execution terminates. That's because an infinite execution could proceed through states in which a weakly decreasing variable remained constant.

[8]Weakly increasing variables are often also called *nondecreasing*. We will avoid this terminology to prevent confusion between nondecreasing variables and variables with the much weaker property of *not* being a decreasing variable.

4.5 Well-Founded Orderings and Termination

4.5.1 Another Robot

Suppose we had a robot positioned at a point in the plane with natural number coordinates, that is, at an integer lattice-point in the Northeast quadrant of the plane. At every second the robot must move a unit distance South or West until it can no longer move without leaving the quadrant. It may also jump *any* integer distance East, but at every point in its travels, the number of jumps East is not allowed to be more than twice the number of previous moves South.

For example, suppose the robot starts at the position (9,8). It can now move South to (9,7) or West to (8,8); it can't jump East because there haven't been any previous South moves.

The robot's moves might continue along the following trajectory: South to (9,7), East to (23,7), South to (23,6), East to (399,6), West to (398,6), East to (511,6), West to (510,6), and East to $(10^5, 6)$. At this point it has moved South twice and East four times, so it can't jump East again until it makes another move South.

Claim 4.5.1. *The robot will always get stuck at the origin.*

If we think of the robot as a nondeterministic state machine, then Claim 4.5.1 is a termination assertion. The Claim may seem obvious, but it really has a different character than the termination results for the algorithms we've considered so far. That's because, even knowing that the starting position was (9,8), for example, there is no way to bound the total number of moves the robot can make before it gets stuck. So we will not be able to prove termination using the natural-number-valued decreasing variable method of Theorem 4.4.9. The robot can delay getting stuck at the origin for as many seconds as it wants; nevertheless, it can't avoid getting stuck eventually.

Does Claim 4.5.1 still seem obvious? Before reading further, it's worth thinking how you might prove it.

We will prove that the robot always gets stuck at the origin by generalizing the decreasing variable method, but with decreasing values that are more general than nonnegative integers. Namely, the traveling robot can be modeled with a state machine with states of the form $((x, y), s, e)$ where

- $(x, y) \in \mathbb{N}^2$ is the robot's position,
- s is the number of moves South the robot took to get to this position, and
- $e \leq 2s$ is the number of moves East the robot took to get to this position.

Now we define a derived variable value : States $\to \mathbb{N}^3$:

$$\text{value}(((x, y), s, e)) \ ::= \quad (y, 2s - e, x),$$

and we order the values of states with the *lexicographic* order, $\preceq_{\text{lex}}$, on $\mathbb{N}^3$:

$$(k, l, m) \preceq_{\text{lex}} (k', l', m') ::= \quad k < k' \text{ or } (k = k' \text{ and } l < l') \text{ or } (k = k' \text{ and } l = l' \text{ and } m \leq m') \quad (4.16)$$

Let's check that values are lexicographically decreasing. Suppose the robot is in state $((x, y), s, e)$.

- If the robot moves West it enters state $((x-1,y),s,e)$, and

$$\text{value}(((x-1,y),s,e)) = (y, 2s-e, x-1) \prec_{\text{lex}} (y, 2s-e, x) = \text{value}(((x,y),s,e)),$$

  as required.

- If the robot jumps East it enters a state $((z,y),s,e+1)$ for some $z > x$. Now

$$\text{value}(((z,y),s,e+1)) = (y, 2s-(e+1), z) = (y, 2s-e-1, z),$$

  but since $2s-e-1 < 2s-e$, the rule (4.16) implies that

$$\text{value}(((z,y),s,e+1)) = (y, 2s-e-1, z) \prec_{\text{lex}} (y, 2s-e, x) = \text{value}(((x,y),s,e)),$$

  as required.

- If the robot moves South it enters state $((x,y-1),s+1,e)$, and

$$\text{value}(((x,y-1),s+1,e)) = (y-1, 2(s+1)-e, x) \prec_{\text{lex}} (y, 2s-e, x) = \text{value}(((x,y),s,e)),$$

  as required.

So indeed state-value is a decreasing variable under lexicographic order. We'll show in the next section that it is impossible for a lexicographically-ordered value to be decreased an infinite number of times. That's just what we need to finish verifying Claim 4.5.1.

4.5.2 Well-founded Partial Orders

Definition 4.5.2. Let $\preceq$ be a partial order and S be a subset of its domain. An element $m \in S$ is minimal in S iff no other element in S is $\preceq m$. A partial order $\preceq$ is *well-founded* iff every nonempty subset of its domain has a minim*al* element.

So saying that the nonnegative integers are well-founded under $\leq$ is equivalent to the Well Ordering Principle, but well-foundedness makes sense much more generally than for just nonnegative integers.

So all finite partial orders are well-founded, since we saw in Week 3 Notes that every partial order in a finite set has a minimal element. (Of course, we can't expect to find a minim*um* element, since even in a finite partial order, there often isn't any minimum.)

There is another helpful way to characterize well-founded partial orders:

Lemma 4.5.3. *A partial order is well-founded iff it has no infinite decreasing chain.*

Saying that the partial order $\preceq$ has no infinite decreasing chain means there is no infinite sequence $p_1, p_2, \ldots, p_n \ldots$ of elements in P such that

$$p_1 \succ p_2 \succ \cdots \succ p_n \ldots.$$

Here we're using the notation "$p \succ q$" to mean [$q \preceq p$ and $q \neq p$]. That's so we can read the decreasing chain left to right, as usual in English.

Proof. (**left to right**) (By contradiction) If there was such an infinite decreasing sequence, then the set of elements in the sequence itself would be a nonempty subset without a minimal element.

(**right to left**) (By contradiction) Suppose $\preceq$ was not well-founded. So there is some subset $S \subseteq$ domain $(\preceq)$ P, such that S has at least one element, s_1, but S has no minimal element. In particular, since s_1 is not minimal, there must be *another* element $s_2 \in S$ such that $s_1 \succ s_2$. Similarly, since s_2 is not minimal, there must be still another element $s_3 \in S$ such that $s_2 \succ s_3$. Continuing in this way, we can construct an infinite decreasing chain $s_1 \succ s_2 \succ s_3 \cdots$ in S. □

An immediate corollary of Lemma 4.5.3 is

Corollary 4.5.4. *Every finite partial order is well-founded.*

Problem 4.5.1. Let D be the usual *dictionary order* on finite sequences of letters of the alphabet. Show that neither D nor D^{-1} is well-founded.

An easy way to construct well-founded partial orders is by taking products of well-founded partial orders. For example, the nonnegative integers are well-founded under $\leq$, so the product partial order $(\leq \times \leq)$ on pairs of nonnegative integers is going to be well-founded.

To prove this, we first generalize coordinatewise and lexicographic partial order to pairs of elements from *any* partial orders, not just nonnegative integers.

Definition 4.5.5. Let $\preceq_1$ and $\preceq_2$ be partial orders with domains A_1 and A_2.

The *coordinatewise partial order*, $\preceq_c$, is defined to be the product relation, $(\preceq_1 \times \preceq_2)$, defined in Week 3 Notes.

The *lexicographic partial order*, $\preceq_{\text{lex}}$, for $\preceq_1$ and $\preceq_2$ is defined by the conditions:

$$\begin{array}{rcl} \text{domain}\,(\preceq_{\text{lex}}) & ::= & A_1 \times A_2 \\ (a_1, a_2) \preceq_{\text{lex}} (b_1, b_2) & \text{iff} & a_1 \prec_1 b_1 \text{ or } (a_1 = b_1 \text{ and } a_2 \preceq_2 b_2). \end{array}$$

Note that calling these relations "partial orders" is accurate: we know from Week 3 Notes that products preserve partial orderings,so $\preceq_c$ is indeed a partial order. It's similarly easy to verify that $\preceq_{\text{lex}}$ is also a partial order.

But not only are these relations really partial orders, but they will also be well-founded providing $\preceq_1$ and $\preceq_2$ are well-founded. Namely,

Theorem 4.5.6. *Let $\preceq_1$ and $\preceq_2$ be well-founded partial orders. Then*

1. *$\preceq_{lex}$ is well-founded,*
2. *$\preceq_c$ is well-founded,*
3. *if $\preceq_1$ and $\preceq_2$ are both total orders, then so is $\preceq_{lex}$.*

Proof. To prove part 1., suppose $\emptyset \neq S \subseteq A_1 \times A_2$. We want to prove that S has a minimal element. The idea of the proof is easy: first find a minimal element among those appearing in the first coordinate of the partially ordered set of pairs, then for that minimal element, find the minimal element among the second coordinates of the pairs where it appears. Now here's a careful proof.

We begin by noting that

$$S_1 ::= \{a_1 \in A_1 \mid (a_1, a_2) \in S \text{ for some } a_2 \in A_2\}$$

is a nonempty subset of A_1, and so has a $\preceq_1$-minimal element, m_1. This means the set

$$S_{12} ::= \{a_2 \in A_2 \mid (m_1, a_2) \in S\}$$

is a nonempty subset of A_2 and so has a $\preceq_2$-minimal element, m_2. We claim that (m_1, m_2) is a minimal element of S under $\preceq_{\text{lex}}$.

To check this, note first that $(m_1, m_2) \in S$ by definition. So to show it is minimal, we need only show that if

$$(a_1, a_2) \in S, \text{ and} \tag{4.17}$$
$$(a_1, a_2) \preceq_{\text{lex}} (m_1, m_2) \tag{4.18}$$

then

$$(a_1, a_2) = (m_1, m_2). \tag{4.19}$$

But

$$a_1 \in S_1 \quad \text{by (4.17) and def. of } S_1, \tag{4.20}$$
$$a_1 \preceq_1 m_1 \quad \text{by (4.18) and def. of } \preceq_{\text{lex}}, \tag{4.21}$$
$$a_1 =_1 m_1 \quad \text{by (4.20), (4.21), and minimality of } m_1 \text{ in } S_1, \tag{4.22}$$
$$a_2 \in S_{12} \quad \text{by (4.17), (4.22) and def. of } S_{12}, \tag{4.23}$$
$$a_2 \preceq_2 m_2 \quad \text{by (4.18), (4.22), and def. of } \preceq_{\text{lex}}, \tag{4.24}$$
$$a_2 = m_2 \quad \text{by (4.23), (4.24), and minimality of } m_1 \text{ in } S_{12}. \tag{4.25}$$

Now (4.19) follows from (4.22) and (4.25), completing the proof of part 1.

Now notice that this argument above also holds if we replace $\preceq_{\text{lex}}$ by $\preceq_{\text{c}}$, allowing us to conclude that (m_1, m_2) is also a minimal element of S under $\preceq_{\text{c}}$. This proves part 2.

Part 3. follows straightforwardly from the definitions, and we leave its proof to the reader. □

Of course, the values of states for the robot in the previous section were triples not pairs, but we can easily define the lexicographic partial order on n-tuples for any $n \geq 1$. Namely,

Definition 4.5.7. Suppose $\preceq_1, \ldots, \preceq_n$ are partial orders with domains $A_1, \ldots, A_n$, and define the partial order, $\preceq_{\text{lex}}$, with domain, $A_1 \times \cdots \times A_n$, recursively in n:

$$\langle a_1 \rangle \preceq_{\text{lex}} \langle b_1 \rangle \quad \text{iff} \quad a_1 \preceq_1 b_1$$
$$\langle a_1, \ldots, a_n, a_{n+1} \rangle \preceq_{\text{lex}} \langle b_1, \ldots, b_n, b_{n+1} \rangle \quad \text{iff} \quad a_1 \preceq_1 b_1 \wedge \langle a_2, \ldots, a_{n+1} \rangle \preceq_{\text{lex}} \langle b_2, \ldots, b_{n+1} \rangle$$

Theorem 4.5.6 now generalizes straightforwardly to n-tuples. In particular, we conclude that since $\leq$ is a well-founded total order on $\mathbb{N}$, lexicographic order is a well-founded total order on $\mathbb{N}^n$ for all $n \geq 1$.

Now notice that the lexicographic order on $\mathbb{N}^3$ defined in the previous section (by the condition (4.16)) is exactly the same as $\preceq_{\text{lex}}$ on $\mathbb{N}^3$ according to Definition 4.5.7.

But we already proved that the value of the robot's state decreases at every step. And we have just proved that the order on these values is well-founded. So Lemma 4.5.3 implies that the values cannot keep decreasing forever. That means the robot cannot keep moving forever: it must always terminate.

Problem 4.5.2. Here is a generalization of the "choose-a-pair" game from Week 4, Friday, Team Problems to "choose-a-tuple". The rules are:

Player 1 chooses any integer $n \geq 1$. Then Player 2 chooses any n-tuple of nonnegative integers. After that, the players alternate moves, choosing as a move any n-tuple, t, of nonnegative integers such that no previous move is $\preceq_{\text{c}} t$. A player wins when the other player chooses the origin (0,. . . ,0).

For example, Player 1 might begin by choosing $n = 3$. Then Player 2 might choose the 3-tuple (8, 9, 10). Possible subsequent choices might then be

$$(7, 8, 9), (0, 1, 67), (83, 0, 0), (1, 0, 0), (0, 0, 1)(0, 1, 0)$$

This finally leaves Player 1 with only the move (0,0,0), and the game now ends with his loss.

(a) Prove that every choose-a-tuple play must end.

(b) Conclude that one of the players has a winning strategy.

(c) Which one? (This is an Open Problem – the 6.042 staff doesn't know the answer.)

4.6 In-Class Problems Week 4, Mon.

Problem 4.6.1. Provide simple *recursive* definitions of the following sets:

(a) The set $S ::= \left\{2^k 3^m 5^n \mid k, m, n \in \mathbb{N}\right\}$.

Solution. We can define the set S recursively as follows:

1. $1 \in S$
2. If $n \in S$, then $2n$, $3n$, and $5n$ are in S.

■

(b) The set $T ::= \left\{2^k 3^{2k+m} 5^{m+n} \mid k, m, n \in \mathbb{N}\right\}$.

Solution. We can define the set T recursively as follows:

1. $1 \in T$
2. If $n \in S$, then $18n$, $15n$, and $5n$ are in T.

■

(c) The set $L ::= \left\{(a, b) \in \mathbb{Z}^2 \mid a + 2b = 3k \text{ for some } k \in \mathbb{Z}\right\}$.

Solution. In other words

$$L = \{(a, b) \mid a + 2b \equiv 0 \pmod 3\} = \{(a, b) \mid a \equiv b \pmod 3\}.$$

So we can define the set L recursively as follows:

1. $(0, 0), (1, 1), (2, 2) \in L$
2. If $(a, b) \in L$, then $(a + 3, b)$, $(a - 3, b)$, $(a, b + 3)$, and $(a, b - 3)$ are in L.

Lots of other definitions are also possible. ■

Problem 4.6.2. The Elementary 18.01 Functions (F18's) are the set of functions of one real variable defined recursively as follows:

Base cases:

1. The identity function, $\mathrm{id}(x) ::= x$ is an F18,
2. any constant function is an F18,
3. the sine function is an F18,

Constructor cases:

If f, g are F18's, then so are

1. $f + g$, fg, e^g (the constant e),
2. the inverse function f^{-1},
3. the composition $f \circ g$.

Prove, by Structural Induction on this definition, that the Elementary 18.01 Functions are *closed under taking derivatives*. That is, show that if f is an F18, then so is df/dx.

Solution. *Proof.* By Structural Induction on def of $f \in$ F18. The induction hypothesis is the above statement to be shown.

Base Cases: We want to show that the derivatives of all the functions mentioned in rules 1., 2., 3. are in F18. This is easy: for example, $\frac{d}{dx}\,\mathrm{id}(x) = 1$ (constant) and so by rule 2., it is in F18. Similarly, $\frac{d}{dx}\sin(x) = \cos(x)$ which is also in F18 since $\cos(x) = \sin(x + \pi/2) \in$ F18 by rules 2, 3 and rule 3.

This proves that the induction hypothesis holds in the Base cases.

Constructor Cases: (f^{-1}). Assume $f, df/dx \in$ F18 to prove $\frac{d}{dx} f^{-1}(x) \in$ F18.

Letting $y = f(x)$, so $x = f^{-1}(y)$, we know from Leibniz's rule in calculus that

$$df^{-1}(y)/dy = dx/dy = \frac{1}{dy/dx}. \tag{4.26}$$

For example,

$$d\sin^{-1}(y)/dy = 1/(d\sin(x)/dx) = 1/\cos(x) = 1/\cos(\sin^{-1}(y)).$$

Stated as in (4.26), this rule is easy to remember, but can easily be misleading because of the variable switching between x and y. It's more clearly stated using variable-free notation:

$$(f^{-1})' = (1/f') \circ f^{-1}. \tag{4.27}$$

Now, since $f' \in$ F18 (by assumption), so is $1/f' = f'^{(-1)}$ (by 1.) and f^{-1} (by 2.), and therefore so is their composition (by 3). Hence the righthand side of equation (4.27) defines a function in F18.

Constructor Case: $(f \circ g)$. Assume $f, g, df/dx, dg/dx \in$ F18 to prove $d(f \circ g)(x)/dx \in$ F18.

The Chain Rule states that

$$\frac{d(f(g(x)))}{dx} = \frac{df(g)}{dg} \cdot \frac{dg}{dx}.$$

Stated more clearly in variable-free notation, this is

$$(f \circ g)' = (f' \circ g) \cdot g'.$$

The righthand side of this equation defines a function in F18 by rules 3. and 1.

The other Constructor cases are similar, so we conclude that the induction hypothesis holds in all Constructor cases.

This completes the proof by structural induction that the statement holds for all $f \in$ F18. □

■

Problem 4.6.3. BAexp's are defined in the Appendix.

(a) The value of flatten(e) for $e \in$ BAexp is the sequence of integers in e obtained by "erasing" everything but the integers that appear within tagged variables and tagged `int`'s. For example,

$$\begin{aligned} e &::= \langle \texttt{sum}, \langle \texttt{var}, 3\rangle, \langle \texttt{sum}, \langle \texttt{var}, 2\rangle, \langle \texttt{int}, 2\rangle\rangle\rangle \\ f &::= \langle \texttt{prod}, \langle \texttt{var}, 4\rangle, \langle \texttt{var}, 5\rangle\rangle \\ g &::= \langle \texttt{prod}, e, \langle \texttt{sum}, \langle \texttt{var}, 7\rangle, f\rangle\rangle \\ \text{flatten}(g) &= \langle 3, 2, 2, 7, 4, 5\rangle . \end{aligned}$$

Give a recursive definition of flatten. (You may use the operation of *concatenation* (append) of two sequences.)

Solution. We define flatten recursively on the definition of BAexp.

- **Base cases:**
  1. flatten$(\langle \texttt{int}, n\rangle) ::= \langle n\rangle$
  2. flatten$(\langle \texttt{var}, n\rangle) ::= \langle n\rangle$
- **Constructor cases:**
  1. flatten$(\langle \texttt{sum}, e_1, e_2\rangle) ::=$ flatten(e_1)flatten(e_2)
  2. flatten$(\langle \texttt{prod}, e_1, e_2\rangle) ::=$ flatten(e_1)flatten(e_2)

■

(b) Prove by structural induction on the definition of BAexp that for all $e \in$ BAexp,

$$2 \cdot \text{length}(\text{flatten}(e)) = |e| + 1$$

Solution. The proof is by structural induction on the definition of $e \in$ BAexp. The induction hypothesis is the equation above.

- **Base cases:**
  1. $2 \cdot \text{length}(\text{flatten}(\langle \texttt{int}, n\rangle)) = 2 \cdot 1 = 2 = 1 + 1 = |\langle \texttt{int}, n\rangle| + 1$.
  2. `var` case the same.

  So the equation holds in the base cases.
- **Constructor cases:**
  1. Say $e = \langle \texttt{sum}, e_1, e_2\rangle$ where we assume the Structural Induction hypothesis that e_1 and e_2 satisfy the equation given above. Now,

$$\begin{aligned} & 2 \cdot \text{length}(\text{flatten}(e)) \\ &= 2 \cdot \text{length}(\text{flatten}(\langle \texttt{sum}, e_1, e_2\rangle)) \\ &= 2 \cdot \text{length}(\text{flatten}(e_1)\text{flatten}(e_2)) && \text{(def of flatten)} \\ &= 2 \cdot \text{length}(\text{flatten}(e_1)) + 2 \cdot \text{length}(\text{flatten}(e_2)) && \text{(length of a string)} \\ &= (|e_1| + 1) + |e_2| + 1 && \text{(structural induction hyp.)} \\ &= |\langle \texttt{sum}, e_1, e_2\rangle| + 1 && \text{(def. of } |\langle \text{sum}, \ \rangle|) \\ &= |e| + 1. \end{aligned}$$

  So the equation holds for e.

2. `prod` the same.

This completes the proof for the Constructor cases.

We conclude by Structural Induction that the equation holds for all $e \in$ BAexp. ■

Appendix

The set, BAexp, of *Basic Arithmetic Expressions* is defined recursively as a tagged data type as follows:

- **Base cases:** If $n \in \mathbb{Z}$, then
  1. $\langle \texttt{int}, n \rangle \in$ BAexp, and
  2. $\langle \texttt{var}, n \rangle \in$ BAexp.
- **Constructor cases:** if $e, e' \in$ BAexp, then
  1. $\langle \texttt{sum}, e, e' \rangle \in$ BAexp, and
  2. $\langle \texttt{prod}, e, e' \rangle \in$ BAexp.

The size, $|e|$, of $e \in$ BAexp is defined recursively on this definition by:

- **Base cases:**
  1. $|\langle \texttt{int}, n \rangle| ::= 1$
  2. $|\langle \texttt{var}, n \rangle| ::= 1$
- **Constructor cases:**
  1. $|\langle \texttt{sum}, e_1, e_2 \rangle| ::= |e_1| + |e_2| + 1$
  2. $|\langle \texttt{prod}, e_1, e_2 \rangle| ::= |e_1| + |e_2| + 1$

4.7 In-Class Problems Week 4, Wed.

Problem 4.7.1. By now you are very familiar with the 6.042 icon that appears on the course webpage and lecture slides. This icon is a picture of a game called the **Fifteen Puzzle**. In this problem you will establish a basic property of the Fifteen Puzzle using the method of invariants, which may help you appreciate why this icon was chosen as the course logo.

The Fifteen Puzzle consists of sliding square tiles numbered $1, \ldots, 15$ held in a 4×4 frame with one empty square. Any tile adjacent to the empty square can slide into it.

The standard initial position is

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	

We would like to reach the target position (known in my youth as "the impossible" — ARM):

15	14	13	12
11	10	9	8
7	6	5	4
3	2	1	

A state machine model of the puzzle has states consisting of a 4×4 matrix with 16 entries consisting of the integers $1, \ldots, 15$ as well as one "empty" entry—like each of the two arrays above.

The state transitions correspond to exchanging the empty square and an adjacent numbered tile. For example, an empty at position $(2, 2)$ can exchange position with tile above it, namely, at position $(1, 2)$:

n_1	n_2	n_3	n_4
n_5		n_6	n_7
n_8	n_9	n_{10}	n_{11}
n_{12}	n_{13}	n_{14}	n_{15}

$\longrightarrow$

n_1		n_3	n_4
n_5	n_2	n_6	n_7
n_8	n_9	n_{10}	n_{11}
n_{12}	n_{13}	n_{14}	n_{15}

We will use the invariant method to prove that there is no way to reach the target state starting from the initial state.

We begin by noting that a state can also be represented as a pair consisting of two things:

1. a list of the numbers $1, \ldots, 15$ in the order in which they appear—reading rows left-to-right from the top row down, ignoring the empty square, and
2. the coordinates of the empty square—where the upper left square has coordinates $(1, 1)$, the lower right $(4, 4)$.

(a) Write out the "list" representation of the start state and the "impossible" state.

Solution. start: ((1 2 ... 15), (4,4)),

impossible: ((15 14 ... 1), (4,4)). ■

Let L be a list of the numbers $1, \ldots, 15$ in some order. A pair of integers is an *out-of-order pair* in L when the first element of the pair both comes *earlier* in the list and *is larger*, than the second element of the pair. For example, the list $1, 2, 4, 5, 3$ has two out-of-order pairs: (4,3) and (5,3). The increasing list $1, 2 \ldots n$ has no out-of-order pairs.

Let a state, S, be a pair $(L, (i, j))$ described above. We define the *parity* of S to be the mod 2 sum of the number, $p(L)$, of out-of-order pairs in L and the row-number of the empty square, that is the parity of S is $p(L) + i \pmod 2$.

(b) Verify that the parity of the start state and the target state are different.

Solution. The parity of the start state is

$$(0 + 4) \bmod 2 = 0.$$

The parity of the target is

$$((15 \cdot 14/2) + 4) \bmod 2 = 1.$$

■

(c) Show that the parity of a state is invariant under transitions. Conclude that "the impossible" is impossible to reach.

Solution. To show that the parity is constant, consider how moves may affect the parity. There are only 4 types of moves: a move to the left, a move to the right, a move to the row above, or a move to the row below.

Note that horizontal moves change nothing, and vertical moves both change i by 1, and move a tile three places forward or back in the list, L. To consider how the parity is changed in this case, we need to consider only the 3 pairs in L that are between the tile's old and new position. (The other pairs are not effected by the tile's move). This reverses the order of three pairs in L, changing the number of inversions by 3 or 1, but always by an odd amount.

To confirm this last remark, note that if the 3 pairs were all out of order or all in order before, the amount is changed by 3. If two pairs were out of order and 1 pair was in order or if one pair was out of order and two were in order, this will change the amount by 1. So the sum of i and the number of out-of-order pairs changes by an even amount (either 1+3 or 1+1), which implies that its parity remains the same. Since the initial state has parity 0 (even), all states reachable from the initial state must have parity 0, so the target state with parity 1 can't be reachable. ■

By the way, if two states have the same parity, then in fact there *is* a way to get from one to the other. If you like puzzles, this is a good one to think about on your own after class.

Problem 4.7.2. The most straightforward way to compute the bth power of a number, a, is to multiply a by itself b times. This of course requires $b-1$ multiplications. There is another way to do it using considerably fewer multiplications. This algorithm is called *Fast Exponentiation*:

Given inputs $a \in \mathbb{R}, b \in \mathbb{N}$, initialize registers x, y, z to $a, 1, b$ respectively, and repeat the following sequence of steps until termination:

- if $z = 0$ **return** y and terminate
- $r :=$ remainder$(z, 2)$
- $z :=$ quotient$(z, 2)$
- if $r = 1$, then $y := xy$
- $x := x^2$

We claim this algorithm always terminates and leaves $y = a^b$.

(a) Model this algorithm with a state machine, carefully defining the states and transitions.

Solution. 1. The set of states is $\mathbb{R} \times \mathbb{R} \times \mathbb{N}$,

2. The start state is $(a, 1, b)$,
3. the transitions are defined by the rule

$$(x, y, z) \to \begin{cases} (x^2, y, \text{quotient}(z, 2)) & \text{if } z \text{ is positive and even,} \\ (x^2, xy, \text{quotient}(z, 2)) & \text{if } z \text{ is positive and odd.} \end{cases}$$

■

(b) Let $d ::= a^b$. Verify that the following predicate, P, is an invariant:

$$P((x, y, z)) ::= \quad [yx^z = d].$$

Solution. We show that P is invariant, namely, assuming $P((x, y, z))$,

$$yx^z = d$$

holds and $(x, y, z) \to (x_t, y_t, z_t)$ is a transition, then $P((x_t, y_t, z_t))$,

$$y_t x_t^{z_t} = d$$

holds.

We consider two cases:

If $z > 0$ and is even, then we have that $x_t = x^2, y_t = y, z_t = \text{quotient}(z, 2)$. Therefore,

$$\begin{aligned} y_t x_t^{z_t} &= y x^{2 \cdot \text{quotient}(z,2)} \\ &= y x^{2 \cdot \frac{z}{2}} \\ &= y x^z \\ &= d \qquad \text{(by } P((x, y, z))) \end{aligned}$$

If $z > 0$ and is odd, then we have that $x_t = x^2, y_t = xy, z_t = \text{quotient}(z, 2)$. Therefore,

$$\begin{aligned} y_t x_t^{z_t} &= xyx^{2\cdot\text{quotient}(z,2)} \\ &= yx^{1+2\cdot\frac{(z-1)}{2}} \\ &= yx^{1+(z-1)} \\ &= yx^z \\ &= d \qquad \text{(by } P((x,y,z))\text{)} \end{aligned}$$

So in both cases, $P((x_t, y_t, z_t))$ holds, proving that P is an invariant. ■

(c) Prove that the algorithm is partially correct: if it halts, it does so with $y = d$.

Solution. P holds for the start state $(a, 1, b)$ since $1 \cdot a^b = a^b = d$ by definition. So by the Invariant Theorem, P holds for all reachable states. But a terminal state must have $z = 0$, so if any terminal state $(x, y, 0)$ is reachable, then $y = yx^0 = d$ as required. ■

(d) Prove that the algorithm terminates.

Solution. Just notice that z is a natural-number-valued variable that gets smaller at every transition. So by the Well-Ordering Principle, when this variable reaches its minimum value, the algorithm terminates. ■

(e) In fact, prove that it requires at most $2\log_2 b$ multiplications for the Fast Exponentiation algorithm to compute a^b for $b > 1$.

Solution. The value of z is initially b and gets halved at least at every step. So it can't be halved more than $\log_2 b$ times before hitting zero. ■

4.8 In-Class Problems Week 4, Fri.

Problem 4.8.1. Define *lexicographic order*, $\prec_{\text{lex}}$, on $\mathbb{N}^2$ by the rule:

$$(a_2, b_2) \prec_{\text{lex}} (a_1, b_1)$$

iff

$$a_2 < a_1 \text{ or } [a_2 = a_1 \text{ and } b_2 < b_1].$$

Prove that there is *no* infinite $\prec_{\text{lex}}$-decreasing sequence

$$(a_1, b_1) \succ_{\text{lex}} (a_2, b_2) \succ_{\text{lex}} \cdots \succ_{\text{lex}} (a_n, b_n) \succ_{\text{lex}} \cdots \tag{4.28}$$

Hint: Consider the smallest a_m.

Solution. By contradiction. Suppose there was such a lexicographically decreasing sequence (4.28). By the Well-ordering Principle (WoP), there must be a minimum value among the elements of the set $\{a_1, a_2, \dots\}$. Also, by definition of lexicographic order, the sequence $a_1, a_2, \dots$ is weakly decreasing. So if a_m is the minimum value, then the sequence of a_i's is constant from the mth element on.

Now for $i \geq m$, we have by (4.28)

$$(a_m, b_i) = (a_i, b_i) \succ_{\text{lex}} (a_{i+1}, b_{i+1}) = (a_m, b_{i+1})$$

and so $b_i > b_{i+1}$ by definition of lexicographic order. Hence, $\{b_m, b_{m+1}, \dots\}$ is an infinite set of natural numbers with no least element, contradicting the WoP. ■

Problem 4.8.2. Consider the following game for two players. The players alternate moves. A move consists of a pair (x, y) of natural numbers, subject to the constraint that none of the previous moves may simultaneously be below and to the left of the current move. That is, if (x, y) is the current move and (x', y') is any previous move, then either $x < x'$ (so the previous move is to the right of the current one) or $y < y'$ (so the previous move is above the previous one). A player who moves to the origin $(0, 0)$ is the loser.

For example, the Player 1 might choose (5,6), after which Player 2 can move to any point (n, m) such that $n < 5$ or $m < 6$, for example, (4,12). Now the players might move successsively to (4,11), (29,5), (1,1), (0,54), (0,1). At this point it's Player 2's turn, and he can move to (1,0). At this point it is Player 1's move, and the only available move is to the origin (0,0), so Player 1 loses this play of the game.

(a) Identify a winning strategy for the first player, and argue its correctness. Does your strategy guarantee any bound on the number of game moves?

Solution. The first player essentially has two winning strategies:

Strategy 1: The first player starts with $(1, 1)$. Then the second player can only pick $(0, n)$ or $(n, 0)$ for some n. The first player then responds with, $(n, 0)$ or $(0, n)$, respectively. By symmetry, the first player will always have a move whenever the second player had a move, except when the second player picks $(0, 0)$ and loses.

Strategy 2: The second winning strategy is slightly more tricky: The first player starts by picking $(2, 0)$,[9] and then:

- If the 2nd player picks $(0, n)$ for any $n \geq 1$, 1st player responds with $(1, n - 1)$.
- If the 2nd player picks $(1, n)$ for any $n \geq 0$, 1st player responds with $(0, n + 1)$.

The above procedure should be repeated for all consecutive moves. Notice that if Player 2 had a valid move, so does Player 1, unless Player 2 picked $(0, 0)$ and lost. ■

(b) Even if the players conspire to keep the game going as long as possible, it will necessarily terminate. Prove this as follows:

i. At any point in the game, let x_m be the minimum of the x coordinates of all of the previous moves, and likewise, y_m be the minimum of the y coordinates of all of the previous moves. Verify that $x_m + y_m$ is a weakly decreasing natural-number valued variable.

Solution. That's because min's cannot increase as more moves are made, so neither can their sum. ■

ii. Suppose a is the least number such that a move (x_m, a) has been made, and likewise b is least such that move (b, y_m) has been made. The *bounded moves* are defined to be the possible moves in the rectangle with corners at (x_m, a) and (b, y_m). Explain why the only moves that do not decrease $x_m + y_m$ must be bounded moves.

Solution. Moves North of the rectangle are not allowed because they would be coordinatewise larger than (x_m, a), moves East of the rectangle are not allowed because they would be coordinatewise larger than (b, y_m), and moves Northeast would be bigger than both. Moves South of the rectangle decrease the minimum value of y, moves West decrease the minimum x, and moves Southwest decrease both. That leaves only moves within the rectangle as possibly allowed moves that do not decrease $x_m + y_m$. ■

iii. Define the *size*, σ, of a game position to be $(x_m + y_m, k)$ where k is the number of bounded moves. Explain why σ is a strictly decreasing variable under lexicographic order, and use this to conclude that the game must terminate.

Solution. (i.) and (ii.) immediately imply that σ is a strictly decreasing variable under lexicographic order on $\mathbb{N}^2$. Since lexicographic order has no infinite decreasing sequences, the sizes of all the states reached in any game must have a minimum value. The state with such a minimum value must be a terminal state. ■

(c) *(Optional)* Is there a winning strategy for the first player that guarantees a bound on the number of game moves?

[9] Of course, picking $(0, 2)$ would be equivalent, and the following argument will hold for it as well, with x and y coordinates replaced.

Solution. No, because neither of the winning strategies described in the solution to part (a) guarantees a bound, and there are no other winning strategies.

To see this, suppose Player 1 starts with any legal move other than the ones above. Player 2 will be able to adopt one of the two winning strategies for himself and win for sure:

- If Player 1 picks some (m, n) where $m, n \geq 1$ and $(m, n) \neq (1, 1)$, then Player 2 can pick $(1, 1)$ next, and then use the first strategy to win against Player 1.
- If Player 1 picks some $(0, n)$ or $(n, 0)$ where $n \geq 3$, then Player 2 can respectively pick $(0, 2)$ or $(2, 0)$ next, and then use the second strategy to win against Player 1.
- If Player 1 picks $(1, 0)$ or $(0, 1)$ then Player 2 picks the other one and wins.

■

(d) Conclude that under the coordinatewise partial order, $\mathbb{N}^2$ has no infinite antichain.

Solution. The elements of an antichain, listed in any order, is an allowed sequence of moves. An infinite antichain would describe a nonterminating game, contradicting the fact that the game must terminate. ■

(e) In $\mathbb{N}^2$ under the coordinatewise partial order, describe an antichain of size n, for each $n > 1$.

Solution. (0,n-1),(1,n-2),(2,n-3),. . . ,(n-1,0) ■

4.9 Problem Set 3

Problem 4.9.1. Let m, n be integers, not both zero. Define a set of integers, $L_{m,n}$, recursively as follows:

- **Base cases**: $m, n \in L_{m,n}$.
- **Constructor steps**: If $j, k \in L_{m,n}$, then
  1. $-j \in L_{m,n}$,
  2. $j + k \in L_{m,n}$.

Let L be an abbreviation for $L_{m,n}$ in the rest of this problem.

(a) Show *by structural induction* that

$$L \subseteq \{mx + ny \mid x, y \in \mathbb{Z}\}.$$

Solution. We need to prove that every number in L is of the form $mx+ny$. We do this by structural induction on the definition of L.

Base cases: The base cases are of the required form because

$$m = m \cdot 1 + n \cdot 0,$$
$$n = m \cdot 0 + n \cdot 1.$$

Constructor steps: We must prove that $-j$ and $j + k$ are of the required form for $j, k \in L$, where by structural induction hypothesis, we may assume j, k are of the required form. But this follows immediately since

$$-j = -(mx + ny) = m(-x) + n(-y)$$
$$j + k = (mx + ny) + (mx' + ny') = m(x + x') + n(y + y')$$

This completes the structural induction. ■

(b) Show that

$$\{mx + ny \mid x, y \in \mathbb{Z}\} \subseteq L.$$

Solution. We must show that $mx + ny \in L$ for all integers x, y. To begin, suppose $j \in L$ and let

$$P_j(k) ::= jk \in L.$$

We prove that $P_j(k)$ holds for all $k \in \mathbb{N}$ by induction on k, with induction hypothesis, P_j.

Base case ($k = 0$): $j0 = 0 = m + (-m) \in L$.

Induction step: Assume $P(k)$, so $jk \in L$. But we are given that $j \in L$, and $j(k + 1) = jk + j \in L$ by the second recursive step in the definition of L. This proves $P(k+1)$, completing the induction proof.

Now since $jk \in L$ implies $-jk \in L$ by the first recursive step in the definition of L, we conclude

Lemma. *If $j \in L$, then $jx \in L$ for all $x \in \mathbb{Z}$.*

But $m, n \in L$ by definition of L, so the Lemma implies that $mx \in L$ for $x \in \mathbb{Z}$, and also $ny \in L$ for all $y \in \mathbb{Z}$. Now by the second recursive step in the definition of L, we conclude that $mx + ny \in L$ for all integers x, y. ■

(c) Conclude that any common divisor of m and n also divides every member of L.

Solution. If $k \in L$, then by part (a), $k = mx + ny$. Now if d is a common divisor of m and n, then it divides mx and ny, and hence divides their sum $mx + ny$. So d also divides k. ■

(d) Show that if $j, k \in L$ and $k \neq 0$, then the remainder of j divided by k is in L.

Solution. Say r is the remainder of j divided by k. That is, $j = qk + r$ for some quotient $q \in \mathbb{Z}$ and remainder, r, where $0 \leq r < |k|$.

Now the Lemma in the solution to part b implies that $(-q)k \in L$, and then the second recursive step in the definition of L implies that $j + (-q)k \in L$. That is, $r \in L$. ■

(e) Show that there is a positive integer $g \in L$ which divides every member of L. *Hint:* The least positive integer in L.

Solution. At least one of the integers $m, -m, n, -n \in L$ must be positive. Hence by the Well Ordering Principle, there is a least positive integer $g \in L$.

Suppose $j \in L$. We must show that $g \mid j$.

We prove this by contradiction: if g does not divide j, then the remainder of j divided by g is a positive number smaller than g, which by the previous part is in L, contradicting the fact that g is the smallest such integer. ■

(f) Conclude that $\gcd(m, n) = mx + ny$ for some $x, y \in \mathbb{Z}$.

Solution. From part (e), we have a positive integer $g \in L$ that divides every element of L. In particular, g is a common divisor of m, n, since $m, n \in L$.

Further, $g = mx + ny$ for some $x, y \in \mathbb{Z}$ by part (a), and any common divisor of m and n divides g, by part (c). So g is a common divisor that is divided by, and hence at least as large as, any common divisors. So g must be the *greatest* common divisor of m and n. ■

Problem 4.9.2. We define full binary trees, FBT's, as a tagged recursive data type with tags from some set of "labels."

Definition. Base case: $\langle l, \texttt{leaf} \rangle$ is an FBT, where l is a label.

Constructor case: If B_1 and B_2 are FBT's, then $\langle l,\ B_1,\ B_2 \rangle$ is an FBT, where l is a label.

The labels and leaf labels *appearing* in an FBT, B, are defined recursively in the obvious way:

Definition. Base case: If $B = \langle l, \texttt{leaf}\rangle$. Then l *appears* in B and is a *leaf label* of B.

Constructor case: If $B = \langle l, B_1,\ B_2\rangle$ is an FBT, then the labels that *appear* in B are the ones that appear in B_1, or in B_2; also, l *appears* in B. The *leaf labels* of B are the union of the leaf labels of B_1 and the leaf labels of B_2.

The FBT's with *unique* labels are also defined recursively:

Definition. Base case: If $B = \langle l, \texttt{leaf}\rangle$. Then B has *unique labels*.

Constructor case: If $B = \langle l, B_1,\ B_2\rangle$ is an FBT, then B has *unique labels* iff l does not appear in B_1 or B_2, and no other label appears in both B_1 and B_2.

If B is an FBT, let n_B be the number of labels appearing in B and f_B be the number of leaf labels of B. Use structural induction to prove that

$$f_B = \frac{n_B + 1}{2} \tag{4.29}$$

for all FBT's *with unique labels*. Also give a counterexample for an FBT that does not have unique labels. So your proof had better use uniqueness of labels at some point; be sure to indicate where.

Solution. Base case: If $B = \langle l, \texttt{leaf}\rangle$, then $f_B = n_B = 1$ and

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

proving equation (4.29) in this case.

Constructor case: If $B = \langle l, B_1,\ B_2\rangle$, has unique labels then

$$\begin{aligned}
f_B &= |\text{leaf-labels}(B_1) \cup \text{leaf-labels}(B_2)| && \text{(by def. of leaf labels)}\\
&= f_{B_1} + f_{B_2} && \text{(no label appears in both } B_1 \text{ and } B_2)\\
&= \frac{n_{B_1}+1}{2} + \frac{n_{B_2}+1}{2} && \text{(by structural induction hypothesis)}\\
&= \frac{(1 + n_{B_1} + n_{B_2}) + 1}{2}\\
&= \frac{|\{l\} \cup \text{labels}(B_1) \cup \text{labels}(B_2)| + 1}{2} && \text{(uniqueness of labels)}\\
&= \frac{n_B + 1}{2} && \text{(by def. of } n_B\text{).}
\end{aligned}$$

This proves (4.29) holds for B, completing the proof of the Constructor case. It follows by structural induction that (4.29) holds for all FBT's with unique labels.

A counterexample is the FBT

$$C ::= \quad \langle 0, \langle 1, \texttt{leaf}\rangle, \langle 1, \texttt{leaf}\rangle\rangle,$$

where $n_C = 2$ and $f_C = 1 \neq 1.5 = (2+1)/2 = (n_C + 1)/2$. ■

Problem 4.9.3. *Search Trees*, ST, are a special case of the FBT's of Problem 4.9.2 with labels that are numbers. They are defined recursively as follows:

Base case: $\langle x, \texttt{leaf}\rangle$ is an ST for any real number, x.

Constructor case: If T_1 and T_2 are ST's, and x is a real number such that every label that appears in T_1 is smaller than x, and every label that appears in T_2 is larger than x, then

$$\langle x, T_1, T_2\rangle$$

is an ST.

(a) Draw all the ST's with labels 1, 2, 3, 4, 5.

Solution. See Figure 4.5.

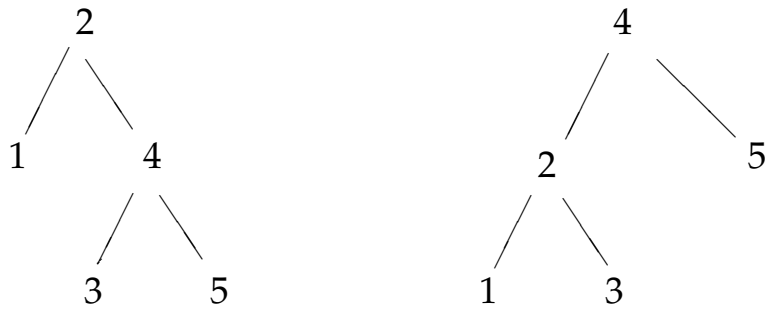


Figure 4.5: The two possible search trees

■

(b) Define a function appears-in : $(\mathbb{R} \times \text{ST}) \to \{0, 1\}$ recursively on the definition of ST's as follows:

Base case:

$$\begin{aligned} &\text{appears-in}(x, \langle x, \texttt{leaf}\rangle) ::= 1; \\ &\text{appears-in}(x, \langle y, \texttt{leaf}\rangle) ::= 0, \qquad\qquad \text{if } x \neq y. \end{aligned}$$

Constructor case:

$$\text{appears-in}(x, \langle x, T_1, T_2\rangle) ::= 1; \tag{4.30}$$

$$\text{appears-in}(x, \langle y, T_1, T_2\rangle) ::= \text{appears-in}(x, T_1) \qquad \text{if } x < y; \tag{4.31}$$

$$\text{appears-in}(x, \langle y, T_1, T_2\rangle) ::= \text{appears-in}(x, T_2) \qquad \text{if } x > y. \tag{4.32}$$

Prove by structural induction on the definition of ST that

$$\text{appears-in}(x, T) = 1 \quad \text{iff} \quad x \text{ appears in } T. \tag{4.33}$$

Solution. *Proof.* The proof will be by structural induction on the definition of ST.

Base case: $T = \langle x, \texttt{leaf}\rangle$. Then y appears in T iff $y = x$ iff appears-in$(x, T) = 1$, proving (4.33) holds for T.

Constructor case: $T = \langle x, T_1, T_2\rangle$.

$$
\begin{array}{lllll}
y & \text{appears in } T & & & \\
 & \text{iff} \quad y = x & \text{or} & [y \text{ appears in } T_1] & \\
 & & \text{or} & [y \text{ appears in } T_2] & \\
 & & & \text{(def. of ``appears in'')} & \\
 & \text{iff} \quad y = x & \text{or} & [x < y \text{ and } y \text{ appears in } T_1] & \\
 & & \text{or} & [x > y \text{ and } y \text{ appears in } T_2] & \\
 & & & \text{(constructor case of def. of ST)} & \\
 & \text{iff} \quad y = x & \text{or} & [x < y \text{ and appears-in}(y, T_1) = 1] & \\
 & & \text{or} & [x > y \text{ and appears-in}(y, T_2) = 1] & \\
 & & & \text{(ind. hyp. (4.33))} & \\
 & \text{iff} \quad y = x & \text{or} & [x < y \text{ and appears-in}(y, \langle x, T_1, T_2\rangle) = 1] & \\
 & & \text{or} & [x > y \text{ and appears-in}(y, \langle x, T_1, T_2\rangle) = 1] & \\
 & & & \text{(constructor cases 4.31 and 4.32 of appears-in def.)} & \\
 & \text{iff} & & \text{appears-in}(y, \langle x, T_1, T_2\rangle) = 1 & \\
 & & & \text{(constructor cases 4.30, 4.31, and 4.32 of appears-in def.)} &
\end{array}
$$

This proves the induction hypothesis (4.33) for T, which completes the proof by structural induction. □

■

Problem 4.9.4. The concatenation st of strings s and t is the string beginning with the symbols in s followed by the symbols in t. Notes 4 gave a recursive definition of concatenation based on the the definition of strings as a tagged recursive data type. As a way to ensure the somewhat technical recursive definition works as intended, we will verify that the recursive definition implies some of the basic properties we expect. So in what follows, you should *not* assume that the recursive definition of concatenation matches the "string followed by string" formulation, because that's what we're trying to check. But you do have a powerful principle for proving properties of recursive definitions, namely structural induction.

(a) By the base case of the recursive definition of concatenation in Notes 4, concatenating a string with the empty string, λ, *on the right* leaves the string unchanged, that is, $s\lambda = s$. From the "string followed by string" formulation, it's clear that concatenating *on the left* with the empty string also leaves a string unchanged. That is,

$$\lambda s = s \tag{4.34}$$

for all strings, s. Use structural induction to prove that the recursively defined concatenation operation satisfies (4.34).

Solution. The induction hypothesis will be equation (4.34). The proof is by structural induction on $s \in A^*$.

- **Base case** $s = \lambda$: We have

$$\begin{aligned} \lambda s &= \lambda\lambda \\ &= \lambda && \text{(base case of concat)} \\ &= s. \end{aligned}$$

  So (4.34) holds in this case.

- **Constructor case:** $s = ra$ where $r \in A^*, a \in A$. By structural induction, we may assume

$$\lambda r = r \tag{4.35}$$

  We have

$$\begin{aligned} \lambda s &= \lambda(ra) \\ &= (\lambda r)a && \text{(constructor case of concat)} \\ &= ra && \text{by (4.35)} \\ &= s. \end{aligned}$$

  Which proves (4.34).

By structural induction, we conclude that (4.34) holds for all strings, s. ■

(b) It's also clear from the "string followed by string" definition of concatenation that it is *associative*. That is,

$$(pq)s = p(qs) \tag{4.36}$$

for all $p, q, s \in A^*$. We should verify this for the recursive definition of concatenation as well. Do so using structural induction.

Solution. The induction hypothesis will be equation (4.36). The proof is by structural induction on s.

- **Base case** $s = \lambda$:

$$\begin{aligned} (pq)s &= (pq)\lambda \\ &= pq && \text{(base case of concat)} \\ &= p(q\lambda) && \text{(base case of concat)} \\ &= p(qs) \end{aligned}$$

  which proves (4.36).

- **Constructor case:** $s = ta$ where $t \in A^*, a \in A$.
  By structural induction, we may assume

$$(pq)t = p(qt). \tag{4.37}$$

  Now we have:

$$\begin{aligned} (pq)s &= (pq)(ta) \\ &= ((pq)t)a && \text{(constructor case of concat)} \\ &= (p(qt))a && \text{by (4.37)} \\ &= p((qt)a) && \text{(constructor case of concat)} \\ &= p(q(ta)) && \text{(constructor case of concat)} \\ &= p(qs) \end{aligned}$$

Which proves (4.36).

By structural induction, we conclude that (4.36) holds for all strings, s. ■

Problem 4.9.5. Define 2-person terminating *value* games of perfect information, VG's, just like 2-person terminating games of perfect information (2PTG's) in Notes 4, except that there is only one

Base case: $\langle \texttt{value}, r\rangle$ is a VG where r is any real number.

If the play of a VG ends at the game $\langle \texttt{value}, r\rangle$, then r is called the value of the play. Now the objective of one player (call him the *max*-player) is to have play end with as high a value as possible, and the other player (called the *min*-player) aims to have play end with as low a value as possible.

Given which of the players moves first, a strategy for the max-player is said to *ensure* the value, r, if play ends with a value of at least r, no matter what moves the min-player makes. Likewise, a strategy for the min-player is said to *hold down* the value to r, if play ends with a value of at most r, no matter what moves the max-player makes.

A VG is said to have *final max value*, r, if the max-player has a strategy that ensures r, and the min-player has a strategy that holds down the value to r, when the *max-player moves first*. Likewise, the VG has *final min value*, r, if the max-player has a strategy that ensures r, and the min-player has a strategy that holds down the value to r, when the *min-player moves first*.

A Fundamental Theorem for *finite* VG's is that every finite VG has a both a final max value and a final min value. (Note: the two values are usually different.)

(a) Prove this Fundamental Theorem for finite VG's by structural induction on the definition of VG's.

Solution. The proof is very similar to the proof of the corresponding theorem for win/lose games in Notes 4, but there are double the cases since there are two final values.

In particular, we will show that $P(G)$ holds for all $G \in$ VG using structural induction on the definition of VG, where $P(G)$ is:

If G is finite, then G has a final max value and a final min value.

Base case: If $G =< \texttt{value}, r >$ for a real number r, then only possible play is G itself. Therefore both the final max value and final min value exist and they are both r.

Constructor case: If $G =< t, \mathcal{G} >\in$ VG, is a finite game, then by structural induction, we may assume that all the games in $\mathcal{G}$ have both final max values and final min values.

We first show that G has final max value, r, where r is the largest final min value among the games in $\mathcal{G}$. There must be such an r because, since G is finite, there are only finitely many games in $\mathcal{G}$, and by structural induction hypothesis, each of them has a final min value.

To prove the final max value of G is r, we must show how the max-player, moving first in G, can ensure r, and how the min-player, moving second in G, can hold down the value to r.

To ensure r, the max-player simply makes her first move to a game $H \in \mathcal{G}$ that has final min value, r. The min-player has the first move in H, so by definition of final min-value, the max-player has a strategy in H that ensures r, which she can now follow. So this first move, combined with the ensuring strategy in H, defines a strategy for the max-player in G that ensures r.

Likewise, there is a simple strategy for the min-player, moving second in G, to hold down the value to r. Namely, suppose the max-player's first move to is $H' \in \mathcal{G}$. Then H' has a final min value of at most r, by definition of r. So by definition of final min value, there is a strategy in H' for the min-player to hold down the value to r, which he can now follow, thereby holding down the value of play on G to r.

The existence of these ensuring and holding down strategies for G implies that the final max value of G is r.

Second, to show that G has a final min value, we can repeat the previous argument with min and max exchanged.

This proves $P(G)$.

Therefore, by structural induction, we can conclude that

$$\forall G \in \text{VG}.\ P(G).$$

■

Note that this Fundamental Theorem may not hold exactly for infinite games. For example, suppose the max player moves first and his only moves are those that terminate with a real number less than 1. Then he can ensure a value as close to 1 as he wants, but he can't ensure 1.

(b) See if you can formulate a good generalization of the Fundamental Theorem that applies to all VG's. You need not prove it.

Solution. Actually, the statement of the Fundamental Theorem holds unchanged for infinite games. The only things that get changed are the definitions of "ensures" and "holds down," which become slightly more complicated because it might only be possible to *approximate* a final value to within ϵ, or the final value might be infinite.

If we require that the possible values of base games are restricted to some bounded interval of real numbers, then final values must be in this interval. In this case, a value, r, is now said to be *ensured* for the max-player if, for every $\epsilon > 0$, there is a strategy for the max-player that ends with a value of at least $r - \epsilon$, no matter what moves the min-player makes. The definition of *holding down* must be modified similarly with an ending value of at most $r + \epsilon$. With these modified definitions, the proof of the Fundamental Theorem for finite VG's carries over directly for infinite VG's, except that there may not be a maximum final min value among the games in $\mathcal{G}$, in which case we choose r to be the lub of these values.

If the base values are not bounded, then the final values may not be finite. It's even easier to extend the definition of "ensures" and "hold down" to encompass infinite values, but there's little to be learned from doing it, so we'll skip it. ■

Problem 4.9.6. In the late 1960s, the military junta that ousted the government of the small republic of Nerdia outlawed multiplication and division by any number other than 3. Fortunately, a young dissident found a way to help the population multiply any two nonnegative integers without risking persecution by the junta. The procedure he taught people appears on the next page:

procedure *multiply*(x, y: nonnegative integers)
$r := x$;
$s := y$;
$a := 0$;
do until $s = 0$
  if $3 \mid s$ **then**
    $r := 3r$;
    $s := s/3$;
  else if $3 \mid (s-1)$ **then**
    $a := a + r$;
    $r := 3r$;
    $s := (s-1)/3$;
  else
    $a := a + 2r$;
    $r := 3r$;
    $s := (s-2)/3$;
return a;

We can model the algorithm as a state machine whose states are triples of nonnegative integers (r, s, a). The initial state is $(x, y, 0)$. The transitions are given by the rule that for $s > 0$:

$$(r,s,a) \to \begin{cases} (3r, s/3, a) & \text{if } 3 \mid s \\ (3r, (s-1)/3, a+r) & \text{if } 3 \mid (s-1) \\ (3r, (s-2)/3, a+2r) & \text{otherwise.} \end{cases}$$

(a) List the sequence of steps that appears in the execution of the algorithm for inputs $x = 5$ and $y = 10$.

Solution. $(5, 10, 0) \to (15, 3, 5) \to (45, 1, 5) \to (135, 0, 50)$ ■

(b) Use the invariant method to prove that the algorithm is partially correct —that is, if $s = 0$, then $a = xy$.

Solution. Let

$$P((r,s,a)) ::= \quad [rs + a = xy].$$

Clearly, P holds for the start state because

$$P((x,y,0)) \quad \text{iff} \quad [xy + 0 = xy].$$

Now, we show that P is indeed invariant, namely, assuming $P((r, s, a))$,

$$rs + a = xy, \tag{4.38}$$

holds and $(r, s, a) \to (r', s', a')$ is a transition, then $P((a', b', p'))$,

$$r's' + a' = xy, \tag{4.39}$$

holds.

We consider three cases:

If $3 \mid s$, then we have that $r' = 3r, s' = s/3, a' = a$. Therefore,

$$\begin{aligned} r's' + a' =& 3r \cdot \frac{s}{3} + a \\ =& rs + a \\ =& xy \qquad \text{(by (4.38)).} \end{aligned}$$

If $3 \mid s - 1$, then $r' = 3r, s' = (s - 1)/3, a = a + r$. So:

$$\begin{aligned} r's' + a' =& 3r \cdot \frac{s-1}{3} + a + r \\ =& r \cdot (s - 1) + a + r \\ =& rs + a \\ =& xy \qquad \text{(by (4.38)).} \end{aligned}$$

Otherwise, we have $r' = 3r, s' = (s - 2)/3, a = a + 2r$. So:

$$\begin{aligned} r's' + a' =& 3r \cdot \frac{s-2}{3} + a + 2r \\ =& r \cdot (s - 2) + a + 2r \\ =& rs + a \\ =& xy \qquad \text{(by (4.38)).} \end{aligned}$$

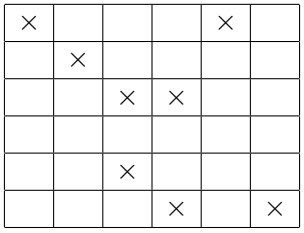
So in all three cases, (4.39) holds, proving that P is indeed an invariant.

Since the procedure's only termination condition is that $s = 0$, partial correctness will follow if we can show that if $s = 0$, then $a = xy$. But this follows immediately from (4.38). ■

(c) Prove that the algorithm terminates after at most $1 + \log_3 y$ executions of the body of the `do` statement.

Solution. We first notice that $s \in \mathbb{N}$ is an invariant. Also, each transition corresponds to an execution of the `do` statement body, and each transition reduces s to at most $s/3$. Hence, after at most $1 + \log_3 y$ executions of the body, the value of s is at most its initial value, y, times $(1/3)^{1+\log_3 y} = 1/3y$. That is, the value of s will at most $1/3$. Since $s \in \mathbb{N}$, it follows that s will be 0 after this many executions of the body, if it wasn't 0 earlier. But with $s = 0$, the procedure terminates. ■

Problem 4.9.7. In a MIT large conference room, n^2 students are sitting in a grid pattern of n rows of length n. A sudden outbreak of beaver flu (a rare variant of bird flu that lasts forever; symptoms include yearning for problem sets and loving mini-quizzes) causes some students to get infected. Here is an example where $n = 6$ and infected students are marked $\times$.

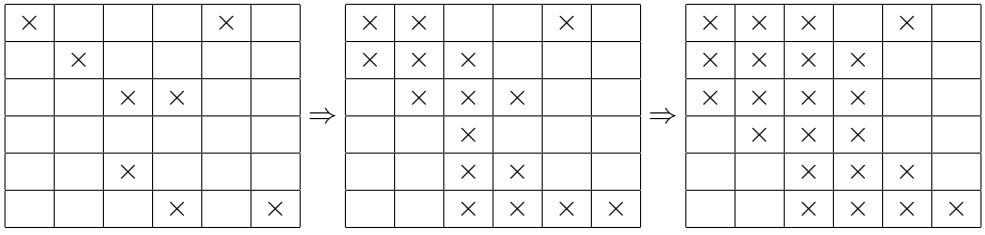


Two students are considered *adjacent* if they share an edge (that is, front, back, left or right, but NOT diagonal). So each student is adjacent to 2, 3 or 4 others, depending on whether the student is at an edge of the grid. An uninfected student can become infected iff

- the student is adjacent to *at least two* already-infected students.

Since beaver flu lasts forever, an infected student always remains infected.

In the example, the infection might spread as shown below.



Here we infected as many students as possible at each step, but the infection might also spread as slowly as one student at a time. In this example the infection can spread to all the students.

(a) It turns out that if fewer than n students in class are initially infected, the infection cannot spread to the whole class. Prove it!

Hint: There is a simple infection-process derived variable that is weakly decreasing, and the value of this variable when fewer than n students are infected is strictly smaller than its value when all students are infected. If you are stuck defining this variable, ask your recitation instructor for the one-word clue that makes it easy.

Solution. *Proof.* Define the *perimeter* of an infected set of students to be the number of edges with infection on exactly one side. Let I denote the perimeter of the initially-infected set of students.

We claim the size of the perimeter is weakly decreasing. This is easy to verify, because each newly-infected square must be adjacent to at least two previously-infected squares. Thus, for each newly-infected square, at least two edges are removed from the perimeter of the infected region, and at most two edges are added to the perimeter. Therefore, the perimeter of the infected region cannot increase.

If an $n \times n$ grid is completely infected, then the perimeter of the infected region is $4n$. Thus, the whole grid can become infected only if the perimeter is initially at least $4n$. Since each square has perimeter 4, at least n squares must be infected initially for the whole grid to be infected. □

■

(b) What is the smallest number of infected students that could spread infection to the entire class?

Solution. The smallest number is at least n by the previous part. Moreover, n students can infect the whole class if they appear on the same diagonal of the grid. ■

(c) (optional, may be hard) What is the *largest* number of students who can be infected without the infection being able to spread to everyone? Briefly explain.

Solution. ARM thinks it's $n(n-1)$ when the first $n-1$ rows are completely infected. **He hasn't found a proof that this is the max.** ■

4.10 Miniquiz Mar. 7

Problem 4.10.1. BAexp's are defined in the Appendix.

(a) The value of flatten(e) for $e \in$ BAexp is the sequence of integers in e obtained by "erasing" everything but the integers that appear within tagged variables and tagged `int`'s. For example,

$$\begin{aligned} e &::= \langle \texttt{sum}, \langle \texttt{var}, 3\rangle, \langle \texttt{sum}, \langle \texttt{var}, 2\rangle, \langle \texttt{int}, 2\rangle\rangle\rangle \\ f &::= \langle \texttt{prod}, \langle \texttt{var}, 4\rangle, \langle \texttt{var}, 5\rangle\rangle \\ g &::= \langle \texttt{prod}, e, \langle \texttt{sum}, \langle \texttt{var}, 7\rangle, f\rangle\rangle \\ \text{flatten}(g) &= \langle 3, 2, 2, 7, 4, 5\rangle . \end{aligned}$$

Give a recursive definition of flatten. (You may use the operation of *concatenation* (append) of two sequences.)

Solution.

- **Base cases:**
  1. flatten($\langle \texttt{int}, n\rangle$) ::= $\langle n\rangle$
  2. flatten($\langle \texttt{var}, n\rangle$) ::= $\langle n\rangle$
- **Constructor cases:**
  1. flatten($\langle \texttt{sum}, e_1, e_2\rangle$) ::= flatten($e_1$)flatten($e_2$)
  2. flatten($\langle \texttt{prod}, e_1, e_2\rangle$) ::= flatten($e_1$)flatten($e_2$)

■

(b) Prove by structural induction on the definition of BAexp that for all $e \in$ BAexp,

$$2 \cdot \text{length}(\text{flatten}(e)) = |e| + 1$$

Solution. The proof is by structural induction on the definition of $e \in$ BAexp. The induction hypothesis is the equation above.

- **Base cases:**
  1. $2 \cdot \text{length}(\text{flatten}(\langle \texttt{int}, n\rangle)) = 2 \cdot 1 = 2 = 1 + 1 = |\langle \texttt{int}, n\rangle| + 1$.
  2. `var` case the same.

  So the equation holds in the base cases.
- **Constructor cases:**
  1. Say $e = \langle \texttt{sum}, e_1, e_2\rangle$ where we assume the Structural Induction hypothesis that e_1 and e_2 satisfy the equation given above. Now,

$$\begin{aligned} & 2 \cdot \text{length}(\text{flatten}(e)) \\ &= 2 \cdot \text{length}(\text{flatten}(\langle \texttt{sum}, e_1, e_2\rangle)) \\ &= 2 \cdot \text{length}(\text{flatten}(e_1)\text{flatten}(e_2)) && \text{(def of flatten)} \\ &= 2 \cdot \text{length}(\text{flatten}(e_1)) + 2 \cdot \text{length}(\text{flatten}(e_2)) && \text{(length of a string)} \\ &= (|e_1| + 1) + |e_2| + 1 && \text{(structural induction hyp.)} \\ &= |\langle \texttt{sum}, e_1, e_2\rangle| + 1 && \text{(def. of } |\langle \text{sum}, \ \rangle|) \\ &= |e| + 1. \end{aligned}$$

  So the equation holds for e.

2. `prod` the same.

This completes the proof for the Constructor cases.

We conclude by Structural Induction that the equation holds for all $e \in$ BAexp. ■

Problem 4.10.2. The following state machine describes an algorithm to multiply any two nonnegative integers, without multiplying or dividing by any number other than 3.

Its states are triples of nonnegative integers (r, s, a). The initial state is $(x, y, 0)$. The transitions are given by the rule that for $s > 0$:

$$(r, s, a) \to \begin{cases} (3r, s/3, a) & \text{if } 3 \mid s \\ (3r, (s-1)/3, a+r) & \text{if } 3 \mid (s-1) \\ (3r, (s-2)/3, a+2r) & \text{otherwise.} \end{cases}$$

(a) Circle the predicate that is invariant for this program:

- $P((5, 10, 0)) ::= \quad [(15, 3, 5)]$
- $P((r, s, a)) ::= \quad [rs + a^2 = xy]$
- $P((r, s, a)) ::= \quad [rs + a = xy]$
- $P((r, s, a)) ::= \quad [sx + rxy = ar]$

Solution. $P((r, s, a)) ::= \quad [rs + a = xy]$ is the correct answer. Notice that the first item is not even a predicate! ■

(b) Use the invariant principle to prove that the algorithm is partially correct —that is, if $s = 0$, then $a = xy$.

Solution. See Pset3, Problem 6.

■

(c) When the state machine reaches a state with $s = 0$, it outputs a and terminates. Prove that the algorithm terminates.

Solution. We'll actually prove something stronger, namely, that the algorithm terminates after at most $1 + \log_3 y$ executions of `do` statement. We first notice that $s \in \mathbb{N}$ is an invariant. Also, each transition corresponds to an execution of the `do` statement body, and at each transition, s is reduced by a factor of at most $1/3$. Hence, after at most $1 + \log_3 y$ executions of the body, the final value of s is at most $1/3^{1+\log_3 y} = 1/3y$ times its initial value, y. This means the value of s will be less than 1, and since the value is a nonnegative integer, s must be 0 at this point if it wasn't 0 earlier. But with $s = 0$, the procedure terminates. ■

Appendix

The set, BAexp, of *Basic Arithmetic Expressions* is defined recursively as a tagged data type as follows:

- **Base cases:** If $n \in \mathbb{Z}$, then
  1. $\langle \texttt{int}, n\rangle \in$ BAexp, and
  2. $\langle \texttt{var}, n\rangle \in$ BAexp.
- **Constructor cases:** if $e, e' \in$ BAexp, then
  1. $\langle \texttt{sum}, e, e'\rangle \in$ BAexp, and
  2. $\langle \texttt{prod}, e, e'\rangle \in$ BAexp.

The size, $|e|$, of $e \in$ BAexp is defined recursively on this definition by:

- **Base cases:**
  1. $|\langle \texttt{int}, n\rangle| ::= 1$
  2. $|\langle \texttt{var}, n\rangle| ::= 1$
- **Constructor cases:**
  1. $|\langle \texttt{sum}, e_1, e_2\rangle| ::= |e_1| + |e_2| + 1$
  2. $|\langle \texttt{prod}, e_1, e_2\rangle| ::= |e_1| + |e_2| + 1$

Some string definitions:

Definition 4.10.1. The *length*, $|s|$, of a string, s, is defined recursively by the rules:

- $|\lambda| ::= 0$
- $|sa| ::= 1 + |s|$.

Definition 4.10.2. The *concatenation*, st, of strings s and t over an alphabet, A, is defined recursively on t by the rules:

- $s\lambda ::= s$.
- $s(ta) ::= (st)a$ for $a \in A$.

For example, consider two strings, $s = \langle 1, 3, 6, 2\rangle$, and $t = \langle 2, 7, 3\rangle$.

The concatenation of s and t is: $st = \langle 1, 3, 6, 2, 2, 7, 3\rangle$. Likewise, the concatenation of t and s is $ts = \langle 2, 7, 3, 1, 3, 6, 2\rangle$.

The length of s is: $length(s) = 4$. The length of t is: $length(t) = 3$.

Chapter 5

Stable Marriages; Simple Graphs

5.1 The Stable Marriage Problem

Okay, frequent public reference to derived variables may not help your mating prospects. But they can help with the analysis!

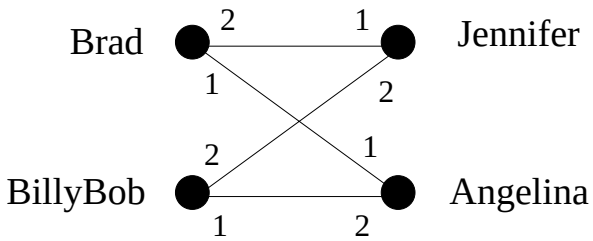
5.1.1 The Problem

Suppose there are a bunch of boys and an equal number of girls that we want to marry off. Each boy has his personal preferences about the girls —in fact, we assume he has a complete list of all the girls ranked according to his preferences, with no ties. Likewise, each girl has her rank list of all of the boys.

The preferences don't have to be symmetric. That is, Jennifer might like Brad best, but Brad doesn't necessarily like Jennifer best. The goal is to marry off boys and girls: every boy must marry exactly one girl and vice-versa —no polygamy. In mathematical terms, we want the mapping from boys to their wives to be a bijection or *perfect matching*. We'll just call this a "matching," for short.

Here's the difficulty: suppose *every* boy likes Angelina best, and every girl likes Brad best, but Brad and Angelina are married to other people, say Jennifer and Billy Bob. Now *Brad and Angelina prefer each other to their spouses*, which puts their marriages at risk: pretty soon, they're likely to start spending late nights doing 6.042 homework together.

This situation is illustrated in the following diagram where the digits "1" and "2" near a boy shows which of the two girls he ranks first and which second, and similarly for the girls:



More generally, in any matching, a boy and girl who are not married to each other and who like each other better than their spouses, is called a *rogue couple*. In the situation above, Brad and Angelina would be a rogue couple.

Having a rogue couple is not a good thing, since it threatens the stability of the marriages. On the other hand, if there are no rogue couples, then for any boy and girl who are not married to each other, at least one likes their spouse better than the other, and so won't be tempted to start an affair.

Definition 5.1.1. A matching is *stable* iff it has no rogue couples.

The question is, given everybody's preferences, how do you find a stable set of marriages? In the example consisting solely of the four people above, we could let Brad and Angelina both have their first choices by marrying each other. Now neither Bard nor Angelina prefers anybody else to their spouse, so neither will be in a rogue couple. This leaves Jen not-so-happily married to Billy Bob, but neither Jen nor Billy Bob can entice somebody else to marry them.

It is something of a surprise that there always is a stable matching among a group of boys and girls, but there is, and we'll shortly explain why. The surprise springs in part from considering the apparently similar "buddy" matching problem. That is, if people can be paired off as buddies, regardless of gender, then a stable matching *may not* be possible. For example, there might be a love triangle with a fourth person who is everyone's last choice:

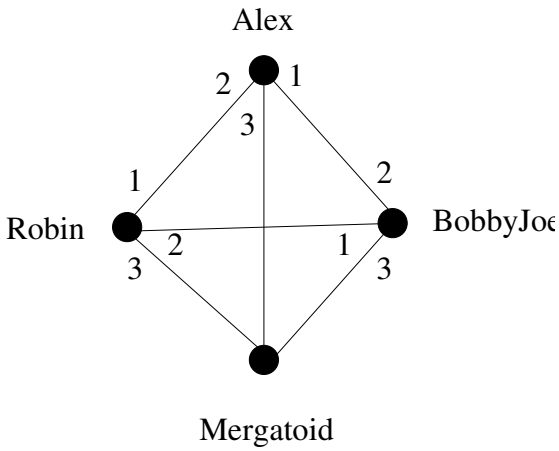


Figure 5.1: Some preferences with no stable buddy matching.

Here, Mergatoid's preferences aren't shown in Figure 5.1 because they don't even matter. Let's see why there is no stable matching:

Lemma. *There is no stable buddy matching among the four people in Figure 5.1.*

Proof. We'll prove this by contradiction.

Assume, for the purposes of contradiction, that there is a stable matching. Then there are two members of the love triangle that are matched. Since preferences in the triangle are symmetric, we may assume in particular, that Robin and Alex are matched. Then the other pair must be Bobby-Joe matched with Mergatoid.

But then there is a rogue couple: Alex likes Bobby-Joe best and Bobby-Joe prefers Alex to his buddy Mergatoid. That is, Alex and Bobby-Joe are a rogue couple, contradicting the assumed stability of the matching. □

So getting a stable *buddy* matching may not only be hard, it may be impossible. But when boys are only allowed to marry girls, and vice versa, then it turns out that a stable matching is not hard to find.

5.1.2 The Mating Ritual

The procedure for finding a stable matching involves a *Mating Ritual* that takes place over several days. The following events happen each day:

Morning: Each girl stands on her balcony. Each boy stands under the balcony of his favorite among the girls on his list, and he serenades her. If a boy has no girls left on his list, he stays home and does his 6.042 homework.

Afternoon: Each girl who has one or more suitors serenading her, says to her favorite suitor, "We might get engaged. Come back tomorrow." To the others, she says, "No. I will never marry you! Take a hike!"

Evening: Any boy who is told by a girl to take a hike, crosses that girl off his list.

Termination condition: When every girl has at most one suitor, the ritual ends with each girl marrying her suitor, if she has one.

There are a number of facts about this Mating Ritual that we would like to prove:

- The Ritual has a last day.
- Everybody ends up married.
- The resulting marriages are stable.

5.1.3 A State Machine Model

Before we can prove anything, we should have clear mathematical definitions of what we're talking about. In this section we sketch how to define a rigorous state machine model of the Marriage Problem.

So let's begin by formally defining the problem.

Definition 5.1.2. A ***Marriage Problem*** consists of two disjoint sets of the same finite size, called the-Boys and the-Girls. The members of the-Boys are called *boys*, and members of the-Girls are called *girls*. For each boy, B, there is a strict total order, $<_B$, on the-Girls, and for each girl, G, there is a strict total order, $<_G$, on the-Boys. If $G_1 <_B G_2$ we say B ***prefers*** girl G_2 to girl G_1. Similarly, if $B_1 <_G B_2$ we say G ***prefers*** boy B_2 to boy B_1.

A *marriage assignment* or ***perfect matching*** is a bijection, w : the-Boys $\to$ the-Girls. If $B \in$ the-Boys, then $w(B)$ is called B's *wife* in the assignment, and if $G \in$ the-Girls, then $w^{-1}(G)$ is called G's *husband*. A ***rogue couple*** is a boy, B, and a girl, G, such that B prefers G to his wife, and G prefers B to her husband. An assignment is ***stable*** if it has no rogue couples. A ***solution*** to a marriage problem is a stable perfect matching.

To model the Mating Ritual with a state machine, we make a key observation: to determine what happens on any day of the Ritual, all we need to know is which girls are on which boys' lists on the morning of that day. So we define a state to be some mathematical data structure providing this information. For example, we could define a state to be the "still-has-on-his-list" relation, R, between boys and girls, where $B\,R\,G$ means girl G is still on boy B's list.

We start the Mating Ritual with no girls crossed off. That is, the start state is the *complete bipartite* relation in which every boy is related to every girl.

According to the Mating Ritual, on any given morning, a boy will *serenade* the girl he most prefers among those he has not as yet crossed out. Mathematically, the girl he is serenading is just the maximum among the girls on B's list, ordered by $<_B$. (If the list is empty, he's not serenading anybody.) A girl's *favorite* is just the maximum, under her preference ordering, among the boys serenading her.

Continuing in this way, we could mathematicallly specify a precise Mating Ritual state machine, but we won't bother. The intended behavior of the Mating Ritual is clear enough that we don't gain much by giving a formal state machine, so we stick to a more memorable description in terms of boys, girls, and their preferences. The point is, though, that it's not hard to define everything using basic mathematical data structures like sets, functions, and relations, if need be.

5.1.4 There is a Marriage Day

It's easy to see why the Mating Ritual has a terminal day when people finally get married. Every day at least one boy will cross a girl off his list. When no girl can be crossed off any list, then the Ritual has terminated. So starting with n boys and n girls, each of the n boys' lists initially has n girls on it, for a total of n^2 list entries. Since no girl ever gets added to a list, the total number of entries on the lists decreases every day that the Ritual continues, and so the Ritual can continue for at most n^2 days.

5.1.5 They All Live Happily Every After...

We still have to prove that the Mating Ritual leaves everyone in a stable marriage. To do this, we note one very useful fact about the Ritual: if a girl has a favorite boy suitor on some morning of the Ritual, then that favorite suitor will still be serenading her the next morning —because his list won't have changed. So she is sure to have today's favorite boy among her suitors tomorrow. That means she will be able to choose a favorite suitor tomorrow who is at least as desirable to her as today's favorite. So day by day, her favorite suitor can stay the same or get better, never worse. In others words, a girl's favorite is a weakly increasing variable with respect to her preference order on the boys.

Now we can verify the Mating Ritual using a simple invariant predicate, P, that captures what's going on:

> For every girl, G, and every boy, B, if G is crossed off B's list, then G has a favorite suitor and she prefers him over B.

Why is P invariant? Well, we know that G's favorite tomorrow will be at least as desirable to her as her favorite today, and since her favorite today is more desirable than B, tomorrow's favorite will be too.

Notice that P also holds on the first day, since every girl is on every list. So by the Invariant Theorem, we know that P holds on every day that the Mating Ritual runs. Knowing the invariant holds when the Mating Ritual terminates will let us complete the proofs.

Theorem 5.1.3. *Everyone is married by the Mating Ritual.*

Proof. Suppose, for the sake of contradiction, that some boy is not married on the last day of the Mating Ritual. So he can't be serenading anybody, that is, his list must be empty. So by invariant P, every girl has a favorite boy whom she prefers to that boy. In particular, every girl has a favorite boy that she marries on the last day. So all the girls are married. What's more there is no bigamy: a boy only serenades one girl, so no two girls have the same favorite.

But there are the same number of girls as boys, so all the boys must be married too. □

Theorem 5.1.4. *The Mating Ritual produces a stable matching.*

Proof. Let Brad be some boy and Jen be any girl that he is *not* married to on the last day of the Mating Ritual. We claim that Brad and Jen are not a rogue couple. Since Brad is an arbitrary boy, it follows that no boy is part of a rogue couple. Hence the marriages on the last day are stable.

To prove the claim, we consider two cases:

Case 1. Jen is not on Brad's list. Then by invariant P, we know that Jen prefers her husband to Brad. So she's not going to run off with Brad: the claim holds in this case.

Case 2. Otherwise, Jen is on Brad's list. But since Brad is not married to Jen, he must be choosing to serenade his wife instead of Jen, so he must prefer his wife. So he's not going to run off with Jen: the claim also holds in this case. □

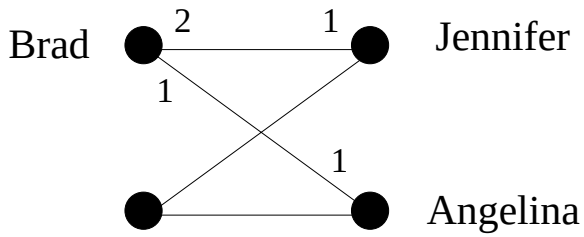
5.1.6 ...Especially the Boys

Who is favored by the Mating Ritual, the boys or the girls? The girls seem to have all the power: they stand on their balconies choosing the finest among their suitors and spurning the rest. What's more, we know their suitors can only change for the better as the Ritual progresses. Similarly, a boy keeps serenading the girl he most prefers among those on his list until he must cross her off, at which point he serenades the next most preferred girl on his list. So from the boy's point of view, the girl he is serenading can only change for the worse. Sounds like a good deal for the girls.

But it's not! The fact is that from the beginning, the boys are serenading their first choice girl, and the desirability of the girl being serenaded decreases only enough to give the boy his most desirable possible spouse. The mating algorithm actually does as well as possible for all the boys and does the worst possible job for the girls.

To explain all this we need some definitions. Let's begin by observing that while the mating algorithm produces one stable matching, there may be other stable matchings among the same set of boys and girls. For example, reversing the roles of boys and girls will often yield a different stable matching among them.

But some spouses might be out of the question in all possible stable matchings. For example, Brad is just not in the realm of possibility for Jennifer, since if you ever pair them, Brad and Angelina will form a rogue couple; here's a picture:



Definition 5.1.5. Given any marriage problem, one person is in another person's *realm of possible spouses* if there is a stable matching in which the two people are married. A person's *optimal spouse* is their most preferred person within their realm of possibility. A person's *pessimal spouse* is their least preferred person in their realm of possibility.

Everybody has an optimal and a pessimal spouse, since we know there is at least one stable matching, namely the one produced by the Mating Ritual. Now here is the shocking truth about the Mating Ritual:

Theorem 5.1.6. *The Mating Ritual marries every boy to his optimal spouse.*

Proof. Assume for the purpose of contradiction that some boy does not get his optimal girl. There must have been a day when he crossed off his optimal girl —otherwise he would still be serenading her or some even more desirable girl.

By the Well Ordering Principle, there must be a *first* day when a boy, call him "Keith," crosses off his optimal girl, Nicole.

According to the rules of the Ritual, Keith crosses off Nicole because Nicole has a favorite suitor, Tom, and

> Nicole prefers Tom to Keith (*)

(remember, this is a proof by contradiction :-)).

Now since this is the first day an optimal girl gets crossed off, we know Tom hasn't crossed off his optimal girl. So

> Tom ranks Nicole at least as high as his optimal girl. (**)

So in the stable set of marriages where Keith gets his optimal girl, Nicole, the preferences given in (*) and (**) imply that Nicole and Tom are a rogue couple, contradicting the stability of this set of marriages. □

Theorem 5.1.7. *The Mating Ritual marries every girl to her pessimal spouse.*

Proof. Say Nicole and Keith marry each other as a result of the Mating Ritual. By the previous Theorem 5.1.6, Nicole is Keith's optimal spouse, and so in any set of stable marriages,

Keith rates Nicole at least as high as his spouse. (+)

Now suppose for the purpose of contradiction that there is another stable marriage set where Nicole does worse than Keith. That is, Nicole is married to Tom, and

Nicole prefers Keith to Tom (++)

Then in this stable set of marriages where Nicole is married to Tom, (+) and (++) imply that Nicole and Keith are a rogue couple, contradicting stability. We conclude that Nicole cannot do worse than Keith. □

5.1.7 Applications

Not surprisingly, a stable matching procedure is used by at least one large dating agency. But although "boy-girl-marriage" terminology is traditional and makes some of the definitions easier to remember (we hope without offending anyone), solutions to the Stable Marriage Problem are widely useful.

The Mating Ritual was first announced in a paper by D. Gale and L.S. Shapley in 1962, but ten years before the Gale-Shapley paper was appeared, and unbeknownst to them, the Ritual was being used to assign residents to hospitals by the National Resident Matching Program (NRMP). The NRMP has, since the turn of the twentieth century, assigned each year's pool of medical school graduates to hospital residencies (formerly called "internships") with hospitals and graduates playing the roles of boys and girls. (In this case there may be multiple boys married to one girl, but there's an easy way to use the Ritual in this situation.) Before the Ritual was adopted, there were chronic disruptions and awkward countermeasures taken to preserve assignments of graduates to residencies. The Ritual resolved these problems so successfully, that it was used essentially without change at least through 1989.[1]

MIT Math Prof. Tom Leighton, who regularly teaches 6.042 and also founded the internet infrastructure company, Akamai, reports another application. Akamai uses a variation of the Gale-Shapley procedure to assign web traffic to servers. In the early days, Akamai used other combinatorial optimization algorithms that got to be too slow as the number of servers and traffic increased. Akamai switched to Gale-Shapley since it is fast and can be run in a distributed manner. In this case, the web traffic corresponds to the boys and the web servers to the girls. The servers have preferences based on latency and packet loss; the traffic has preferences based on the cost of bandwidth.

5.2 Simple Graphs

5.2.1 Introduction

Graphs are an incredibly useful structure in Computer Science! They arise in all sorts of applications, including scheduling, optimization, communications, and the design and analysis of algorithms. Two Stanford students even used graph theory to become multibillionaires.

[1] Much more about the Stable Marriage Problem can be found in the very readable mathematical monograph by Dan Gusfield and Robert W. Irving, The Stable Marriage Problem: Structure and Algorithms, MIT Press, Cambridge, Massachusetts, 1989, 240 pp.

But first we are going to talk about something else. Namely, sex. The question that we'll address is, on average, who has more opposite-gender partners, men or women? This has been the subject of many studies. In one of the largest, researchers from the University of Chicago interviewed a "random sample" of 2500 people over several years to try to get an answer to this question. Their study, published in 1994, and entitled *The Social Organization of Sexuality* found that on average men have 74% more opposite-gender partners than women, which fits right in with stereotype of male promiscuity versus female sexual selectiveness.

Other studies have found that the disparity is even larger. In particular, ABC News claims that the average man has 20 partners over his lifetime, and the average woman has 6, for a percentage disparity of 233%. The ABC News study, aired on Primetime Live in 2004, claimed to be one of the most scientific ever done with only a 2.5% margin of error. It was called "American Sex Survey: A peak between the sheets"—- hmmmmmm, doesn't sound so scientific. The promotion for the study is even better:

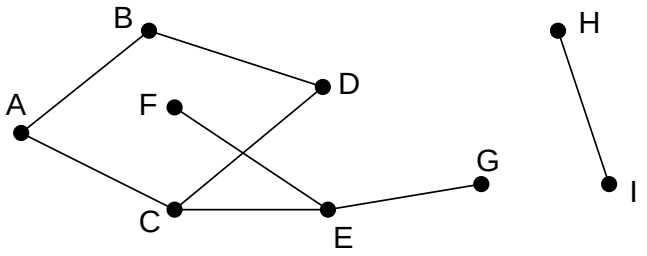
> A ground breaking ABC News "Primetime Live" survey finds a range of eye-popping sexual activities, fantasies and attitudes in this country, confirming some conventional wisdom, exploding some myths – and venturing where few scientific surveys have gone before.

Probably that last part about going where few scientific surveys have gone before is pretty accurate!

Anyway, whose numbers do you think are more accurate, the University of Chicago or ABC News? Don't answer: this is a setup question like "When did you stop beating your wife?" Using a little graph theory, we'll explain why neither finding can be anywhere near the truth. Whoever reported these findings was either confused or was intentionally misinterpreting the data!

5.2.2 Definition of Simple Graph

Informally, an graph is a bunch of dots with lines connecting some of them. Here is an example:



For many Mathematical purposes, we don't really care how the points and lines are laid out — only which points are connected by lines. The definition of *simple graphs* aims to capture just this connection data.

Definition 5.2.1. A *simple graph*, G, consists of a nonempty set, V, called the *vertices* of G, and a collection, E, of two-element subsets of V. The members of E are called the *edges* of G.

The vertices correspond to the dots in the picture, and the edges correspond to the lines. For example, the connection data in dots-and-lines diagram above is a captured by the following simple graph:

$$\begin{aligned} V &= \{A, B, C, D, E, F, G, H, I\} \\ E &= \{\{A, B\}, \{A, C\}, \{B, D\}, \{C, D\}, \{C, E\}, \{E, F\}, \{E, G\}, \{H, I\}\}. \end{aligned}$$

It will be helpful to use the notation A—B for the edge $\{A, B\}$. Note that A—B and B—A are different descriptions of the same edge, since sets are unordered.

Two vertices in a simple graph are said to be ***adjacent*** if they are joined by an edge, and an edge is said to be ***incident*** to the vertices it joins. The number of edges incident to a vertex is called the ***degree*** of the vertex. For example, in the simple graph above, A is adjacent to B and B is adjacent to D, and the edge A—C is incident to vertices A and C. Vertex H has degree 1, D has degree 2, and E has degree 3.

Graph Synonyms

A synonym for "vertices" is "nodes," and we'll use these words interchangeably.

Simple graphs are sometimes called *networks*, edges are sometimes called *arcs*, and adjacent vertices are sometimes called *neighbors*. We mention this as a "heads up" in case you look at other graph theory literature; we won't use these words.

Some technical consequences of Definition 5.2.1 are worth noting right from the start:

1. Simple graphs do not have edges going from a vertex back around to itself (called a *self-loop*).
2. There is at most one edge between two vertices of a simple graph.
3. Simple graphs have at least one vertex, though they might not have any edges.

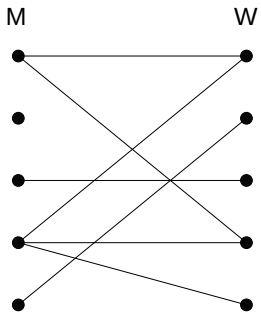
There's no harm in relaxing these conditions, and some authors do, but we don't need self-loops, multiple edges between the same two vertices, or graphs with no vertices, and it's simpler not to have them around.

For the rest of these Notes and much of next week's Notes, we'll only be considering simple graphs, so we'll just call them "graphs" from now on.

5.2.3 Sex in America

Let's model the question of heterosexual partners in graph theoretic terms. To do this, we'll let G be the graph whose vertices, V, are all the people in America. Now each vertex is either a male or a female, so we can split V into two separate [2] subsets: M, which contains all the male vertices, and a set, F, which contains all the female, and we'll put an edge between a male and female vertex iff they have ever been sexual partners. We can illustrate this with a drawing with M vertices on the left and the F vertices on the right:

[2] For simplicity, we'll ignore the possibility of someone being both male and female.



Actually, this is a pretty hard graph to figure out, let alone draw. The graph is *enormous*: there are about 300 million nodes, that is, $|V| \approx 300$meg! Of these, approximately 50.8% are female and 49.2% are male, so $|M| \approx 147.6$meg, and $|F| \approx 152.4$meg. And we don't even have trustworthy estimates of how many edges there are, let alone exactly which couples are adjacent.

But it turns out that we don't need to know any of this —we just need to figure out the relationship between the average number of partners per male and partners per female. To do this, we note that every edge is incident to exactly one M vertex (remember, we're only considering male-female relationships); so the sum of the degrees of the M vertices equals the number of edges. For the same reason, the sum of the degrees of the F vertices equals the number of edges. So these sums are equal:

$$\sum_{x \in M} \deg(x) = \sum_{y \in F} \deg(y).$$

Now suppose we divide both sides of this equation by the product of the sizes of the two sets, $|M| \cdot |F|$:

$$\left(\frac{\sum_{x \in M} \deg(x)}{|M|}\right) \cdot \frac{1}{|F|} = \left(\frac{\sum_{y \in F} \deg(y)}{|F|}\right) \cdot \frac{1}{|M|}$$

The terms above in parentheses are the *average degree of an M vertex* and the *average degree of a F* vertex. So we know:

$$\text{Avg. deg in } M = \frac{|F|}{|M|} \cdot \text{Avg. deg in } F$$

In other words, we've proved that the average number of female partners of males in the population compared to the average number of males per female is *determined solely by the relative number of males and females in the population*.

Now the Census Bureau reports that there are slightly more females than males in America; in particular $|F| / |M|$ is about 1.035. So we know that on average, males have 3.5% more opposite-gender partners than females, and this tells us nothing about any sex's promiscuity or selectiveness. Rather, it just has to do with the relative number of males and females. Collectively, males and females have the same number of opposite gender partners, since it takes one of each set for every partnership, but there are fewer men, so they have a higher ratio. So the University of Chicago study was way off. After a huge effort, they gave a totally wrong answer.

As it turns out, there have been numerous other studies that have missed the same underlying issue. For example, a couple of years ago, the Boston Globe ran a story on a survey of the study habits of students on Boston area campuses. Their survey showed that on average, minority students tended to study with non-minority students more than the other way around. They went

on at great length to explain why this "remarkable phenomenon" might be true. But it's not remarkable at all —using our graph theory formulation, we can see that all it says is that there are fewer minority students than non-minority students. Well, that just follows from the definition of "minority"!

5.2.4 Handshaking Lemma

The previous argument hinged on the connection between a sum of degrees and the number edges. There is a simple connection between these in any graph:

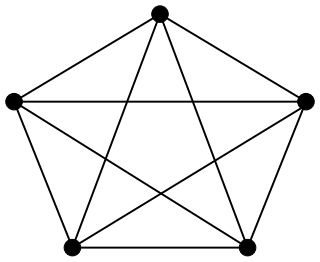
Theorem 5.2.2. *The sum of the degrees of the vertices in a graph equals twice the number of edges.*

Proof. Every edge contributes two to the sum of the degrees, one for each of its endpoints. □

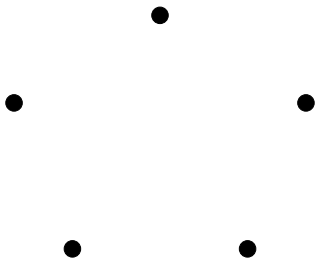
Theorem 5.2.2 is sometimes called the *Handshake Theorem*: if we total up the number of people each person at a party shakes hands with, the total will be twice the number of handshakes that occurred.

5.2.5 Some Common Graphs

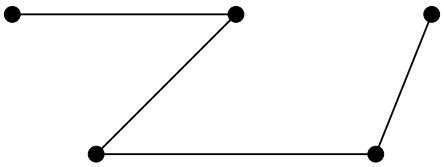
Some graphs come up so frequently that they have names. The *complete graph* on n vertices, also called K_n, has an edge between every two vertices. Here is K_5:



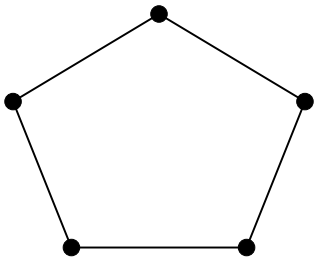
The *empty graph* has no edges at all. Here is the empty graph on 5 vertices:



Another 5 vertex graph is L_4, the *line graph* of length four:

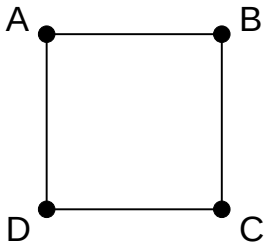


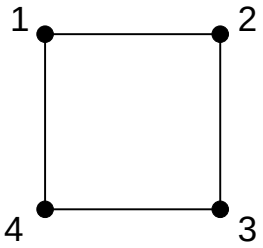
And here is C_5, a *simple cycle* with 5 vertices:



5.2.6 Isomorphism

Two graphs that look the same might actually be different in a formal sense. For example, the two graphs below are both simple cycles with 4 vertices:





But one graph has vertex set $\{A, B, C, D\}$ while the other has vertex set $\{1, 2, 3, 4\}$. If so, then the graphs are different mathematical objects, strictly speaking. But this is a frustrating distinction; the graphs *look the same*!

Fortunately, we can neatly capture the idea of "looks the same." Graphs G_1 and G_2 are *isomorphic* if there exists a bijection between the vertices in G_1 and the vertices in G_2 such that there is an edge between two vertices in G_1 if and only if there is an edge between the two corresponding vertices in G_2. For example, take the following bijection between vertices in the two graphs above:

A corresponds to 1	B corresponds to 2
D corresponds to 4	C corresponds to 3.

Now there is an edge between two vertices in the graph on the left if and only if there is an edge between the two corresponding vertices in the graph on the right. Therefore, the two graphs are isomorphic. The bijection itself is called an *isomorphism*.

Definition 5.2.3. If G_1 is a graph with vertices, V_1, and edges, E_1, and likewise for G_2, then G_1 is *isomorphic* to G_2 iff there exists a **bijection**, $f : V_1 \to V_2$, such that for every pair of vertices $u, v \in V_1$:

$$u\text{—}v \in E_1 \quad \text{iff} \quad f(u)\text{—}f(v) \in E_2.$$

The function f is called an *isomorphism* between G_1 and G_2.

Two isomorphic graphs may be drawn very differently. For example, here are two different ways of drawing C_5:

Isomorphism captures all the connection properties of a graph, abstracting out what the vertices are called, what they are made out of, or where they appear in a drawing of the graph. So a property like "having three vertices of degree 4" is preserved under isomorphism, while "having a vertex that is an integer" is not preserved. In particular, if one graph has three vertices of degree 4 and another does not, they can't be isomorphic. Similarly, if one graph has an edge that is incident to a degree 8 vertex and a degree 3 vertex, then any isomorphic graph must also have such an edge.

Looking for properties like these can make it easy to determine that two graphs are not isomorphic, or to actually find an isomorphism between them, if there is one. In practice, it's frequently easy to decide whether two graphs are isomorphic. However, no one has yet found a *general* procedure for determining whether two graphs are isomorphic which is *guaranteed* to run much faster than an exhaustive (and exhausting) search through all possible bijections between their vertices.

Having an efficient procedure to detect isomorphic graphs would, for example, make it easy to search for a particular molecule in a database given the molecular bonds. On the other hand, knowing there is no such efficient procedure would also be valuable: secure protocols for encryption and remote authentication can be built on the hypothesis that graph isomorphism is computationally exhausting.

5.3 Connectedness

5.3.1 Paths and Simple Cycles

A *path* in a graph describes how to get from one vertex to another following edges of the graph. Formally,

Definition 5.3.1. A ***path*** in a graph, G, is a sequence of $k \geq 0$ vertices

$$v_0, \ldots, v_k$$

such that v_i—v_{i+1} is an edge of G for all i where $0 \leq i < k$. The path is said to ***start*** at v_0, to ***end*** at v_k, and ***length*** of the path is defined to be k. An edge, e, is ***traversed*** n ***times*** by the path if there are n different values of i such that edge v_i—v_{i+1} is e.

The path is ***simple***[3] iff all the v_i's are different, that is, $v_i = v_j$ only if $i = j$.

For example, the graph in Figure 5.2 has a length 6 simple path A,B,C,D,E,F,G. This is the longest simple path in the graph.

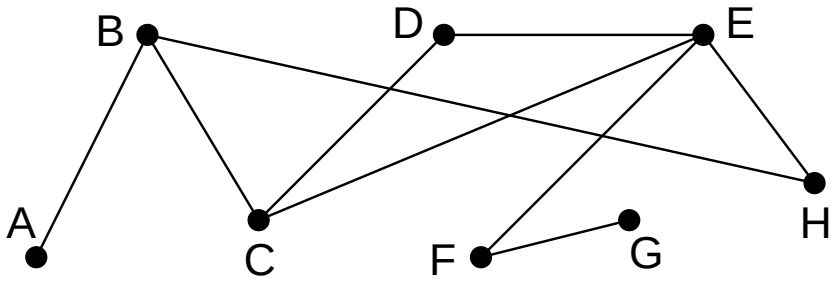


Figure 5.2: *A graph with 3 simple cycles.*

Notice that the *length* of a path is the total number of times it traverses edges, which is *one less* than its length as a sequence of vertices. The length 6 path A,B,C,D,E,F,G is actually a sequence of seven vertices.

A *cycle* can described by a path that begins and ends with the same vertex. For example, B,C,D,E,C,B is a cycle in the graph in Figure 5.2. This path suggests that the cycle begins and ends at vertex B, but a cycle isn't intended to have a beginning and end, and can be described by *any* of the paths that go around it. For example, D,E,C,B,C,D describes this same cycle as though it started and ended at D, and D,C,B,C,E,D describes the same cycle as though it started and ended at D but went in the opposite direction. (By convention, a single vertex is a length 0 cycle beginning and ending at the vertex.)

All the paths that describe the same cycle have the same length which is defined to be the *length* of the cycle. (Note that this implies that going around the same cycle twice is considered to be different than going around it once.)

A *simple* cycle is a *positive length* cycle that doesn't cross or backtrack on itself. For example, the graph in Figure 5.2 has three simple cycles B,H,E,C,B and C,D,E,C and B,C,D,E,H,B. More precisely, a simple cycle is a cycle that can be described by a positive length path whose vertices are all different except for the beginning and end vertices. So in contrast to simple *paths*, the length of a simple *cycle* is the *same* as the number of distinct vertices that appear in it.

[3]Some authors refer to paths as "walks,"and then use "path" to mean simple path.

From now on we'll stop being picky about distinguishing a cycle from a path that describes it, and we'll just refer to the path as a cycle. [4]

5.3.2 Connected Components

Definition 5.3.2. Two vertices in a graph are said to be ***connected*** if there is a path that begins at one and ends at the other.

By convention, every vertex is considered to be connected to itself by a path of length zero.

The diagram in Figure 5.3 looks like a picture of three graphs, but is intended to be a picture of *one* graph. This graph consists of three pieces (subgraphs). Each piece by itself is connected, but there are no paths between vertices in different pieces.

Figure 5.3: *A graph with 3 connected components.*

Definition 5.3.3. A graph is said to be ***connected*** if every pair of vertices are connected.

These connected pieces of a graph are called its ***connected components***. A rigorous definition is easy: a connected component is the set of all the vertices connected to some single vertex. So a graph is connected iff it has exactly one connected component. The empty graph on n vertices has n connected components.

5.3.3 How Well Connected?

If we think of a graph as modelling cables in a telephone network, or oil pipelines, or electrical power lines, then we not only want connectivity, but we want connectivity that survives component failure. This leads to the following definition:

Definition 5.3.4. Two vertices in a graph are k-connected if they remain connected in any subgraph obtained by deleting $k-1$ edges. A graph is k-connected if every pair of its vertices are k-connected.

[4] Technically speaking, we haven't ever defined what a cycle *is*, only how to describe it with paths. But we won't need an abstract definition of cycle, since all that matters about a cyle is which paths describe it.

So 1-connected is the same as connected for both vertices and graphs. Another way to say that a graph is k-connected is that every subgraph obtained from it by deleting at most $k-1$ edges is connected.

For example, in the graph in Figure 5.2, vertices B and E are 2-connected, G and E are 1-connected, and no vertices are 3-connected. The graph as a whole is only 1-connected.

More generally, any simple cycle is 2-connected, and the complete graph, K_n, is $(n-1)$-connected.

If two vertices are connected by k edge-disjoint paths (that is, no two paths traverse the same edge), then they are obviously k-connected. A fundamental fact, whose ingenious proof we omit, is Menger's theorem which confirms that the converse is also true: if two vertices are k-connected, then there are k edge-disjoint paths connecting them. It takes some ingenuity to prove this even for the case $k = 2$.

5.3.4 Connection by Simple Path

Where there's a path, there's a simple path. This is sort of obvious, but proving it carefully provides a nice illustration of the Well-ordering Principle.

Lemma 5.3.5. *If vertex u is connected to vertex v in a graph, then there is a simple path from u to v.*

Proof. Since there is a path from u to v, there must, by the Well-ordering Principle, be a minimum length path from u to v. If the minimum length is zero or one, this minimum length path is itself a simple path from u to v.

Otherwise, there is a minimum length path

$$v_0, v_1, \ldots, v_k$$

from $u = v_0$ to $v = v_k$ where $k \geq 2$. We claim this path must be simple.

To prove the claim, suppose to the contrary that the path is not simple, that is, some vertex on the path occurs twice. This means that there are integers i, j such that $0 \leq i < j \leq k$ with $v_i = v_j$. Then deleting the subsequence

$$v_{i+1}, \ldots v_j$$

yields a strictly shorter path

$$v_0, v_1, \ldots, v_i, v_{j+1}, v_{j+2}, \ldots, v_k$$

from u to v, contradicting the minimality of the given path. □

Actually, we proved something stronger:

Corollary 5.3.6. *For any path of length k in a graph, there is a simple path of length* at most *k with the same endpoints.*

5.3.5 The Minimum Number of Edges in a Connected Graph

The following theorem says that a graph with few edges must have many connected components.

Theorem 5.3.7. *Every graph with v vertices and e edges has at least $v - e$ connected components.*

Of course for Theorem 5.3.7 to be of any use, there must be fewer edges than vertices.

Proof. We use induction on the number of edges, e. Let $P(e)$ be the proposition that

> for every v, every graph with v vertices and e edges has at least $v - e$ connected components.

Base case:($e = 0$). In a graph with 0 edges and v vertices, each vertex is itself a connected component, and so there are exactly $v = v - 0$ connected components. So $P(e)$ holds.

Inductive step: Now we assume that the induction hypothesis holds for every e-edge graph in order to prove that it holds for every $(e + 1)$-edge graph, where $e \geq 0$.

Consider a graph, G, with $e+1$ edges and k vertices. We want to prove that G has at least $v-(e+1)$ connected components.

To do this, remove an arbitrary edge a—b and call the resulting graph G'. By the induction assumption, G' has at least $v - e$ connected components.

Now add back the edge a—b to obtain the original graph G. If a and b were in the same connected component of G', then G has the same connected components as G', so G has at least $v - e > v - (e + 1)$ components. Otherwise, if a and b were in different connected components of G', then these two components are merged into one in G, but all other components remain unchanged, reducing the number of components by 1. Therefore, G has at least $(v - e) - 1 = v - (e + 1)$ connected components. So in either case, $P(e + 1)$ holds. This completes the Induction step.

The theorem now follows by induction. □

Corollary 5.3.8. *Every connected graph with e vertices has at least $e - 1$ edges.*

A couple of points about the proof of Theorem 5.3.7 are worth noting. First, notice that we used induction on the number of edges in the graph. This is very common in proofs involving graphs, and so is induction on the number of vertices. When you're presented with a graph problem, these two approaches should be among the first you consider.

The second point is more subtle. Notice that in the inductive step, we took an arbitrary $(n + 1)$-edge graph, threw out an edge so that we could apply the induction assumption, and then put the edge back. You'll see this shrink-down, grow-back process very often in the inductive steps of proofs related to graphs. This might seem like needless effort; why not start with an e-edge graph and add one more to get an $(n + 1)$-edge graph? That would work fine in this case, but opens the door to a nasty logical error called *buildup* error. You'll see an example in class. Always use shrink-down, grow-back arguments, and you'll never fall into this trap.

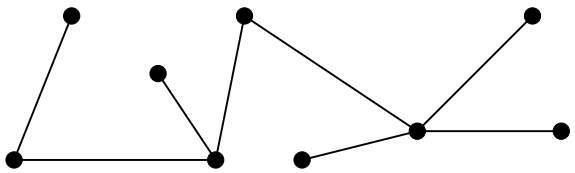
5.4 Trees

Trees are a fundamental data structure in Computer Science, and there are many kinds, for example rooted, ordered, or binary trees. In this section we focus on the purest kind of tree. Namely, we use the ***tree*** to mean a connected graph without simple cycles.

A graph with no simple cycles is called ***acyclic***; so trees are acyclic connected graphs.

5.4.1 Tree Properties

Here is an example of a tree:



A vertex of degree one is called a ***leaf***. In this example, there are 5 leaves.

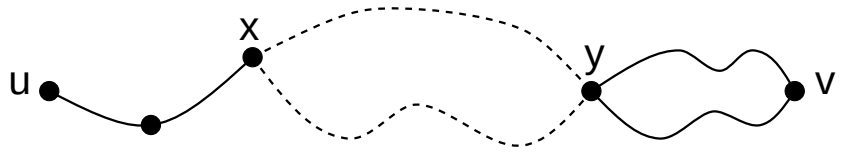
The graph shown above would no longer be a tree if any edge were removed, because it would no longer be connected. The graph would also not remain a tree if any edge were added between two of its vertices, because then it would contain a simple cycle. Furthermore, note that there is a unique path between every pair of vertices. These features of the example tree are actually common to all trees.

Theorem 5.4.1. *Every tree has the following properties:*

1. *Any connected subgraph is a tree.*
2. *There is a unique simple path between every pair of vertices.*
3. *Adding an edge between two vertices creates a cycle.*
4. *Removing any edge disconnects the graph.*
5. *If it has at least two vertices, then it has at least two leaves.*
6. *The number of vertices is one larger than the number of edges.*

Proof. 1. A simple cycle in a subgraph is also a simple cycle in the whole graph, so any subgraph of an acyclic graph must also be acyclic. If the subgraph is also connected, then by definition, it is a tree.

2. There is at least one path, and hence one simple path, between every pair of vertices, because the graph is connected. Suppose that there are two different simple paths between vertices u and v. Beginning at u, let x be the first vertex where the paths diverge, and let y be the next vertex they share. Then there are two simple paths from x to y with no common edges, which defines a simple cycle. This is a contradiction, since trees are acyclic. Therefore, there is exactly one simple path between every pair of vertices.



3. An additional edge u—v together with the unique path between u and v forms a simple cycle.

4. Suppose that we remove edge u—v. Since a tree contained a unique path between u and v, that path must have been u—v. Therefore, when that edge is removed, no path remains, and so the graph is not connected.

5. Let $v_1, \ldots, v_m$ be the sequence of vertices on a longest simple path in the tree. Then $m \geq 2$, since a tree with two vertices must contain at least one edge. There cannot be an edge v_1—v_i for $2 < i \leq m$; otherwise, vertices $v_1, \ldots, v_i$ would from a simple cycle. Furthermore, there cannot be an edge u—v_1 where u is not on the path; otherwise, we could make the path longer. Therefore, the only edge incident to v_1 is v_1—v_2, which means that v_1 is a leaf. By a symmetric argument, v_m is a second leaf.

6. We use induction on the number of vertices. For a tree with a single vertex, the claim holds since it has no edges and $0 + 1 = 1$.

   Now suppose that the claim holds for all n-vertex trees and consider an $(n + 1)$-vertex tree, T. Let v be a leaf of the tree.

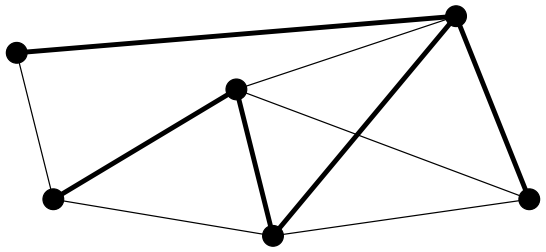
   We will let the reader verify that deleting a vertex of degree 1 (and its incident edge) from any connected graph leaves a connected subgraph. So by (1), deleting v and its incident edge gives a smaller tree, and this smaller tree has one more vertex than edge by induction. If we reattach the vertex, v, and its incident edge, then the equation still holds because the number of vertices and number of edges both increase by 1. Thus, the claim holds for T and, by induction, for all trees.

□

Various subsets of these properties provide alternative characterizations of trees, though we won't prove this. For example, a *connected* graph with a number of vertices one larger than the number of edges is necessarily a tree. Also, a graph with unique paths between every pair of vertices is necessarily a tree.

5.4.2 Spanning Trees

Trees are everywhere. In fact, every connected graph contains a a subgraph that is a tree with the same vertices as the graph. This is a called a ***spanning tree*** for the graph. For example, here is a connected graph with a spanning tree highlighted.



Theorem 5.4.2. *Every connected graph contains a spanning tree.*

Proof. Let T be a connected subgraph of G, with the same vertices as G, and with the smallest number of edges possible for such a subgraph. We show that T is acyclic by contradiction. So suppose that T has a cycle with the following edges:

$$v_0\text{—}v_1, v_1\text{—}v_2, \ldots, v_n\text{—}v_0$$

Suppose that we remove the last edge, v_n—v_0. If a pair of vertices x and y was joined by a path not containing v_n—v_0, then they remain joined by that path. On the other hand, if x and y were joined by a path containing v_n—v_0, then they remain joined by a path containing the remainder of the cycle. So all the vertices of G are still connected after we remove an edge from T. This is a contradiction, since T was defined to be a minimum size connected subgraph with all the vertices of G. So T must be acyclic. □

5.5 In-Class Problems Week 5, Mon.

Problem. See if you can come up with a stable marriage assignment given the following preferences. You are not expected to know/remember the Dating Protocol that solves this problem and which is about to be covered in lecture. (And if you do remember the protocol, don't spoil your teammates' fun by telling them.)

$boys$	$girls$
$1 : CBEAD$	$A : 35214$
$2 : ABECD$	$B : 52143$
$3 : DCBAE$	$C : 43512$
$4 : ACDBE$	$D : 12345$
$5 : ABDEC$	$E : 23415$

Solution. See Slides 5M (pdf). ■

Problem 5.5.1. Four Students want separate assignments to four VI-A Companies. Here are their preference rankings:

Student	Companies
Albert:	HP, Bellcore, AT&T, Draper
Jeff:	AT&T, Bellcore, Draper, HP
Tina:	HP, Draper, AT&T, Bellcore
Jay:	Draper, AT&T, Bellcore, HP

Company	Students
AT&T:	Jay, Albert, Tina, Jeff
Bellcore:	Tina, Jeff, Albert, Jay
HP:	Jay, Tina, Albert, Jeff
Draper:	Jeff, Jay, Tina, Albert

(a) Use the Mating Ritual (given in the Appendix) to find *two* stable assignments of Students to Companies.

Solution. Treat Students as Boys and the result is the following assignment:

Student	Companies	Rank in the original list
Albert:	Bellcore	2
Jeff:	AT&T	1
Tina:	HP	1
Jay:	Draper	1

Treat Companies as Boys and the result is the following assignment:

Company	Students	Rank in the original list
AT&T:	Albert	2
Bellcore:	Jeff	2
HP:	Tina	2
Draper:	Jay	2

■

(b) Describe a simple procedure to determine whether any given stable marriage problem has a unique solution, that is, only one possible stable matching.

Solution. See if the Mating Ritual with Boys as suitors yields the same solution as the algorithm with Girls as suitors. These two marriage assignments are boy-optimal and boy-pessimal, respectively, so they agree iff there is a unique solution.

To see why exactly why this is so, suppose there are two stable matchings. Then at least one boy, call him Brad, must have different wives, Angelina and Jen, in these two matchings. Since Brad will prefer one of these spouses, say Angelina over Jen, then in his optimal matching he must have a wife at least as desirable as Angelina and in his pessimal matching a wife at most as desirable as Jen. So Brad must have different wives in the boy-optimal and boy-pessimal matchings. It follows that if the boy-pessimal and boy-optimal matchings agree, there can't be any other stable matching. ■

Problem 5.5.2. You heard that an invariant of the Mating ritual is:

> For every girl, G, and every boy, B, if G is crossed off B's list, then G has a favorite suitor and she prefers him over B.

Use the invariant to prove that the Mating Algorithm produces stable marriages. (Don't look up the proof in the Course Notes.)

Solution. *Proof.* Let Brad be some boy and Jen be any girl that he is *not* married to on the last day of the Mating Ritual. We claim that Brad and Jen are not a rogue couple. Since Brad is an arbitrary boy, it follows that no boy is part of a rogue couple. Hence the marriages on the last day are stable.

To prove the claim, we consider two cases:

Case 1. Jen is not on Brad's list. Then by invariant P, we know that Jen prefers her husband to Brad. So she's not going to run off with Brad: the claim holds in this case.

Case 2. Otherwise, Jen is on Brad's list. But since Brad is not married to Jen, he must be choosing to serenade his wife instead of Jen, so he must prefer his wife. So he's not going to run off with Jen: the claim also holds in this case. □

■

Problem 5.5.3. Suppose there are more boys than girls.

(a) Define what a stable matching should mean in this case.

Solution. The definition of rogue couple in a matching should now include an unmarried boy and a married girl that likes the boy better than her husband. ■

(b) Explain why applying the Mating Ritual in this case will yield a stable matching in which every girl is married.

Solution. The same invariant holds, and the same proof applies, in this case, except at the very last step: since there are more boys than girls, the fact that all the girls are married does not imply that all the boys are (and of course there have to be some unmarried boys at the end, since polyandry can't happen). The proof of stability then goes through unchanged for married boys, and for an unmarried boy, Billy Bob, we need only observe that the Invariant implies all the girls like their spouses better than Billy Bob, so he can't be in a rogue couple either. ■

Appendix: The Mating Ritual

The *Mating Ritual* takes place over several days. The following events happen each day:

Morning: Each girl stands on her balcony. Each boy stands under the balcony of his favorite among the girls on his list, and he serenades her. If a boy has no girls left on his list, he stays home and does his 6.042 homework.

Afternoon: Each girl who has one or more suitors serenading her, says to her favorite suitor, "We might get engaged. Come back tomorrow." To the others, she says, "No. I will never marry you! Take a hike!"

Evening: Any boy who is told by a girl to take a hike, crosses that girl off his list.

Termination condition: When every girl has at most one suitor, the ritual ends with each girl marrying her suitor, if she has one.

5.6 In-Class Problems Week 5, Wed.

Problem 5.6.1. **(a)** Prove that in every graph, there are an even number of vertices of odd degree.

Hint: The Handshaking Theorem.

Solution. *Proof.* Partitioning the vertices into those of even degree and those of odd degree, we know

$$\sum_{v \in V} d(v) \;=\; \sum_{d(v) \text{ is even}} d(v) \;+\; \sum_{d(v) \text{ is odd}} d(v)$$

By the Handshaking Theorem, the value of the lefthand side of this equation equals twice the number of edges, and so is even. The first summand on the righthand side is even since it is a sum of even values. So the second summand on the righthand side must also be even. But since it is entirely a sum of odd values, it must must contain an even number of terms. That is, there must be an even number of vertices with odd degree. □

■

(b) Conclude that at a party where some people shake hands, the number of people who shake hands an odd number of times is an even number.

Solution. We can represent the people at the party by the vertices of a graph. If two people shake hands, then there is an edge between the corresponding vertices. So the degree of a vertex is the number of handshakes the corresponding person performed. The result in the first part of this problem now implies that there are an even number of odd-degree vertices, which translates into an even number of people who shook an odd number of hands. ■

Problem 5.6.2. For each of the following pairs of graphs, either define an isomorphism between them, or prove that there is none. (We write ab as shorthand for a—b.)

(a)

$$G_1 \text{ with } V_1 = \{1,2,3,4,5,6\},\ E_1 = \{12,23,34,14,15,35,45\}$$
$$G_2 \text{ with } V_2 = \{1,2,3,4,5,6\},\ E_2 = \{12,23,34,45,51,24,25\}$$

Solution. Not isomorphic: G_2 has a node, 2, of degree 4, but the maximum degree in G_1 is 3. ■

(b)

$$G_1 \text{ with } V_1 = \{1,2,3,4,5,6\},\ E_1 = \{12,23,34,14,45,56,26\}$$
$$G_2 \text{ with } V_2 = \{a,b,c,d,e,f\},\ E_2 = \{ab,bc,cd,de,ae,ef,cf\}$$

Solution. Isomorphic with the vertex correspondence: $1f, 2c, 3d, 4e, 5a, 6b$ ■

(c)

$$G_1 \text{ with } V_1 = \{a, b, c, d, e, f, g, h\}\,,\ E_1 = \{ab, bc, cd, ad, ef, fg, gh, he, dh, bf\}$$
$$G_2 \text{ with } V_2 = \{s, t, u, v, w, x, y, z\}\,,\ E_2 = \{st, tu, uv, sv, wx, xy, yz, wz, sw, vz\}$$

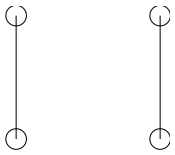
Solution. Not isomorphic: they have the same number of vertices, edges, and set of vertex degrees. But the degree 2 vertices of G_1 are all adjacent to two degree 3 vertices, while the degree 2 vertices of G_2 are all adjacent to one degree 2 vertex and one degree 3 vertex. ■

Problem 5.6.3. A graph is ***connected*** when for every pair of vertices, u and v, there is a path between u and v.

False Claim. *If every vertex in a graph has positive degree, then the graph is connected.*

(a) Prove that this Claim is indeed false by providing a counterexample.

Solution. There are many counterexamples; here is one:



■

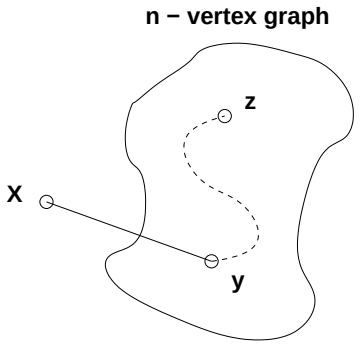
(b) Since the claim is false, there must be an logical mistake in the following "proof." Pinpoint the *first* logical mistake (unjustified step) in the proof.

False proof. We use induction. Let $P(n)$ be the proposition that if every vertex in an n-vertex graph has positive degree, then the graph is connected.

Base cases: ($n \leq 2$). In a graph with 1 vertex, that vertex cannot have positive degree, so $P(1)$ holds vacuously.

$P(2)$ holds because there is only one graph with two vertices of positive degree, namely, the graph with an edge between the vertices, and this graph is connected.

Inductive step: We must show that $P(n)$ implies $P(n+1)$ for all $n \geq 2$. Consider an n-vertex graph in which every vertex has positive degree. By the assumption $P(n)$, this graph is connected; that is, there is a path between every pair of vertices. Now we add one more vertex x to obtain an $(n+1)$-vertex graph:



All that remains is to check that there is a path from x to every other vertex z. Since x has positive degree, there is an edge from x to some other vertex; call it y. Thus, we can obtain a path from x to z by adjoining the edge x—y to the path from y to z. This proves $P(n+1)$.

By the principle of induction $P(n)$ is true for all $n \geq 0$, which proves the theorem.

□

Solution. This one is tricky: the proof is actually a good proof of something else. The first error in the proof is only in the final statement of the inductive step: "This proves $P(n+1)$".

The issue is that to prove $P(n+1)$, *every* $(n+1)$-vertex positive-degree graph must be shown to be connected. But the proof doesn't show this. Instead, it shows that every $(n+1)$-vertex positive-degree graph *that can be built up by adding a vertex of positive degree to an n-vertex connected graph*, is connected.

The problem is that *not every* $(n+1)$-vertex positive-degree graph can be built up in this way. The counterexample above illustrates this: there is no way to build that 4-vertex positive-degree graph from a 3-vertex positive-degree graph.

More generally, this is an example of "buildup error". This error arises from a faulty assumption that every size $n+1$ graph with some property can be "built up" in some particular way from a size n graph with the same property. (This assumption is correct for some properties, but incorrect for others— such as the one in the argument above.)

One way to avoid an accidental build-up error is to use a "shrink down, grow back" process in the inductive step: start with a size $n+1$ graph, remove a vertex (or edge), apply the inductive hypothesis $P(n)$ to the smaller graph, and then add back the vertex (or edge) and argue that $P(n+1)$ holds. Let's see what would have happened if we'd tried to prove the claim above by this method:

Inductive step: We must show that $P(n)$ implies $P(n+1)$ for all $n \geq 1$. Consider an $(n+1)$-vertex graph G in which every vertex has degree at least 1. Remove an arbitrary vertex v, leaving an n-vertex graph G' in which every vertex has degree... uh-oh!

The reduced graph G' might contain a vertex of degree 0, making the inductive hypothesis $P(n)$ inapplicable! We are stuck— and properly so, since the claim is false! ■

Problem 5.6.4. **(a)** For any vertex, v, in a graph, let $\widehat{v}$ be the set of vertices adjacent to v, that is,

$$\widehat{v} ::= \{v' \mid v\text{—}v' \text{ is an edge of the graph}\}.$$

Suppose f is an isomorphism from graph G to graph H. Carefully prove that $f(\widehat{v}) = \widehat{f(v)}$.

Solution. We first show that $\widehat{f(v)} \subseteq f(\widehat{v})$ by showing that if $w \in \widehat{f(v)}$, then $w \in f(\widehat{v})$.

Now $w \in \widehat{f(v)}$ means that w—$f(v)$ is an edge of H. Since f is an isomorphism, there must be some v' such that $w = f(v')$. So $f(v')$—$f(v)$ is an edge of H, and therefore v'—v is an edge of G, by definition of isomorphism. This means $v' \in \widehat{v}$, and so $f(v') \in f(\widehat{v})$ by definition of $f(\widehat{v})$. So $w = f(v') \in f(\widehat{v})$, as required.

Conversely, we show that $f(\widehat{v}) \subseteq \widehat{f(v)}$ by showing that if $w \in f(\widehat{v})$, then $w \in \widehat{f(v)}$.

But $w \in f(\widehat{v})$ means that $w = f(v')$ for some v' adjacent to v in G. This means v—v' is an edge of G, and so $f(v)$—$f(v')$ is an edge of H by definition of isomorphism. So $w = f(v')$ is adjacent to $f(v)$; in other words, $w \in \widehat{f(v)}$, as required. ■

(b) Conclude that if G and H are isomorphic graphs, then for each $k \in \mathbb{N}$, they have the same number of degree k vertices.

Solution. By definition, $\deg(v) = |\widehat{v}|$. Since an isomorphism is a bijection, a set and its image will be the same size (by the Mapping Rule from Week 2 Notes), so the Lemma of part (a) implies that an isomorphism, f, maps degree k vertices to degree k vertices. This means that the image under f of the set of degree k vertices of G is precisely the set of degree k vertices of H. So by the Mapping Rule again, there are the same number of degree k vertices in G and H. ■

Definitions

A *path* in a graph, G, is a sequence of $k \geq 0$ vertices

$$v_0, \ldots, v_k$$

such that v_i—v_{i+1} is an edge of G for all i where $0 \leq i < k$. The path is said to *start* at v_0, to *end* at v_k, and *length* of the path is defined to be k. An edge, e, is ***traversed*** n ***times*** by the path if there are n different values of i such that edge v_i—v_{i+1} is e.

The path is *simple* iff all the v_i's are different, that is, $v_i = v_j$ only if $i = j$.

Two vertices in a graph are *connected* iff there is a path that begins with one of the vertices and ends with the other. The *shortest* path between two vertices is always simple.

A graph is *connected* iff every pair of vertices is connected.

Cycles are paths that begin and end with the same vertex. *Simple* cycles are cycles that don't cross themselves.[5]

[5]The technical definition of simple cycle appears in Week 5 Notes.

5.7 In-Class Problems Week 5, Fri.

Problem 5.7.1. (From Wednesday's class problems.)

(a) Prove that in every graph, there are an even number of vertices of odd degree.

Hint: The Handshaking Theorem.

Solution. See Class Problems 5W solutions. ■

(b) Conclude that at a party where some people shake hands, the number of people who shake hands an odd number of times is an even number.

Solution. See Class Problems 5W solutions. ■

Problem 5.7.2. Procedure $Mark$ starts with a connected, simple graph with all edges unmarked and then marks some edges. At any point in the procedure a path that traverses only marked edges is called a *fully marked* path, and an edge that has no fully marked path between its endpoints is called *eligible*.

Procedure $Mark$ simply keeps marking eligible edges, and terminates when there are none.

Prove that $Mark$ terminates, and that when it does, the set of marked edges forms a spanning tree of the original graph.

Solution. As a state machine, the start state of $Mark$ is some given connected graph, G. The rest of the states are copies of G with some edges marked.

$Mark$ terminates because the number of unmarked edges decreases by one at each transition, so this number is a strictly decreasing nonnegative integer-valued variable, which we know implies termination. (A common mistake in arguing termination of $Mark$ was to instead say that the number of eligible edges was strictly decreasing, without any additional reasoning. It turns out that the fact is true, but it is not obvious; you would have to prove that removing an eligible edge did not result in new edges becoming eligible.)

To prove partial correctness, we show if $Mark$ terminates, the marked edges of the final state form a spanning tree of G. So we must show that the marked edges form an acyclic connected graph with the same set of vertices as G.

To do this we verify the invariant:

> The marked edges form an acyclic graph. (*)

To verify (*) is an invariant, consider a step $H \to H'$, where H satisfies (*). This means that H has no fully marked cycles, and H' is the same as H with an edge, e, that was one eligible in H, now marked in H'. But in H', the only fully marked path between the endpoints of e must be the edge e itself, by definition of "eligible." So e is not on a fully marked simple cycle in H'. Since H

and H' are otherwise the same, there is no fully marked simple cycle elsewhere in H'. That is, H' satisfies (*).

Since the start state G has no marked edges, it satisfies (*) trivially. Hence, by the Invariant Principle, any final state of $Mark$ satisfies (*).

We also claim that in any final state, there is a fully marked path between any two vertices. To prove this assume to the contrary that there were two vertices, u and v, with no fully marked path between them. Since there is a path in G between u and v, there must be a path between u and v traversing the smallest number of unmarked edges, and this path must contain at least one unmarked edge, e. Now if there was a fully marked path between the endpoints of e, we could replace e by this path to obtain a path between u and v with fewer unmarked edges. So there can't be a fully marked path between the endpoints of e, which means that e is eligible, contradicting the fact the state was final.

So in any final state, the marked edges determine an acyclic, connected graph on the vertices of G. That is, the marked edges determine a spanning tree of G. ■

Problem 5.7.3. Prove that K_n is $(n-1)$-connected for $n > 1$.

Solution. *Proof.* Consider any two distinct vertices u and v in K_n. For each of the other $n-2$ vertices in K_n, there is a length 2 path from u to v via that vertex. This gives us a total of $n-1$ edge disjoint paths between u and v, namely the the $n-2$ length 2 paths above and the length 1 path directly from u to v. This implies it takes at least $n-1$ cuts to disconnect any 2 vertices, meaning that K_n is $(n-1)$-connected. □

■

Problem 5.7.4. Prove that a graph is a tree iff it has a unique simple path between any two vertices.

Solution. The proof in Week 5 Notes of Theorem 4.1.2 shows that in a tree there are unique simple paths between any two vertices:

> There is at least one path, and hence one simple path, between every pair of vertices, because the tree is connected. Suppose that there are two different simple paths between vertices u and v. Beginning at u, let x be the first vertex where the paths diverge, and let y be the next vertex they share, illustrated in the figure below. Then there are two simple paths from x to y with no common edges, and this defines a simple cycle. This is a contradiction, since trees are acyclic. Therefore, there is exactly one simple path between every pair of vertices.

Conversely, suppose we have a graph, G, with unique paths. Now G is connected since there is a path between any two vertices. So we need only show that G is acyclic. But if there was a simple cycle in G, there are two paths between any two vertices on the cycle (going one way around the cycle or the other way around), a violation of uniqueness. So G must cannot have any simple cycles. ■

5.8 Problem Set 4

Problem 5.8.1. The Mating Ritual described in Week 5 Notes yields a stable perfect match when there are an equal number of boys and girls.

(a) Suppose there are more girls than boys. Define what a stable matching should mean in this case, and explain why applying the Mating Ritual will yield a stable matching in which every boy is married. Then briefly explain why each boy gets married to his optimal spouse. (You can give complete proofs like those in the Notes, but it's also OK just to explain how the proofs in the Notes should be modified to handle the case with more boys.)

Solution. The definition of rogue couple in a matching should now include a boy and unmarried girl that the boy likes better than his spouse. Nothing much else changes. ■

(b) The Notes also described an early and valuable application of the Mating Ritual to assigning medical students to hospital residencies. In this case, each student has a preference ranking of all the hospitals, each hospital has a preference ranking of all the students, and in addition, each hospital has a certain number of resident positions to fill. The number of resident positions to be filled typically differs between hospitals, and the total number to be filled may not equal the number of students.

Carefully define what a *stable assignment* of residents to hospitals should mean in the case that the total number of available positions is at least as large as the number of students. Then explain how to extend the Mating Ritual to find such a stable assignment.

Solution. A student and hospital will be a rogue couple in an assignment if the student prefers the hospital to his assigned hospital and the hospital prefers the student to one of the students assigned to it. In addition, rogue couples include a student who prefers a hospital with an unfilled position to his assigned hospital.

To apply the Mating Ritual, replace each hospital with $k > 1$ positions to fill by k new hospitals each with the same preference list as the original hospital. Also, replace each occurrence of the hospital on a boy's list by these k new hospitals, in any order. Now treat the hospitals as girls and run the Mating Ritual from the previous part. Then assign a student to a hospital when the Mating Ritual matches him with a copy of that hospital. ■

Problem 5.8.2. In a set of stable marriages between an equal number of boys and girls, call a person *lucky* if their spouse appears in the top half of their preference list.

Claim. *The Mating Algorithm produces a set of marriages with at least one lucky person.*

To prove the Claim, for each girl, G, define a "rejection count" derived variable, $r(G)$, to be the number of boys she has rejected. Similarly, for each boy, B, define a "rejected count" variable, $r(B)$, to be the number of times he has been rejected by girls.

(a) Define the predicate $L(B)$ meaning "B is a lucky boy," in terms of the final value of $r(B)$.

Solution. Since a boy works down his list of favorite girls, he will be in the top half of his list if he has been rejected by fewer than half the girls:

$$L(B) ::= \text{ the value of } r(B) \text{ on the final day is less than half the number of girls.}$$

■

(b) Suppose that on the final day, the value of $r(G)$, averaged over all the girls, is at *least* half the number of boys. Explain why there must be a lucky girl.

Solution. Suppose there was no lucky girl. Then the value of $r(G)$ for each girl would be less than $n/2$, since if any girl rejected at least $n/2$ boys, she would be married to a boy in the top half of her list, and therefore she would be lucky. But if the value of $r(G)$ is less than $n/2$ for each girl, then the average of $r(G)$ would be less than $n/2$, which contradicts the fact that the average of $r(G)$ is at least $n/2$. ■

(c) The rejection counts in the Mating Algorithm satisfy an obvious invariant. Use this invariant and the previous problem parts to prove the Claim.

Solution. At any stage of the Mating Algorithm, the total number of rejections by girls must equal the total number of times boys get rejected:

$$\sum_G r(G) = \sum_B r(B).$$

Now if no boy is lucky, then by part (a), the final value (on the wedding day) of $\sum_B r(B)$ is at least $n(n/2)$ where n is the number of boys. So the invariant implies that on the wedding day $\sum_G r(G) > n(n/2)$. So the average value of $r(G)$ on the wedding day is at least $n/2$, and part (b) implies there must be a lucky girl. ■

Problem 5.8.3. **(a)** Construct a set of marriage preferences in which there are at least three different stable marriage assignments.

Solution. Consider a stable marriage problem with 2 boys and 3 girls and the following partial information about their preferences:

B1:	G1	G2	–	–
B2:	G2	G1	–	–
B3:	–	–	G3	G4
B4:	–	–	G4	G3
G1:	B2	B1	–	–
G2:	B1	B2	–	–
G3:	–	–	B3	B4
G4:	–	–	B4	B3

The assignment

$$(B1, G1), (B2, G2), (B3, G3)(B4, G4)$$

will be a stable matching whatever the unspecified preferences may be:

- B1 and B2 get their 1st choice, so won't be in a rogue couple.
- G1 and G2 get their 2nd choices, so won't be in a rogue couple with the other two boys, B3 or B4. So G1 and G2 won't be in any rogue couple, either.
- G3 and G4 get their best remaining choices, so will never be in a rogue couple.
- This leaves no possible rogue partners for B3 and B4.

So the marriages are sure to be stable.

Moreover, this set of marriages could not be produced by the Mating Ritual described in class, because B1, B2, G3, and G4 get their optimal choices, but G1, G2, B3, and B4 do not. So this set of marriages is optimal for neither all the boys nor all the girls, and so would not be produced by the Mating Ritual. Moreover, it's easy to run the Mating Ritual and check that it yields two additional stable matchings, for a total of three stable matchings.

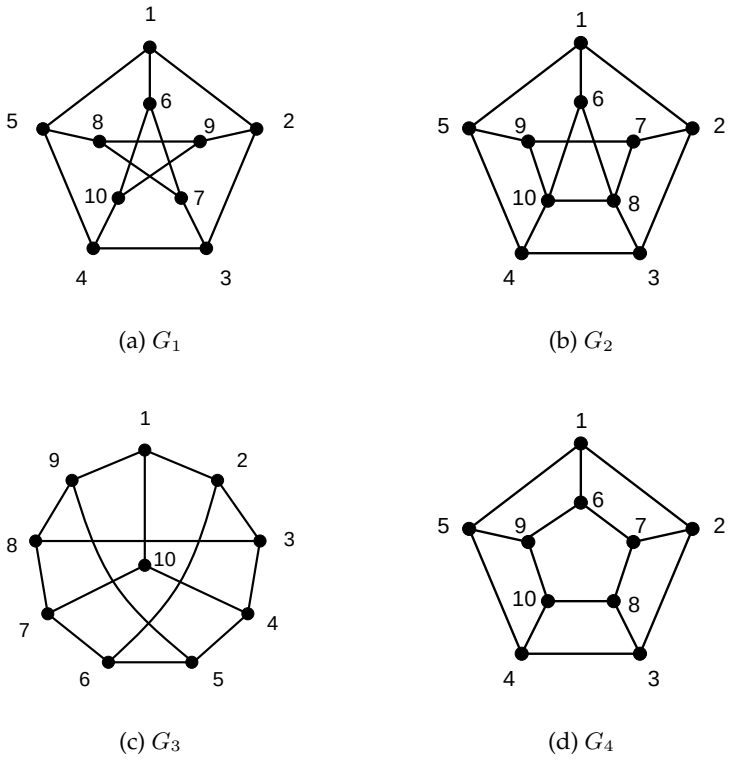
In fact, we can deduce that the Mating Ritual must produce two more matchings without actually running it. Namely, since the stable matching above is not an optimal matching, somebody must have different spouses in this matching and an optimal matching. This person will prefer one of these spouses to the other, and so will have different spouses in the boy-optimal/girl-pessimal and boy-pessimal/girl-optimal matchings. Since the Mating Ritual produces these different optimal matchings, it will yield two more stable matchings for a total of three. ■

(b) (Optional) Describe how to define a set of marriage preferences among n boys and n girls which have more than $2^{n/4}$ stable assignments.

Solution. To find a set of preferences which have many stable matchings, arrange the boys into a list of $n/2$ pairs, and likewise arrange the girls into a list of $n/2$ pairs of girls. Choose preferences so that the kth pair of boys ranks the kth pair of girls just below the previous pairs of girls, and likewise so the kth pair of girls. This ensures that any matching in which the corresponding pairs of boys and girls are matched will be stable, as long as the submatches involving each pair separately is stable. But the preferences of the kth pairs for each other can be chosen so that either of the two ways of matching them to each other will be stable in isolation. So there will be at least $2^{n/2}$ such stable matchings for this set of preferences. ■

Problem 5.8.4. A property of a graph is said to be *preserved under isomorphism* if whenever G has that property, every graph isomorphic to G also has that property. For example, the property of having five vertices is preserved under isomorphism: if G has five vertices then every graph isomorphic to G also has five vertices.

Determine which among the four graphs pictured in Figure 5.4 are isomorphic. If two of these graphs are isomorphic, describe an isomorphism between them. If they are not, give a property that is preserved under isomorphism such that one graph has the property, but the other does not. For at least one of the properties you choose, *prove* that it is indeed preserved under isomorphism (you only need prove one of them).



(a) G_1 (b) G_2

(c) G_3 (d) G_4

Figure 5.4: Which graphs are isomorphic?

Solution. G_1 and G_3 are isomorphic. In particular, the function $f : V_1 \to V_3$ is an isomomorphism, where

$$f(1) = 1 \qquad f(2) = 2 \qquad f(3) = 3 \qquad f(4) = 8 \qquad f(5) = 9$$
$$f(6) = 10 \qquad f(7) = 4 \qquad f(8) = 5 \qquad f(9) = 6 \qquad f(10) = 7$$

G_1 and G_4 are not isomorphic to G_2: G_2 has a vertex of degree four and neither G_1 nor G_4 has one.

G_1 and G_4 are not isomorphic: G_4 has a simple cycle of length four and G_1 does not. ■

Problem 5.8.5. Given a simple graph G, we apply the following operation to the graph: pick two vertices $u \neq v$ such that either

1. there is an edge of G between u and v and there is also a path from u to v which does *not* include this edge; in this case, delete the edge $\{u, v\}$.
2. or, there is no path from u to v; in which case, add the edge $\{u, v\}$.

We keep repeating these operations until it is no longer possible to find two vertices $u \neq v$ to which an operation applies.

Assume the vertices of G are the integers $1, 2, \ldots, n$ for some $n \geq 2$. This procedure can be modelled as a state machine whose states are all possible simple graphs with vertices $1, 2, \ldots, n$. The start state is G, and the final states are the graphs on which no operation is possible.

(a) Let G be the graph with vertices $\{1, 2, 3, 4\}$ and edges

$$\{\{1, 2\}, \{3, 4\}\}$$

What are the possible final states reachable from start state G? Draw them.

Solution. It's not possible to delete any edge. The procedure can only add an edge connecting exactly one of vertices 1 or 2 to exactly one of vertices 3 or 4, and then terminate. So there are four possible final states. ■

(b) For any state, G', let e be the number of edges in G', c be the number of connected components it has, and s be the number of simple cycles. For each of the derived variables below, indicate the *strongest* of the properties that it is guaranteed to satisfy, no matter what the starting graph G is, and briefly explain your answer.

The choices for properties are: *constant*, *strictly increasing*, *strictly decreasing*, *weakly increasing*, *weakly decreasing*, *none of these*. The derived variables are

(i) e

Solution. none of these ■

(ii) c

Solution. weakly decreasing ■

(iii) s

Solution. weakly decreasing ■

(iv) $e - s$

Solution. weakly increasing ■

(v) $c + e$

Solution. weakly decreasing ■

(vi) $3c + 2e$

Solution. strictly decreasing ■

(vii) $c + s$

Solution. strictly decreasing ■

(viii) (c, e), partially ordered coordinatewise (the *product* partial order).

Solution. none of these ■

(ix) (c, e), ordered lexicographically

Solution. strictly decreasing ■

(c) Conclude that the procedure terminates.

Solution. To show that the variable (vi) strictly decreases, note that the rule for deleting an edge ensures that the connectedness relation does not change, so neither does the number of connected components c. Meanwhile the number of edges e decreases by one when an edge is deleted. Therefore the variable $3c + 2e$ decreases by 2. The rule for adding an edge ensures that the number of connected components c decreases by one and the number of edges e increases by one. Therefore the variable $3c + 2e$ decreases by 1.

To show that the variable (vii) strictly decreases, note that the rule for deleting an edge ensures that the number of connected components c does not change and the number of simple cycles s decreases by n, where $n \geq 1$. Therefore the variable $c + s$ decreases by n. The rule for adding an edge ensures that the number of connected components c decreases by one and the number of simple cycles s does not change. Therefore the variable $c + s$ decreases by one.

To show that the lexicographically ordered (c, e) strictly decreases, note that the rule for deleting an edge ensures that the number of connected components c does not change and the number of edges e decreases by one. The rule for adding an edge ensures that the number of connected components c decreases by one. ■

(d) Prove that any final state must be a tree on the vertices.

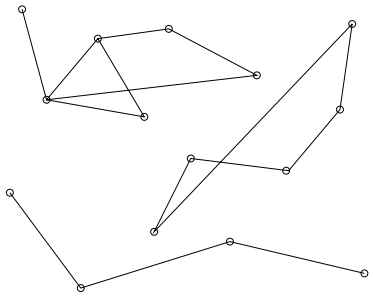
Solution. We use the characterization of a tree as a cycle-free, connected, simple graph.

A final state must be connected, because otherwise there would be two vertices with no path between them, and then a transition adding the edge between them would be possible, contradicting finality of the state.

A final state can't have a cycle, because deleting any edge on the cycle would be a possible transition. ■

Problem 5.8.6. A set, M, of vertices of a graph is a *maximal connected* set if every pair of vertices in the set are connected, and any set of vertices properly containing M will contain two vertices that are not connected.

(a) What are the maximal connected subsets of the following (unconnected) graph?



Solution. There are three maximal subsets, each one equal to the vertices of one of the connected components of the graph. ■

(b) Explain the connection between maximal connected sets and connected components. Prove it.

Solution. They are the same.

We first show that a connected component is a maximal set. A connected component consists of *all* the vertices connected to some vertex v. A larger set of vertices would, by definition, contain both v and a vertex, w, that is not in the connected component. This means that w is not connected to v, and therefore the larger set is not connected. So a connected component is maximal.

Conversely, suppose we have a maximal connected set, M. Since M is connected, any vertex, v, in M is connected to all the other vertices in M. If there was any vertex, w, connected to v, that was not in M, then $M \cup \{w\}$ would be connected and properly contain M, contradicting the maximality of M. So M consists of exactly the vertices connected to v, proving that it is a connected component. ■

Problem 5.8.7. **(a)** Describe a connected graph such that every vertex is on a simple cycle, but the graph is not 2-connected.

Solution. Two cycles connected by an edge. ■

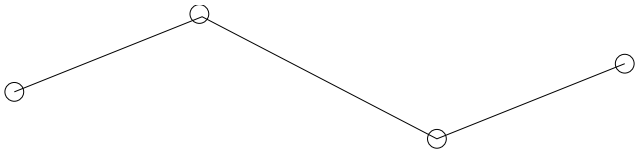
(b) Prove that a graph is 2-connected iff it is connected and every edge is traversed by a simple cycle.

Solution. To prove the iff from left to right, we assume some edge, v—w, is not traversed by any simple cycle and then show that the graph is not 2-connected.

Now if there was a simple path between v and w that did not traverse this edge, then this path together with the edge, would be a simple cycle that traversed the edge, contradicting our assumption. So every simple path from w to v must traverse this edge. This implies that if this edge is removed, no simple path will connect v and w, which implies that no path whatsoever connects them. This proves the graph is not 2-connected.

Conversely, suppose the graph is connected but not 2-connected. So there must be a "cut" edge, v—w, whose removal leaves a disconnected graph. Now if there was a simple cycle that traversed this edge, then the cycle minus this edge would be a simple path, P, between v and w that did not traverse the edge. So any path connecting two points that traversed this edge could be replaced by a path containing P in place of the edge. Therefore, all edges that were connected before the edge was removed would still be connected after it is removed. So since the original graph was connected, it remains connected after the edge is removed. This would imply the graph was 2-connected, a contradiction. So there cannot be a simple cycle that traverses a cut edge. ■

Problem 5.8.8. Let's say that a graph has "two ends" if it has exactly two vertices of degree 1 and all its other vertices have degree 2. For example, here is one such graph:



(a) A *line graph* is a graph whose edges can all be traversed by a simple path. So the two-ended graph above is also a line graph of length 4.

Prove that the following theorem is false by drawing a counterexample.

False Theorem. *Every two-ended graph is a line graph.*

Solution. A graph consisting of a path together with a simple cycle is a counterexample. ■

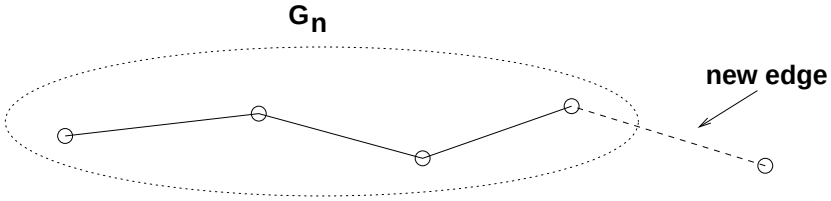
(b) Point out the first erroneous statement in the following alleged proof of the false theorem. Describe the error as best you can.

False proof. We use induction. The induction hypothesis is that every two-ended graph with n edges is a path.

Base case ($n = 1$): The only two-ended graph with a single edge consists of two vertices joined by an edge:

Sure enough, this is a line graph.

Inductive case: We assume that the induction hypothesis holds for some $n \geq 1$ and prove that it holds for $n + 1$. Let G_n be any two-ended graph with n edges. By the induction assumption, G_n is a line graph. Now suppose that we create a two-ended graph G_{n+1} by adding one more edge to G_n. This can be done in only one way: the new edge must join an endpoint of G_n to a new vertex; otherwise, G_{n+1} would not be two-ended.



Clearly, G_{n+1} is also a line graph. Therefore, the induction hypothesis holds for all graphs with $n + 1$ edges, which completes the proof by induction.

□

Solution. Actually, this is a correct proof of something else. That is, the first erroneous statement is the last one claiming that the induction hypothesis holds for *all* $(n + 1)$-edge two-ended graphs.

The proof doesn't show this; rather, it only shows that the induction hypothesis holds for those two-ended $(n + 1)$-edge graphs *that can be obtained by adding one more edge to an n-edge two-ended graph*. This is not all two-ended graphs, as the counterexample demonstrates.

This is an example of "buildup" error, where you assume that a size $n+1$ object is built up in some particular way from similar objects of smaller size. (This assumption is correct for some kinds of objects, but incorrect for others— such as the one in the argument above.)

One way to avoid an accidental buildup error is to use a "shrink down, grow back" process in the inductive step: start with a size $n+1$ object, say a graph, remove a vertex (or edge), apply the inductive hypothesis $P(n)$ to the smaller graph, and then add back the vertex (or edge) and argue that $P(n+1)$ holds. Let's see what would have happened if we'd tried to prove the claim above by this method:

Inductive step: We must show that $P(n)$ implies $P(n+1)$ for all $n \geq 1$. Consider an $(n+1)$-vertex graph G in which every vertex has degree at least 1. Remove an arbitrary vertex v, leaving an n-vertex graph G' in which every vertex has degree... uh-oh!

The reduced graph G' might contain a vertex of degree 0, making the inductive hypothesis $P(n)$ inapplicable! We are stuck— and properly so, since the claim is false! ■

5.9 Miniquiz Mar. 14

Problem 5.9.1. Circle all of the properties below that are invariants of the *Mating Ritual* for finding a stable matching (see the Appendix if you need to look up the Ritual). Assume that the numbers of boys and girls are the same.

a. If a girl is crossed off a boy's list, she has a suitor that she prefers to that boy.

b. If a girl is crossed off a boy's list, the girl and the boy will not be a rogue couple.

c. There is a girl with no suitor (boy who is serenading her).

d. If a girl is crossed off a boy's list, he prefers that girl to the girl he is serenading.

e. All the boys have the same number of girls left uncrossed in their list.

f. If a boy has only one girl left on his list, she will be his wife on the last day of the Ritual.

Solution. a. Invariant; this is the basic invariant used to verify the Ritual.

b. Invariant; if a girl is crossed off a boy's list, that means the girl has a better suitor. Therefore those two cannot be a rogue couple.

c. Not invariant; a girl won't have a suitor on the first day if she's not at the top of any boy's list, but every girl is guaranteed to have one at the end, namely, her husband.

d. Invariant; boys work down their lists in order of their preference.

e. Not invariant; it is true only at the start state.

f. Invariant; no one else is left who could be his wife. ■

Problem 5.9.2. Prove that in every graph, there are an even number of vertices of odd degree. You may assume the Handshaking Theorem (see Appendix). Your proof will be graded primarily on the clarity of your argument.

Solution. *Proof.* Partitioning the vertices into those of even degree and those of odd degree, we know

$$\sum_{v \in V} d(v) = \sum_{d(v) \text{ is even}} d(v) + \sum_{d(v) \text{ is odd}} d(v)$$

By the Handshaking Theorem, the value of the lefthand side of this equation equals twice the number of edges, and so is even. The first summand on the righthand side is even since it is a sum of even values. So the second summand on the righthand side must also be even. But since it is entirely a sum of odd values, it must must contain an even number of terms. That is, there must be an even number of vertices with odd degree. □

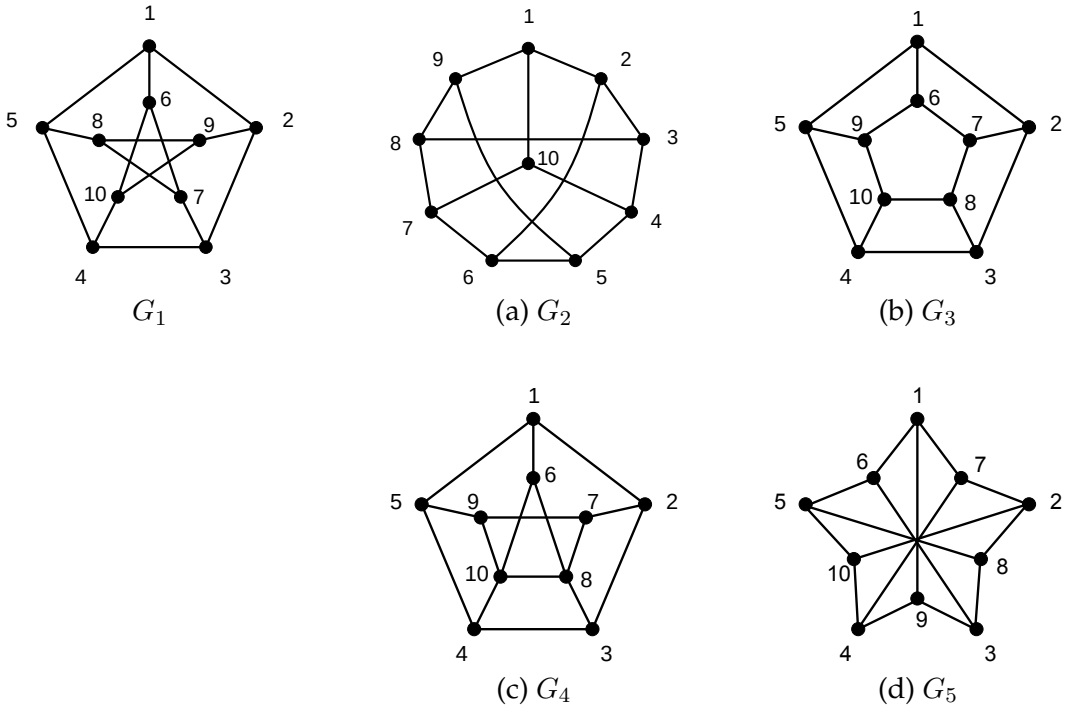
■

Problem 5.9.3. Circle all the properties below that are preserved under isomorphism.

- Some vertex is unlabeled.
- There is an edge incident to a degree 8 vertex and a degree 4 vertex.
- Two of the edges form a right angle.
- The negation of a property that is preserved under isomorphism.
- There are two simple cycles that share at least one vertex.
- There are two connected components.

Solution. "Some vertex is unlabeled" and "two edges form a right angle" are not preserved under isomorphism. ■

Problem 5.9.4. Circle all the graphs from G_2 to G_5 below that are isomorphic to G_1.



G_1 (a) G_2 (b) G_3 (c) G_4 (d) G_5

Figure 5.5: Which graphs are isomorphic to G_1?

Solution. G_1 and G_2 are isomorphic. In particular, the function $f : V_1 \to V_2$ is an isomomorphism, where

$f(1) = 1$	$f(2) = 2$	$f(3) = 3$	$f(4) = 8$	$f(5) = 9$
$f(6) = 10$	$f(7) = 4$	$f(8) = 5$	$f(9) = 6$	$f(10) = 7$

G_1 and G_4 are not isomorphic: G_4 has a vertex of degree four and G_1 does not have one.

G_3 and G_5 are not isomorphic to G_1: G_3 and G_5 have simple cycles of length four and G_1 does not.

■

Problem 5.9.5. One of the properties below is not a property of all trees. Circle that property.

- Any connected subgraph is a tree.
- Adding an edge between two vertices creates a cycle.
- The number of vertices is one less than twice the number of leaves.
- Removing any edge disconnects the graph.
- If it has at least two vertices, then it has at least two leaves.

Solution. "The number of all the vertices is one less than twice the number of leaf nodes" is not true. This property holds for full binary trees, but not in general. A tree with two vertices is a counterexample. ■

Appendix

Stable marriage

A *marriage assignment* or *perfect matching* is a bijection, w : Boys $\rightarrow$ Girls.

A *rogue couple* is a boy, B, and a girl, G, such that B prefers G to his wife, and G prefers B to her husband.

An assignment is *stable* if it has no rogue couples.

The Mating Ritual

The Mating Ritual takes place over several days. The following events happen each day:

Morning: Each girl stands on her balcony. Each boy stands under the balcony of his favorite among the girls on his list, and he serenades her. If a boy has no girls left on his list, he stays home and does his 6.042 homework.

Afternoon: Each girl who has one or more suitors serenading her, says to her favorite suitor, "We might get engaged. Come back tomorrow." To the others, she says, "No. I will never marry you! Take a hike!"

Evening: Any boy who is told by a girl to take a hike, crosses that girl off his list.

Termination condition: When every girl has at most one suitor, the ritual ends with each girl marrying her suitor, if she has one.

Graph theory

A *simple graph*, G, consists of a nonempty set, V, called the *vertices* of G, and a collection, E, of two-element subsets of V. The members of E are called the *edges* of G.

Two vertices in a simple graph are said to be *adjacent* if they are joined by an edge, and an edge is said to be *incident* to the vertices it joins. The number of edges incident to a vertex is called the *degree* of the vertex.

Theorem. *(Handshaking) The sum of the degrees of the vertices in a graph equals twice the number of edges.*

A *path* in a graph, G, is a sequence of $k \geq 0$ vertices, $v_0, \ldots, v_k$, such that v_i—v_{i+1} is an edge of G for all i where $0 \leq i < k$. The path is said to *start* at v_0, to *end* at v_k, and *length* of the path is defined to be k. An edge, e, is *traversed n times* by the path if there are n different values of i such that edge v_i—v_{i+1} is e.

The path is *simple* iff all the v_i's are different, that is, $v_i = v_j$ only if $i = j$.

A *cycle* is a path that begins and ends with the same vertex. A *simple cycle* is a positive length cycle without repeated vertices except for the beginning and end vertices.

Two vertices in a graph are *connected* iff there is a path that begins with one of the vertices and ends with the other. The *shortest* path between two vertices is always simple.

A graph is *connected* iff every pair of vertices is connected.

A graph with no simple cycles is called *acyclic*. *Trees* are acyclic connected graphs.

A *subgraph*, G', of a graph, G, is a graph whose vertices, V', are a subset of the vertices of G and whose edges are a subset of the edges of G.

A vertex of degree one is called a *leaf*.

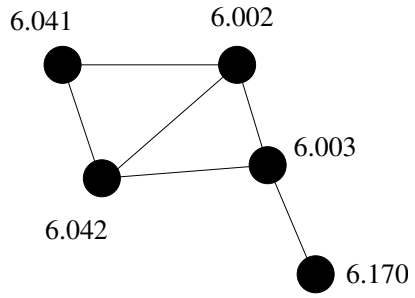
Chapter 6

Graphs and Digraphs

6.1 Coloring Graphs

In the "Sex in America" graph In Week 5 Notes, we used edges to indicate an affinity between two nodes, but having an edge represent a *conflict* between two nodes also turns out to be really useful. For example, each term the MIT Schedules Office must assign a time slot for each final exam. This is not easy, because some students are taking several classes with finals, and a student can take only one test during a particular time slot. The Schedules Office wants to avoid all conflicts. Of course, you can make such a schedule by having every exam in a different slot, but then you would need hundreds of slots for the hundreds of courses, and exam period would run all year! So, the Schedules Office would also like to keep exam period short.

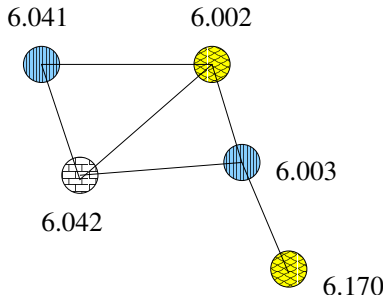
The Schedules Office's problem is easy to describe as a graph. There will be a vertex for each course with a final exam, and two vertices will be adjacent exactly when some student is taking both courses. For example, suppose we need to schedule exams for 6.041, 6.042, 6.002, 6.003 and 6.170. The scheduling graph might look like this:



6.002 and 6.042 cannot have an exam at the same time since there are students in both courses, so there is an edge between their nodes. On the other hand, 6.042 and 6.170 can have an exam at the same time if they're taught at the same time (which they sometimes are), since no student can be enrolled in both (that is, no student *should* be enrolled in both when they have a timing conflict). Next, identify each time slot with a color. For example, Monday morning is red, Monday afternoon is blue, Tuesday morning is green, etc.

Assigning an exam to a time slot is now equivalent to coloring the corresponding vertex. The main constraint is that *adjacent vertices must get different colors* —otherwise, some student has two exams

at the same time. Furthermore, in order to keep the exam period short, we should try to color all the vertices using as *few different colors as possible*. For our example graph, three colors suffice:



This coloring corresponds to giving one final on Monday morning (white), two Monday afternoon (blue), and two Tuesday morning (green).

Can we use fewer than three colors? No! We can't use only two colors since there is a triangle in the graph, and three vertices in a triangle must all have different colors.

This is an example of what is a called a *graph coloring problem*: given a graph G, assign colors to each node such that adjacent nodes have different colors. A color assignment with this property is called a *valid coloring* of the graph —a "coloring," for short. A graph G is ***k*-colorable** if it has a coloring that uses at most k colors. The minimum value of k for which a coloring exists is called the ***chromatic number***, $\chi(G)$, of G.

In general, trying to figure out if you can color a graph with a fixed number of colors can take a long time. It's a classic example of a problem for which no fast algorithms are known. In fact, it is easy to check if a coloring works, but it seems really hard to find it (if you figure out how, then you can get a $1 million Clay prize).

6.1.1 Degree-bounded Coloring

There are some simple graph properties that give useful upper bounds on colorings. For example, if we have a bound on the degrees of all the vertices in a graph, then we can easily find a coloring with only one more color than the degree bound.

Theorem 6.1.1. *A graph with maximum degree at most k is $(k+1)$-colorable.*

Unfortunately, if you try induction on k, it will lead to disaster. It is not that it is impossible, just that it is extremely painful and would ruin you if you tried it on an exam. Another option, especially with graphs, is to change what you are inducting on. In graphs, some good choices are n, the number of nodes, or e, the number of edges.

Proof. We use induction on the number of vertices in the graph, which we denote by n. Let $P(n)$ be the proposition that an n-vertex graph with maximum degree at most k is $(k+1)$-colorable.

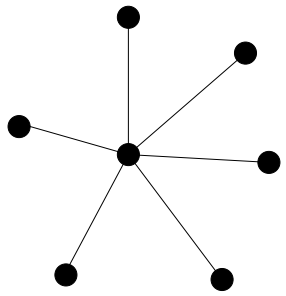
Base case: $(n = 1)$ A 1-vertex graph has maximum degree 0 and is 1-colorable, so $P(1)$ is true.

Inductive step: Now assume that $P(n)$ is true, and let G be an $(n+1)$-vertex graph with maximum degree at most k. Remove a vertex v (and all edges incident to it), leaving an n-vertex subgraph,

H. The maximum degree of H is at most k, and so H is $(k+1)$-colorable by our assumption $P(n)$. Now add back vertex v. We can assign v a color different from all its adjacent vertices, since there are at most k adjacent vertices and $k+1$ colors are available. Therefore, G is $(k+1)$-colorable. This completes the Inductive step, and the theorem follows by induction. □

Sometimes $k+1$ colors is the best you can do. For example, in the complete graph, K_n, every one of its n vertices is adjacent to all the others, so all n must be assigned different colors. Of course n colors is also enough, so $\chi(K_n) = n$. So K_{k+1} is an example where Theorem 6.1.1 gives the best possible bound. This means that Theorem 6.1.1 also gives the best possible bound for *any* graph with degree bounded by k that has K_{k+1} as a subgraph.

But sometimes $k+1$ colors is far from the best that you can do. Here's an example of an n-node star graph for $n = 7$:



In the n-node star graph, the maximum degree is $n-1$, but the star only needs 2 colors!

6.1.2 Why coloring?

Coloring problems come up in all sorts of applications. For example, at Akamai, a new version of software is deployed over each of 20,000 servers every few days. The updates cannot be done at the same time since the servers need to be taken down in order to deploy the software. Also, the servers cannot be handled one at a time, since it would take forever to update them all (each one takes about an hour). Moreover, certain pairs of servers cannot be taken down at the same time since they have common critical functions. This problem was eventually solved by making a 20,000 node conflict graph and coloring it with 8 colors – so only 8 waves of install are needed!

Another example comes from the need to assign frequencies to radio stations. If two stations have an overlap in their broadcast area, they can't be given the same frequency. Frequencies are precious and expensive, so you want to minimize the number handed out. This amounts to finding the minimum coloring for a graph whose vertices are the stations and whose edges are between stations with overlapping areas.

Coloring also comes up allocating registers for program variables. While a variable is in use, its value needs to be saved in a register, but registers can often be reused for different variables. But two variables need different registers if they are referenced during overlapping intervals of program execution. So register allocation is the coloring problem for a graph whose vertices are the variables; vertices are adjacent if their intervals overlap, and the colors are registers.

Finally, there's the famous map coloring problem mentioned in Week 1 Notes. The question is how many colors are needed to color a map so that adjacent territories get different colors? This is

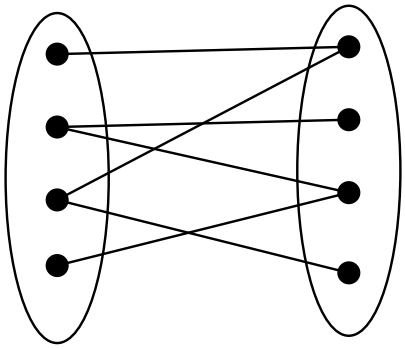
the same as the number of colors needed to color a graph that can be drawn in the plane without edges crossing. A proof that four colors are enough for the *planar* graphs was acclaimed when it was discovered about thirty years ago. Implicit in that proof was a 4-coloring procedure that takes time proportional to the number of vertices in the graph (countries in the map). On the other hand, it's another of those million dollar prize questions to find an efficient procedure to tell if a planar graph really *needs* four colors or if three will actually do the job. But it's always easy to tell if an *arbitrary* graph is 2-colorable, as we show next. Later we'll develop enough planar graph theory to show that planar graphs are 6-colorable.

6.1.3 Bipartite Graphs

There were two kinds of vertices in the "Sex in America" graph —males and females, and edges only went between the two kinds. Graphs like this come up so frequently they have earned a special name —they are called ***bipartite graphs***.

Definition 6.1.2. A *bipartite graph* is a graph together with a partition of its vertices into two sets, L and R, such that every edge is incident to a vertex in L and to a vertex in R.

So every bipartite graph looks something like this:



Now we can immediately see how to color a bipartite graph using only two colors: let all the L vertices be black and all the R vertices be white. Conversely, if a graph is 2-colorable, then it is bipartite with L being the vertices of one color and R the vertices of the other color. In other words,

> "bipartite" is a synonym for "2-colorable."

The following Lemma gives another useful characterization of bipartite graphs.

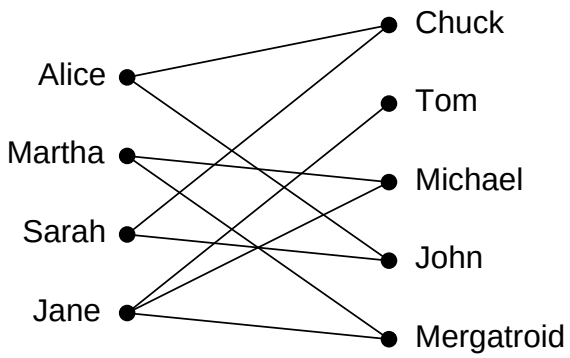
Theorem 6.1.3. *A graph is bipartite iff it has no odd-length cycle.*

The proof of Theorem 6.1.3 will be on a problem set.

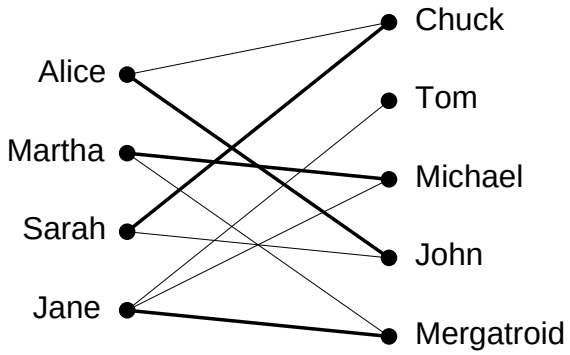
6.2 Bipartite Matchings

The *bipartite matching problem* resembles the stable Marriage Problem in that it concerns a set of girls and a set of at least as many boys. There are no preference lists, but each girl does have some boys she likes and others she does not like. In the bipartite matching problem, we ask whether every girl can be paired up with a boy that she likes.

Any particular matching problem can be specified by a bipartite graph with a vertex for each girl, a vertex for each boy, and an edge between a boy and a girl iff the girl likes the boy. For example, we might obtain the following graph:



Now a *matching* will mean a way of assigning every girl to a boy so that different girls are assigned to different boys, and a girl is always assigned to a boy she likes. For example, here is one possible matching for the girls:



Hall's Matching Theorem states necessary and sufficient conditions for the existence of a matching in a bipartite graph. It turns out to be a remarkably useful mathematical tool.

6.2.1 The Matching Condition

We'll state and prove Hall's Theorem using girl-likes-boy terminology. Define *the set of boys liked by a given set of girls* to consist of all boys liked by at least one of those girls. For example, the set of boys liked by Martha and Jane consists of Tom, Michael, and Mergatroid.

For us to have any chance at all of matching up the girls, the following ***matching condition*** must hold:

Every subset of girls likes at least as large a set of boys.

For example, we can not find a matching if some 4 girls like only 3 boys. Hall's Theorem says that this necessary condition is actually sufficient; if the matching condition holds, then a matching exists.

Theorem 6.2.1. *A matching for a set of girls G with a set of boys B can be found if and only if the matching condition holds.*

Proof. First, let's suppose that a matching exists and show that the matching condition holds. Consider an arbitrary subset of girls. Each girl likes at least the boy she is matched with. Therefore, every subset of girls likes at least as large a set of boys. Thus, the matching condition holds.

Next, let's suppose that the matching condition holds and show that a matching exists. We use strong induction on $|G|$, the number of girls. If $|G| = 1$, then the matching condition implies that the lone girl likes at least one boy, and so a matching exists. Now suppose that $|G| \geq 2$. There are two possibilities:

1. Every proper subset of girls likes a *strictly larger* set of boys. In this case, we have some latitude: we pair an arbitrary girl with a boy she likes and send them both away. The matching condition still holds for the remaining boys and girls, so we can match the rest of the girls by induction.

2. Some proper subset of girls $X \subset G$ likes an *equal-size* set of boys $Y \subset B$. We match the girls in X with the boys in Y by induction and send them all away. We will show that the matching condition holds for the remaining boys and girls, and so we can match the rest of the girls by induction as well.

   To that end, consider an arbitrary subset of the remaining girls $X' \subseteq G - X$, and let Y' be the set of remaining boys that they like. We must show that $|X'| \leq |Y'|$. Originally, the combined set of girls $X \cup X'$ liked the set of boys $Y \cup Y'$. So, by the matching condition, we know:

   $$\left|X \cup X'\right| \leq \left|Y \cup Y'\right|$$

   We sent away $|X|$ girls from the set on the left (leaving X') and sent away an equal number of boys from the set on the right (leaving Y'). Therefore, it must be that $|X'| \leq |Y'|$ as claimed.

In both cases, there is a matching for the girls. The theorem follows by induction. □

The proof of this theorem gives an algorithm for finding a matching in a bipartite graph, albeit not a very efficient one. However, efficient algorithms for finding a matching in a bipartite graph do exist. Thus, if a problem can be reduced to finding a matching, the problem is essentially solved from a computational perspective.

6.2.2 A Formal Statement

Let's restate Hall's Theorem in abstract terms so that you'll not always be condemned to saying, "Now this group of little girls likes at least as many little boys..."

In any graph, the set $N(S)$, of *neighbors*[1] of some set, S, of vertices is the set of all vertices adjacent to any vertex in S. That is,

$$N(S) ::= \{r \mid s\text{—}r \text{ is an edge for some } s \in S\}.$$

S is called a *bottleneck* if

$$|S| > |N(S)|.$$

A *matching* in a graph is an injection on the set of vertices that only maps a vertex to an adjacent vertex.

Theorem 6.2.2 (Hall's Theorem). *Let G be a bipartite graph with vertex partition L, R. There is total matching from L to R iff no subset of L is a bottleneck.*

An Easy Matching Condition

The bipartite matching condition requires that *every* subset of girls has a certain property. In general, verifying that every subset has some property, even if it's easy to check any particular subset for the property, quickly becomes overwhelming because the number of subsets of even relatively small sets is enormous —over a billion subsets for a set of size 30.

However, there is a simple property of vertex degrees in a bipartite graph that guarantees a match and is very easy to check. Namely, call a bipartite graph *degree-constrained* if vertex degrees on the left are at least as large as those on the right. More precisely,

Definition 6.2.3. A bipartite graph G with vertex partition L, R is *degree-constrained* if $\deg(l) \geq \deg(r)$ for every $l \in L$ and $r \in R$.

Now we can always find a matching in a degree-constrained bipartite graph.

Lemma 6.2.4. *Every degree-constrained bipartite graph satisifies the matching condition.*

Proof. Let S be any set of vertices in L. The number of edges incident to vertices in S is exactly the sum of the degrees of the vertices in S. Each of these edges is incident to a vertex in $N(S)$ by definition of $N(S)$. So the sum of the degrees of the vertices in $N(S)$ is at least as large as the sum for S. But since the degree of every vertex in $N(S)$ is at most as large as the degree of every vertex in S, there would have to be at least as many terms in the sum for $N(S)$ as in the sum for S. So there have to be at least as many vertices in $N(S)$ as in S, proving that S is not a bottleneck. So there are no bottlenecks, proving that the degree-constrained graph satisifies the matching condition. □

[1] An equivalent definition of $N(S)$ uses relational notation: $N(S)$ is simply SJ, where J is the adjacency relation of the graph.

6.3 Digraphs

A *directed graph* (*digraph* for short) is formally the same as a binary relation, R, on a set, A, but we picture the digraph geometrically by representing elements of A as points on the plane, with an arrow from the point for a to the point for b exactly when $(a, b) \in \text{graph}(R)$. The elements of A are referred to as the *vertices* of the digraph, and the pairs $(a, b) \in \text{graph}(R)$ are called its *directed edges*. We use the notation $a \to b$ as an alternative notation for the pair (a, b).

For example, the divisibility relation on $\{1, 2, \ldots, 12\}$ is represented by the digraph:

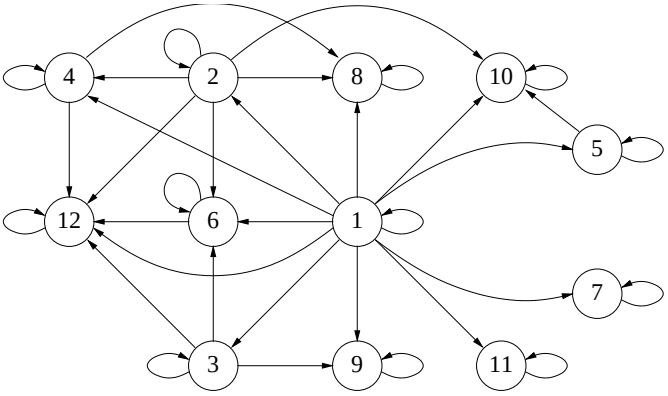


Figure 6.1: Divisibility Digraph on $\{1, 2, \ldots, 12\}$.

6.3.1 Paths in Digraphs

Pictured with points and arrows, a length k path in a digraph looks like a line that starts at a point, a_0, and traverses k arrows between successive points, $a_1, a_2, \ldots$ to end at a point, a_k. Note that k may be 0 —a single vertex counts as length zero path to itself, just as for simple graphs. The precise definitions are very similar to those for simple graphs:

Definition 6.3.1. A *path in a digraph* is a sequence of vertices $a_0, \ldots, a_k$ with $k \geq 0$ such that $a_i \to a_{i+1}$ is an edge for every $i \geq 0$ such that $i < k$. The path is said to *start* at a_0, to *end* at a_k, and the *length* of the path is defined to be k. The path is *simple* iff all the a_i's are different, that is, $a_i = a_j$ only if $i = j$.

Many of the relational properties have geometric descriptions in terms of digraphs. For example:

Reflexivity: All vertices have self-loops (a *self-loop* at a vertex is an arrow going from the vertex back to itself).

Irreflexivity: No vertices have self-loops.

Asymmetry: No self-loops and at most one (directed) edge between any two vertices.

Symmetry: A binary relation R is ***symmetric*** iff aRb implies bRa for all a, b in the domain of R. So if there is an edge from a to b, there is also one in the reverse direction. So edges may as well be represented without arrows, indicating that they can be followed in either direction.

Transitivity: Short-circuits—for any path through the graph, there is an arrow from the first vertex to the last vertex on the path.

We can define some new relations based on paths. Let R be the edge relation of a digraph. Define relations R^* and R^+ on the vertices by the conditions that for all vertices a, b:

$$\begin{aligned} a\, R^*\, b &::= \text{there is a path in } R \text{ from } a \text{ to } b, \\ a\, R^+\, b &::= \text{there is a positive length path in } R \text{ from } a \text{ to } b. \end{aligned}$$

R^* is called the *path relation*[2] of R. It follows from the definition of path that R^* is transitive. It is also reflexive (because of the length-zero paths) and it contains the graph of R (because of the length-one paths). R^+ is called the *positive-length path relation*; it also contains graph (R) and is transitive.

6.3.2 Directed Acyclic Graphs

Definition 6.3.2. A *cycle* in a digraph is a path that begins and ends at the same vertex. Note that by convention, a single vertex is considered to be a cycle of length 0 that begins and ends at the vertex. A *directed acyclic graph (DAG)* is a directed graph with no positive length cycles.

A *simple cycle* in a digraph is a cycle whose vertices are distinct except for the beginning and end vertices.

DAG's are an economical way to represent partial orders. For example, the direct prerequisite relation between MIT subjects described in Week 3 Notes was used to determine the partial order of indirect prerequisites on subjects. This indirect prerequisite partial order is precisely the positive length path relation of the direct prerequisites.

Lemma 6.3.3. *If D is a DAG, then D^+ is a strict partial order.*

Proof. We know that D^+ is transitive. Also, a positive length path from a vertex to itself would be a cycle, so there are no such paths. This means D^+ is irreflexive, which implies it is a strict partial order (see Week 3, Tuesday, Class Problem 4). □

It's easy to check that conversely, the graph of any strict partial order is a DAG.

Problem 6.3.1. Verify that any strict partial order is a DAG.

The divisibility partial order can also be more economically represented by the path relation in a DAG. The DAG for divisibility on $\{1, 2, \ldots, 12\}$ is shown in Figure 6.2; the arrowheads are omitted in the Figure, and edges are understood to point upwards.

The minimum edge DAG representing a finite partial order is unique, and is easy to find. This is not hard to verify:

Problem 6.3.2. If a and b are distinct nodes of a digraph, then a is said to *cover* b if there is an edge from a to b and there is no other path from a to b. If a covers b, the edge from a to b is called a covering edge.

[2] In many texts, R^* is called the *transitive closure* of R.

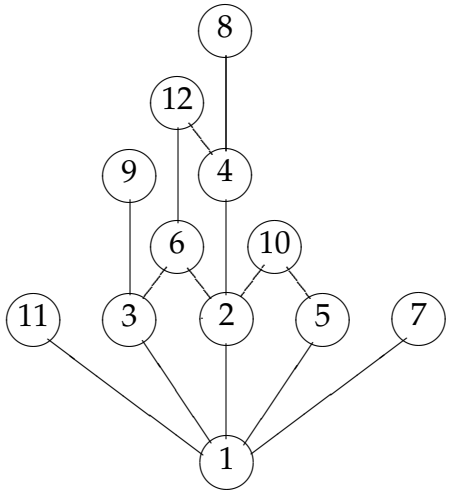


Figure 6.2: DAG for Divisibility on $\{1, 2, \ldots, 12\}$.

(a) Show that if two DAG's have the same positive path relation, then they have the same set of covering edges.

(b) For any DAG, D, let $\widehat{D}$ be the subgraph of D consisting of only the covering edges. Show that if D is finite and has no self-loops, then D and $\widehat{D}$ have the same positive path relation, that is $D^+ = \widehat{D}^+$.

(c) Conclude that if D is a finite DAG, then $\widehat{D}$ is the unique DAG with the smallest number of edges among all digraphs with the same positive path relation.

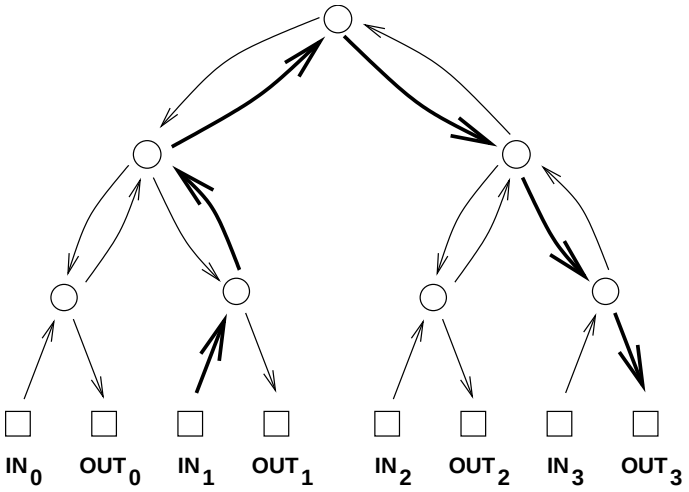
(d) Show that the previous result is not true in general for the infinite DAG corresponding to the total order on the rational numbers.

6.4 Communication Networks

Modeling communication networks is an important application of graphs in Computer Science. Here, vertices represent computers, processors, and switches; edges will represent wires, fiber, or other transmission lines through which data flows. For some communication networks, like the internet, the corresponding graph is enormous and largely chaotic. However, there do exist more organized networks, such as certain telephone switching networks and the communication networks inside parallel computers. For these, the corresponding graphs are highly structured. In these Notes, we'll look at some of the nicest and most commonly used communication networks.

6.4.1 Complete Binary Tree

Let's start with a ***complete binary tree***. Here is an example with 4 inputs and 4 outputs.



The kinds of communication networks we consider aim to transmit packets of data between computers, processors, telephones, or other devices. The term ***packet*** refers to some roughly fixed-size quantity of data— 256 bytes or 4096 bytes or whatever. In this diagram and many that follow, the squares represent ***terminals***, sources and destinations for packets of data. The circles represent ***switches***, which direct packets through the network. A switch receives packets on incoming edges and relays them forward along the outgoing edges. Thus, you can imagine a data packet hopping through the network from an input terminal, through a sequence of switches joined by directed edges, to an output terminal.

Recall that there is a unique simple path between every pair of vertices in a tree. So the natural way to route a packet of data from an input terminal to an output in the complete binary tree is along the corresponding directed path. For example, the route of a packet traveling from input 1 to output 3 is shown in bold.

6.4.2 Latency and Diameter

Latency is a critical issue in communication networks. This is the largest delay between the time a packet is sent from an input until it arrives at its designated output. Assuming it takes one time unit to travel across a wire with no delays at switches, the delay of a packet is the number of wires it crosses going from input to output, that is, the packet delay is the *length* of the path the packet follows.

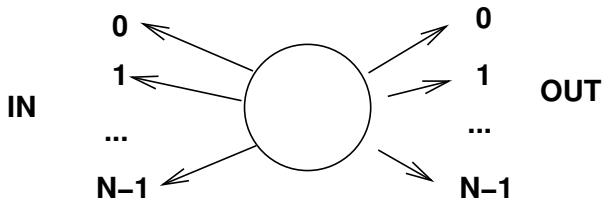
The latency of a network will depend on how packets are routed, but generally packets are routed to go from input to output by the shortest path possible. With a shortest path routing, the worst case delay is the distance between the input and output that are farthest apart. This is called the ***diameter*** of the network. In other words, the diameter of a network[3] is the maximum length of any shortest path between an input and an output. For example, in the complete binary tree above, the distance from input 1 to output 3 is six. No input and output are farther apart than this, so the diameter of this tree is also six.

[3]The usual definition of *diameter* for a general *graph* (simple or directed) is the largest distance between *any* two vertices, but in the context of a communication network we're only interested in the distance between inputs and outputs, not between arbitrary pairs of vertices.

We're going to consider several different communication networks. For a fair comparison, let's assume that each network has N inputs and N outputs, where N is a power of two. For example, the diameter of a complete binary tree with N inputs and outputs is $2 \log N + 2$. (All logarithms in this lecture— and in most of Computer Science —are base 2.) This is quite good, because the logarithm function grows very slowly. We could connect up $2^{10} = 1024$ inputs and outputs using a complete binary tree and still have a latency of only $2 \log(2^{10}) + 2 = 22$.

6.4.3 Switch Size

One way to reduce the diameter of a network is to use larger switches. For example, in the complete binary tree, most of the switches have three incoming edges and three outgoing edges, which makes them 3×3 switches. If we had 4×4 switches, then we could construct a complete *ternary* tree with an even smaller diameter. In principle, we could even connect up all the inputs and outputs via a single monster switch:



This isn't very productive, however, since we've just concealed the original network design problem inside this abstract switch. Eventually, we'll have to design the internals of the monster switch using simpler components, and then we're right back where we started. So the challenge in designing a communication network is figuring out how to get the functionality of an $N \times N$ switch using elementary devices, like 3×3 switches. Following this approach, we can build arbitrarily large networks just by adding in more building blocks.

6.4.4 Switch Count

Another goal in designing a communication network is to use as few switches as possible since routing hardware has a cost. The number of switches in a complete binary tree is $1 + 2 + 4 + 8 + \ldots + N$, since there is 1 switch at the top (the "root switch"), 2 below it, 4 below those, and so forth. By the formula for the sum of a geometric series (see Slides 3W, Feb 21, 2007) the total number of switches is $2N - 1$, which is nearly the best possible with 3×3 switches.

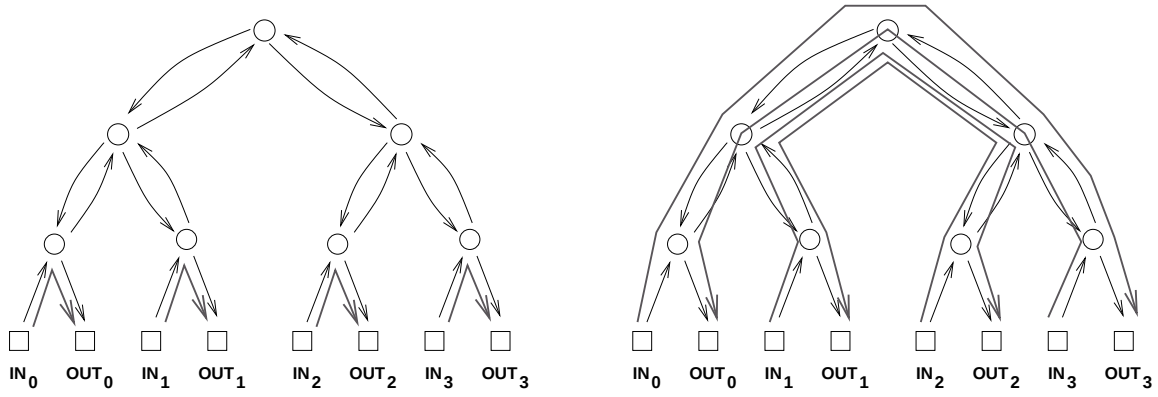
6.4.5 Congestion

The complete binary tree has a fatal drawback: the root switch is a bottleneck. At best, this switch must handle an enormous amount of traffic: every packet traveling from the left side of the network to the right or vice-versa. Passing all these packets through a single switch could take a long time. At worst, if this switch fails, the network is broken into two equal-sized pieces. The *max congestion* of a network is a measure of its bottlenecks; defining max congestion requires some preliminary definitions.

A ***permutation*** is a function π that maps each number in the set $\{0, 1, \ldots, N-1\}$ to another number in the set such that no two numbers are mapped to the same value. In other words, π is a bijection from $\{0, 1, \ldots, N-1\}$ to itself.

For each permutation π, there is a corresponding ***permutation routing problem***. In this problem, one packet starts out at each input; in particular, the packet starting at input i is called packet i. The challenge is to direct each packet i through the network from input i to output $\pi(i)$.

A solution to a permutation routing problem is a specification of the path taken by each of the N packets. In particular, the path taken by packet i from input i to output $\pi(i)$ is denoted $P_{i,\pi(i)}$. For example, if $\pi(i) = i$, then there is an easy solution: let $P_{i,\pi(i)}$ be the path from input i up through one switch and back down to output i. On the other hand, if $\pi(i) = (N-1) - i$, then each path $P_{i,\pi(i)}$ must begin at input i, loop all the way up through the root switch, and then travel back down to output $(N-1) - i$. These two situations are illustrated below.



We can distinguish between a "good" set of paths and a "bad" set based on congestion. The ***congestion*** of a set of paths $P_{0,\pi(0)}, \ldots, P_{N-1,\pi(N-1)}$ is equal to the largest number of paths that pass through a single switch. For example, the congestion of the set of paths in the diagram at left is 1, since at most 1 path passes through each switch. However, the congestion of the paths on the right is 4, since 4 paths pass through the root switch (and the two switches directly below the root). Generally, lower congestion is better since packets can be delayed at an overloaded switch.

By extending the notion of congestion, we can also distinguish between "good" and "bad" networks with respect to bottleneck problems. The ***max congestion*** of a network is the *maximum* over all permutations π of the *minimum* over all paths $P_{i,\pi(i)}$ of the congestion of the paths.

You may find it helpful to think about max congestion in terms of a value game. You design your spiffy, new communication network; this defines the game. Your opponent makes the first move in the game: she inspects your network and specifies a permutation routing problem that will strain your network.You move second: given her specification, you choose the precise paths that the packets should take through your network; you're trying to avoid overloading any one switch. Then her next move is to pick a switch with as large as possible a number of packets passing through it; this number is her score in the competition. The max congestion of your network is the largest score she can ensure; in other words, it is precisely the max-value of this game.

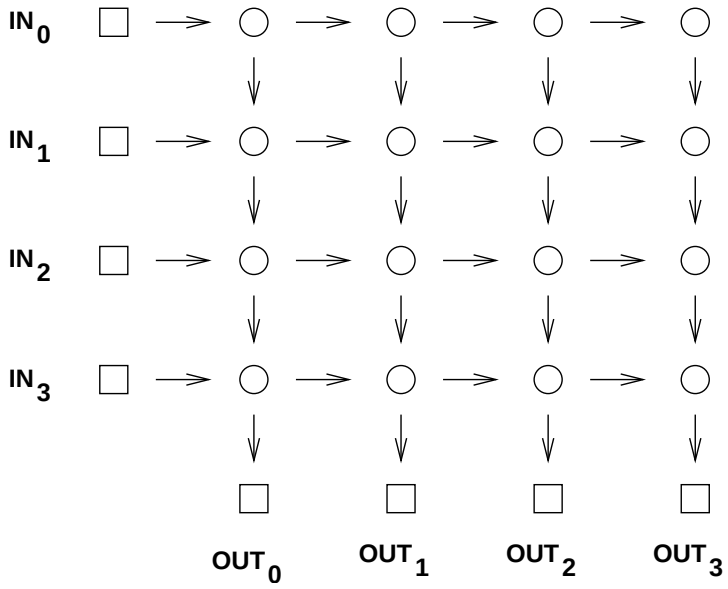
For example, if your enemy were trying to defeat the complete binary tree, she would choose a permutation like $\pi(i) = (N-1) - i$. Then for *every* packet i, you would be forced to select a path $P_{i,\pi(i)}$ passing through the root switch. Thus, the max congestion of the complete binary tree is N— which is horrible!

Let's tally the results of our analysis so far:

network	diameter	switch size	# switches	congestion
complete binary tree	$2\log N + 2$	3×3	$2N - 1$	N

6.4.6 2-D Array

Let's look at an another communication network. This one is called a ***2-dimensional array*** or ***grid***.



Here there are four inputs and four outputs, so $N = 4$.

The diameter in this example is 8, which is the number of edges between input 0 and output 3. More generally, the diameter of an array with N inputs and outputs is $2N$, which is much worse than the diameter of $2\log N + 2$ in the complete binary tree. On the other hand, replacing a complete binary tree with an array almost eliminates congestion.

Theorem 6.4.1. *The congestion of an N-input array is 2.*

Proof. First, we show that the congestion is at most 2. Let π be any permutation. Define $P_{i,\pi(i)}$ to be the path extending from input i rightward to column $\pi(i)$ and then downward to output $\pi(i)$. Thus, the switch in row i and column $\pi(i)$ transmits at most two packets: the packet originating at input i and the packet destined for column $\pi(i)$.

Next, we show that the congestion is at least 2. In any permutation routing problem where $\pi(0) = 0$ and $\pi(N-1) = N-1$, two packets must pass through the lower left switch. □

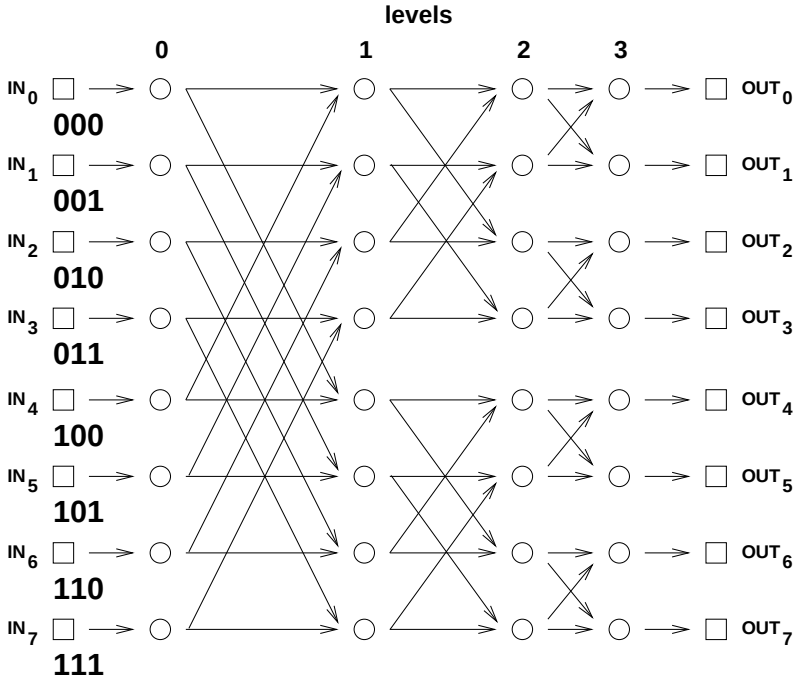
Now we can record the characteristics of the 2-D array.

network	diameter	switch size	# switches	congestion
complete binary tree	$2\log N + 2$	3×3	$2N - 1$	N
2-D array	$2N$	2×2	N^2	2

The crucial entry here is the number of switches, which is N^2. This is a major defect of the 2-D array; a network of size $N = 1000$ would require a *million* 2×2 switches! Still, for applications where N is small, the simplicity and low congestion of the array make it an attractive choice.

6.4.7 Butterfly

The Holy Grail of switching networks would combine the best properties of the complete binary tree (low diameter, few switches) and of the array (low congestion). The ***butterfly*** is a widely-used compromise between the two. Here is a butterfly network with $N = 8$ inputs and outputs.



The structure of the butterfly is certainly more complicated than that of the complete binary tree or 2-D array! Let's work through the various parts of the butterfly.

All the terminals and switches in the network are arranged in N rows. In particular, input i is at the left end of row i, and output i is at the right end of row i. Now let's label the rows in *binary*; thus, the label on row i is the binary number $b_1 b_2 \ldots b_{\log N}$ that represents the integer i.

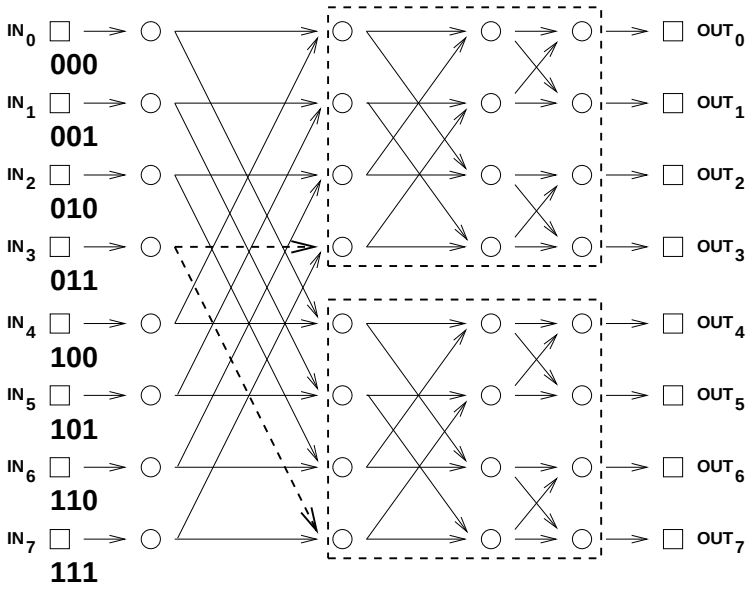
Between the inputs and the outputs, there are $\log(N) + 1$ levels of switches, numbered from 0 to $\log N$. Each level consists of a column of N switches, one per row. Thus, each switch in the network is uniquely identified by a sequence $(b_1, b_2, \ldots, b_{\log N}, l)$, where $b_1 b_2 \ldots b_{\log N}$ is the switch's row in binary and l is the switch's level.

All that remains is to describe how the switches are connected up. The basic connection pattern is expressed below in a compact notation:

$$(b_1, b_2, \ldots, b_{l+1}, \ldots, b_{\log N}, l) \nearrow (b_1, b_2, \ldots, b_{l+1}, \ldots, b_{\log N}, l+1)$$
$$(b_1, b_2, \ldots, b_{l+1}, \ldots, b_{\log N}, l) \searrow (b_1, b_2, \ldots, \overline{b_{l+1}}, \ldots, b_{\log N}, l+1)$$

This says that there are directed edges from switch $(b_1, b_2, \ldots, b_{\log N}, l)$ to two switches in the next level. One edge leads to the switch in the *same* row, and the other edge leads to the switch in the row obtained by *inverting* bit $l+1$. For example, referring back to the illustration of the size $N = 8$ butterfly, there is an edge from switch $(0, 0, 0, 0)$ to switch $(0, 0, 0, 1)$, which is in the same row, and to switch $(1, 0, 0, 1)$, which is the row obtained by inverting bit $l + 1 = 1$.

The butterfly network has a recursive structure; specifically, a butterfly of size $2N$ consists of two butterflies of size N, which are shown in dashed boxes below, and one additional level of switches. Each switch in the new level has directed edges to a pair of corresponding switches in the smaller butterflies; one example is dashed in the figure.



Despite the relatively complicated structure of the butterfly, there is a simple way to route packets. In particular, suppose that we want to send a packet from input $x_1 x_2 \ldots x_{\log N}$ to output $y_1 y_2 \ldots y_{\log N}$. (Here we are specifying the input and output numbers in binary.) Roughly, the plan is to "correct" the first bit by level 1, correct the second bit by level 2, and so forth. Thus, the sequence of switches visited by the packet is:

$$\begin{aligned}(x_1, x_2, x_3, \ldots, x_{\log N}, 0) &\to (y_1, x_2, x_3, \ldots, x_{\log N}, 1)\\ &\to (y_1, y_2, x_3, \ldots, x_{\log N}, 2)\\ &\to (y_1, y_2, y_3, \ldots, x_{\log N}, 3)\\ &\to \quad \ldots\\ &\to (y_1, y_2, y_3, \ldots, y_{\log N}, \log N)\end{aligned}$$

In fact, this is the *only* path from the input to the output!

The congestion of the butterfly network turns out to be around $\sqrt{N}$; more precisely, the congestion is $\sqrt{N}$ if N is an even power of 2 and $\sqrt{N/2}$ if N is an odd power of 2. (You'll prove this fact for homework.)

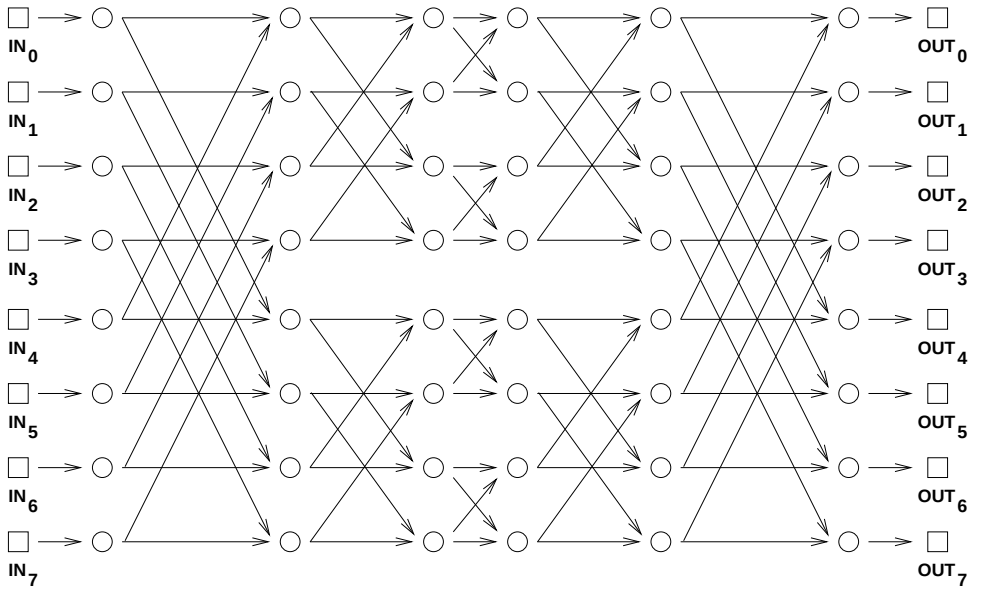
Let's add the butterfly data to our comparison table:

network	diameter	switch size	# switches	congestion
complete binary tree	$2\log N + 2$	3×3	$2N - 1$	N
2-D array	$2N$	2×2	N^2	2
butterfly	$\log N + 2$	2×2	$N(\log(N) + 1)$	$\sqrt{N}$ or $\sqrt{N/2}$

The butterfly has lower congestion than the complete binary tree. And it uses fewer switches and has lower diameter than the array. However, the butterfly does not capture the best qualities of each network, but rather is a compromise somewhere between the two. So our quest for the Holy Grail of routing networks goes on.

6.4.8 Beneš Network

In the 1960's, a researcher at Bell Labs named Beneš had a remarkable idea. He noticed that by placing *two* butterflies back-to-back, he obtained a marvelous communication network:



This doubles the number of switches and the diameter, of course, but completely eliminates congestion problems! The proof of this fact relies on a clever induction argument that we'll come to in a moment. Let's first see how the Beneš network stacks up:

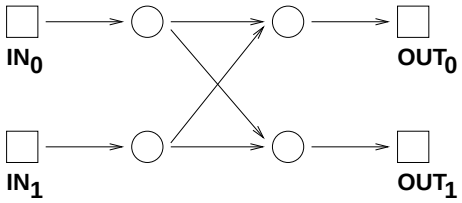
network	diameter	switch size	# switches	congestion
complete binary tree	$2\log N + 2$	3×3	$2N - 1$	N
2-D array	$2N$	2×2	N^2	2
butterfly	$\log N + 2$	2×2	$N(\log(N) + 1)$	$\sqrt{N}$ or $\sqrt{N/2}$
Beneš	$2\log N + 1$	2×2	$2N \log N$	1

The Beneš network has small size and diameter, and completely eliminates congestion. The Holy Grail of routing networks is in hand!

Theorem 6.4.2. *The congestion of the N-input Beneš network is 1.*

Proof. We use induction. Let $P(a)$ be the proposition that the congestion of the size 2^a Beneš network is 1.

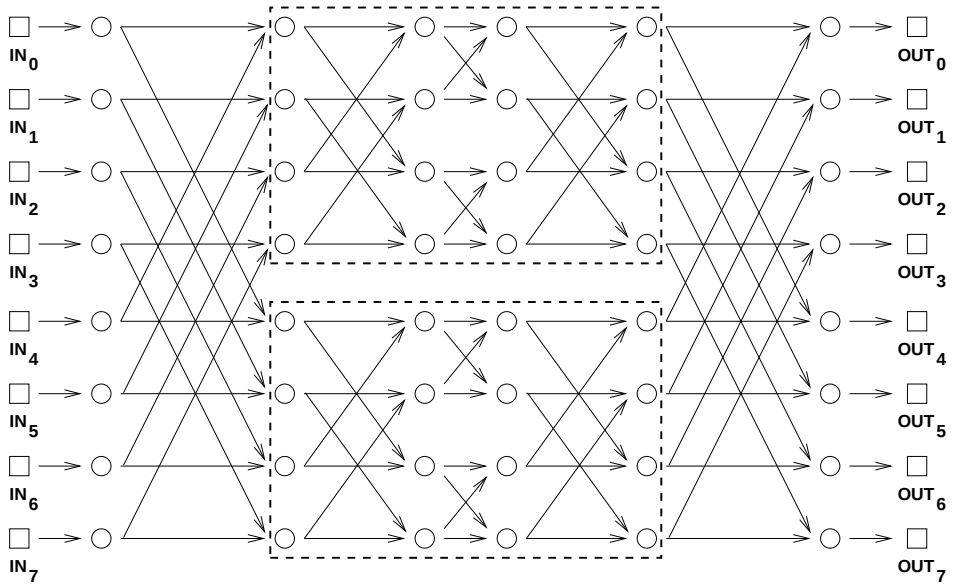
Base case. We must show that the congestion of the size $N = 2^1 = 2$ Beneš network is 1. This network is shown below:



There are only two possible permutation routing problems for a 2-input network. If $\pi(0) = 0$ and $\pi(1) = 1$, then we can route both packets along the straight edges. On the other hand, if $\pi(0) = 1$ and $\pi(1) = 0$, then we can route both packets along the diagonal edges. In both cases, a single packet passes through each switch.

Inductive step. We must show that $P(a)$ implies $P(a + 1)$, where $a \geq 1$. Thus, we assume that the congestion of an N-input Beneš network is 1 in order to prove that the congestion of a $2N$-input Beneš network is also 1.

Digression. Time out! Let's work through an example, develop some intuition, and then complete the proof. Notice that inside a Beneš network of size $2N$ lurk two Beneš subnetworks of size N. (This follows from our earlier observation that a butterfly of size $2N$ contains two butterflies of size N.) In the Beneš network shown below, the two subnetworks are in dashed boxes.

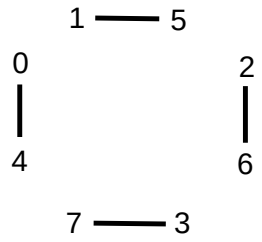


By the inductive assumption, the subnetworks can each route an arbitrary permutation with congestion 1. So if we can guide packets safely through just the first and last levels, then we can rely on induction for the rest! Let's see how this works in an example. Consider the following

permutation routing problem:

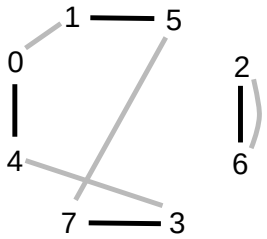
$$\begin{aligned} \pi(0) &= 1 & \pi(4) &= 3 \\ \pi(1) &= 5 & \pi(5) &= 6 \\ \pi(2) &= 4 & \pi(6) &= 0 \\ \pi(3) &= 7 & \pi(7) &= 2 \end{aligned}$$

We can route each packet to its destination through either the upper subnetwork or the lower subnetwork. However, the choice for one packet may constrain the choice for another. For example, we can not route both packet 0 *and* packet 4 through the same network since that would cause two packets to collide at a single switch, resulting in congestion. So one packet must go through the upper network and the other through the lower network. Similarly, packets 1 and 5, 2 and 6, and 3 and 7 must be routed through different networks. Let's record these constraints in a graph. The vertices are the 8 packets. If two packets must pass through different networks, then there is an edge between them. Thus, our constraint graph looks like this:



Notice that at most one edge is incident to each vertex.

The output side of the network imposes some further constraints. For example, the packet destined for output 0 (which is packet 6) and the packet destined for output 4 (which is packet 2) can not both pass through the same network; that would require both packets to arrive from the same switch. Similarly, the packets destined for outputs 1 and 5, 2 and 6, and 3 and 7 must also pass through different switches. We can record these additional constraints in our graph with gray edges:



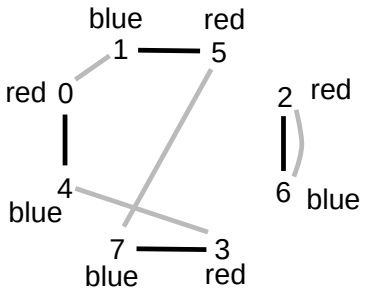
Notice that at most one new edge is incident to each vertex. The two lines drawn between vertices 2 and 6 reflect the two different reasons why these packets must be routed through different networks. However, we intend this to be a simple graph; the two lines still signify a single edge.

Now here's the key insight: *a 2-coloring of the graph corresponds to a solution to the routing problem*. In particular, suppose that we could color each vertex either red or blue so that adjacent vertices are colored differently. Then all constraints are satisfied if we send the red packets through the upper network and the blue packets through the lower network.

The only remaining question is whether the constraint graph is 2-colorable, which is easy to verify:

Problem 6.4.1. Prove that if graphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ both have maximum degree 1, then the graph $G = (V, E_1 \cup E_2)$ is 2-colorable.

For example, here is a 2-coloring of the constraint graph:



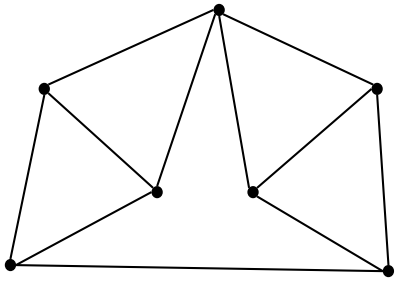
The solution to this graph-coloring problem provides a start on the packet routing problem:

We can complete the routing in the two smaller Beneš networks by induction! Back to the proof. **End of Digression.**

Let π be an arbitrary permutation of $\{0, 1, \ldots, 2N-1\}$. Let $G_1 = (V, E_1)$ be a graph where the vertices are packets 0, 1, ..., $2N-1$ and there is an edge u—v if $|u - v| = N$. Let $G_2 = (V, E_2)$ be a graph with the same vertices and an edge u—v if $|\pi(u) - \pi(v)| = N$. But according to Problem 6.4.1, the graph $G = (V, E_1 \cup E_2)$ is 2-colorable, so color the vertices red and blue. Route red packets through the upper subnetwork and blue packets through the lower subnetwork. Since for each edge in E_1, one vertex goes to the upper subnetwork and the other to the lower subnetwork, there will not be any conflicts in the first level. Since for each edge in E_2, one vertex comes from the upper subnetwork and the other from the lower subnetwork, there will not be any conflicts in the last level. We can complete the routing within each subnetwork by the induction hypothesis $P(a)$. □

6.5 In-Class Problems Week 6, Mon.

Problem 6.5.1. Let G be the graph below[4]. Carefully explain why $\chi(G) = 4$.



Solution. Four colors are sufficient:

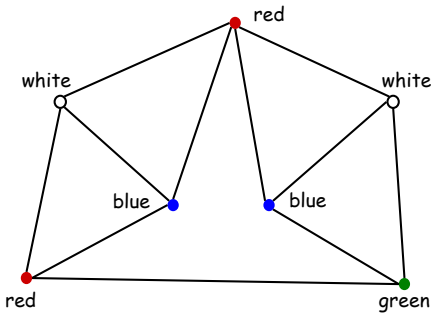


Figure 6.3: A 4-coloring of the Graph

so

$$\chi(G) \leq 4. \tag{6.1}$$

Now assume $\chi(G) = 3$. We may assume the top vertex is colored red. The top two triangles require 3 colors each, and since they share the top red vertex, they must have the other two colors, white and blue, at their bases, as in Figure 6.3. Now the bottom two vertices are both adjacent to vertices colored white and blue, and cannot have the same color since they are adjacent, so there is no alternative but to color one with a third color and the other with a fourth color, contradicting the assumption that 3 colors are enough. Hence, $\chi(G) > 3$. This together with (6.1) implies that $\chi(G) = 4$. ■

[4]From *Discrete Mathematics*, Lovász, Pelikan, and Vesztergombi. Springer, 2003. Exercise 13.3.1

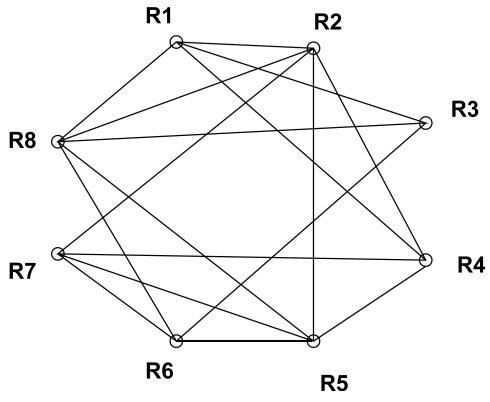
Problem 6.5.2. 6.042 is often taught using recitations. Suppose it happened that 8 recitations were needed, with two or three staff members running each recitation. The assignment of staff to recitation sections is as follows:

- R1: Tina, Jay, Jessica
- R2: Tina, Tom, Albert
- R3: Jay, Jeff
- R4: Chiyoun, Tom, Jessica
- R5: Chiyoun, Srini, Albert
- R6: Srini, Jeff
- R7: Srini, Tom
- R8: Jay, Jeff, Albert

Two recitations can not be held in the same 90-minute time slot if some staff member is assigned to both recitations. The problem is to determine the minimum number of time slots required to complete all the recitations.

(a) Recast this problem as a question about coloring the vertices of a particular graph. Draw the graph and explain what the vertices, edges, and colors represent.

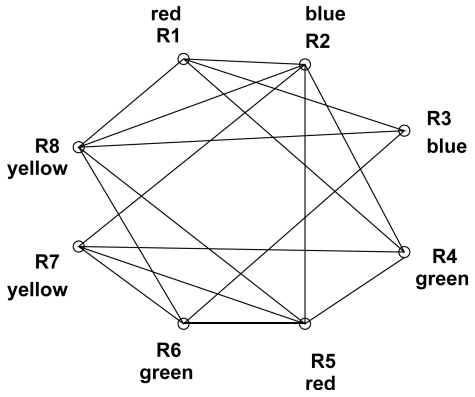
Solution. Each vertex in the graph below represents a recitation section. An edge connects two vertices if the corresponding recitation sections share a staff member and thus can not be scheduled at the same time. The color of a vertex indicates the time slot of the corresponding recitation.



■

(b) Show a coloring of this graph using the fewest possible colors. What schedule of recitations does this imply?

Solution. Four colors are necessary and sufficient. To see why they are *sufficient*, consider the coloring:



This corresponds to the following assignment of recitations to four time slots:

1. R1, R5
2. R2, R3
3. R4, R6
4. R7, R8

Other schedules are also possible.

To see why 4 colors are *necessary*, look at the subgraph defined by the nodes for R2, R4, R5, and R7. This is K_4, the complete graph on 4 vertices, and it obviously needs 4 colors. ■

Problem 6.5.3. A portion of a computer program consists of a sequence of calculations where the results are stored in variables, like this:

$$\begin{array}{rrcl} & \text{Inputs:} & & a, b \\ \text{Step } 1. & c & = & a + b \\ 2. & d & = & a * c \\ 3. & e & = & c + 3 \\ 4. & f & = & c - e \\ 5. & g & = & a + f \\ 6. & h & = & f + 1 \\ & \text{Outputs:} & & d, g, h \end{array}$$

A computer can perform such calculations most quickly if the value of each variable is stored in a *register*, a chunk of very fast memory inside the microprocessor. Computers usually have few registers, however, so they must be used wisely and reused often. The problem of assigning each variable in a program to a register is called *register allocation*.

In the example above, variables a and b must be assigned different registers, because they hold distinct input values. Furthermore, c and d must be assigned different registers; if they used the same one, then the value of c would be overwritten in the second step and we'd get the wrong answer in the third step. On the other hand, variables b and d may use the same register; after the first step, we no longer need b and can overwrite the register that holds its value. Also, f and h

may use the same register; once $f + 1$ is evaluated in the last step, the register holding the value of f can be overwritten.

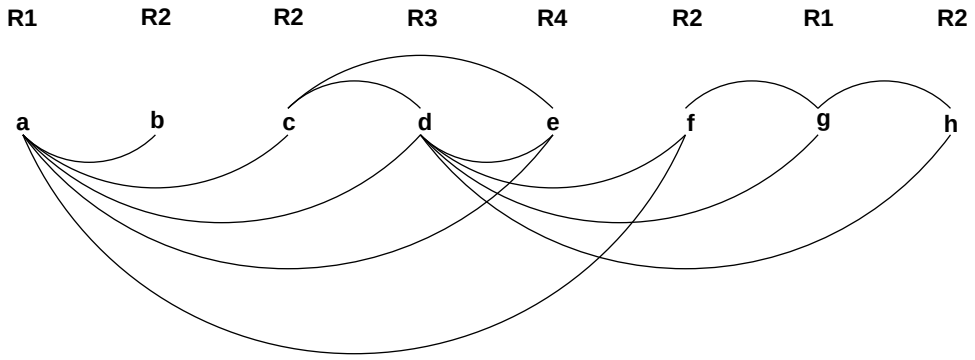
(Assume that the computer carries out each step in the order listed and that each step is completed before the next is begun.)

(a) Recast the register allocation problem as a question about graph coloring. What do the vertices correspond to? Under what conditions should there be an edge between two vertices? Construct the graph corresponding to the example above.

Solution. There is one vertex for each variable. An edge between two vertices indicates that the values of the variables must be stored in different registers.

We can classify each appearance of a variable in the program as either an *assignment* or a *use*. In particular, an appearance is an assignment if the variable is on the left side of an equation or on the "Inputs" line. An appearance of a variable is a use if the variable is on the right side of an equation or on the "Outputs" line.

The *lifetime* of a variable is the segment of code extending from the initial assignment of the variable until the last use. There is an edge between two variables if their lifetimes overlap. This rule generates the following graph:



■

(b) Color your graph using as few colors as you can. Call the computer's registers $R1$, $R2$, etc. Describe the assignment of variables to registers implied by your coloring. How many registers do you need?

Solution. Four registers are needed. One possible assignment of variables to registers is indicated in the figure above.

In general, coloring a graph using the minimum number of colors is quite difficult; no efficient procedure is known. However, the register allocation problem always leads to an *interval graph*. For interval graphs, there are efficient coloring procedures, which can be incorporated into a compiler. ■

(c) Suppose that a variable is assigned a value more than once, as in the code snippet below:

$$\begin{aligned} &\ldots \\ t &= r + s \\ u &= t * 3 \\ t &= m - k \\ v &= t + u \\ &\ldots \end{aligned}$$

How might you cope with this complication?

Solution. Each time a variable is reassigned, we could regard it as a completely new variable. Then we would regard the example as equivalent to the following:

$$\begin{aligned} &\ldots \\ t &= r + s \\ u &= t * 3 \\ t' &= m - k \\ v &= t' + u \\ &\ldots \end{aligned}$$

We can now proceed with graph construction and coloring as before. ■

6.6 In-Class Problems Week 6, Wed.

Problem 6.6.1. MIT has a lot of student clubs loosely overseen by the MIT Student Association. Each eligible club would like to delegate one of its members to appeal to the Dean for funding, but the Dean will not allow a student to be the delegate of more than one club. Fortunately, one of the officers of the Association took 6.042 and so is able to guarantee that there is a way for each club to select a delegate to appeal for funds. What the Association officer noticed is that the Association's data shows that no student is a member of more than 9 clubs. The officer also knows that to be eligible for support from the Dean's office, a club must have at least 13 members.

(a) Explain how to model the delegate selection problem as a bipartite matching problem.

Solution. Define a bipartite graph with the student clubs as one set of vertices and everybody who belongs to some club as the other set of vertices. Let a club and a student be adjacent exactly when the student belongs to the club. Now a matching of clubs to students will give a proper selection of delegates: every club will have a delegate, and every delegate will represent exactly one club. ■

(b) Explain why the Student officer can guarantee there is a proper delegate selection. (If only the Association officer had taken 6.046, *Algorithms*, they could, without an excessive computation, even have found a possible delegate selection for all the clubs.)

Solution. The degree of every club is at least 13, and the degree of every student is at most 9, so the graph is *degree-constrained* (see the Appendix) which implies there will be no bottlenecks to prevent a matching. Hall's Theorem then guarantees a matching. ■

(c) The Student Association officer used the fact that the condition for a matching to exist, namely, that every set of c clubs had a total of at least c members, was sure to hold because the degree of every club was greater than the degree of every student. See if you can come up with a proof of this fact (without looking it up in the Notes).

Solution. Suppose a set of c clubs has a total of s members. We want to show that this set is not a bottleneck to a match, that is, $c \geq s$.

Now the number, e, of edges leaving this set of clubs is at least 13 times the number of clubs, so $e \geq 13c$. All these edges go, by definition, to students who are members of one of these clubs, so the total number of edges going to students in the clubs is at least e. But at most 9 edges go to a student, so the total number of edges going to the students in these clubs is at most $9s$. That is,

$$13c \leq e \leq 9s,$$

which implies that $c \leq s$. ■

Problem 6.6.2. Through a series of acquisitions, the CellTel corporation has obtained a nationwide network of cell phone towers. Each tower can support up to 2110 callers, but a CellTel customer can only contact a tower that is within 10 miles. As a consultant to CellTel, your job is to explain why, at any given moment, their network can service all callers iff, at that moment, the following condition holds:

> *The size of any set of callers is at most 2110 times the number of towers within range of at least one caller in the set.* (*)

(a) Describe how to model this situation as a bipartite graph matching problem.

Solution. There will be one vertex for each caller, and 2110 vertices for each tower. There is an edge between a tower-vertex and a caller when the caller is within the 10 mile range of the tower. So a matching for the callers assigns callers to within-range towers and assigns at most 2110 callers to any tower. Conversely, any such assignment of callers to towers yields a matching of callers to tower vertices that assigns at most 2110 callers per tower.

It's important to notice that a tower is in range of a caller iff *all* its tower-vertices are adjacent to the caller-vertex. So the number of tower-vertices in the neighborhood of a set of callers is exactly 2110 times the number of towers within range of at least one caller in the set. ■

(b) Explain why the network can service all the callers iff condition (*) holds.

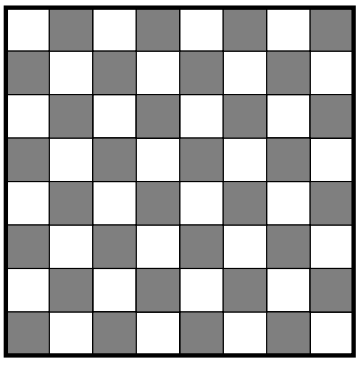
Solution. By Hall's Theorem, there is a matching of callers to tower-vertices iff

> *The size of any set of callers is at most the total number of tower-vertices in the neighborhood of the set.* (**)

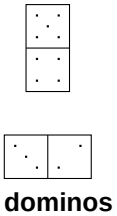
But the total number of tower-vertices in the neighborhood is exactly 2110 times the number of towers in range of at least one caller in the set. It follows that (**) is equivalent to (*). ■

Problem 6.6.3. Suppose that one domino can cover exactly two squares on a chessboard, either vertically or horizontally.

(a) Can you tile an 8×8 chessboard with 32 dominos?



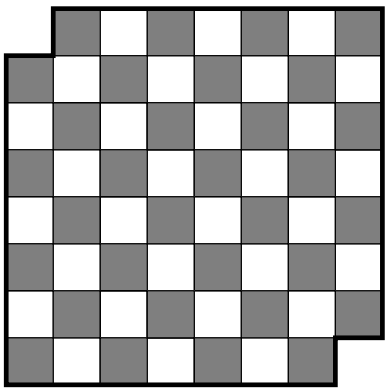
chess board



dominos

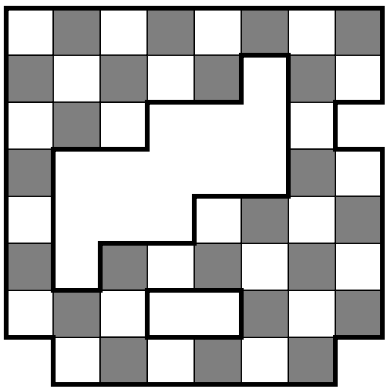
Solution. Yes. Place 4 vertical dominos in each column. ■

(b) Can you tile an 8×8 chessboard with 31 dominos if opposite corners are removed?



Solution. No! Opposing corners are the same color. Therefore, removing opposite corners leaves an unequal number of white and black squares. Since every domino covers one black square and one white square, no tiling is possible. ■

(c) Now suppose that an assortment of squares are removed from a chessboard. An example is shown below.



Explain why a board can be tiled with dominos iff there are an equal number of black squares and white squares, and every set of white squares is adjacent to at least as many black squares. *Hint:* Given a truncated chessboard with an equal number of white and black squares, show how to construct a bipartite graph that has a perfect matching iff the chessboard can be tiled with dominos.

Solution. Create a vertex for every white square and a vertex for every black square. Put an edge between squares that share an edge. This graph is bipartite, since the coloring of the squares defines a valid 2-coloring of the vertices.

If a perfect matching exists in this graph, then a tiling exists: put a domino over each pair of matched vertices. On the other hand, if a tiling exists, then a perfect matching exists: match squares covered by the same domino.

By Hall's Theorem, a matching of all the white squares with black squares will exist iff there are no bottlenecks, namely, iff every set of white squares is adjacent to at least as many black squares. Since the number of white and black squares is the same, any matching of the white squares will also match the black squares, that is, it will be a perfect matching. ■

Bipartite Matching

Suppose S is a set of vertices in a graph. Define $N(S)$ to be the set of all neighbors of S; that is, all vertices that are adjacent to a vertex in S. S is called a *bottleneck* if

$$|S| > |N(S)|\,.$$

A *matching* for S is a set of edges such that

- the edges are non-overlapping —no two edges are incident to the same vertex,
- every edge is incident to exactly one vertex in S, and
- every vertex in S is incident to an edge.

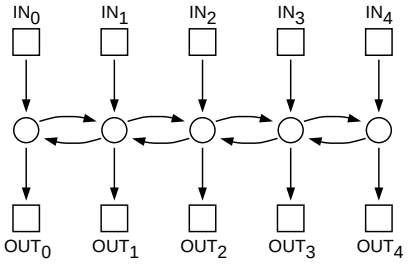
Theorem 6.6.1 (Hall's Theorem). *Let G be a bipartite graph, that is, a graph whose vertices can be separated (partitioned) into two sets, L and R, such that every edge has one endpoint in L and one endpoint in R. There is a matching for L iff no subset of L is a bottleneck.*

Definition 6.6.2. A bipartite graph G with vertex partition L, R is *degree-constrained* if $\deg(l) \geq \deg(r)$ for every $l \in L$ and $r \in R$.

Lemma 6.6.3. *Every degree-constrained bipartite graph satisifies the matching condition.*

6.7 In-Class Problems Week 6, Fri.

Problem 6.7.1. A *5-path* communication network is shown below. From this, it's easy to see what an n-path network would be. Fill in the table of properties below, and be prepared to justify your answers.



5-Path

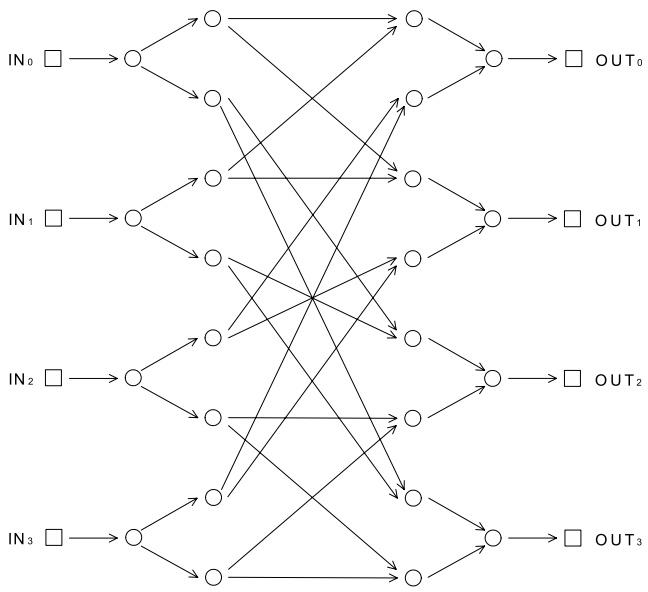
network	# switches	switch size	diameter	max congestion
5-path	5	3×3	6	5
n-path	n	3×3	$n+1$	n

Solution. The congestion of the n-path is at least n, because every path must contain the central switch when $\pi(i) = n - 1 - i$. The congestion is at most n, because there are only n simple paths in total. The longest path is from input 0 to output $n - 1$ of length $n + 1$, so this is the diameter. ■

Problem 6.7.2. A *binary-tree network* has n inputs and n outputs, where n is a power of 2. Each input is connected to the root of a binary tree with $n/2$ leaves and with edges pointing away from the root. Likewise, each output is connected to the root of a binary tree with $n/2$ leaves and with edges pointing toward the root.

Two edges point from each leaf of an input tree, and each of these edges points to a leaf of an output tree. The matching of leaf edges is arranged so that for every input and output tree, there is an edge from a leaf of the input tree to a leaf of the output tree, and every output tree leaf has exactly two edges pointing to it.

(a) Draw such a binary-tree net for $n = 4$.



Solution. ■

(b) Fill in the table, and explain your entries.

# switches	switch size	diameter	max congestion

Solution.

# switches	switch size	diameter	max congestion
$2n(n-1)$	$1 \times 2, 2 \times 1$	$1 + 2\log n$	1

These formulas were gotten as follows: a binary tree with $n/2$ leaves has $n - 1$ nodes (switches), and there are $2n$ trees.

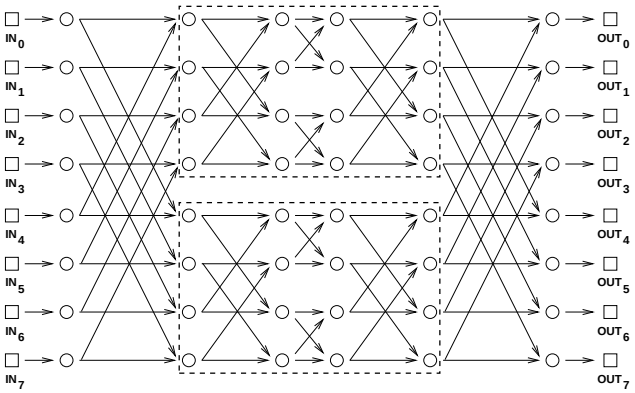
Each node of an input tree has one edge in and two out; the opposite for nodes of output trees.

The distance from any input to any output is 1 from input to tree root, $(\log n) - 1$ from root to leaf, 1 from input leaf to output leaf, $(\log n) - 1$ from output leaf to output root, and 1 to output, for a total of $1 + 2\log n$.

The path from any input to any output is unique, and paths from two inputs to different outputs don't overlap, so at most one packet goes through any switch. ■

Problem 6.7.3. The Beneš network has a max congestion of 1; that is, every permutation can be routed in such a way that a single packet passes through each switch. Let's work through an example. A Beneš network of size $N = 8$ is attached.

(a) Within the Beneš network of size $N = 8$, there are two subnetworks of size $N = 4$. Put boxes around these. Hereafter, we'll refer to these as the *upper* and *lower* subnetworks.

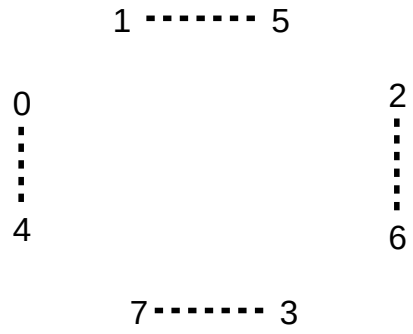


Solution. ■

(b) Now consider the following permutation routing problem:

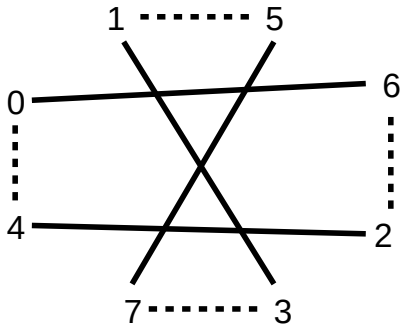
$$\begin{aligned} \pi(0) &= 3 & \qquad \pi(4) &= 2 \\ \pi(1) &= 1 & \qquad \pi(5) &= 0 \\ \pi(2) &= 6 & \qquad \pi(6) &= 7 \\ \pi(3) &= 5 & \qquad \pi(7) &= 4 \end{aligned}$$

Each packet must be routed through either the upper subnetwork or the lower subnetwork. Construct a graph with vertices 0, 1, . . . , 7 and draw a *dashed* edge between each pair of packets that can not go through the same subnetwork because a collision would occur in the second column of switches.



Solution. ■

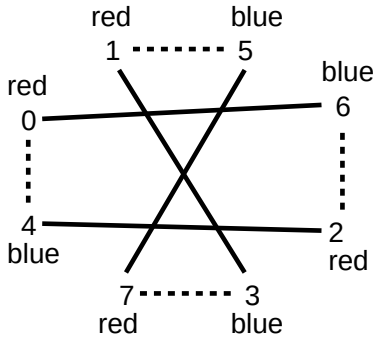
(c) Add a *solid* edge in your graph between each pair of packets that can not go through the same subnetwork because a collision would occur in the next-to-last column of switches.



Solution. ■

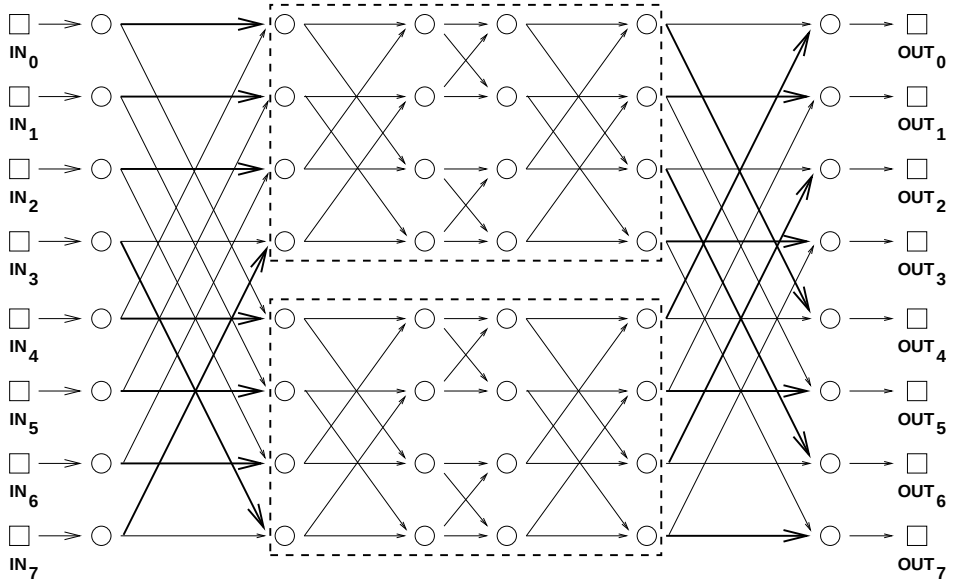
(d) Color the vertices of your graph red and blue so that adjacent vertices get different colors. Why must this be possible, regardless of the permutation π?

Solution. This must be possible, because edges in a cycle are alternately dashed and solid. Thus, every cycle has even length, which implies that the graph is bipartite or, equivalently, 2-colorable.



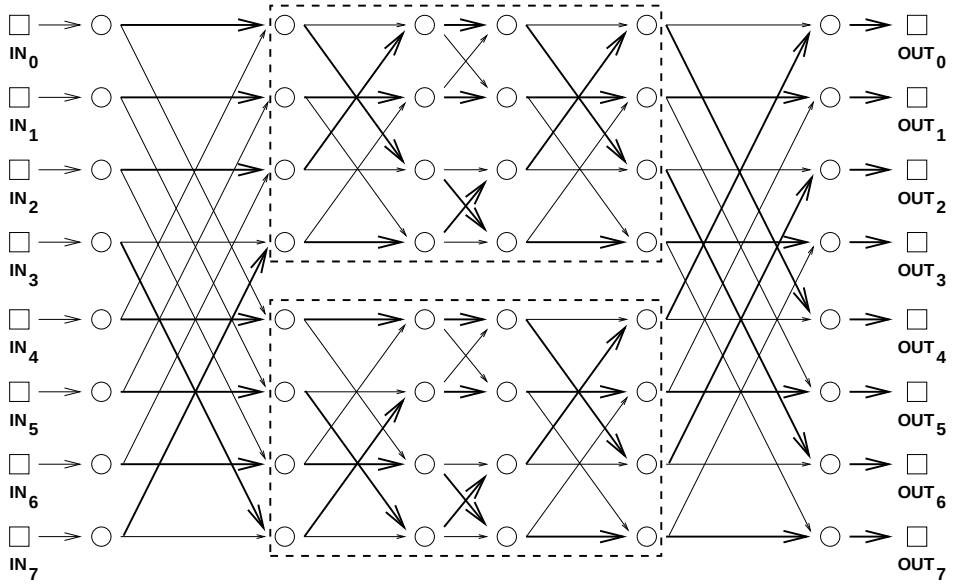
■

(e) Suppose that red vertices correspond to packets routed through the upper subnetwork and blue vertices correspond to packets routed through the lower subnetwork. On the attached copy of the Beneš network, highlight the first and last edge traversed by each packet.

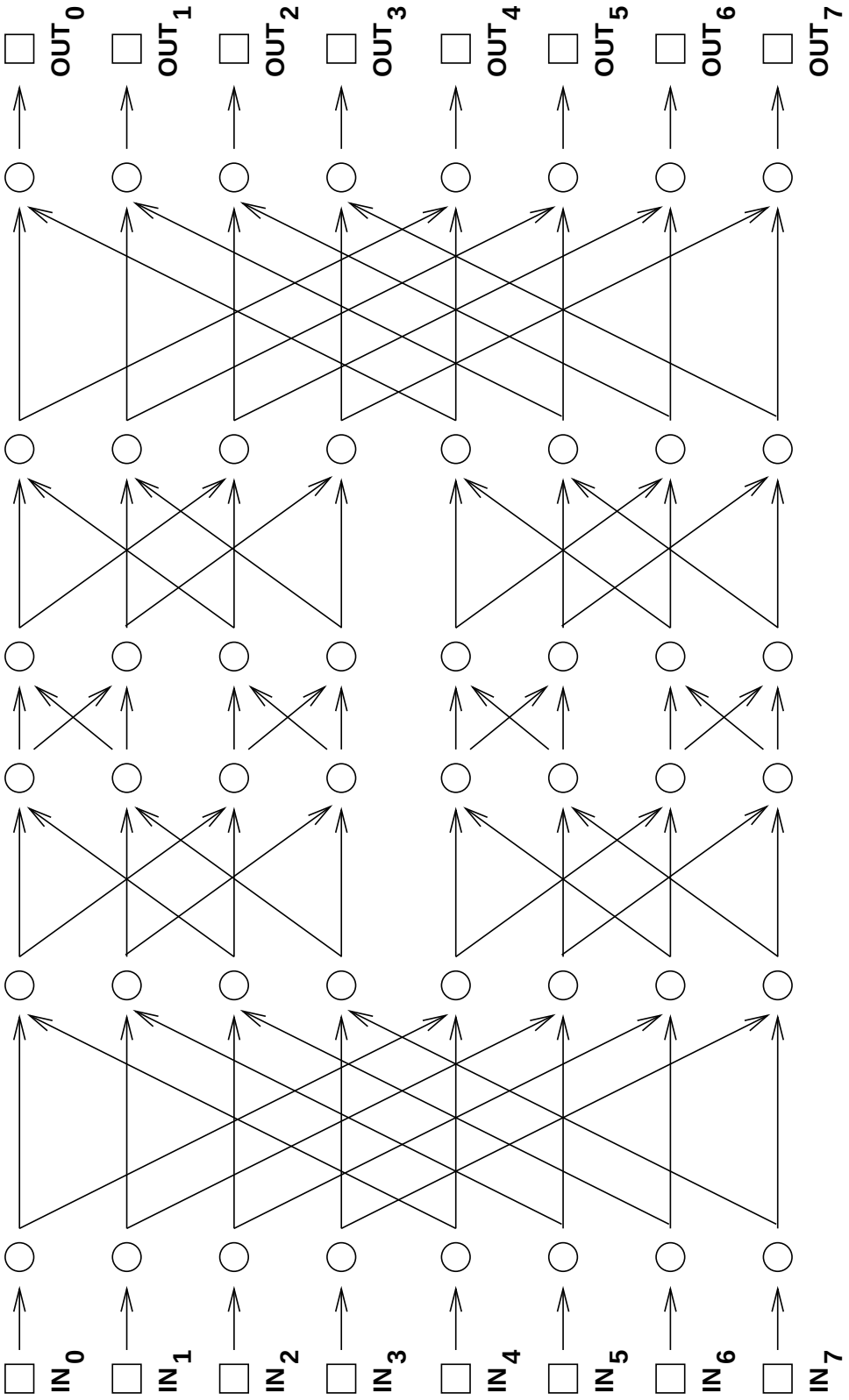


Solution. ■

(f) All that remains is to route packets through the upper and lower subnetworks. One way to do this is by applying the procedure described above recursively on each subnetwork. However, since the remaining problems are small, see if you can complete all the paths on your own.



Solution. ■


OUT0
OUT1
OUT2
OUT3
OUT4
OUT5
OUT6
OUT7
IN0
IN1
IN2
IN3
IN4
IN5
IN6
IN7

6.8 Problem Set 5

Problem 6.8.1. This problem generalizes the result proved in Week 5 Notes that any graph with maximum degree at most w is $(w+1)$-colorable.

A simple graph, G, is said to have ***width***, w, iff its vertices can be arranged in a sequence such that each vertex is adjacent to at most w vertices that precede it in the sequence. If the degree of every vertex is at most w, then the graph obviously has width at most w —just list the vertices in any order.

(a) Describe an example of a graph with 100 vertices, width 3, but *average* degree more than 5. *Hint:* Don't get stuck on this; if you don't see it after five minutes, ask the staff for a hint.

Solution. The hint is to line up the 100 vertices and have each vertex be adjacent to the 3 immediately preceding vertices, if any. By definition of width, this graph has width 3. All vertices other than the first three are now adjacent to three preceding vertices, and all vertices except the last three are also adjacent to the three following vertices. So vertices 4 through 97 all have degree 6; this alone ensures that the average degree is at least $6 \cdot 94/100 = 5.76$. ■

(b) Prove that every graph with width at most w is $(w+1)$-colorable.

Solution. We use induction on n, the number of vertices. Let $P(n)$ be the proposition that for all w, every n-vertex graph with width w is $(w+1)$-colorable.

Base case: ($n = 1$) Every graph with 1 vertex has width 0 and is $0 + 1 = 1$ colorable. Therefore, $P(1)$ is true.

Inductive step: Now we assume $P(n)$ in order to prove $P(n+1)$. Let G be an $(n+1)$-vertex graph with width at most w. This means that the $n+1$ vertices can be arranged in a sequence, S, such that each vertex is connected to at most w preceding vertices. Removing the last vertex, v, and all edges incident to it gives a subgraph G' with n vertices. The subgraph G' also has width at most w, since the sequence S with its last vertex removed is a sequence of all the vertices of G' with each vertex adjacent to exactly the same previous vertices. So by Induction Hypothesis, G' is $(w+1)$-colorable. But any $(w+1)$-coloring of G' can be extended to a $(w+1)$-coloring of G by assigning a color to v that differs from the colors of its adjacent vertices. Since there are at most w colors among the w vertices adjacent to v, there will always be a different one of the $w+1$ colors to assign to v. So G is $(w+1)$-colorable, which proves $P(n+1)$. This completes the proof of the Induction step.

The result now follows for all G by the Principle of Induction. ■

(c) Prove that the average degree of a graph of width w is at most $2w$.

Solution. If we line up the vertices, we can define the *backdegree* of a vertex to be the number of preceding vertices it is adjacent to. The sum of the back degrees equals the number, e, of edges. Since there is a sequence in which all the back degrees are at most w, the total number of edges is at most w times the number, n, of vertices. But by the Handshaking Lemma, the sum of all the degrees is $2e$, so the average degree is $2e/n \leq 2wn/n = 2w$. ■

Problem 6.8.2. In this problem you will prove:

Theorem. *A graph G is 2-colorable iff it contains no odd length cycle.*

As usual with "iff" assertions, the proof splits into two proofs: part (a) asks you to prove that the left side of the "iff" implies the right side. The other problem parts prove that the right side implies the left.

(a) Assume the left side and prove the right side. Three to five sentences should suffice.

Solution. Assume G is 2-colorable and select a 2-coloring of G. Consider an arbitrary cycle with successive vertices $v_1, v_2, \ldots, v_k, v_1$. Then the vertices v_i must be one color for all even i and the other color for all odd i. (This is obvious, but could of course, be proved by induction.) Since v_1 and v_k must be colored differently, k must be even. Thus, the cycle has even length. ■

(b) Now assume the right side. As a first step toward proving the left side, explain why we can focus on a single connected component H within G.

Solution. If we can 2-color every connected component of G, then we can 2-color all of G. Thus, it suffices to show that an arbitrary connected component H of G is 2-colorable. ■

(c) As a second step, explain how to 2-color any tree.

Solution. A 2-coloring of a tree can be defined by selecting any fixed vertex v, and coloring a vertex one color if the (unique) path to it from v has odd length, and coloring it with the other color if the path has even length.

To verify that adjacent vertices in the tree get different colors, let $e ::= x—y$ be an edge in the tree. There is a unique path from v to x. If this path traverses e, it must consist of a path from v to y followed by the e traversal to x. If this path does not traverse e, then it can be extended to a path to y by adding a final traversal of e. In either case, the paths to these vertices from v differ by a single traversal of e, and so the lengths of the paths differ by 1; in particular, one is of odd length and the other is of even length, so x and y are differently-colored. ■

(d) Choose any 2-coloring of a spanning tree, T, of H. Prove that H is 2-colorable by showing that any edge *not* in T must also connect different-colored vertices.

Solution. Let $x—y$ be an edge not in T, and consider the unique paths from v to x and from v to y in T. Exactly one of these two paths must have odd length; otherwise, these two paths together with the edge $x—y$ would form an odd length cycle. But this means x and y are colored differently. ■

Problem 6.8.3. Prove that if graphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ both have maximum degree 1, then the graph $G = (V, E_1 \cup E_2)$ is 2-colorable.

Solution. We need only show G has no odd length cycle.

Suppose to the contrary that G had an odd length cycle. If the same edge was traversed twice in a row in this cycle, these two traversals could be eliminated from the path to yield a shorter odd length cycle. So if u, v, w are three consecutive vertices in the cycle, then u—v and v—w will be different. This implies that one of these edges is in E_1 and the other is in E_2, since otherwise v would have degree two in either G_1 or G_2. So successively traversed edges in the cycle must alternate between E_1 and E_2, which is only possible if the cycle is of even length, a contradiction. ■

Problem 6.8.4. At the Guiness brewery in the eary 1900's, W. S. Gosset (a chemist) and E. S. Beaven (a "maltster") were working to improve barley. Gosset and Beavan planned to grow several varieties of barley in a field and compare the yields. However, local variations in the fertility of the field might skew the results. Their solution was to divide the field into many small plots and grow each crop in several different places.

Similar thinking led to the use of *Latin squares* in experiment design. A Latin square is an $n \times n$ array of numbers such that each row and each column contains every number from 1 to n. For example, here is a 4×4 Latin square:

1	2	3	4
3	4	2	1
2	1	4	3
4	3	1	2

You can imagine that this array is an agricultural field where each square is a small plot, and the number inside indicates the variety of barley planted there.

There are some nice connections between Latin squares and graph theory.

(a) Construct a graph G_n such that there is a one-to-one correspondence between $n \times n$ Latin squares and valid n-colorings of G_n.

Solution. Create a vertex in G for each entry in the Latin Square. Then connect each vertex to every other vertex in the same row and to every other vertex in the same column. Now color the graph with n colors, each corresponding to a number between 1 and n.

Notice that every pair of vertices in the same row are connected, so no two vertices in the same row can get the same color. Similarly, since every pair of vertices in the same column are connected, no two vertices in the same column can get the same color either. These coloring constraints match the constraints on Latin squares, so there is a one-to-one correspondence between colorings of G and $n \times n$ Latin squares. ■

(b) Suppose your teammate wrote down a 3×5 "Latin rectangle":

2	4	5	3	1
4	1	3	2	5
3	2	1	5	4

To fill in a possible next line of a complete Latin Square, construct a bipartite graph with vertices numbered 1 to 5 on the left and on the right. Let i and j be adjacent iff column i does not contain j. Find a matching in this graph and use it to fill the 4th row.

Solution. Here is one possible solution completion:

2	4	5	3	1
4	1	3	2	5
3	2	1	5	4
1	5	2	4	3
5	3	4	1	2

SHOW GRAPH AND MATCHING GIVING 4TH ROW. ■

(c) Show that filling in the $k+1$st row of a $k \times n$ Latin rectangle is equivalent to finding a matching in some bipartite graph with n vertices on the left and right.

Solution. Construct a bipartite graph as follows. The vertices on the left are the columns of the Latin rectangle, and the vertices on the right are the numbers 1 to n. Put an edge between a column and a number if the number has *not yet appeared* in the column. Thus, a matching in this graph would associate each column with a distinct number that has not yet appeared in that column. These numbers would form the next row of the Latin rectangle. ■

(d) Prove that a matching must exist in this bipartite graph and, consequently, a Latin rectangle can always be extended to a Latin square.

Solution. First, we show that the bipartite graph described above has a matching. Each column-vertex on the left has degree $n - k$ and each number-vertex on the right has degree $n - k$ as well. So this is a degree-constrained bipartite graph, and therefore has a matching. ■

Problem 6.8.5. If a and b are distinct nodes of a digraph, then a is said to *cover* b if there is an edge from a to b and there is no other path from a to b. If a covers b, the edge from a to b is called a covering edge.

(a) Show that if two DAG's have the same positive path relation, then they have the same set of covering edges.

Solution. Suppose two DAG's have the same positive path relation, and consider any covering edge in the first DAG. By definition of covering, it is the unique path between its endpoints in *both* DAG's, since they agree on positive length paths. But this implies it is a covering edge in the second DAG as well. ■

(b) For any DAG, D, let $\widehat{D}$ be the subgraph of D consisting of only the covering edges. Show that if D is finite and has no self-loops, then D and $\widehat{D}$ have the same positive path relation, that is $D^+ = \widehat{D}^+$.

Solution. What we need to show is that if there is a path in D between vertices $a \neq b$, then there is a path consisting only of covering edges from a to b. But since D is a finite DAG, there must be a *longest* path from a to b. Now every edge on this path must be a covering edge or it could be replaced by a path of length 2 or more, yielding a longer path from a to b. ■

(c) Conclude that if D is a finite DAG, then $\widehat{D}$ is the unique DAG with the smallest number of edges among all digraphs with the same positive path relation.

Solution. By part a, any DAG with the same positive path relation as D must contain all the edges of $\widehat{D}$, so the unique minimality of $\widehat{D}$ follows immediately from part b. ■

(d) Show that the previous result is not true in general infinite DAG's.

Hint: Consider the DAG for the total order on the rational numbers.

Solution. In the DAG for $<$ on the $\mathbb{Q}$, there are no covering edges, so $\widehat{<}$ has no edges. ■

Problem 6.8.6. Let B_n denote the butterfly network with $N = 2^n$ inputs and N outputs, as defined in Week 5 Notes. Show that the congestion of B_n is exactly $\sqrt{N}$ when n is even.

Hints:

- For the butterfly network, there is a unique path from each input to each output, so the congestion is the maximum number of messages passing through a vertex for any matching of inputs to outputs.
- If v is a vertex at level i of the butterfly network, there is a path from exactly 2^i input vertices to v and a path from v to exactly 2^{n-i} output vertices.
- At which level of the butterfly network must the congestion be worst? What is the congestion at the node whose binary representation is all 0s at that level of the network?

Solution. First we will show that the congestion is at most $\sqrt{N}$.

Let v be an arbitrary vertex at some level i. Let S_v be the set of inputs that can reach vertex v. Let T_v be the set of outputs that are reachable from vertex v.

By the hint, we have $|S_v| = 2^i$ and $|T_v| = 2^{n-i}$. The number of inputs in S_v that are matched with outputs in T_v is at most $\min\left\{2^i, 2^{n-i}\right\}$. To obtain an upper-bound on the congestion of the network, we need to find the maximum value of $\min\left\{2^i, 2^{n-i}\right\}$, where the maximum is taken over

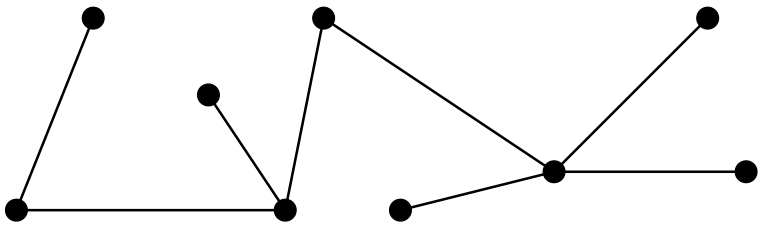
all i. The maximum value is achieved when 2^i and 2^{n-i} are as equal as possible. Since n is even, these two quantities are equal when $i = n/2$, hence the maximum congestion is

$$2^{n/2} = N^{1/2} = \sqrt{N}.$$

Now we need to show that the congestion achieves $\sqrt{N}$ somewhere in the network. We concluded that the congestion of $\sqrt{N}$ can be achieved only at a node at level $\frac{n}{2}$. Consider the node at that level whose binary representation is all 0s. Any packet from the input in the form $z\underbrace{0\ldots000}_{n/2\ bits}$ with destination $\underbrace{000\ldots0}_{n/2\ bits}z'$, where z and z' are any $\frac{n}{2}$-bit numbers, must pass through this node. But there are $2^{n/2} = \sqrt{N}$ of them, giving the node load $\sqrt{N}$. Therefore, we can conclude that the congestion of B_n is exactly $\sqrt{N}$ when n is even.

■

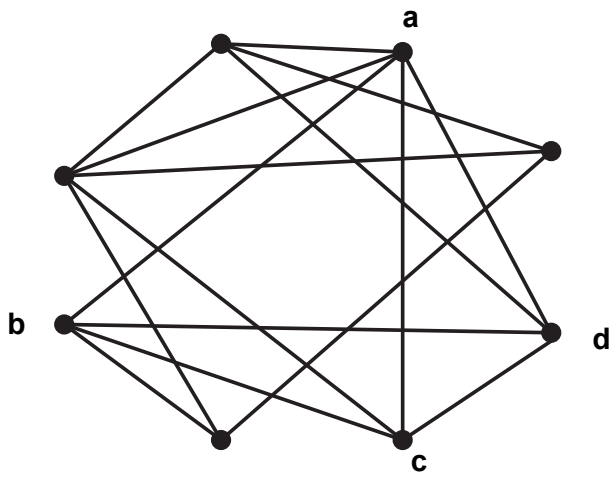
6.9 Miniquiz Mar. 21



Problem 6.9.1. (a)

1. Provide a valid coloring of this graph, by labelling each of the vertices with a color. Use as few colors as possible.
2. Explain why the chromatic number of the graph is at least 2.

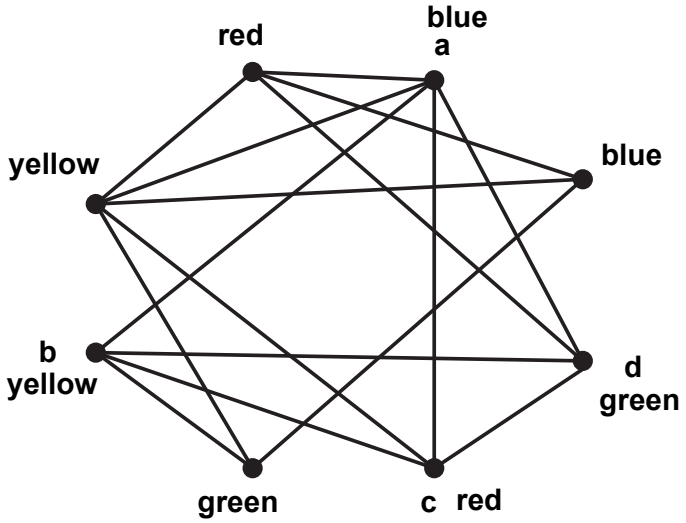
Solution. Since this is a tree with at least one edge, we know the chromatic number is 2. A coloring using 2 colors can be obtained by starting at any node that we call the root, giving it and all nodes an even distance away (on the unique path from the root) the first color, and coloring nodes an odd distance away from it the other color. ■



(b)

1. Provide a valid coloring of this graph, by labelling each of the vertices with a color. Use as few colors as possible.
2. Explain why the chromatic number of the graph is at least 4.

Solution. Four is the chromatic number of this graph. To see why four colors are *sufficient*, consider the coloring:



To see why 4 colors are *necessary*, look at the subgraph defined by the labelled vertices. This is K_4, the complete graph on 4 vertices, and it obviously needs 4 colors. ■

Problem 6.9.2. The team of LA's and TA's think they are being overworked and have decided to oust Albert, and teach their own recitations. They will offer five separate recitations, each at different time slots. Each recitation is only allowed to have up to 20 students.

Each student only has some subset of schedule time slots open.

The LA's and TA's want to determine whether a chosen set of five recitation time slots will allow each student to be assigned to a recitation they can attend, and, if so, which recitation each student should be assigned to.

(a) What should we use to model this problem? Circle the best option below:

- Stable matching
- Vertex coloring
- Bipartite matching

Solution. Bipartite matching would be the best solution. We have to assign students to time slots that are free for them. ■

(b) Describe how to model the problem, based on your choice above. Be sure to describe whatever may be relevant: preference lists, vertices, edges, or partitions, for example, *as well as* a brief description of how to interpret the output stable matching/ vertex coloring/ bipartite matching. (This is a *modeling problem*; we are not looking for conditions under which solutions exist nor descriptions of algorithms that solve the problems in part (a).)

Solution. There will be one vertex for each student, and 20 vertices for each recitation time slot. There is an edge between a student and a recitation time slot vertex if the student can attend a recitation at that time. A matching for the students assigns a student to a recitation that s/he can attend and assigns at most 20 students to any recitation. Conversely, any assignment of students to recitations yields a matching of students to recitation time slot vertices that assigns at most 20 students per recitation.

Notice that the number of recitation time slot vertices in the neighborhood of a set of students is exactly 20 times the number of recitations that could be attended by at least one student in the set. ■

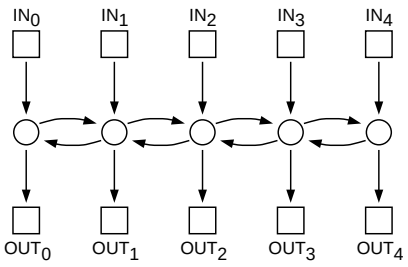
Problem 6.9.3. Prove that every degree-constrained bipartite graph, G, with vertex partition L, R has a matching from L to R. You may assume Hall's Theorem. Hint: Consider the sum of degrees of the vertices in any subset S of L, and $N(S)$.

Your proof will be graded primarily on the clarity of your argument.

Solution. *Proof.* Let S be any set of vertices in L. The number of edges incident to vertices in S is exactly the sum of the degrees of the vertices in S. Each of these edges is incident to a vertex in $N(S)$ by definition of $N(S)$. So the sum of the degrees of the vertices in $N(S)$ is at least as large as the sum for S. But since the degree of every vertex in $N(S)$ is at most as large as the degree of every vertex in S, there would have to be at least as many terms in the sum for $N(S)$ as in the sum for S. So there have to be at least as many vertices in $N(S)$ as in S, proving that S is not a bottleneck. So there are no bottlenecks in L, and, by Hall's Theorem, G has a total matching from L to R. □

■

Problem 6.9.4. A *5-path* communication network is shown below. From this, it's easy to see what an n-path network would be.



5-Path

(a) Circle all the permutations, mapping inputs to outputs, that ensure that the max congestion in a n-path network is at least n:

- $\pi(i) = n - 1 - i$
- $\pi(i) = \begin{cases} \lfloor n/2 \rfloor + i & \text{if } i < n/2, \\ i - \lceil n/2 \rceil & \text{if } i \geq n/2. \end{cases}$
- $\pi(i) = \begin{cases} i + 1 & \text{if } i < n - 1, \\ 0 & \text{if } i = n - 1. \end{cases}$
- $\pi(i) = i$

Solution. The first and second options both force each path to pass through the central switch, making the max congestion at least n. The max congestion for the third choice is 3, and for the fourth choice is 1. ■

(b) Explain why the max congestion in a n-path network is at most n.

Solution. The congestion is at most n, because there are only n inputs and hence, n simple paths in total. ■

Appendix

$\lceil x \rceil$ is the smallest integer y such that $y \geq x$.

$\lfloor x \rfloor$ is the largest integer y such that $y \leq x$.

Stable Marriage

A *marriage assignment* or *perfect matching* is a bijection, w : Boys $\rightarrow$ Girls.

A *rogue couple* is a boy, B, and a girl, G, such that B prefers G to his wife, and G prefers B to her husband.

An assignment is *stable* if it has no rogue couples.

Graph coloring

The *graph coloring problem* is as follows: Given a graph G, assign colors to each node such that adjacent nodes have different colors. A color assignment with this property is called a *valid coloring* of the graph —a "coloring," for short. A graph G is ***k*-colorable** if it has a coloring that uses at most k colors. The minimum value of k for which a coloring exists is called the ***chromatic number***, $\chi(G)$, of G.

Bipartite graphs

A *bipartite graph* is a graph together with a partition of its vertices into two sets, L and R, such that every edge is incident to a vertex in L and to a vertex in R.

A *bipartite matching* from L to R is a way of assigning each vertex in L to some vertex in R with which it shares an edge, and such that each vertex in R gets assigned to at most one vertex in L.

In any graph, the set $N(S)$, of *neighbors* of some set, S, of vertices is the set of all vertices adjacent to any vertex in S.

S is called a *bottleneck* if

$$|S| > |N(S)|\,.$$

Theorem 6.9.1 (Hall's Theorem). *Let G be a bipartite graph with vertex partition L, R. There is total matching from L to R iff no subset of L is a bottleneck.*

A bipartite graph G with vertex partition L, R is *degree-constrained* if $\deg(l) \geq \deg(r)$ for every $l \in L$ and $r \in R$.

Communication networks

The ***congestion*** of a set of paths $P_{0,\pi(0)}, \ldots, P_{N-1,\pi(N-1)}$ is equal to the largest number of paths that pass through a single switch.

The ***max congestion*** of a network is the *maximum* over all permutations π of the *minimum* over all paths $P_{i,\pi(i)}$ of the congestion of the paths.

Chapter 7

Planar Graphs

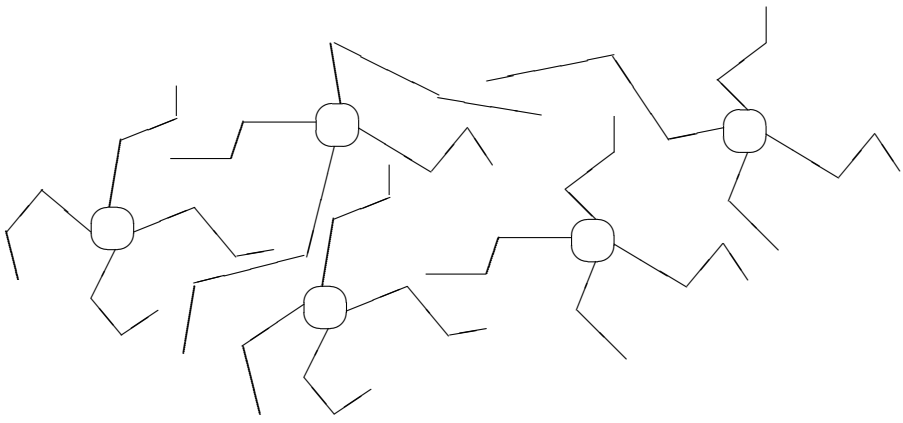
7.1 Drawing Graphs in the Plane

Here are three dogs and three houses.

Dog Dog Dog

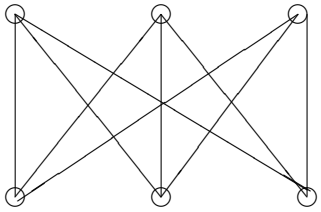
Can you find a path from each dog to each house such that no two paths intersect?

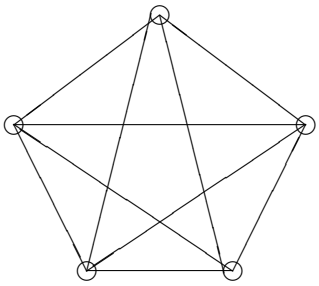
A *quadapus* is a little-known animal similar to an octopus, but with four arms. Here are five quadapi resting on the seafloor:



Can each quadapus simultaneously shake hands with every other in such a way that no arms cross?

Informally, a ***planar graph*** is a graph that can be drawn in the plane so that no edges cross, as in a map of showing the borders of countries or states. Thus, these two puzzles are asking whether the graphs below are planar; that is, whether they can be redrawn so that no edges cross. The first graph is called the ***complete bipartite graph***, $K_{3,3}$, and the second is K_5.





In each case, the answer is, "No— but almost!" In fact, each drawing *would* be possible if any single edge were removed.

Planar graphs have applications in circuit layout and are helpful in displaying graphical data, for example, program flow charts, organizational charts, and scheduling conflicts. We will use them to prove a wonderful mathematical fact that was first proved by the ancient Greeks.

When wires are arranged on a surface, like a circuit board or microchip, crossings require troublesome three-dimensional structures. When Steve Wozniak designed the disk drive for the early Apple II computer, he struggled mightly to achieve a nearly planar design:

> For two weeks, he worked late each night to make a satisfactory design. When he was finished, he found that if he moved a connector he could cut down on feedthroughs, making the board more reliable. To make that move, however, he had to start over in his design. This time it only took twenty hours. He then saw another feedthrough that could be eliminated, and again started over on his design. "The final design was generally recognized by computer engineers as brilliant and was by engineering aesthetics beautiful. Woz later said, 'It's something you can only do if you're the engineer and the PC board layout person yourself. That was an artistic layout. The board has virtually no feedthroughs.'"[a]

[a] From apple2history.org which in turn quotes *Fire in the Valley* by Freiberger and Swaine.

7.2 Continuous & Discrete Faces

Planar graphs are graphs that can be drawn in the plane —like familiar maps of countries or states. "Drawing" the graph means that each vertex of the graph corresponds to a distinct point in the plane, and if two vertices are adjacent, their vertices are connected by a smooth, non-self-intersecting curve. None of the curves may "cross" —the only points that may appear on more

than one curve are the vertex points. These curves are the boundaries of connected regions of the plane called the *continuous faces* of the drawing.

For example, the drawing in Figure 7.1 has four continuous faces. Face IV, which extends off to

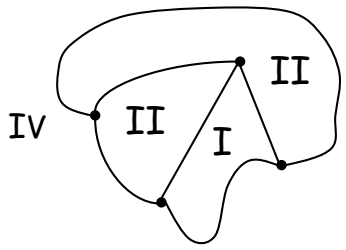


Figure 7.1: A Planar Drawing with Four Faces.

infinity in all directions, is called the ***outside face***.

This definition of planar graphs is perfectly precise, but completely unsatisfying: it invokes smooth curves and continuous regions of the plane to define a property of a discrete data type. So the first thing we'd like to find is a discrete data type that represents planar drawings.

The clue to how to do this is to notice that the vertices along the boundary of each of the faces in Figure 7.1 form a simple cycle. For example, labeling the vertices as in Figure 7.2, the simple cycles for the face boundaries are

`abca` `abda` `bcdb` `acda`.

Since every edge in the drawing appears on the boundaries of exactly two continuous faces, every edge of the simple graph appears on exactly two of the simple cycles.

Vertices around the boundaries of states and countries in an ordinary map are always simple cycles, but oceans are slightly messier. The ocean boundary is the set of all boundaries of islands

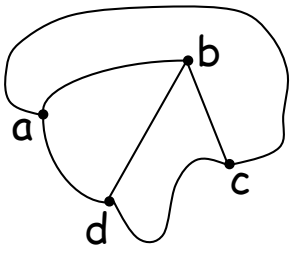


Figure 7.2: The Drawing with Labelled Vertices.

and continents in the ocean; it is a *set* of simple cycles (this can happen for countries too —like Bangladesh). But this happens because islands (and the two parts of Bangladesh) are not connected to each other. So we can dispose of this complication by treating each connected component separately.

But general planar graphs, even when they are connected, may be a bit more complicated than maps. For example a planar graph may have a "bridge," as in Figure 7.3. Now the cycle around

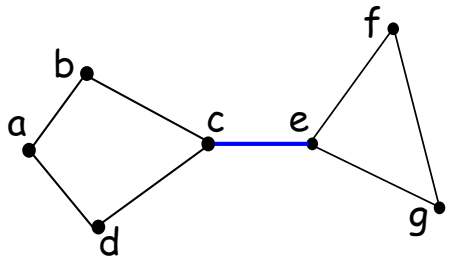


Figure 7.3: A Planar Drawing with a *Bridge*.

the outer face is

`abcefgecda`.

This is not a simple cycle, since it has to traverse the bridge c—e twice.

Planar graphs may also have "dongles," as in Figure 7.4. Now the cycle around the inner face is

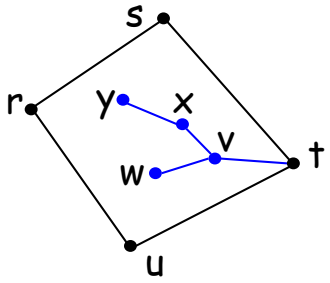


Figure 7.4: A Planar Drawing with a *Dongle*.

`rstvxyxvwvtur`,

because it has to traverse *every* edge of the dongle twice —once "coming" and once "going."

But bridges and dongles are really the only complications, which leads us to the discrete data type of *planar embeddings* that we can use in place of continuous planar drawings. Namely, we'll define a planar embedding recursively to be the set of boundary-tracing cycles we could get drawing one edge after another.

7.3 Planar Embeddings

By thinking of the process of drawing a planar graph edge by edge, we can give a useful recursive definition of planar embeddings.

Definition 7.3.1. A *planar embedding* of a *connected* graph consists of a nonempty set of cycles of the graph called the *discrete faces* of the embedding. Planar embeddings are defined recursively as follows:

- **Base case:** If G is a graph consisting of a single vertex, v, then a planar embedding of G has one discrete face, namely the length zero cycle, v.
- **Constructor Case:** (split a face) Suppose G is a connected graph with a planar embedding, and suppose a and b are distinct, nonadjacent vertices of G that appear on some discrete face, γ, of the planar embedding. That is, γ is a cycle of the form

  $$a \ldots b \cdots a.$$

  Then the graph obtained by adding the edge a—b to the edges of G has a planar embedding with the same discrete faces as G, except that face γ is replaced by the two discrete faces[1]

  $$a \ldots ba \quad \text{and } ab \cdots a,$$

  as illustrated in Figure 7.5.

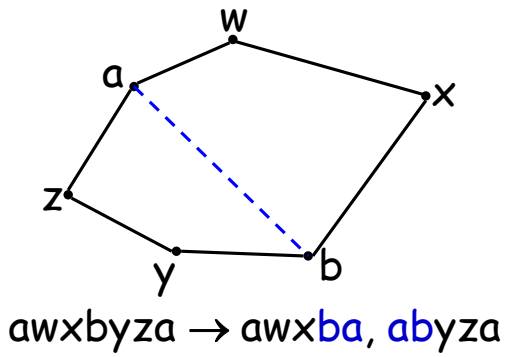


Figure 7.5: The Split a Face Case.

[1] There is one exception to this rule. If G is a line graph beginning with a and ending with b, then the cycles into which γ splits are actually the same. That's because adding edge a—b creates a simple cycle graph, C_n, that divides the plane into an "inner" and an "outer" region with the same border. In order to maintain the correspondence between continuous faces and discrete faces, we have to allow two "copies" of this same cycle to count as discrete faces. But since this is the only situation in which two faces are actually the same cycle, this exception is better explained in a footnote than mentioned explicitly in the definition.

- **Constructor Case:** (add a bridge) Suppose G and H are connected graphs with planar embeddings and disjoint sets of vertices. Let a be a vertex on a discrete face, γ, in the embedding of G. That is, γ is of the form

  $$a \ldots a.$$

  Similarly, let b be a vertex on a discrete face, δ, in the embedding of H, so δ is of the form

  $$b \cdots b.$$

  Then the graph obtained by connecting G and H with a new edge, a—b, has a planar embedding whose discrete faces are the union of the discrete faces of G and H, except that faces γ and δ are replaced by one new face

  $$a \ldots ab \cdots ba.$$

  This is illustrated in Figure 7.6, where the faces of G and H are:

  $$G : \{\texttt{axyza}, \texttt{axya}, \texttt{ayza}\} \qquad H : \{\texttt{btuvwb}, \texttt{btvwb}, \texttt{tuvt}\},$$

  and after adding the bridge a—b, there is a single connected graph with faces

  $$\{\texttt{axyzabtuvwba}, \texttt{axya}, \texttt{ayza}, \texttt{btvwb}, \texttt{tuvt}\}.$$

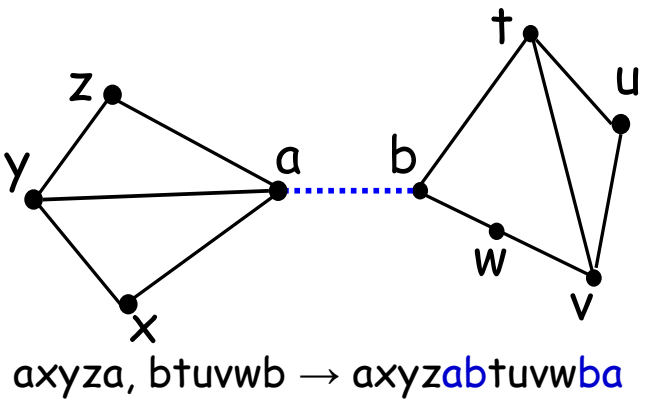


Figure 7.6: The Add Bridge Case.

An arbitrary graph is *planar* iff each of its connected components has a planar embedding.

7.3.1 What outer face?

Notice that the definition of planar embedding does not distinguish an "outer" face. There really isn't any need to distinguish one.

In fact, a planar embedding could be drawn with any given face on the outside. An intuitive explanation of this is to think of drawing the embedding on a *sphere* instead of the plane. Then any face can be made the outside face by "puncturing" that face of the sphere, stretching the puncture hole to an circle around the rest of the faces, and flattening the circular drawing onto the plane.

So pictures that show different "outside" boundaries may actually be illustrations of the same planar embedding.

This is what justifies the "add bridge" case in a planar embedding: whatever face is chosen in the embeddings of each of the disjoint planar graphs, we can draw a bridge between them without needing to cross any other edges in the drawing, because we can assume the bridge connects two "outer" faces.

7.4 Euler's Formula

The value of the recursive definition is that it provides a powerful technique for proving properties of planar graphs, namely, structural induction.

One of the most basic properties of a connected planar graph is that its number of vertices and edges determines the number of faces in every possible planar embedding:

Theorem 7.4.1 (Euler's Formula). *If a connected graph has a planar embedding, then*

$$v - e + f = 2$$

where v is the number of vertices, e is the number of edges, and f is the number of faces.

For example, in Figure 7.1, $|V| = 4$, $|E| = 6$, and $f = 4$. Sure enough, $4 - 6 + 4 = 2$, as Euler's Formula claims.

Proof. The proof is by structural induction on the definition of planar embeddings. Let $P(\mathcal{E})$ be the proposition that $v - e + f = 2$ for an embedding, $\mathcal{E}$.

Base case: ($\mathcal{E}$ is the one vertex planar embedding). By definition, $v = 1$, $e = 0$, and $f = 1$, so $P(\mathcal{E})$ indeed holds.

Constructor case: (split a face) Suppose G is a connected graph with a planar embedding, and suppose a and b are distinct, nonadjacent vertices of G that appear on some discrete face, $\gamma = a \ldots b \cdots a$, of the planar embedding.

Then the graph obtained by adding the edge a—b to the edges of G has a planar embedding with one more face and one more edge than G. So the quantity $v - e + f$ will remain the same for both graphs, and since by structural induction this quantity is 2 for G's embedding, it's also 2 for the embedding of G with the added edge. So P holds for the constructed embedding.

Constructor case: (add bridge) Suppose G and H are connected graphs with planar embeddings and disjoint sets of vertices. Then connecting these two graphs with a cross yields a planar embedding of a connected graph with $v_G + v_H$ vertices, $e_G + e_H + 1$ edges, and $f_G + f_H - 1$ faces. But

$$\begin{aligned}
&(v_G + v_H) - (e_G + e_H + 1) + (f_G + f_H - 1) \\
&= (v_G - e_G + f_G) + (v_H - e_H + f_H) - 2 \\
&= (2) + (2) - 2 && \text{(by structural induction hypothesis)} \\
&= 2.
\end{aligned}$$

So $v - e + f$ remains equal to 2 for the constructed embedding. That is, P also holds in this case.

This completes the proof of the constructor cases, and the theorem follows by structural induction. □

7.4.1 Number of Edges versus Vertices

Like Euler's formula, the following lemmas follow by structural induction directly from the definition of planar embedding.

Lemma 7.4.2. *In a planar embedding of a connected graph, each edge is traversed once by each of two different faces, or is traversed exactly twice by one face.*

Lemma 7.4.3. *In a planar embedding of a graph with at least three vertices, each face is of length at least three.*

Corollary 7.4.4. *Suppose a connected planar graph has $v \geq 3$ vertices and e edges. Then*

$$e \leq 3v - 6.$$

Proof. By definition, a connected graph is planar iff it has a planar embedding. So suppose a connected graph with v vertices and e edges has a planar embedding with f faces. By Lemma 7.4.2, every edge is traversed exactly twice by the face boundaries. So the sum of the lengths of the face boundaries is exactly $2e$. Also by Lemma 7.4.3, when $v \geq 3$, each face boundary is of length at least three, so this sum is at least $3f$. This implies that

$$3f \leq 2e. \tag{7.1}$$

But $f = e - v + 2$ by Euler's formula, and substituting into (7.1) gives

$$\begin{aligned}
3(e - v + 2) &\leq 2e \\
e - 3v + 6 &\leq 0 \\
e &\leq 3v - 6
\end{aligned}$$

□

Corollary 7.4.4 lets us prove that the quadapi can't all shake hands without crossing. Representing quadapi by vertices and the necessary handshakes by edges, we get the complete graph, K_5. Shaking hands without crossing amounts to showing that K_5 is planar. But K_5 is connected, has

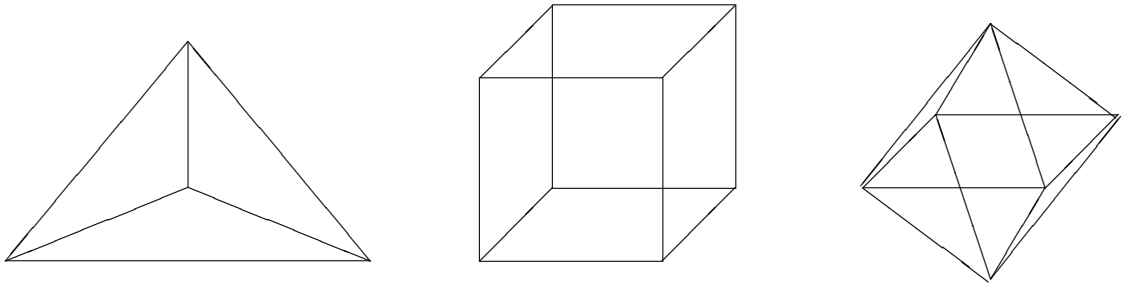
5 vertices and 10 edges, and $10 > 3 \cdot 5 - 6$. This violates the condition of Corollary 7.4.4 required for K_5 to be planar.

Corollary 7.4.4 also leads to an easy proof that every planar graph is 6-colorable, as shown in Problem Set 6. Building on this, it's not too hard to prove that every planar graph is 5-colorable, but we won't go into it. The proof that every planar graph is 4-colorable is one of the celebrated Mathematical results of the latter half of the Twentieth Century. Even after three decades of simplification, it requires graduate-level training and computer support to follow the proof.

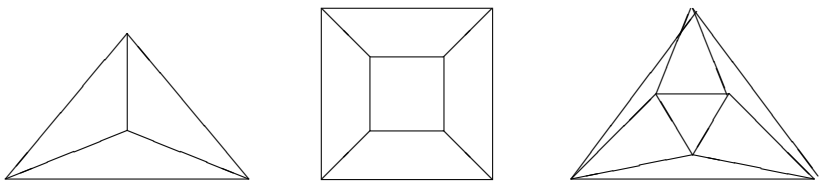
7.5 Classifying Polyhedra

The Pythagoreans had two great mathematical secrets, the irrationality of 2 and a geometric construct that we're about to rediscover!

A ***polyhedron*** is a convex, three-dimensional region bounded by a finite number of polygonal faces. If the faces are identical regular polygons and an equal number of polygons meet at each corner, then the polyhedron is ***regular***. Three examples of regular polyhedra are shown below: the tetrahedron, the cube, and the octahedron.



How many more polyhedra are there? Imagine putting your eye very close to one face of a translucent polyhedron. The edges of that face would ring the periphery of your vision and all other edges would be visible within. For example, the three polyhedra above would look something like this:



Thus, we can regard the corners and edges of these polyhedra as the vertices and edges of a connected planar graph. (This is another logical leap based on geometric intuition.) This means Euler's formula for planar graphs can help guide our search for regular polyhedra.

Let m be the number of faces that meet at each corner of a polyhedron, and let n be the number of sides on each face. In the corresponding planar graph, there are m edges incident to each of the v vertices. Since each edge is incident to two vertices, we know:

$$mv = 2e$$

Also, each face is bounded by n edges. Since each edge is on the boundary of two faces, we have:

$$nf = 2e$$

Solving for v and f in these equations and then substituting into Euler's formula gives:

$$\frac{2e}{m} - e + \frac{2e}{n} = 2$$
$$\frac{1}{m} + \frac{1}{n} = \frac{1}{e} + \frac{1}{2}$$

The last equation places strong restrictions on the structure of a polyhedron. Every nondegenerate polygon has at least 3 sides, so $n \geq 3$. And at least 3 polygons must meet to form a corner, so $m \geq 3$. On the other hand, if either n or m were 6 or more, then the left side of the equation could be at most $1/3 + 1/6 = 1/2$, which is less than the right side. Checking the finitely-many cases that remain turns up only five solutions. For each valid combination of n and m, we can compute the associated number of vertices v, edges e, and faces f. And polyhedra with these properties do actually exist:

n	m	v	e	f	polyhedron
3	3	4	6	4	tetrahedron
4	3	8	12	6	cube
3	4	6	12	8	octahedron
3	5	12	30	20	icosahedron
5	3	20	30	12	dodecahedron

The last polyhedron in this list, the dodecahedron, was the other great mathematical secret of the Pythagorean sect. These five, then, are the only possible regular polyhedra.

So if you want to put more than 20 geocentric satellites in orbit so that they *uniformly* blanket the globe —tough luck!

7.6 In-Class Problems Week 7, Mon.

Problem 7.6.1. Figures 1–4 show different pictures of planar graphs.

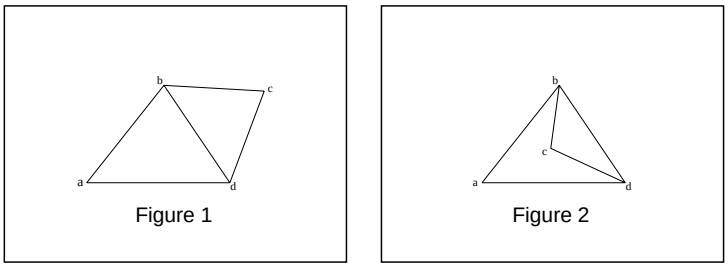


Figure 1

Figure 2

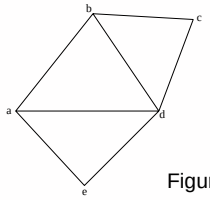


Figure 3

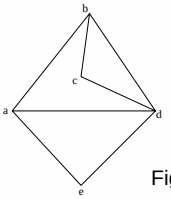


Figure 4

1

(a) For each picture, describe its discrete faces (simple cycles that define the region borders).

Solution. Figs 1 & 2: abda, bcdb, abcda. Fig 3: abcdea, adea,abda,bcdb. Fig 4: abcda, abdea, bdcb, adea. ■

(b) Which of the pictured graphs are isomorphic? Which pictures represent the same *planar embedding*? – that is, they have the same discrete faces.

Solution. Figs 1 & 2 have the same faces, so are different pictures of the *same* planar drawing. Figs 3 & 4 both have four faces, but they are different, for example, Fig 3 has a face with 5 edges, but the longest face in Fig 4 has 4 edges. ■

(c) Describe a way to construct the embedding in Figure 4 according to the recursive definition of planar embedding (in the Appendix).

Solution. Here's one way. By PS6, Problem 1, these steps could be done in any order.

recursive step		faces
vertex a	(base case)	a
vertex b	(base)	b
a—b	(bridge)	aba
vertex c	(base)	c
b—c	(bridge)	$abcba$
vertex d	(base)	d
c—d	(bridge)	$abcdcba$
a—d	(split)	$dabcd$, $dabcd$
b—d	(split)	$dabd$, $dbcd$, $abcda$
vertex e	(base)	e
d—e	(bridge)	$dedabd$, $dbcd$, $abcda$
a—e	(split)	$abdea$, $adea$, $dbcd$, $abcda$

■

Problem 7.6.2. Prove the following assertions by structural induction on the definition of planar embedding.

(a) In a planar embedding of a graph, each edge is traversed a total of two times by the faces of the embedding.

Solution. In the bridge case, the only change is that some face now makes two traversals of a new edge. In the face-splitting case, the only change is that one face splits into two new faces, each traversing the same new edge once. ■

(b) In a planar embedding of a graph with at least three vertices, each face is of length at least three.

Solution. Base case: Check all possible embeddings of 3 vertex graphs.

Constructor case: (bridge) Two faces are replaced by a longer face.

Constructor case: (face-splitting) The new faces are of the from $ab...a$ where the three dots indicate at least one vertex since a and b are not adjacent, so both new faces are of length at least 3; no other faces change. ■

Problem 7.6.3. **(a)** Show that if a connected planar graph with more than two vertices is bipartite, then

$$e \leq 2v - 4. \tag{7.2}$$

Hint: Similar to the proof that $e \leq 3v - 6$ (see the Appendix).

Solution. By Problem7.6.2.b, every face is of length at least 3. But all cycles in a bipartite graph are of even length, and so every face of an embedding must be of length at least 4.

Each edge is traversed by exactly two faces, so

$$2e = \sum_{f \in \text{ faces}} \text{length}(f) \geq \sum_{f \in \text{ faces}} 4 = 4f. \tag{7.3}$$

By Euler's formula, $f = e - v + 2$, so substituting for f in (7.3), yields

$$2e \geq 4(e - v + 2),$$

which simplifies to (7.2). ■

(b) Conclude that that $K_{3,3}$ is not planar. ($K_{3,3}$ is the graph with six vertices and an edge from each of the first three vertices to each of the last three.)

Solution. $K_{3,3}$ is bipartite and connected. Also, it has 9 edges and 6 vertices, and since $9 > 8 = 2 \cdot 6 - 4$, it does not satisfy (7.2), and so cannot be planar. ■

Appendix

Definition 7.6.1. A *planar embedding* of a *connected* graph consists of a nonempty set of cycles of the graph called the *discrete faces* of the embedding. Planar embeddings are defined recursively as follows:

- **Base case:** If G is a graph consisting of a single vertex, v, then a planar embedding of G has one discrete face, namely the length zero cycle, v.
- **Constructor Case:** (split a face) Suppose G is a connected graph with a planar embedding, and suppose a and b are distinct, nonadjacent vertices of G that appear on some discrete face, γ, of the planar embedding. That is, γ is a cycle of the form

  $$a \ldots b \cdots a.$$

  Then the graph obtained by adding the edge a—b to the edges of G has a planar embedding with the same discrete faces as G, except that face γ is replaced by the two discrete faces[2]

  $$a \ldots ba \quad \text{and } ab \cdots a,$$

  as illustrated in Figure 7.7.

[2] There is one exception to this rule. If G is a line graph beginning with a and ending with b, then the cycles into which γ splits are actually the same. That's because adding edge a—b creates a simple cycle graph, C_n, that divides the plane into an "inner" and an "outer" region with the same border. In order to maintain the correspondence between continuous faces and discrete faces, we have to allow two "copies" of this same cycle to count as discrete faces. But since this is the only situation in which two faces are actually the same cycle, this exception is better explained in a footnote than mentioned explicitly in the definition.

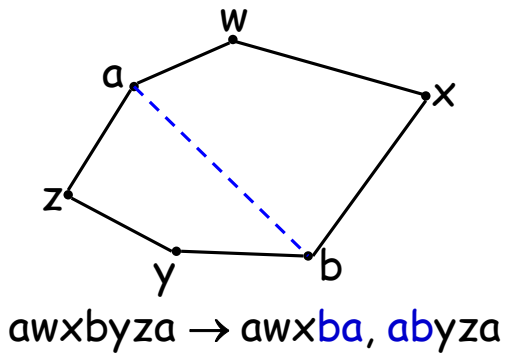


Figure 7.7: The Split a Face Case.

- **Constructor Case:** (add a bridge) Suppose G and H are connected graphs with planar embeddings and disjoint sets of vertices. Let a be a vertex on a discrete face, γ, in the embedding of G. That is, γ is of the form

  $$a \ldots a.$$

  Similarly, let b be a vertex on a discrete face, δ, in the embedding of H, so δ is of the form

  $$b \cdots b.$$

  Then the graph obtained by connecting G and H with a new edge, a—b, has a planar embedding whose discrete faces are the union of the discrete faces of G and H, except that faces γ and δ are replaced by one new face

  $$a \ldots ab \cdots ba,$$

  as illustrated in Figure 7.8.

An arbitrary graph is *planar* iff each of its connected components has a planar embedding.

Theorem 7.6.2 (Euler's Formula). *If a connected graph has a planar embedding, then*

$$v - e + f = 2$$

where v is the number of vertices, e is the number of edges, and f is the number of faces.

Corollary 7.6.3. *Suppose a connected planar graph has $v \geq 3$ vertices and e edges. Then*

$$e \leq 3v - 6.$$

Proof. By definition, a connected graph is planar iff it has a planar embedding. So suppose a connected graph with v vertices and e edges has a planar embedding with f faces. By Problem7.6.2.a, every edge is traversed exactly twice by the face boundaries. So the sum of the lengths of the face

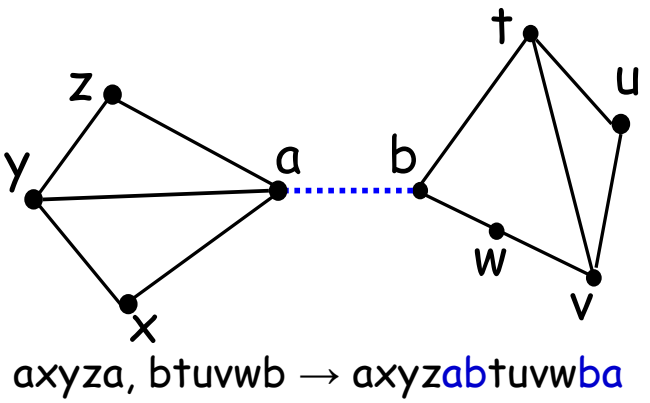


Figure 7.8: The Add Bridge Case.

boundaries is exactly $2e$. Also by Problem7.6.2.b, when $v \geq 3$, each face boundary is of length at least three, so this sum is at least $3f$. This implies that

$$3f \leq 2e. \tag{7.4}$$

But $f = e - v + 2$ by Euler's formula, and substituting into (7.4) gives

$$\begin{aligned} 3(e - v + 2) &\leq 2e \\ e - 3v + 6 &\leq 0 \\ e &\leq 3v - 6 \end{aligned}$$

□

Corollary 7.6.4. *K_5 is not planar.*

Proof.

$$e = 10 > 9 = 3v - 6.$$

□

Chapter 8

Introduction to Number Theory

Number theory is the study of the integers. *Why* anyone would want to study the integers is not immediately obvious. First of all, what's to know? There's 0, there's 1, 2, 3, and so on, along with their negatives. Which one don't you understand? After all, the mathematician G. H. Hardy wrote:

> [Number theorists] may be justified in rejoicing that there is one science, at any rate, and that their own, whose very remoteness from ordinary human activities should keep it gentle and clean.

What most concerned Hardy was that number theory not be used in warfare; he was a pacifist. Good for him, but if number theory is remote from *all* human activity, then why study it? We'll come back to that question later on, but ironically, we'll see that poor Hardy must be turning in his grave.

8.1 Divisibility

We'll be examining integer properties in these notes, so we'll adopt the convention that variables range over integers.

The nature of number theory emerges as soon as we consider the ***divides*** relation

$$a \textbf{ ***divides*** } b \quad \text{iff} \quad ak = b \text{ for some } k.$$

The notation, $a \mid b$, is an abbreviation for "a divides b." If $a \mid b$, then we also say that b is a ***multiple*** of a. As we've seen, a consequence of this definition is that every number divides zero.

This seems simple enough, but let's play with this definition. The Pythagoreans, an ancient sect of mathematical mystics, said that a number is ***perfect*** if it equals the sum of its positive integral divisors, excluding itself. For example, $6 = 1 + 2 + 3$ and $28 = 1 + 2 + 4 + 7 + 14$ are perfect numbers. On the other hand, 10 is not perfect because $1 + 2 + 5 = 8$, and 12 is not perfect because $1+2+3+4+6 = 16$. Euclid characterized all the *even* perfect numbers around 300 BC. But is there an *odd* perfect number? More than two thousand years later, we still don't know! All numbers up

to about 10^{300} have been ruled out, but no one has proved that there isn't an odd perfect number waiting just over the horizon.

So a half-page into number theory, we've strayed past the outer limits of human knowledge! This is pretty typical; number theory is full of questions that are easy to pose, but incredibly difficult to answer. Interestingly, computer scientists have found ways to turn these difficulties to their advantage. Every time you buy a book from Amazon, check your grades on WebSIS, or use a PayPal account, you are relying on number theoretic algorithms.

DON'T PANIC— we're going to stick to some relatively benign parts of number theory. We won't put any of these super-hard unsolved problems on exams!

8.1.1 Facts About Divisibility

The lemma below states some basic facts about divisibility that are *not* difficult to prove:

Lemma 8.1.1. *The following statements about divisibility hold.*

1. *If $a \mid b$, then $a \mid bc$ for all c.*
2. *If $a \mid b$ and $b \mid c$, then $a \mid c$.*
3. *If $a \mid b$ and $a \mid c$, then $a \mid sb + tc$ for all s and t.*
4. *For all $c \neq 0$, $a \mid b$ if and only if $ca \mid cb$.*

Proof. We'll prove only part 2.; the other proofs are similar.

Proof of 2.: Since $a \mid b$, there exists an integer k_1 such that $ak_1 = b$. Since $b \mid c$, there exists an integer k_2 such that $bk_2 = c$. Substituting ak_1 for b in the second equation gives $ak_1k_2 = c$, which implies that $a \mid c$.

□

A number $p > 1$ with no positive divisors other than 1 and itself is called a ***prime***. Every other number greater than 1 is called ***composite***. For example, 2, 3, 5, 7, 11, and 13 are all prime, but 4, 6, 8, and 9 are composite. The number 1 is considered neither prime nor composite. This is just a matter of definition, but reflects the fact that 1 does not behave like a prime in many contexts, such as the Fundamental Theorem of Arithmetic, which we'll come to shortly.

8.1.2 When Divisibility Goes Bad

As you learned in elementary school, if one number does *not* evenly divide another, then there is a "remainder" left over. More precisely, if you divide n by d, then you get a quotient q and a remainder r:

Theorem 8.1.2 (Division Theorem). *Let n and d be integers such that $d > 0$. Then there exists a unique pair of integers q and r such that $n = qd + r$ and $0 \leq r < d$.*[1]

[1] This theorem is often called the "Division Algorithm," even though it is not an algorithm in the modern sense.

Famous Problems in Number Theory

Fermat's Last Theorem Do there exist positive integers x, y, and z such that

$$x^n + y^n = z^n$$

for some integer $n > 2$? In a book he was reading around 1630, Fermat claimed to have a proof, but not enough space in the margin to write it down. Wiles finally gave a proof of the theorem in 1994, after seven years of working in secrecy and isolation in his attic. His proof did not fit in any margin.

Goldbach Conjecture Is every even integer greater than or equal to 4 the sum of two primes? For example, $4 = 2 + 2$, $6 = 3 + 3$, $8 = 3 + 5$, etc. The conjecture holds for all numbers up to 10^{16}. In 1939 Schnirelman proved that every even number can be written as the sum of not more than 300,000 primes, which was a start. Today, we know that every even number is the sum of at most 6 primes.

Twin Prime Conjecture Are there infinitely many primes p such that $p+2$ is also a prime? In 1966 Chen showed that there are infinitely many primes p such that $p + 2$ is the product of at most two primes. So the conjecture is known to be *almost* true!

Primality Testing Is there an efficient way to determine whether n is prime? A naive search for factors of n takes a number of steps exponential in $\log n$, which is the size of n in bits. All known procedures for prime checking blew up like this on various inputs. Finally in 2002, an amazingly simple, new method was discovered by Agrawal, Kayal, and Saxena, which showed that prime testing only required a polynomial number of steps. Their paper began with a quote from Gauss emphasizing the importance and antiquity of the problem even in his time— two centuries ago. So prime testing is definitely not in the category of infeasible problems requiring an exponentially growing number of steps in bad cases.

Factoring Given the product of two large primes $n = pq$, is there an efficient way to recover the primes p and q? The best known algorithm is the "number field sieve", which runs in time proportional to:

$$e^{1.9(\ln n)^{1/3}(\ln \ln n)^{2/3}}$$

This is infeasible when n has 300 digits or more.

As an example, suppose that $a = 10$ and $b = 2716$. Then the quotient is $q = 271$ and the remainder is $r = 6$, since $2716 = 271 \cdot 10 + 6$.

The remainder r in the Division Theorem is denoted $\text{rem}(n, d)$. In other words, $\text{rem}(n, d)$ is the remainder when n is divided by d. For example, $\text{rem}(32, 5)$ is the remainder when 32 is divided by 5, which is 2. Similarly, $\text{rem}(-11, 7) = 3$, since $-11 = (-2) \cdot 7 + 3$. There is a remainder operator built into many programming languages. For example, the expression "32 % 5" evaluates to 2 in Java, C, and C++. However, all these languages treat negative numbers strangely.

We'll take this familiar Division Theorem for granted without proof.

8.2 Die Hard

We've previously looked at the Die Hard water jug problem with jugs of sizes 3 and 5, and 3 and 9. It would be nice if we could solve all these silly water jug questions at once. In particular, how can one form g gallons using jugs with capacities a and b? Here's where number theory comes in handy.

8.2.1 Finding an Invariant Property

Suppose that we have water jugs with capacities a and b. The state of the system is described below with a pair of numbers (x, y), where x is the amount of water in the jug with capacity a and y is the amount in the jug with capacity b. Let's carry out sample operations and see what happens, assuming the b-jug is big enough:

$$\begin{aligned}
(0,0) &\to (a,0) && \text{fill first jug}\\
&\to (0,a) && \text{pour first into second}\\
&\to (a,a) && \text{fill first jug}\\
&\to (2a-b,b) && \text{pour first into second (assuming } 2a \geq b)\\
&\to (2a-b,0) && \text{empty second jug}\\
&\to (0,2a-b) && \text{pour first into second}\\
&\to (a,2a-b) && \text{fill first}\\
&\to (3a-2b,b) && \text{pour first into second (assuming } 3a \geq 2b)
\end{aligned}$$

What leaps out is that at every step, the amount of water in each jug is of the form

$$s \cdot a + t \cdot b \tag{8.1}$$

for some integers s and t. An expression of the form (8.1) is called an ***integer linear combination*** of a and b, but in these notes we'll just call it a ***linear combination***, since we're only talking integers. So we're suggesting:

Lemma 8.2.1. *Suppose that we have water jugs with capacities a and b. Then the amount of water in each jug is always a linear combination of a and b.*

Lemma 8.2.1 is easy to prove by induction on the number of pourings.

But Lemma 8.2.1 isn't very satisfying. We've just managed to recast a pretty understandable question about water jugs into a complicated question about linear combinations. This might not seem like progress. Fortunately, linear combinations are closely related to something more familiar and that will help us solve the water jug problem.

8.3 The Greatest Common Divisor

We've already examined the Euclidean Algorithm for computing $\gcd(a, b)$, the greatest common divisor of a and b. This quantity turns out to be a very valuable piece of information about the relationship between a and b. We'll be making arguments about greatest common divisors all the time.

8.3.1 Linear Combinations and the GCD

The theorem below relates the greatest common divisor to linear combinations. This theorem is *very* useful; take the time to understand it and then remember it!

Theorem 8.3.1. *The greatest common divisor of a and b is equal to the smallest positive linear combination of a and b.*

For example, the greatest common divisor of 52 and 44 is 4. And, sure enough, 4 is a linear combination of 52 and 44:

$$6 \cdot 52 + (-7) \cdot 44 = 4$$

Furthermore, no linear combination of 52 and 44 is equal to a smaller positive integer.

Proof. By the well-ordering principle, there is a smallest positive linear combination of a and b; call it m. We'll prove that $m = \gcd(a, b)$ by showing both $\gcd(a, b) \leq m$ and $m \leq \gcd(a, b)$.

First, we show that $\gcd(a, b) \leq m$. Now any common divisor of a and b, that is, any c such that $c \mid a$ and $c \mid b$ will divide both sa and tb, and therefore also divides $sa + tb$. The $\gcd(a, b)$ is by definition a common divisor of a and b, so

$$\gcd(a, b) \mid sa + tb$$

every s and t. In particular, $\gcd(a, b) \mid m$, which implies that $\gcd(a, b) \leq m$.

Now, we show that $m \leq \gcd(a, b)$. We do this by showing that $m \mid a$. A symmetric argument shows that $m \mid b$, which means that m is a common divisor of a and b. Thus, m must be less than or equal to the *greatest* common divisor of a and b.

All that remains is to show that $m \mid a$. By the Division Algorithm, there exists a quotient q and remainder r such that:

$$a = q \cdot m + r \qquad \text{(where } 0 \leq r < m\text{)}$$

Recall that $m = sa + tb$ for some integers s and t. Substituting in for m and rearranging terms gives:

$$a = q \cdot (sa + tb) + r$$
$$r = (1 - qs)a + (-qt)b$$

We've just expressed r as a linear combination of a and b. However, m is the *smallest* positive linear combination and $0 \leq r < m$. The only possibility is that the remainder r is not positive; that is, $r = 0$. This implies $m \mid a$. □

The proof notes that every linear combination of a and b is a multiple of $\gcd(a, b)$. Conversely, since $\gcd(a, b)$ is a linear combination of a and b, every multiple of $\gcd(a, b)$ is as well. This establishes a corollary:

Corollary 8.3.2. *Every linear combination of a and b is a multiple of $\gcd(a, b)$ and vice versa.*

Now we can restate the water jugs lemma in terms of the greatest common divisor:

Corollary 8.3.3. *Suppose that we have water jugs with capacities a and b. Then the amount of water in each jug is always a multiple of $\gcd(a, b)$.*

For example, there is no way to form 2 gallons using 1247 and 899 gallon jugs, because 2 is not a multiple of $\gcd(1247, 899) = 29$.

8.3.2 Properties of the Greatest Common Divisor

We'll often make use of some basic $\gcd$ facts:

Lemma 8.3.4. *The following statements about the greatest common divisor hold:*

1. *Every common divisor of a and b divides $\gcd(a, b)$.*
2. $\gcd(ka, kb) = k \cdot \gcd(a, b)$ *for all* $k > 0$.
3. *If* $\gcd(a, b) = 1$ *and* $\gcd(a, c) = 1$, *then* $\gcd(a, bc) = 1$.
4. *If* $a \mid bc$ *and* $\gcd(a, b) = 1$, *then* $a \mid c$.
5. $\gcd(a, b) = \gcd(b, \text{rem}(a, b))$.

Here's the trick to proving these statements: translate the $\gcd$ world to the linear combination world using Theorem 8.3.1, argue about linear combinations, and then translate back using Theorem 8.3.1 again.

Proof. We prove only parts 3. and 4.

Proof of 3.: The assumptions together with Theorem 8.3.1 imply that there exist integers s, t, u, and v such that:

$$sa + tb = 1$$
$$ua + vc = 1$$

Multiplying these two equations gives:

$$(sa + tb)(ua + vc) = 1$$

The left side can be rewritten as $a \cdot (asu + btu + csv) + b(ctv)$. This is a linear combination of a and bc that is equal to 1, so $\gcd(a, bc) = 1$ by Theorem 8.3.1.

Proof of 4.: Theorem 8.3.1 says that $\gcd(ac, bc)$ is equal to a linear combination of ac and bc. Now $a \mid ac$ trivially and $a \mid bc$ by assumption. Therefore, a divides *every* linear combination of ac and bc. In particular, a divides $\gcd(ac, bc) = c \cdot \gcd(a, b) = c \cdot 1 = c$. The first equality uses part 2. of this lemma, and the second uses the assumption that $\gcd(a, b) = 1$. □

Lemma 8.3.4, part 5. is the fact we assumed when we proved correctness of the Euclidean Algorithm.

Now let's see if it's possible to make 3 gallons using 21 and 26-gallon jugs. Using Euclid's algorithm:

$$\gcd(26, 21) = \gcd(21, 5) = \gcd(5, 1) = 1.$$

Now 3 is a multiple of 1, so we can't *rule out* the possibility that 3 gallons can be formed. On the other hand, we don't know it can be done.

8.3.3 One Solution for All Water Jug Problems

Can Bruce form 3 gallons using 21 and 26-gallon jugs? This question is not so easy to answer without some number theory.

Corollary 8.3.2 says that 3 can be written as a linear combination of 21 and 26, since 3 is a multiple of $\gcd(21, 26) = 1$. In other words, there exist integers s and t such that:

$$3 = s \cdot 21 + t \cdot 26$$

We don't know what the coefficients s and t are, but we do know that they exist.

Now the coefficient s could be either positive or negative. However, we can readily transform this linear combination into an equivalent linear combination

$$3 = s' \cdot 21 + t' \cdot 26$$

where the coefficient s' is positive. The trick is to notice that if we increase s by 26 in the original equation and decrease t by 21, then the value of the expression $s \cdot 21 + t \cdot 26$ is unchanged overall. Thus, by repeatedly increasing the value of s (by 26 at a time) and decreasing the value of t (by 21 at a time), we get a linear combination $s' \cdot 21 + t' \cdot 26 = 3$ where the coefficient s' is positive. Notice that t' must be negative; otherwise, this expression would be much greater than 3.

Now here's how to form 3 gallons using jugs with capacities 21 and 26:

Repeat s' times:

1. Fill the 21-gallon jug.
2. Pour all the water in the 21-gallon jug into the 26-gallon jug. Whenever the 26-gallon jug becomes full, empty it out.

At the end of this process, there must be exactly 3 gallons in the 26-gallon jug! Here's why: we've taken $s' \cdot 21$ gallons of water from the fountain, we've poured out some multiple of 26 gallons, and in the end the 26-gallon jug holds somewhere between 0 and 26 gallons. Furthermore, we know:

$$s' \cdot 21 + t' \cdot 26 = 3$$

Thus, we must have emptied the 26-gallon jug exactly $-t'$ times; if we had emptied it fewer times, then there would be more than 26 gallons left. And we did not withdraw enough water from the fountain to empty the 26-gallon jug more than $-t'$ times. Thus, by the equation above, there must be exactly 3 gallons left.

Remarkably, we don't even need to know the coefficients s' and t' in order to use this strategy! Instead of repeating the outer loop s' times, we could just repeat *until we obtain 3 gallons*, since that must happen eventually. Of course, we have to keep track of the amounts in the two jugs so we know when we're done. Here's the solution that approach gives:

$$
\begin{array}{llllllll}
(0,0) & \xrightarrow{\text{fill 21}} & (21,0) & \xrightarrow{\text{pour 21 into 26}} & (0,21) & & & \\
 & \xrightarrow{\text{fill 21}} & (21,21) & \xrightarrow{\text{pour 21 into 26}} & (16,26) & \xrightarrow{\text{empty 26}} & (16,0) & \xrightarrow{\text{pour 21 into 26}} \quad (0,16) \\
 & \xrightarrow{\text{fill 21}} & (21,16) & \xrightarrow{\text{pour 21 into 26}} & (11,26) & \xrightarrow{\text{empty 26}} & (11,0) & \xrightarrow{\text{pour 21 into 26}} \quad (0,11) \\
 & \xrightarrow{\text{fill 21}} & (21,11) & \xrightarrow{\text{pour 21 into 26}} & (6,26) & \xrightarrow{\text{empty 26}} & (6,0) & \xrightarrow{\text{pour 21 into 26}} \quad (0,6) \\
 & \xrightarrow{\text{fill 21}} & (21,6) & \xrightarrow{\text{pour 21 into 26}} & (1,26) & \xrightarrow{\text{empty 26}} & (1,0) & \xrightarrow{\text{pour 21 into 26}} \quad (0,1) \\
 & \xrightarrow{\text{fill 21}} & (21,1) & \xrightarrow{\text{pour 21 into 26}} & (0,22) & & & \\
 & \xrightarrow{\text{fill 21}} & (21,22) & \xrightarrow{\text{pour 21 into 26}} & (17,26) & \xrightarrow{\text{empty 26}} & (17,0) & \xrightarrow{\text{pour 21 into 26}} \quad (0,17) \\
 & \xrightarrow{\text{fill 21}} & (21,17) & \xrightarrow{\text{pour 21 into 26}} & (12,26) & \xrightarrow{\text{empty 26}} & (12,0) & \xrightarrow{\text{pour 21 into 26}} \quad (0,12) \\
 & \xrightarrow{\text{fill 21}} & (21,12) & \xrightarrow{\text{pour 21 into 26}} & (7,26) & \xrightarrow{\text{empty 26}} & (7,0) & \xrightarrow{\text{pour 21 into 26}} \quad (0,7) \\
 & \xrightarrow{\text{fill 21}} & (21,7) & \xrightarrow{\text{pour 21 into 26}} & (2,26) & \xrightarrow{\text{empty 26}} & (2,0) & \xrightarrow{\text{pour 21 into 26}} \quad (0,2) \\
 & \xrightarrow{\text{fill 21}} & (21,2) & \xrightarrow{\text{pour 21 into 26}} & (0,23) & & & \\
 & \xrightarrow{\text{fill 21}} & (21,23) & \xrightarrow{\text{pour 21 into 26}} & (18,26) & \xrightarrow{\text{empty 26}} & (18,0) & \xrightarrow{\text{pour 21 into 26}} \quad (0,18) \\
 & \xrightarrow{\text{fill 21}} & (21,18) & \xrightarrow{\text{pour 21 into 26}} & (13,26) & \xrightarrow{\text{empty 26}} & (13,0) & \xrightarrow{\text{pour 21 into 26}} \quad (0,13) \\
 & \xrightarrow{\text{fill 21}} & (21,13) & \xrightarrow{\text{pour 21 into 26}} & (8,26) & \xrightarrow{\text{empty 26}} & (8,0) & \xrightarrow{\text{pour 21 into 26}} \quad (0,8) \\
 & \xrightarrow{\text{fill 21}} & (21,8) & \xrightarrow{\text{pour 21 into 26}} & (3,26) & \xrightarrow{\text{empty 26}} & (3,0) & \xrightarrow{\text{pour 21 into 26}} \quad (0,3)
\end{array}
$$

The same approach works regardless of the jug capacities and even regardless the amount we're trying to produce! Simply repeat these two steps until the desired amount of water is obtained:

1. Fill the smaller jug.
2. Pour all the water in the smaller jug into the larger jug. Whenever the larger jug becomes full, empty it out.

By the same reasoning as before, this method eventually generates every multiple of the greatest common divisor of the jug capacities —all the quantities we can possibly produce. No ingenuity is needed at all!

8.3.4 The Pulverizer

We saw that no matter which pair of integers a and b we are given, there is always a pair of integer coefficients s and t such that

$$\gcd(a, b) = sa + tb.$$

The previous subsection gives a roundabout and not very efficient method of finding such coefficients s and t. In the Notes on State Machines we defined and verified the "Extended Euclidean GCD algorithm" which is a much more efficient way to find these coefficients. In this section we give a more straightforward description of this procedure for finding s and t that dates to sixth-century India, where it was called *kuttak*, which means "The Pulverizer."

Suppose we use Euclid's Algorithm to compute the GCD of 259 and 70, for example:

$$\begin{array}{rcll} \gcd(259, 70) & = & \gcd(70, 49) & \text{since } \mathrm{rem}(259, 70) = 49 \\ & = & \gcd(49, 21) & \text{since } \mathrm{rem}(70, 49) = 21 \\ & = & \gcd(21, 7) & \text{since } \mathrm{rem}(49, 21) = 7 \\ & = & \gcd(7, 0) & \text{since } \mathrm{rem}(21, 7) = 0 \\ & = & 7. & \end{array}$$

The Pulverizer goes through the same steps, but requires some extra bookkeeping along the way: as we compute $\gcd(a, b)$, we keep track of how to write each of the remainders (49, 21, and 7, in the example) as a linear combination of a and b (this is worthwhile, because our objective is to write the last nonzero remainder, which is the GCD, as such a linear combination). For our example, here is this extra bookkeeping:

x	y	$(\mathrm{rem}(x, y))$	$=$	$x - q \cdot y$
259	70	49	$=$	$259 - 3 \cdot 70$
70	49	21	$=$	$70 - 1 \cdot 49$
			$=$	$70 - 1 \cdot (259 - 3 \cdot 70)$
			$=$	$-1 \cdot 259 + 4 \cdot 70$
49	21	7	$=$	$49 - 2 \cdot 21$
			$=$	$(259 - 3 \cdot 70) - 2 \cdot (-1 \cdot 259 + 4 \cdot 70)$
			$=$	$\boxed{3 \cdot 259 - 11 \cdot 70}$
21	7	0		

We began by initializing two variables, $x = a$ and $y = b$. In the first two columns above, we carried out Euclid's algorithm. At each step, we computed $\mathrm{rem}(x, y)$, which can be written in the form $x - q \cdot y$. (Remember that the Division Algorithm says $x = q \cdot y + r$, where r is the remainder. We get $r = x - q \cdot y$ by rearranging terms.) Then we replaced x and y in this equation with equivalent linear combinations of a and b, which we already had computed. After simplifying, we were left with a linear combination of a and b that was equal to the remainder as desired. The final solution is boxed.

8.4 The Fundamental Theorem of Arithmetic

We now have almost enough tools to prove something that you probably already know.

Theorem (Fundamental Theorem of Arithmetic). *Every positive integer n can be written in a unique way as a product of primes:*

$$n = p_1 \cdot p_2 \cdots p_j \qquad (p_1 \leq p_2 \leq \cdots \leq p_j)$$

Notice that the theorem would be false if 1 were considered a prime; for example, 15 could be written as $3 \cdot 5$ or $1 \cdot 3 \cdot 5$ or $1^2 \cdot 3 \cdot 5$. Also, we're relying on a standard convention: the product of an empty set of numbers is defined to be 1, much as the sum of an empty set of numbers is defined to be 0. Without this convention, the theorem would be false for $n = 1$.

There is a certain wonder in the Fundamental Theorem, even if you've known it since you were in a crib. Primes show up erratically in the sequence of integers. In fact, their distribution seems almost random:

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, \ldots$$

Basic questions about this sequence have stumped humanity for centuries. And yet we know that every natural number can be built up from primes in *exactly one way*. These quirky numbers are the building blocks for the integers. The Fundamental Theorem is not hard to prove, but we'll need a couple of preliminary facts.

Lemma 8.4.1. *If p is a prime and $p \mid ab$, then $p \mid a$ or $p \mid b$.*

Proof. The greatest common divisor of a and p must be either 1 or p, since these are the only divisors of p. If $\gcd(a, p) = p$, then the claim holds, because a is a multiple of p. Otherwise, $\gcd(a, p) = 1$ and so $p \mid b$ by part (4) of Lemma 8.3.4. □

A routine induction argument extends this statement to:

Lemma 8.4.2. *Let p be a prime. If $p \mid a_1 a_2 \cdots a_n$, then p divides some a_i.*

Now we're ready to prove the Fundamental Theorem of Arithmetic.

Theorem 8.4.3 (Fundamental Theorem of Arithmetic). *Every positive integer n can be written in a unique way as a product of primes:*

$$n = p_1 \cdot p_2 \cdots p_j \qquad (p_1 \leq p_2 \leq \cdots \leq p_j)$$

Proof. We proved earlier using the well-ordering principle that every positive integer can be expressed as a product of primes. So we just have to prove this expression is unique. We will use the well-ordering principle to prove this too.

The proof is by contradiction: assume, contrary to the claim, that there exist positive integers that can be written as products of primes in more than one way. By the well-ordering principle, there is a smallest integer with this property. Call this integer n, and let

$$\begin{aligned} n &= p_1 \cdot p_2 \cdots p_j \\ &= q_1 \cdot q_2 \cdots q_k \end{aligned}$$

be two of the (possibly many) ways to write n as a product of primes. Then $p_1 \mid n$ and so $p_1 \mid q_1 q_2 \cdots q_k$. Lemma 8.4.2 implies that p_1 divides one of the primes q_i. But since q_i is a prime, it must be that $p_1 = q_i$. Deleting p_1 from the first product and q_i from the second, we find that n/p_1 is a positive integer *smaller* than n that can also be written as a product of primes in two distinct ways. But this contradicts the definition of n as the smallest such positive integer. □

The Prime Number Theorem

Let $\pi(x)$ denote the number of primes less than or equal to x. For example, $\pi(10) = 4$ because 2, 3, 5, and 7 are the primes less than or equal to 10. Primes are very irregularly distributed, so the growth of π is similarly erratic. However, the Prime Number Theorem gives an approximate answer:

$$\lim_{x\to\infty} \frac{\pi(x)}{x/\ln x} = 1$$

Thus, primes gradually taper off. As a rule of thumb, about 1 integer out of every $\ln x$ in the vicinity of x is a prime.

The Prime Number Theorem was conjectured by Legendre in 1798 and proved a century later by de la Vallee Poussin and Hadamard in 1896. However, after his death, a notebook of Gauss was found to contain the same conjecture, which he apparently made in 1791 at age 15. (You sort of have to feel sorry for all the otherwise "great" mathematicans who had the misfortune of being contemporaries of Gauss.)

In late 2004 a billboard appeared in various locations around the country:

$$\left\{ \begin{array}{c} \text{first 10-digit prime found} \\ \text{in consecutive digits of } e \end{array} \right\} \textbf{.com}$$

Substituting the correct number for the expression in curly-braces produced the URL for a Google employment page. The idea was that Google was interested in hiring the sort of people that could and would solve such a problem.

How hard is this problem? Would you have to look through thousands or millions or billions of digits of e to find a 10-digit prime? The rule of thumb derived from the Prime Number Theorem says that among 10-digit numbers, about 1 in

$$\ln 10^{10} \approx 23$$

is prime. This suggests that the problem isn't really so hard! Sure enough, the first 10-digit prime in consecutive digits of e appears quite early:

$$\begin{aligned} e = & 2.7182818284590452353602874713526624977572470936999595749669 \\ & 6762772407663035354759457138217852516642\mathbf{7427466391}9320030 \\ & 5992181741359662904357290033429526059563073813232862794 34\ldots \end{aligned}$$

8.5 Alan Turing

The man pictured above is Alan Turing, the most important figure in the history of computer science. For decades, his fascinating life story was shrouded by government secrecy, societal taboo, and even his own deceptions.

At 24 Turing wrote a paper entitled *On Computable Numbers, with an Application to the Entscheidungsproblem*. The crux of the paper was an elegant way to model a computer in mathematical terms. This was a breakthrough, because it allowed the tools of mathematics to be brought to bear on questions of computation. For example, with his model in hand, Turing immediately proved that there exist problems that no computer can solve— no matter how ingenius the programmer. Turing's paper is all the more remarkable because he wrote it in 1936, a full decade before any electronic computer actually existed.

The word "Entscheidungsproblem" in the title refers to one of the 28 mathematical problems posed by David Hilbert in 1900 as challenges to mathematicians of the 20th century. Turing knocked that one off in the same paper. And perhaps you've heard of the "Church-Turing thesis"? Same paper. So Turing was obviously a brilliant guy who generated lots of amazing ideas. But this lecture is about one of Turing's less-amazing ideas. It involved codes. It involved number theory. And it was sort of stupid.

8.6 Turing's Code

Let's look back to the fall of 1937. Nazi Germany was rearming under Adolf Hitler, world-shattering war looked imminent, and— like us— Alan Turing was pondering the usefulness of number theory. He foresaw that preserving military secrets would be vital in the coming conflict and proposed a way *to encrypt communications using number theory*. This is an idea that has ricocheted up to our own time. Today, number theory is the basis for numerous public-key cryptosystems, digital signature schemes, cryptographic hash functions, and digital cash systems. Every time you buy a book from Amazon, check your grades on WebSIS, or use a PayPal account, you

are relying on number theoretic algorithms. Furthermore, military funding agencies are among the biggest investors in cryptographic research. Sorry Hardy!

Soon after devising his code, Turing disappeared from public view, and half a century would pass before the world learned the full story of where he'd gone and what he did there. We'll come back to Turing's life in a little while; for now, let's investigate the code Turing left behind. The details are uncertain, since he never formally published the idea, so we'll consider a couple of possibilities.

8.6.1 Turing's Code (Version 1.0)

The first challenge is to translate a text message into an integer so we can perform mathematical operations on it. This step is not intended to make a message harder to read, so the details are not too important. Here is one approach: replace each letter of the message with two digits ($A = 01$, $B = 02$, $C = 03$, etc.) and string all the digits together to form one huge number. For example, the message "victory" could be translated this way:

	"v	i	c	t	o	r	y"
$\rightarrow$	22	09	03	20	15	18	25

Turing's code requires the message to be a prime number, so we may need to pad the result with a few more digits to make a prime. In this case, appending the digits 13 gives the number 2209032015182513, which is prime.

Now here is how the encryption process works. In the description below, m is the unencoded message (which we want to keep secret), m^* is the encrypted message (which the Nazis may intercept), and k is the key.

Beforehand The sender and receiver agree on a secret key, which is a large prime k.

Encryption The sender encrypts the message m by computing:

$$m^* = m \cdot k$$

Decryption The receiver decrypts m^* by computing:

$$\frac{m^*}{k} = \frac{m \cdot k}{k} = m$$

For example, suppose that the secret key is the prime number $k = 22801763489$ and the message m is "victory". Then the encrypted message is:

$$\begin{aligned} m^* &= m \cdot k \\ &= 2209032015182513 \cdot 22801763489 \\ &= 50369825549820718594667857 \end{aligned}$$

There are a couple of questions that one might naturally ask about Turing's code.

1. How can the sender and receiver ensure that m and k are prime numbers, as required?

   The general problem of determining whether a large number is prime or composite has been studied for centuries, and reasonably good primality tests were known even in Turing's time. In 2002, Manindra Agrawal, Neeraj Kayal, and Nitin Saxena announced a primality test that is guaranteed to work on a number n in about $(\log n)^{12}$ steps, that is, a number of steps bounded by a twelfth degree polynomial in the length (in bits) of the input, n. This definitively places primality testing way below the problems of exponentially difficulty. Amazingly, the description of their breakthrough algorithm was only thirteen lines long!

   Of course, a twelfth degree polynomial grows pretty fast, so the Agrawal, *et al.* procedure is of no practical use. Still, good ideas have a way of breeding more good ideas, so there's certainly hope further improvements will lead to a procedure that is useful in practice. But the truth is, there's no practical need to improve it, since very efficient *probabilistic* procedures for prime-testing have been known since the early 1970's. These procedures have some probability of giving a wrong answer, but their probability of being wrong is so tiny that betting on their answers is the best bet you'll ever make.

2. Is Turing's code secure?

   The Nazis see only the encrypted message $m^* = m \cdot k$, so recovering the original message m requires factoring m^*. Despite immense efforts, no really efficient factoring algorithm has ever been found. It appears to be a fundamentally difficult problem, though a breakthrough someday is not impossible. In effect, Turing's code puts to practical use his discovery that there are limits to the power of computation. Thus, provided m and k are sufficiently large, the Nazis seem to be out of luck!

This all sounds promising, but there is a major flaw in Turing's code.

8.6.2 Breaking Turing's Code

Let's consider what happens when the sender transmits a *second* message using Turing's code and the same key. This gives the Nazis two encrypted messages to look at:

$$m_1^* = m_1 \cdot k \qquad \text{and} \qquad m_2^* = m_2 \cdot k$$

The greatest common divisor of the two encrypted messages, m_1^* and m_2^*, is the secret key k. And, as we've seen, the $\gcd$ of two numbers can be computed very efficiently. So after the second message is sent, the Nazis can recover the secret key and read *every* message!

It is difficult to believe a mathematician as brilliant as Turing could overlook such a glaring problem. One possible explanation is that he had a slightly different system in mind, one based on *modular* arithmetic.

8.7 Modular Arithmetic

On page 1 of his masterpiece on number theory, *Disquisitiones Arithmeticae*, Gauss introduced the notion of "congruence". Now, Gauss is another guy who managed to cough up a half-decent idea

every now and then, so let's take a look at this one. Gauss said that a ***is congruent to*** b ***modulo*** n iff $n \mid (a-b)$. This is denoted $a \equiv b \pmod{n}$. For example:

$$29 \equiv 15 \pmod{7} \quad \text{because } 7 \mid (29-15).$$

There is a close connection between congruences and remainders:

Lemma 8.7.1 (Congruences and Remainders).

$$a \equiv b \pmod{n} \quad \textit{iff} \quad \text{rem}(a,n) = \text{rem}(b,n).$$

Proof. By the Division Theorem, there exist unique pairs of integers q_1, r_1 and q_2, r_2 such that:

$$\begin{aligned} a &= q_1 n + r_1 && \text{where } 0 \le r_1 < n, \\ b &= q_2 n + r_2 && \text{where } 0 \le r_2 < n. \end{aligned}$$

In these terms, $\text{rem}(a,n) = r_1$ and $\text{rem}(b,n) = r_2$. Subtracting the second equation from the first gives:

$$a - b = (q_1 - q_2)n + (r_1 - r_2) \qquad \text{where } -n < r_1 - r_2 < n.$$

Now $a \equiv b \pmod{n}$ if and only if n divides the left side. This is true if and only if n divides the right side, which holds if and only if $r_1 - r_2$ is a multiple of n. Given the bounds on $r_1 - r_2$, this happens precisely when $r_1 = r_2$, which is equivalent to $\text{rem}(a,n) = \text{rem}(b,n)$. □

So we can also see that

$$29 \equiv 15 \pmod{7} \quad \text{because } \text{rem}(29,7) = 1 = \text{rem}(15,7).$$

This formulation explains why the congruence relation has properties like an equality relation. Notice that even though (mod 7) appears over on the right side the $\equiv$ symbol, it is in no sense more strongly associated with the 15 than the 29. It would really be clearer to write $29 \equiv_{\text{mod } 7} 15$ for example, but the notation with the modulus at the end is firmly entrenched and we'll stick to it.

We'll make frequent use of the following immediate Corollary of Lemma 8.7.1:

Corollary 8.7.2.

$$a \equiv \text{rem}(a,n) \pmod{n}$$

Still another way to think about congruence modulo n is that it *defines a partition of the integers into n sets so that congruent numbers are all in the same set*. For example, suppose that we're working modulo 3. Then we can partition the integers into 3 sets as follows:

$$\begin{array}{llllllll} \{ & \ldots, & -6, & -3, & 0, & 3, & 6, & 9, & \ldots & \} \\ \{ & \ldots, & -5, & -2, & 1, & 4, & 7, & 10, & \ldots & \} \\ \{ & \ldots, & -4, & -1, & 2, & 5, & 8, & 11, & \ldots & \} \end{array}$$

according to whether their remainders on division by 3 are 0, 1, or 2. The upshot is that when arithmetic is done modulo n there are really only n different kinds of numbers to worry about, because there are only n possible remainders. In this sense, modular arithmetic is a simplification of ordinary arithmetic and thus is a good reasoning tool.

There are many useful facts about congruences, some of which are listed in the lemma below. The overall theme is that *congruences work a lot like equations*, though there are a couple of exceptions.

Lemma 8.7.3 (Facts About Congruences). *The following hold for $n \geq 1$:*

1. $a \equiv a \pmod{n}$
2. $a \equiv b \pmod{n}$ *implies* $b \equiv a \pmod{n}$
3. $a \equiv b \pmod{n}$ *and* $b \equiv c \pmod{n}$ *implies* $a \equiv c \pmod{n}$
4. $a \equiv b \pmod{n}$ *implies* $a + c \equiv b + c \pmod{n}$
5. $a \equiv b \pmod{n}$ *implies* $ac \equiv bc \pmod{n}$
6. $a \equiv b \pmod{n}$ *and* $c \equiv d \pmod{n}$ *imply* $a + c \equiv b + d \pmod{n}$
7. $a \equiv b \pmod{n}$ *and* $c \equiv d \pmod{n}$ *imply* $ac \equiv bd \pmod{n}$

Proof. Parts 1.–3. follow immediately from Lemma 8.7.1. Part 4. follows immediately from the definition that $a \equiv b \pmod{n}$ iff $n \mid (a - b)$. Likewise, part 5. follows because if $n \mid (a - b)$ then it divides $(a - b)c = ac - bc$. To prove part 6., assume

$$a \equiv b \pmod{n} \tag{8.2}$$

and

$$c \equiv d \pmod{n}. \tag{8.3}$$

Then

$$\begin{aligned}
a + c &\equiv b + c \pmod{n} && \text{(by part 4. and (8.2)),}\\
c + b &\equiv d + b \pmod{n} && \text{(by part 4. and (8.3)), so}\\
b + c &\equiv b + d \pmod{n} && \text{and therefore}\\
a + c &\equiv b + d \pmod{n} && \text{(by part 3.)}
\end{aligned}$$

Part 7. has a similar proof. □

8.8 Turing's Code (Version 2.0)

In 1940 France had fallen before Hitler's army, and Britain alone stood against the Nazis in western Europe. British resistance depended on a steady flow of supplies brought across the north Atlantic from the United States by convoys of ships. These convoys were engaged in a cat-and-mouse game with German "U-boats" —submarines —which prowled the Atlantic, trying to sink supply ships and starve Britain into submission. The outcome of this struggle pivoted on a balance of information: could the Germans locate convoys better than the Allies could locate U-boats or vice versa?

Germany lost.

But a critical reason behind Germany's loss was made public only in 1974: the British had broken Germany's naval code, Enigma. Through much of the war, the Allies were able to route convoys around German submarines by listening into German communications. The British government didn't explain *how* Enigma was broken until 1996. When the analysis was finally released (by

the US), the author was none other than Alan Turing. In 1939 he had joined the secret British codebreaking effort at Bletchley Park. There, he played a central role in cracking the German's Enigma code and thus in preventing Britain from falling into Hitler's hands.

Governments are always tight-lipped about cryptography, but the half-century of official silence about Turing's role in breaking Enigma and saving Britain may be related to some disturbing events after the war.

Let's consider an alternative interpretation of Turing's code. Perhaps we had the basic idea right (multiply the message by the key), but erred in using *conventional* arithmetic instead of *modular* arithmetic. Maybe this is what Turing meant:

Beforehand The sender and receiver agree on a large prime p, which may be made public. (This will be the modulus for all our arithmetic.) They also agree on a secret key $k \in \{1, 2, \ldots, p-1\}$.

Encryption The message m can be any integer in the set $\{0, 1, 2, \ldots, p-1\}$; in particular, the message is no longer required to be a prime. The sender encrypts the message m to produce m^* by computing:

$$m^* = \text{rem}(mk, p) \tag{8.4}$$

Decryption (Uh-oh.)

The decryption step is a problem. We might hope to decrypt in the same way as before: by dividing the encrypted message m^* by the key k. The difficulty is that $m*$ is the *remainder* when mk is divided by p. So dividing m^* by k might not even give us an integer!

This decoding difficulty can be overcome with a better understanding of arithmetic modulo a prime.

8.8.1 Multiplicative Inverses

The ***multiplicative inverse*** of a number x is another number x^{-1} such that:

$$x \cdot x^{-1} = 1$$

Generally, multiplicative inverses exist over the real numbers. For example, the multiplicative inverse of 3 is $1/3$ since:

$$3 \cdot \frac{1}{3} = 1$$

The sole exception is that 0 does not have an inverse.

On the other hand, inverses generally do not exist over the integers. For example, 7 can not be multiplied by another integer to give 1.

Surprisingly, multiplicative inverses do exist when we're working *modulo a prime number*. For example, if we're working modulo 5, then 3 is a multiplicative inverse of 7, since:

$$7 \cdot 3 \equiv 1 \pmod 5$$

(All numbers congruent to 3 modulo 5 are also multiplicative inverses of 7; for example, $7 \cdot 8 \equiv 1 \pmod 5$ as well.) The only exception is that numbers congruent to 0 modulo 5 (that is, the multiples of 5) do not have inverses, much as 0 does not have an inverse over the real numbers. Let's prove this.

Lemma 8.8.1. *If p is prime and k is not a multiple of p, then k has a multiplicative inverse.*

Proof. Since p is prime, it has only two divisors: 1 and p. And since k is not a multiple of p, we must have $\gcd(p, k) = 1$. Therefore, there is a linear combination of p and k equal to 1:

$$sp + tk = 1$$

Rearranging terms gives:

$$sp = 1 - tk$$

This implies that $p \mid (1 - tk)$ by the definition of divisibility, and therefore $tk \equiv 1 \pmod{p}$ by the definition of congruence. Thus, t is a multiplicative inverse of k. □

Multiplicative inverses are the key to decryption in Turing's code. Specifically, we can recover the original message by multiplying the encoded message by the *inverse* of the key:

$$\begin{aligned} m^* \cdot k^{-1} &= \operatorname{rem}(mk, p) \cdot k^{-1} && \text{(def. (8.4) of } m^*) \\ &\equiv (mk)k^{-1} \pmod{p} && \text{(by Cor. 8.7.2)} \\ &\equiv m \pmod{p}. \end{aligned}$$

This shows that m^*k^{-1} is congruent to the original message m. Since m was in the range $0, 1, \ldots, p-1$, we can recover it exactly by taking a remainder:

$$m = \operatorname{rem}(m^*k^{-1}, p)$$

So now we can decrypt!

8.8.2 Cancellation

Another sense in which real numbers are nice is that one can cancel multiplicative terms. In other words, if we know that $m_1k = m_2k$, then we can cancel the k's and conclude that $m_1 = m_2$, provided $k \neq 0$. In general, cancellation is *not* valid in modular arithmetic. For example, this congruence is correct:

$$2 \cdot 3 \equiv 4 \cdot 3 \pmod{6}$$

But if we cancel the 3's, we reach a false conclusion:

$$2 \equiv 4 \pmod{6}$$

The fact that multiplicative terms can not be cancelled is the most significant sense in which congruences differ from ordinary equations. However, this difference goes away if we're working modulo a *prime*; then cancellation is valid.

Lemma 8.8.2. *Suppose p is a prime and k is not a multiple of p. Then*

$$ak \equiv bk \pmod{p} \qquad \textit{implies} \qquad a \equiv b \pmod{p}$$

Proof. Multiply both sides of the congruence by k^{-1}. □

We can use this lemma to get a bit more insight into how Turing's code works. In particular, the encryption operation in Turing's code *permutes the set of possible messages*. This is stated more precisely in the following corollary.

Corollary 8.8.3. *Suppose p is a prime and k is not a multiple of p. Then the sequence:*

$$\text{rem}((0 \cdot k), p), \quad \text{rem}((1 \cdot k), p), \quad \text{rem}((2 \cdot k), p), \quad \ldots, \quad \text{rem}(((p-1) \cdot k), p)$$

is a permutation[2] *of the sequence:*

$$0, \quad 1, \quad 2, \quad \ldots, \quad (p-1)$$

This remains true if the first term is deleted from each sequence.

Proof. The first sequence contains p numbers, which are all in the range 0 to $p-1$ by the definition of remainder. Furthermore, the numbers in the first sequence are all different; by Lemma 8.8.2, $ik \equiv jk \pmod{p}$ if and only if $i \equiv j \pmod{p}$, and no two numbers in the range 0, 1, ..., p - 1 are congruent modulo p. Thus, the first sequence must contain *all* of the numbers from 0 to $p-1$ in some order. The claim remains true if the first terms are deleted, because both sequences begin with 0. □

For example, suppose $p = 5$ and $k = 3$. Then the sequence:

$$\underbrace{\text{rem}((0 \cdot 3), 5)}_{=0}, \quad \underbrace{\text{rem}((1 \cdot 3), 5)}_{=3}, \quad \underbrace{\text{rem}((2 \cdot 3), 5)}_{=1}, \quad \underbrace{\text{rem}((3 \cdot 3), 5)}_{=4}, \quad \underbrace{\text{rem}((4 \cdot 3), 5)}_{=2}$$

is a permutation of 0, 1, 2, 3, 4 and the last four terms are a permutation of 1, 2, 3, 4. As long as the Nazis don't know the secret key k, they don't know how the set of possible messages are permuted by the process of encryption and thus can't read encoded messages.

8.8.3 Fermat's Theorem

A remaining challenge in using Turing's code is that decryption requires the inverse of the secret key k. An effective way to calculate k^{-1} follows from the proof of Lemma 8.8.1: $k^{-1} = \text{rem}(t, p)$ where s, t are coefficients such that $sp + tk = 1$. Notice that t is easy to find using the Pulverizer

An alternative approach, about equally efficient and probably more memorable, is to rely on Fermat's Theorem, which is much easier than his famous Last Theorem —and more useful.

Theorem 8.8.4 (Fermat's Theorem). *Suppose p is a prime and k is not a multiple of p. Then:*

$$k^{p-1} \equiv 1 \pmod{p}$$

[2] A *permutation* of a sequence of elements is a sequence with the same elements (including repeats) possibly in a different order. More formally, if

$$\vec{e} ::= e_1, e_2, \ldots, e_n$$

is a length n sequence, and $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ is a bijection, then

$$e_{\pi(1)}, e_{\pi(2)}, \ldots, e_{\pi(n)},$$

is a *permutation* of $\vec{e}$.

Proof. We reason as follows:

$$\begin{aligned} 1 \cdot 2 \cdots (p-1) &= \text{rem}(k,p) \cdot \text{rem}(2k,p) \cdots \text{rem}((p-1)k,p) && \text{(by Cor 8.8.3)} \\ &\equiv k \cdot 2k \cdots (p-1)k \pmod{p} && \text{(by Cor 8.7.2)} \\ &\equiv (p-1)! \cdot k^{p-1} \pmod{p} && \text{(rearranging terms)} \end{aligned}$$

Now $(p-1)!$ can not be a multiple of p, because the prime factorizations of $1, 2, \ldots, (p-1)$ contain only primes smaller than p. Therefore, we can cancel $(p-1)!$ from the first expression and the last by Lemma 8.8.2, which proves the claim. □

Here is how we can find inverses using Fermat's Theorem. Suppose p is a prime and k is not a multiple of p. Then, by Fermat's Theorem, we know that:

$$k^{p-2} \cdot k \equiv 1 \pmod{p}$$

Therefore, k^{p-2} must be a multiplicative inverse of k. For example, suppose that we want the multiplicative inverse of 6 modulo 17. Then we need to compute $\text{rem}(6^{15}, 17)$, which we can do by successive squaring. All the congruences below hold modulo 17.

$$\begin{aligned} 6^2 &\equiv 36 \equiv 2 \\ 6^4 &\equiv (6^2)^2 \equiv 2^2 \equiv 4 \\ 6^8 &\equiv (6^4)^2 \equiv 4^2 \equiv 16 \\ 6^{15} &\equiv 6^8 \cdot 6^4 \cdot 6^2 \cdot 6 \equiv 16 \cdot 4 \cdot 2 \cdot 6 \equiv 3 \end{aligned}$$

Therefore, $\text{rem}(6^{15}, 17) = 3$. Sure enough, 3 is the multiplicative inverse of 6 modulo 17, since:

$$3 \cdot 6 \equiv 1 \pmod{17}$$

In general, if we were working modulo a prime p, finding a multiplicative inverse by trying every value between 1 and $p-1$ would require about p operations. However, the approach above requires only about $\log p$ operations, which is far better when p is large.

8.8.4 Breaking Turing's Code— Again

The Germans didn't bother to encrypt their weather reports with the highly-secure Enigma system. After all, so what if the Allies learned that there was rain off the south coast of Iceland? But, amazingly, this practice provided the British with a critical edge in the Atlantic naval battle during 1941.

The problem was that some of those weather reports had originally been transmitted from U-boats out in the Atlantic. Thus, the British obtained both unencrypted reports and the same reports encrypted with Enigma. By comparing the two, the British were able to determine which key the Germans were using that day and could read all other Enigma-encoded traffic. Today, this would be called a ***known-plaintext attack***.

Let's see how a known-plaintext attack would work against Turing's code. Suppose that the Nazis know both m and m^* where:

$$m^* \equiv mk \pmod{p}$$

Now they can compute:

$$\begin{aligned} m^{p-2} \cdot m^* &= m^{p-2} \cdot \text{rem}(mk, p) && \text{(def. (8.4) of } m^*) \\ &\equiv m^{p-2} \cdot mk \pmod{p} && \text{(by Cor 8.7.2)} \\ &\equiv m^{p-1} \cdot k \pmod{p} \\ &\equiv k \pmod{p} && \text{(Fermat's Theorem)} \end{aligned}$$

Now the Nazis have the secret key k and can decrypt any message!

This is a huge vulnerability, so Turing's code has no practical value. Fortunately, Turing got better at cryptography after devising this code; his subsequent cracking of Enigma surely saved thousands of lives, if not the whole of Britain.

8.9 Turing Postscript

A few years after the war, Turing's home was robbed. Detectives soon determined that a former homosexual lover of Turing's had conspired in the robbery. So they arrested him —that is, they arrested Alan Turing —because, at that time, homosexuality was a crime in Britain, punishable by up to two years in prison. Turing was sentenced to a humiliating hormonal "treatment" for his homosexuality: he was given estrogen injections. He began to develop breasts.

Three years later, Alan Turing, the founder of computer science, was dead. His mother explained what happened in a biography of her own son. Despite her repeated warnings, Turing carried out chemistry experiments in his own home. Apparently, her worst fear was realized: by working with potassium cyanide while eating an apple, he poisoned himself.

However, Turing remained a puzzle to the very end. His mother was a devoutly religious woman who considered suicide a sin. And, other biographers have pointed out, Turing had previously discussed committing suicide by eating a poisoned apple. Evidently, Alan Turing, who founded computer science and saved his country, took his own life in the end, and in just such a way that his mother could believe it was an accident.

8.10 Arithmetic with an Arbitrary Modulus

Turing's code did not work as he hoped. However, his essential idea— using number theory as the basis for cryptography— succeeded spectacularly in the decades after his death.

In 1977, Ronald Rivest, Adi Shamir, and Leonard Adleman at MIT proposed a highly secure cryptosystem (called **RSA**) based on number theory. Despite decades of attack, no significant weakness has been found. Moreover, RSA has a major advantage over traditional codes: the sender and receiver of an encrypted message need not meet beforehand to agree on a secret key. Rather, the receiver has both a ***secret key***, which she guards closely, and a ***public key***, which she distributes as widely as possible. To send her a message, one encrypts using her widely-distributed public

The Riemann Hypothesis

Turing's last project before he disappeared from public view in 1939 involved the construction of an elaborate mechanical device to test a mathematical conjecture called the Riemann Hypothesis. This conjecture first appeared in a sketchy paper by Berhard Riemann in 1859 and is now one of the most famous unsolved problem in mathematics. The formula for the sum of an infinite geometric series says:

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

Substituting $x = \frac{1}{2^s}$, $x = \frac{1}{3^s}$, $x = \frac{1}{5^s}$, and so on for each prime number gives a sequence of equations:

$$1 + \frac{1}{2^s} + \frac{1}{2^{2s}} + \frac{1}{2^{3s}} + \cdots = \frac{1}{1-1/2^s}$$
$$1 + \frac{1}{3^s} + \frac{1}{3^{2s}} + \frac{1}{3^{3s}} + \cdots = \frac{1}{1-1/3^s}$$
$$1 + \frac{1}{5^s} + \frac{1}{5^{2s}} + \frac{1}{5^{3s}} + \cdots = \frac{1}{1-1/5^s}$$

etc.

Multiplying together all the left sides and all the right sides gives:

$$\sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{p \in \text{primes}} \left(\frac{1}{1-1/p^s} \right)$$

The sum on the left is obtained by multiplying out all the infinite series and applying the Fundamental Theorem of Arithmetic. For example, the term $1/300^s$ in the sum is obtained by multiplying $1/2^{2s}$ from the first equation by $1/3^s$ in the second and $1/5^{2s}$ in the third. Riemann noted that every prime appears in the expression on the right. So he proposed to learn about the primes by studying the equivalent, but simpler expression on the left. In particular, he regarded s as a complex number and the left side as a function, $\zeta(s)$. Riemann found that the distribution of primes is related to values of s for which $\zeta(s) = 0$, which led to his famous conjecture:

> *The Riemann Hypothesis*: Every nontrivial zero of the zeta function $\zeta(s)$ lies on the line $s = 1/2 + ci$ in the complex plane.

Researchers continue to work intensely to settle this conjecture, as they have for over a century. A proof would immediately imply, among other things, a strong form of the Prime Number Theorem— and earn the prover a $1 million prize! (We're not sure what the cash would be for a counter-example, but the discoverer would be wildly applauded by mathematicians everywhere.)

key. Then she decrypts the message using her closely-held private key. The use of such a ***public key cryptography*** system allows you and Amazon, for example, to engage in a secure transaction without meeting up beforehand in a dark alley to exchange a key.

Interestingly, RSA does not operate modulo a prime, as Turing's scheme may have, but rather modulo the product of *two* large primes. Thus, we'll need to know a bit about how arithmetic works modulo a composite number in order to understand RSA. Arithmetic modulo an arbitrary positive integer is really only a little more painful than working modulo a prime, in the same sense that a doctor says "This is only going to hurt a little" before he jams a big needle in your arm.

8.10.1 Relative Primality and Phi

First, we need a new definition. Integers a and b are ***relatively prime*** iff $\gcd(a, b) = 1$. For example, 8 and 15 are relatively prime, since $\gcd(8, 15) = 1$. Note that every integer is relatively prime to a genuine prime number p, except for multiples of p.

We'll also need a certain function that is defined using relative primality. Let n be a positive integer. Then $\phi(n)$ denotes the number of integers in $\{1, 2, \ldots, n-1\}$ that are relatively prime to n. For example, $\phi(7) = 6$, since 1, 2, 3, 4, 5, and 6 are all relatively prime to 7. Similarly, $\phi(12) = 4$, since only 1, 5, 7, and 11 are relatively prime to 12. If you know the prime factorization of n, then computing $\phi(n)$ is a piece of cake, thanks to the following theorem.

Theorem 8.10.1. *The function ϕ obeys the following relationships:*

(a) If a and b are relatively prime, then $\phi(ab) = \phi(a)\phi(b)$.

(b) If p is a prime, then $\phi(p^k) = p^k - p^{k-1}$ for $k \geq 1$.

A proof of Theorem 8.10.1 appears in Problem Set 7, and we'll give another proof in a few weeks after we've developed a few principles for counting things. In the meanwhile, here's an example of using Theorem 8.10.1 to compute $\phi(300)$:

$$\begin{aligned}
\phi(300) &= \phi(2^2 \cdot 3 \cdot 5^2) \\
&= \phi(2^2) \cdot \phi(3) \cdot \phi(5^2) && \text{(by Theorem 8.10.1.(a))} \\
&= (2^2 - 2^1)(3^1 - 3^0)(5^2 - 5^1) && \text{(by Theorem 8.10.1.(b))} \\
&= 80.
\end{aligned}$$

8.10.2 Generalizing to an Arbitrary Modulus

Let's generalize what we know about arithmetic modulo a prime. Now, instead of working modulo a prime p, we'll work modulo an arbitrary positive integer n. The basic theme is that arithmetic modulo n may be complicated, but the integers *relatively prime* to n remain fairly well-behaved. For example, the proof of Lemma 8.8.1 of an inverse for k modulo p extends to an inverse for k relatively prime to n:

Lemma 8.10.2. *Let n be a positive integer. If k is relatively prime to n, then there exists an integer k^{-1} such that:*

$$k \cdot k^{-1} \equiv 1 \pmod{n}$$

As a consequence of this lemma, we can cancel a multiplicative term from both sides of a congruence if that term is relatively prime to the modulus:

Corollary 8.10.3. *Suppose n is a positive integer and k is relatively prime to n. If*

$$ak \equiv bk \pmod{n}$$

then

$$a \equiv b \pmod{n}$$

This holds because we can multiply both sides of the first congruence by k^{-1} and simplify to obtain the second.

8.10.3 Euler's Theorem

RSA essentially relies on Euler's Theorem, a generalization of Fermat's Theorem to an arbitrary modulus. The proof is much like the proof of Fermat's Theorem, except that we focus on integers relatively prime to the modulus. Let's start with a lemma.

Lemma 8.10.4. *Suppose n is a positive integer and k is relatively prime to n. Let $k_1, \ldots, k_r$ denote all the integers relatively prime to n in the range $0 \leq k_i < n$. Then the sequence:*

$$\text{rem}(k_1 \cdot k, n), \quad \text{rem}(k_2 \cdot k, n), \quad \text{rem}(k_3 \cdot k, n), \quad \ldots \quad , \text{rem}(k_r \cdot k, n)$$

is a permutation of the sequence:

$$k_1, \quad k_2, \quad \ldots \quad , k_r.$$

Proof. We will show that the numbers in the first sequence are all distinct and all appear in the second sequence. Since the two sequences have the same length, the first must be a permutation of the second.

First, we show that the numbers in the first sequence are all distinct. Suppose that $\text{rem}(k_i k, n) = \text{rem}(k_j k, n)$. This is equivalent to $k_i k \equiv k_j k \pmod{n}$, which implies $k_i \equiv k_j \pmod{n}$ by Corollary 8.10.3. This, in turn, means that $k_i = k_j$ since both are between 1 and $n-1$. Thus, a term in the first sequence is not equal to any other term.

Next, we show that each number in the first sequence appears in the second. By assumption, $\gcd(k_i, n) = 1$ and $\gcd(k, n) = 1$, which means that

$$\begin{aligned} \gcd(n, \text{rem}(k_i k, n)) &= \gcd(k_i k, n) && \text{(by Lemma 8.3.4.5)} \\ &= 1 && \text{(by Lemma 8.3.4.3).} \end{aligned}$$

So $\text{rem}(k_i k, n)$ is relatively prime to n and is in the range from 0 to $n-1$ by the definition of remainder. The second sequence is defined to consist of all such integers. □

We can now prove Euler's Theorem:

Theorem 8.10.5 (Euler's Theorem). *Suppose n is a positive integer and k is relatively prime to n. Then*

$$k^{\phi(n)} \equiv 1 \pmod{n}$$

Proof. Let $k_1, \ldots, k_r$ denote all integers relatively prime to n such that $0 \leq k_i < n$. Then $r = \phi(n)$, by the definition of the function ϕ. Now we can reason as follows:

$$\begin{aligned}
&k_1 \cdot k_2 \cdots k_r \\
&\quad = \operatorname{rem}(k_1 \cdot k, n) \cdot \operatorname{rem}(k_2 \cdot k, n) \cdots \operatorname{rem}(k_r \cdot k, n) && \text{(by Lemma 8.10.4)} \\
&\quad \equiv (k_1 \cdot k) \cdot (k_2 \cdot k) \cdot \cdots \cdot (k_r \cdot k) \pmod{n} && \text{(by Cor 8.7.2)} \\
&\quad \equiv (k_1 \cdot k_2 \cdots k_r) \cdot k^r \pmod{n} && \text{(rearranging terms)}
\end{aligned}$$

Lemma 8.3.4.3. implies that $k_1 \cdot k_2 \cdots k_r$ is prime relative to n. Therefore, we can cancel this product from the first expression and the last by Corollary 8.10.3. This proves the claim. □

We can find multiplicative inverses using Euler's theorem as we did with Fermat's theorem: if k is relatively prime to n, then $k^{\phi(n)-1}$ is a multiplicative inverse of k modulo n. However, this approach requires computing $\phi(n)$. Our best method for doing so requires factoring n, which can be quite difficult in general. Fortunately, when we know how to factor n, we can use Theorem 8.10.1 to compute $\phi(n)$ efficiently!

8.10.4 RSA

Finally, we are ready to see how the RSA public key encryption scheme works:

RSA Public Key Encryption

Beforehand The receiver creates a public key and a secret key as follows.

1. Generate two distinct primes, p and q.
2. Let $n = pq$.
3. Select an integer e such that $\gcd(e, (p-1)(q-1)) = 1$.
   The *public key* is the pair (e, n). This should be distributed widely.
4. Compute d such that $de \equiv 1 \pmod{(p-1)(q-1)}$.
   The *secret key* is the pair (d, n). This should be kept hidden!

Encoding The sender encrypts message m to produce m' using the public key:

$$m' = \operatorname{rem}(m^e, n).$$

Decoding The receiver decrypts message m' back to message m using the secret key:

$$m = \operatorname{rem}((m')^d, n).$$

We'll explain in class why this way of Decoding works!

8.11 In-Class Problems Week 7, Wed.

Problem 8.11.1. A number is *perfect* if it is equal to the sum of its positive divisors, other than itself. For example, 6 is perfect, because $6 = 1{+}2{+}3$. Similarly, 28 is perfect, because $28 = 1{+}2{+}4{+}7{+}14$. Explain why $2^{k-1}(2^k - 1)$ is perfect if $2^k - 1$ is prime.

Solution. If $2^k - 1$ is prime, then the only divisors of $2^{k-1}(2^k - 1)$ are:

$$1, \quad 2, \quad 4, \quad \ldots, \quad 2^{k-1}$$

which sum to $2^k - 1$ (using the formula for a geometric series; in this case there's a direct "Computer Science" proof: think about the binary representations of 2^j and $2^k - 1$), and also

$$1 \cdot (2^k - 1), \quad 2 \cdot (2^k - 1), \quad 4 \cdot (2^k - 1), \quad \ldots, \quad 2^{k-2} \cdot (2^k - 1)$$

which sum to $(2^{k-1} - 1) \cdot (2^k - 1)$. Adding these two sums gives $2^{k-1}(2^k - 1)$, so the number is perfect.

Euclid knew this, and also was able to show that all *even* perfect numbers are of exactly this form. To do this day, no one knows if there are any odd perfect numbers! ■

Problem 8.11.2. **(a)** Use the Pulverizer (see the Appendix) to find integers x, y such that

$$x50 + y21 = \gcd(50, 21).$$

Solution. Here is the table produced by the Pulverizer:

x	y	$\text{rem}(x,y)$	$=$	$x - q \cdot y$
50	21	8	$=$	$50 - 2 \cdot 21$
21	8	5	$=$	$21 - 2 \cdot 8$
			$=$	$21 - 2 \cdot (50 - 2 \cdot 21)$
			$=$	$-2 \cdot 50 + 5 \cdot 21$
8	5	3	$=$	$8 - 1 \cdot 5$
			$=$	$(50 - 2 \cdot 21) - 1 \cdot (-2 \cdot 50 + 5 \cdot 21)$
			$=$	$3 \cdot 50 - 7 \cdot 21$
5	3	2	$=$	$5 - 1 \cdot 3$
			$=$	$(-2 \cdot 50 + 5 \cdot 21) - 1 \cdot (3 \cdot 50 - 7 \cdot 21)$
			$=$	$-5 \cdot 50 + 12 \cdot 21$
3	2	1	$=$	$3 - 1 \cdot 2$
			$=$	$(3 \cdot 50 - 7 \cdot 21) - 1 \cdot (-5 \cdot 50 + 12 \cdot 21)$
			$=$	$\boxed{8 \cdot 50 - 19 \cdot 21}$
2	1	0		

■

(b) Now find integer x', y' with $y' > 0$ such that

$$x'50 + y'21 = \gcd(50, 21)$$

Solution. since $(x, y) = (8, -19)$ works, so does $(8 - 21n, -19 + 50n)$ for any $n \in \mathbb{Z}$, so letting $n = 1$, we have

$$-13 \cdot 50 + 31 \cdot 21 = 1$$

■

Problem 8.11.3. Use the fact that $\gcd(a, b)$ is an integer linear combination of a and b to prove:

(a) Every common divisor of a and b divides $\gcd(a, b)$.

Solution. For some s and t, $\gcd(a, b) = sa + tb$. Let c be a common divisor of a and b. Since $c \mid a$ and $c \mid b$, we have $a = kc, b = k'c$ so

$$sa + tb = skc + tk'c = c(sk + tk')$$

so $c \mid sa + tb$. ■

(b) $\gcd(ka, kb) = k \cdot \gcd(a, b)$ for all $k > 0$.

Solution. We prove that each divides the other, which implies that one is $\pm$ the other. Since both are nonnegative, this implies they are equal.

Since $k \gcd(a, b) = k(sa + tb) = s(ka) + t(kb)$ for some s, t, any common divisor of ka and kb will divide $k \gcd(a, b)$, so in particular,

$$\gcd(ka, kb) \mid k \gcd(a, b)$$

Conversely, $\gcd(ka, kb) = s'(ka) + t'(kb) = (s'k)a + (t'k)b$ for some s', t', so is a linear combination of a and b and therefore,

$$\gcd(a, b) \mid \gcd(ka, kb)$$

■

(c) If $a \mid bc$ and $\gcd(a, b) = 1$, then $a \mid c$.

Solution. Since $\gcd(a, b) = 1$, we have $sa + tb = 1$ for some s, t. Multiplying by c, we have

$$sac + tbc = c$$

but a divides the second term of the sum since $a \mid bc$, and it obviously divides the first term, and therefore divides the sum, which equals c. ■

(d) If $p \mid ab$ for some prime, p, then $p \mid a$ or $p \mid b$.

Solution. If p does not divide a, then since p is prime, $\gcd(p, a) = 1$. By the previous part, we conclude that $p \mid b$. ■

(e) $\gcd(a, b) = \gcd(b, \text{rem}(a, b))$

Solution. Let $r = \text{rem}(a, b)$.

Since $r = a - qb$ for some q, we have that r is a linear combination of a and b and is therefore divisible by $\gcd(a, b)$. So any linear combination of r and b is divisible by $\gcd(a, b)$. Hence,

$$\gcd(a, b) \mid \gcd(b, r)$$

Conversely, $a = qb + r$ is a linear combination of b and r and is therefore divisible by $\gcd(b, r)$. Since $\gcd(b, r)$ divides both a and b, we conclude from part (a) that

$$\gcd(b, r) \mid \gcd(a, b).$$

■

Appendix: The Pulverizer

Euclid's algorithm for finding the GCD of two numbers relies on repeated application of the equation:

$$\gcd(a, b) = \gcd(b, \text{rem}(a, b))$$

For example, we can compute the GCD of 259 and 70 as follows:

$$\begin{aligned}
\gcd(259, 70) &= \gcd(70, 49) && \text{since } \text{rem}(259, 70) = 49 \\
&= \gcd(49, 21) && \text{since } \text{rem}(70, 49) = 21 \\
&= \gcd(21, 7) && \text{since } \text{rem}(49, 21) = 7 \\
&= \gcd(7, 0) && \text{since } \text{rem}(21, 7) = 0 \\
&= 7.
\end{aligned}$$

The Pulverizer goes through the same steps, but requires some extra bookkeeping along the way: as we compute $\gcd(a, b)$, we keep track of how to write each of the remainders (49, 21, and 7, in the example) as a linear combination of a and b (this is worthwhile, because our objective is to write the last nonzero remainder, which is the GCD, as such a linear combination). For our example, here is this extra bookkeeping:

x	y	$\text{rem}(x, y)$	$=$	$x - q \cdot y$
259	70	49	$=$	$259 - 3 \cdot 70$
70	49	21	$=$	$70 - 1 \cdot 49$
			$=$	$70 - 1 \cdot (259 - 3 \cdot 70)$
			$=$	$-1 \cdot 259 + 4 \cdot 70$
49	21	7	$=$	$49 - 2 \cdot 21$
			$=$	$(259 - 3 \cdot 70) - 2 \cdot (-1 \cdot 259 + 4 \cdot 70)$
			$=$	$\boxed{3 \cdot 259 - 11 \cdot 70}$
21	7	0		

We began by initializing two variables, $x = a$ and $y = b$. In the first two columns above, we carried out Euclid's algorithm. At each step, we computed $\text{rem}(x, y)$, which can be written in the form $x - q \cdot y$. (Remember that the Division Algorithm says $x = q \cdot y + r$, where r is the remainder. We get $r = x - q \cdot y$ by rearranging terms.) Then we replaced x and y in this equation with equivalent linear combinations of a and b, which we already had computed. After simplifying, we were left with a linear combination of a and b that was equal to the remainder as desired. The final solution is boxed.

8.12 In-Class Problems Week 8, Mon.

Problem 8.12.1. **(a)** Let $m = 2^9 5^{24} 11^7 17^{12}$ and $n = 2^3 7^{22} 11^{211} 13^1 17^9 19^2$. What is the $\gcd(m,n)$? What is the *least common multiple*, $\text{lcm}(m,n)$, of m and n? Verify that

$$\gcd(m,n) \cdot \text{lcm}(m,n) = mn. \tag{8.5}$$

Solution.

$$\begin{aligned} g &= 2^3 11^7 17^9, \\ l &= 2^9 5^{24} 7^{22} 11^{211} 13^1 17^{12} 19^2 \\ gl &= 2^{12} 5^{24} 7^{22} 11^{218} 13^1 17^{21} 19^2 = mn \end{aligned}$$

■

(b) Describe in general how to find the $\gcd(m,n)$ and $\text{lcm}(m,n)$ from the prime factorizations of m and n. Conclude that equation (8.5) holds for all positive integers m, n.

Solution. The divisors of m correspond to subsequences of the weakly increasing sequence of primes in the factorization of m, and likewise for n. So the factorization $\gcd(m,n)$ is the largest common subsequence of the two factorizations. This can be calculated by taking all the primes that appear in both factorizations raised to the *minimum* of the powers of that prime in each factorization.

Likewise, the factorization of $\text{lcm}(m,n)$ is the shortest sequence that has the factorizations of m and n as subsequences. So the factorization of $\text{lcm}(m,n)$ can be calculated by taking all the primes that appear in either factorization raised to the *maximum* of the powers of that prime in each factorization.

So in the factorization of $\gcd(m,n) \cdot \text{lcm}(m,n)$ each prime appears raised to a power equal to the sum of its powers in the factorizations of m and n, which is precisely its power in the factorization of mn. ■

Problem 8.12.2. The following properties of equivalence mod n follow directly from its definition and simple properties of divisibility. See if you can prove them without looking up the proofs in the notes.

If $a \equiv b \pmod{n}$, then $ac \equiv bc \pmod{n}$.

Solution. The condition $a \equiv b \pmod{n}$ is equivalent to the assertion $n \mid (a-b)$. This implies that $n \mid (a-b)c$, and so $n \mid (ac - bc)$. This is equivalent to $ac \equiv bc \pmod{n}$. ■

(a) **(b)** If $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$, then $a \equiv c \pmod{n}$.

Solution. Assume $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$, that is, $n \mid (a-b)$ and $n \mid (b-c)$. Then $n \mid (a-b) + (b-c) = (a-c)$, so $a \equiv c \pmod{n}$. ■

(c) If $a \equiv b \pmod{n}$ and $c \equiv d \pmod{n}$, then $ac \equiv bd \pmod{n}$.

Solution. $a \equiv b \pmod{n}$ implies $ac \equiv bc \pmod{n}$ by part (a); likewise, $c \equiv d \pmod{n}$ implies $bc \equiv bd \pmod{n}$. So $ac \equiv bd \pmod{n}$ by part (b). ■

(d) $\text{rem}(a, n) \equiv a \pmod{n}$.

Solution. The remainder $\text{rem}(a, n)$ is equal to $a - qn$ for some integer q. However, for every integer q:

$$\begin{aligned} n \mid qn &\longleftrightarrow n \mid ((a - qn) - a) \\ &\longrightarrow n \mid (\text{rem}(a, n) - a) \\ &\longleftrightarrow \text{rem}(a, n) \equiv a \pmod{n}. \end{aligned}$$

■

Problem 8.12.3. **(a)** Why is a number written in decimal evenly divisible by 9 if and only if the sum of its digits is a multiple of 9? *Hint:* $10 \equiv 1 \pmod{9}$.

Solution. Since $10 \equiv 1 \pmod{9}$, so is

$$10^k \equiv 1^k \equiv 1 \pmod{9}. \tag{8.6}$$

Now a number in decimal has the form:

$$d_k \cdot 10^k + d_{k-1} \cdot 10^{k-1} + \ldots + d_1 \cdot 10 + d_0.$$

From (8.6), we have

$$d_k \cdot 10^k + d_{k-1} \cdot 10^{k-1} + \ldots + d_1 \cdot 10 + d_0 \equiv d_k + d_{k-1} + \ldots + d_1 + d_0 \pmod{9}$$

This shows something stronger than what we were asked to show, namely, it shows that the remainder when the original number is divided by 9 is equal to the remainder when the sum of the digits is divided by 9. In particular, if one is zero, then so is the other. ■

(b) Take a big number, such as 37273761261. Sum the digits, where every other one is negated:

$$3 + (-7) + 2 + (-7) + 3 + (-7) + 6 + (-1) + 2 + (-6) + 1 = -11$$

Explain why the original number is a multiple of 11 if and only if this sum is a multiple of 11. *Hint:* $10 \equiv -1 \pmod{11}$.

Solution. A number in decimal has the form:

$$d_k \cdot 10^k + d_{k-1} \cdot 10^{k-1} + \ldots + d_1 \cdot 10 + d_0$$

From the observation above, we know:

$$\begin{aligned} & d_k \cdot 10^k + d_{k-1} \cdot 10^{k-1} + \ldots + d_1 \cdot 10 + d_0 \\ & \equiv d_k \cdot (-1)^k + d_{k-1} \cdot (-1)^{k-1} + \ldots \cdot + d_1 \cdot (-1)^1 + d_0 \cdot (-1)^0 \pmod{11} \\ & \equiv d_k - d_{k-1} + \ldots \cdot - d_1 + d_0 \pmod{11} \end{aligned}$$

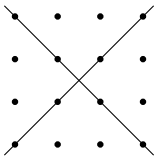
assuming k is even. The case where k is odd is the same with signs reversed.

The procedure given in the problem computes $\pm$ this alternating sum of digits, and hence yields a number divisible by 11 ($\equiv 0 \pmod{11}$) iff the original number was divisible by 11. ■

Problem 8.12.4. Two nonparallel lines in the real plane intersect at a point. Algebraically, this means that the equations

$$\begin{aligned} y &= m_1 x + b_1 \\ y &= m_2 x + b_2 \end{aligned}$$

have a unique solution (x, y), provided $m_1 \neq m_2$. This statement would be false if we restricted x and y to the integers, since the two lines could cross at a noninteger point:



However, an analogous statement holds if we work over the integers *modulo a prime*, p. Find a solution to the congruences

$$\begin{aligned} y &\equiv m_1 x + b_1 \pmod{p} \\ y &\equiv m_2 x + b_2 \pmod{p} \end{aligned}$$

when $m_1 \not\equiv m_2 \pmod{p}$. Express your solution in the form $x \equiv? \pmod{p}$ and $y \equiv?? \pmod{p}$ where the ?'s denote expressions involving m_1, m_2, b_1, and b_2. You may find it helpful to solve the original equations over the reals first.

Solution. Subtracting the second congruence from the first, we have:

$$\begin{aligned} 0 &\equiv m_1 x + b_1 - (m_2 x + b_2) \pmod{p} \\ (m_1 - m_2)x &\equiv b_2 - b_1 \pmod{p} \\ x &\equiv (m_1 - m_2)^{-1} \cdot (b_2 - b_1) \pmod{p} \end{aligned}$$

Substituting this value of x into the first congruence, we have

$$y \equiv m_1 \cdot (m_1 - m_2)^{-1} \cdot (b_2 - b_1) + b_1 \pmod{p}$$

Here $(m_1 - m_2)^{-1} \pmod{p}$ exists because $m_1 \not\equiv m_2 \pmod{p}$ and hence p does not divide $(m_1 - m_2)$.

Further exercise: Show that (x, y) are unique modulo p. ■

Appendix

Definition. $a \equiv b \pmod{n}$ iff $n \mid a - b$.

Lemma 8.12.1. *[Facts About Congruences] The following hold for $n \geq 1$:*

1. $a \equiv a \pmod{n}$
2. $a \equiv b \pmod{n}$ *implies* $b \equiv a \pmod{n}$
3. $a \equiv b \pmod{n}$ *and* $b \equiv c \pmod{n}$ *implies* $a \equiv c \pmod{n}$
4. $a \equiv \text{rem}(a, n) \pmod{n}$
5. $a \equiv b \pmod{n}$ *implies* $a + c \equiv b + c \pmod{n}$
6. $a \equiv b \pmod{n}$ *implies* $ac \equiv bc \pmod{n}$
7. $a \equiv b \pmod{n}$ *and* $c \equiv d \pmod{n}$ *imply* $a + c \equiv b + d \pmod{n}$
8. $a \equiv b \pmod{n}$ *and* $c \equiv d \pmod{n}$ *imply* $ac \equiv bd \pmod{n}$

Lemma 8.12.2 (Inverses mod n). *If k and $n > 1$ are relatively prime, then there is a positive integer $k^{-1} < n$ called the* modulo n inverse *of k, such that*

$$k \cdot k^{-1} \equiv 1 \pmod{n}.$$

Proof. That integers k and n are relatively prime means that $\gcd(k, n) = 1$. But $\gcd(k, n) = 1$ implies that $1 = ak + bn$ for some integers a, b, and so $1 \equiv ak \pmod{n}$. So the positive integer less than n that is equivalent to $a \pmod{n}$ is k^{-1}, namely, $k^{-1} = \text{rem}(a, n)$. □

8.13 Problem Set 6

Problem 8.13.1. **(a)** Show that if a planar embedding can be constructed by adding two successive edges to an embedding for a graph, G, then the same planar embedding can be constructed by adding the edges in the reverse order.

Solution. Check 4 cases: two possible rules for edge 1, and two possible rules for edge 2. ■

(b) Let G be a graph with a planar embedding. Conclude that the same planar embedding can be built up recursively by adding the edges of G in any order.

Solution. Routine structural induction using previous part for base case. ■

(c) Conclude that any subgraph of a planar graph is planar.

Solution. Let H be a connected component of a subgraph of G. Assume the embedding of the connected component of G containing H was constructed so all the edges not in H were added last. Then deleting these edge-adding steps leaves a recursive construction of an embedding for H. It follows that every connected component of a subgraph is planar, and so by definition, the subgraph is planar. ■

Problem 8.13.2. **(a)** Prove that every planar graph has a vertex of degree at most 5. *Hint:* $e \leq 3v - 6$.

Solution. The sum of the degrees is $2e$. Suppose to the contrary that every vertex had degree at least 6. Then the sum of degrees is at least $6v$. So $2e \geq 6v$, and hence $e \geq 3v$. But this contradicts the inequality given in the hint, which is satisfied by every connected planar graph with $v \geq 3$. ■

(b) Conclude that every planar graph has width at most 5 and therefore is 6-colorable.[3]

Hint: Use the result of Problem 8.13.1(c).

Solution. By induction on number of vertices with hypothesis:

$$P(n) ::= \text{every planar graph with } n \text{ vertices has width at most 5.}$$

Base case: ($n = 1$) Trivial.

Inductive step: Let G be a planar graph with $n + 1$ vertices. By Problem 8.13.2(a), G has a vertex, v, of degree at most 5. Remove v to obtain a subgraph, H, with n vertices. By Problem 8.13.1 H is planar, and so by induction, H has width at most 5, so its vertices can be arranged in a sequence such that each vertex is adjacent to at most 5 vertices that precede it. Adding v to the end of this sequence gives a width 5 sequence for G. So G is of width at most 5, and since G was arbitrary, we conclude that $P(n + 1)$ holds, completing the proof.

6-colorability now follows from Pset 5, Problem 1. ■

[3]From Pset 5: A simple graph, G, is said to have ***width***, w, iff its vertices can be arranged in a sequence such that each vertex is adjacent to at most w vertices that precede it in the sequence.

Problem 8.13.3. Here is a *very, very fun* game. We start with two distinct, positive integers written on a blackboard. Call them a and b. You and I now take turns. (I'll let you decide who goes first.) On each player's turn, he or she must write a new positive integer on the board that is the difference of two numbers that are already there. If a player can not play, then he or she loses.

For example, suppose that 12 and 15 are on the board initially. Your first play must be 3, which is $15 - 12$. Then I might play 9, which is $12 - 3$. Then you might play 6, which is $15 - 9$. Then I can not play, so I lose.

(a) Show that every number on the board at the end of the game is a multiple of $\gcd(a, b)$.

Solution. Thinking of the game as a state machine, we observe that the property that $\gcd(a, b)$ divides all the numbers on the board is an invariant. This follows because the next state (board) is the same as the previous state, except for an additional number which is the difference of two numbers already there. Assuming these two numbers are divisible by $\gcd(a, b)$, we know that their difference will be as well, which proves that the next state satisfies the invariant.

Since the start state, with just a and b on the board, satisfies the invariant, so will any final state. ■

(b) Show that every positive multiple of $\gcd(a, b)$ up to $\max(a, b)$ is on the board at the end of the game.

Solution. Assume without loss of generality that $a > b$. Let s be the smallest number on the board at the end of the game. So $a = qs + r$ where $0 \leq r < s$ by the division algorithm. Then $a - s$ must be on the board and thus so must $a - 2s$, $a - 3s$, ..., $a - (q - 1)s$. However, $r = a - qs$ cannot be on the board, since $r < s$ and s is defined to be the smallest number there. The only explanation is that $r = 0$, which implies that $s \mid a$. By the same argument, $s \mid b$. Therefore, s is a common divisor of a and b. Since s is a multiple of the greatest common divisor of a and b by the preceding problem part, s must actually be the greatest common divisor. We already argued that a, $a - s$, $a - 2s$, ..., $a - (q - 1)s$ must be on the board, and these are all the positive multiples of $\gcd(a, b)$ up to $\max(a, b)$.

■

(c) Describe a strategy that lets you win this game every time.

Solution. Assume without loss of generality that $a = \max(a, b)$. By the previous parts, the numbers that appear on the final board are precisely all the multiples $\leq a$ of $gcd(a, b)$. Thus, for each game, we know *exactly* how many values will be placed on the board before the game ends. So if an odd number of values will appear on the final board (which happens precisely when a is an even multiple of $\gcd(a, b)$), then choose to go first. Otherwise, choose to go second. ■

Problem 8.13.4. **(a)** Use the Fundamental Theorem of Arithmetic (that the factorization into primes of an integer greater than 1 is unique) to give simple proofs of a few of the properties of gcd and divisibility listed in Lemma 3.4 of the Number Theory.[4]

[4]These properties were proved in the Notes strictly from the definitions, which took more care. But it would have been "cheating" to prove them in the first place using the Fundamental Theorem, since the proof of the Fundamental Theorem builds on these properties.

Solution. In what follows, let the unique prime factorizations of a, b, and c be as follows:

$$\begin{aligned} a &= p_1^{\alpha_1} p_2^{\alpha_2} \dots p_n^{\alpha_n} \\ b &= q_1^{\beta_1} q_2^{\beta_2} \dots q_m^{\beta_m} \\ c &= r_1^{\gamma_1} r_2^{\gamma_2} \dots r_l^{\gamma_l} \end{aligned}$$

3. If $\gcd(a, b) = 1$ and $\gcd(a, c) = 1$, then $\gcd(a, bc) = 1$.
   $\gcd(a, b) = 1$ implies that $p_i \neq q_j$ for any i, j, and $\gcd(a, c) = 1$ implies that $p_i \neq r_j$ for any i, j. The primes in the unique prime factorization of bc consist of the union of the primes that occur in the prime factorizations of b and c, i.e. q_j and r_j. But since we know that $p_i \neq q_j$ and $p_i \neq r_j$ for any i, j, we can conclude that a and bc have no common factors, and $\gcd(a, bc) = 1$.
4. If $a \mid bc$ and $\gcd(a, b) = 1$, then $a \mid c$.
   If $a \mid bc$, then $p_1^{\alpha_1} p_2^{\alpha_2} \dots p_n^{\alpha_n}$ appears in the prime factorization of bc. Now, if $\gcd(a, b) = 1$, then $p_i \neq q_j$, for any j. The prime factorization of bc is:

   $$bc = q_1^{\beta_1} q_2^{\beta_2} \dots q_m^{\beta_m} r_1^{\gamma_1} r_2^{\gamma_2} \dots r_l^{\gamma_l}.$$

   Since $a \mid bc$ and $p_i \neq q_j$, for any j, it must be that $p_1^{\alpha_1} p_2^{\alpha_2} \dots p_n^{\alpha_n}$ appears in $r_1^{\gamma_1} r_2^{\gamma_2} \dots r_l^{\gamma_l}$, which means that $a \mid c$.

■

Suppose m and n are relatively prime. Use the Fundamental Theorem of Arithmetic (that the factorization into primes of an integer greater than 1 is unique) to give simple proofs of:

(b) $mn \mid a$ iff $m \mid a$ and $n \mid a$.

Solution. In what follows, let the unique prime factorizations of m and n be as follows:

$$\begin{aligned} m &= p_1^{a_1} p_2^{a_2} \cdots p_s^{a_s} \\ n &= q_1^{b_1} q_2^{b_2} \cdots q_t^{b_t} \end{aligned}$$

Since m and n are relatively prime, we know that $p_i \neq q_j$, for any i, j. Furthermore, the unique prime factorization of mn is the "concatenation" of the two disjoint prime factorizations of m and n:

$$mn = (p_1^{a_1} p_2^{a_2} \cdots p_s^{a_s})(q_1^{b_1} q_2^{b_2} \cdots q_t^{b_t}).$$

If $mn \mid a$, then the entire sequence $(p_1^{a_1} p_2^{a_2} \cdots p_s^{a_s})(q_1^{b_1} q_2^{b_2} \cdots q_t^{b_t})$ appears in the prime factorization of a. But this implies that the prime factorization of m and n both appear in the prime factorization of a, so both m and n divide a.

Conversely, if $m \mid a$ and $n \mid a$, then both $p_1^{a_1} p_2^{a_2} \cdots p_s^{a_s}$ and $q_1^{b_1} q_2^{b_2} \cdots q_t^{b_t}$ appear in the prime factorization of a. Since these two sequences are disjoint (i.e. do not share any common terms), their "concatenation" $(p_1^{a_1} p_2^{a_2} \cdots p_s^{a_s})(q_1^{b_1} q_2^{b_2} \cdots q_t^{b_t})$ also appears in the prime factorization, and this implies that mn divides a. ■

(c)

x is relatively prime to mn

iff x is relatively prime to m and x is relatively prime to n.

Solution. Let the unique prime factorization of x be $r_1^{c_1} r_2^{c_2} \cdots r_u^{c_u}$. Suppose x is relatively prime to mn. This implies that $r_i \neq p_j$ for any i, j, i.e. x does not share any prime factors with m, so x and m are relatively prime. Furthermore, it also implies that $r_i \neq q_j$, for any i, j, i.e. x does not share any prime factors with n, so x and n are relatively prime.

Conversely, suppose x is relatively prime to m and x is relatively prime to n. Again, $r_i \neq p_j$ for any i, j, and $r_i \neq q_j$ for any i, j. Since the prime factorization of $mn = (p_1^{a_1} p_2^{a_2} \cdots p_s^{a_s})(q_1^{b_1} q_2^{b_2} \cdots q_t^{b_t})$, none of the prime factors of x appear in the prime factorization of mn, so x is relatively prime to mn. ■

Problem 8.13.5. Albert decides to entertain the class with a magic trick. He says:

1. Pick any 5 digit number containing at least two different digits.
2. Shuffle the digits to obtain a different number.
3. Subtract the smaller number from the larger.
4. Now sum the digits of the result.
5. Repeat step 4 until you have only one digit, and write down your answer.

He announces the right answer without seeing the paper, but the 6.042 students are not impressed. This problem demonstrates why they were not impressed with Albert's "magic."

(a) Show that taking *any* nonnegative integer (not necessarily a 5-digit number), rearranging its digits to form a new number, and finding the difference between the two numbers, will always result in a multiple of 9.

Solution. Let n be the nonnegative integer, and let m be the number obtained after rearrangement of the digits of n. We want to show that $n - m$ is divisible by 9.

To this end, let

$$n = 10^0 \cdot a_0 + 10^1 \cdot a_1 + \ldots + 10^k \cdot a_k$$
$$m = 10^{\pi(0)} \cdot a_0 + 10^{\pi(1)} \cdot a_1 + \ldots + 10^{\pi(k)} \cdot a_k$$

where π is a permutation mapping the k digits of n to k digits of m. So,

$$n - m = \sum_i a_i \left(10^i - 10^{\pi(i)}\right).$$

Since $10^k \equiv 1^k = 1 \pmod 9$ for any k, we have that $10^i - 10^{\pi(i)} \equiv 0 \pmod 9$. That is, each term, $a_i \left(10^i - 10^{\pi(i)}\right)$, in the sum is equivalent $0 \pmod 9$, and so the whole sum is also $\equiv 0 \pmod 9$. That is,

$$n - m \equiv 0 \pmod 9,$$

which means that $n - m$ is divisible by 9. ■

(b) Show that summing the digits of a positive integer results in an integer that is congruent to it modulo 9.

Solution. Let $n = 10^0 \cdot a_0 + 10^1 \cdot a_1 + \ldots + 10^k \cdot a_k$. We want to show that:

$$\sum_i a_i \equiv \sum_i 10^i a_i \pmod 9.$$

Since $10 \equiv 1 \pmod 9$, we have that $10^i \equiv 1^i = 1 \pmod 9$. Therefore, $10^i a_i \equiv a_i \pmod 9$, and we have our desired result: $\sum_i a_i \equiv \sum_i 10^i a_i \pmod 9$. ■

(c) Show that for any 5 digit number, this procedure always terminates with the same digit. What would happen if the starting number had more than 5 digits?

Solution. After completing steps 1-3, we have arrived at a number that is divisible by 9, as shown in part (a) of this problem. Moreover, this number will be positive, since the original and shuffled numbers are different. Then, by part (b), summing the digits of this number will still yield a positive number divisible by 9. Furthermore, summing the digits of a number results in a smaller positive number, so the procedure is guaranteed to terminate with the number 9.

All of the previous reasoning holds no matter how many digits the original number had. ■

Problem 8.13.6. Suppose that p is a prime and $0 < k < p$.

(a) k is *self-inverse* if $k^2 \equiv 1 \pmod p$. Prove that k is self-inverse iff either $k = 1$ or $k = p - 1$.

Hint: $k^2 - 1 = (k - 1)(k + 1)$

Solution. By definition of $\equiv \pmod p$, the integer k is self-inverse iff $p \mid k^2 - 1$. But $k^2 - 1 = (k - 1)(k + 1)$, and since p is a prime, we conclude that either $p \mid k - 1$ or $p \mid k + 1$. But $0 < k < p$, so $p \mid k - 1$ iff $k - 1 = 0$, and $p \mid k + 1$ iff $k + 1 = p$, so we must have $k = 1$ or $k = p - 1$.

Conversely, $1 \cdot 1 \equiv 1 \pmod p$ and $(p - 1) \cdot (p - 1) = p^2 + 2p + 1 \equiv 1 \pmod p$. ■

(b) Wilson's Theorem asserts

Theorem 8.13.1 (Wilson's Theorem). *If p is a prime, then*

$$(p - 1)! \equiv -1 \pmod p$$

The English mathematician Edward Waring said that this theorem would probably be very difficult to prove because there was no adequate notation for primes. Gauss proved it while standing (on one foot, it is rumored). He suggested that Waring failed for lack of notions, not notations. Prove Wilson's Theorem. *Hint:* While standing on one foot, think about pairing each term in $(p - 1)!$ with its multiplicative inverse.

Solution. If $p = 2$, then the theorem holds, because $1 \equiv -1 \pmod 2$. If $p > 2$, then $p - 1$ and 1 are distinct terms in the product $1 \cdot 2 \cdot \cdots (p - 1)$, and these are the only self-inverses. Consequently,

we can pair each of the remaining terms with its multiplicative inverse. Since the product of a number and its inverse is congruent to 1, all of these remaining terms cancel. Therefore, we have:

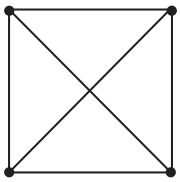
$$\begin{aligned}(p-1)! &\equiv 1 \cdot (p-1) \pmod{p} \\ &\equiv -1 \pmod{p}\end{aligned}$$

■

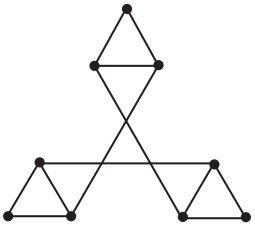
8.14 Miniquiz Apr. 6

Problem 8.14.1. **(a)** Circle the graphs below that are planar (that *can be drawn* in the plane so that no edges cross).

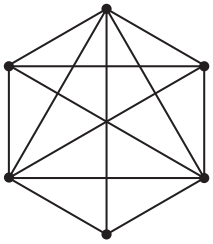
(a) G_1



(b) G_2



(c) G_3



(d) G_4

(b) For each of the nonplanar graphs above, briefly explain why it is not planar.

Solution. G_1 is planar (base case of the recursive definition).

G_2 is planar (a tetrahedron is one of the regular polyhedra).

G_3 can be constructed from three 3-cycles by adding two bridges followed by a split face.

G_4 does not have a planar embedding. One acceptable explanation is that it contains as a subgraph K_5, the complete graph on 5 vertices, which is not planar (see PS.6.1c). Another is that it has 6 vertices and 13 edges and so $e > 3v - 6$. ■

Problem 8.14.2.

(a) Circle all the valid statements about the greatest common divisor:

- Every common divisor of a and b divides $\gcd(a, b)$.
- $\gcd(ka, kb) = k \cdot \gcd(a, b)$ for all $k > 0$.
- If $\gcd(a, b) = 1$ and $\gcd(b, c) = 1$, then $\gcd(a, c) = 1$.
- If $a \mid bc$ and $\gcd(a, b) = 1$, then $a \mid c$.
- $\gcd(a, b) = \gcd(b, \text{rem}(a, b))$.

Solution. All but the third are properties of the gcd (Lemma 3.4 in Notes 7). The third statement is not universally true. Consider the following example where $a = c$: $\gcd(2, 3) = 1$ and $\gcd(3, 2) = 1$ but $\gcd(2, 2) = 2 \neq 1$. ■

(b) Circle all the valid statements about equivalence mod n for $n \geq 1$:

- If $ac \equiv bc \pmod{n}$, then $a \equiv b \pmod{n}$.
- If $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$, then $a \equiv c \pmod{n}$.
- If $a \equiv b \pmod{n}$ and $c \equiv d \pmod{n}$, then $ac \equiv bd \pmod{n}$.
- If $a \equiv b \pmod{n}$ and $c \equiv d \pmod{n}$, then $a + c \equiv b + d \pmod{n}$.
- $\text{rem}(a, n) \equiv a \pmod{n}$.

Solution. All but the first are properties of congruences and remainders (Corollary 7.2 and Lemma 7.3 in Notes 7). The first statement does not hold in general if c and n are not relatively prime (Corollary 10.3 in Notes 7). ■

Problem 8.14.3.

(a) Let $x = 13^2 17^5 23^{88} 31^{1000}$ and $y = 11^{53} 13^{12} 29^{35} 37^{28}$. What is the $\gcd(x, y)$?

Solution. $\gcd(x, y) = 13^2 = 169$ ■

(b) For a given prime p, find *all* k in the range $\{0, 1, \ldots, p-1\}$ such that $k^2 \equiv 1 \pmod{p}$.

Solution. $k^2 \equiv 1 \pmod{p}$ holds for $k = 1$ and $k = p - 1$ (See solution for PS6, Problem 6a). ■

(c) Find a value of k that makes the following statement true.

$$13^k \text{ is a multiplicative inverse of } 13 \pmod{17}.$$

Solution. $k = 15$ by Fermat's Theorem: $13^{15} \cdot 13 = 13^{16} \equiv 1 \pmod{17}$. ■

(d) What is the multiplicative inverse of 13 modulo 17? (We want the number between 0 and 16. If you get stuck working out the math, we will award partial credit if you explain how you would calculate this.)

Solution. The inverse of 13 modulo 17 is 4.

This can be calculated easily in several ways:

- Compute $\text{rem}(13^{15}, 17)$ by successive squaring (see section 8.3 of Notes 7)
- use the pulverizer to find coefficients x, y so that $1 = \gcd(17, 13) = x13 + y17$ and compute $\gcd(x, 17)$, or
- notice that

$$13^{15} \equiv (-4)^{15} \equiv ((-4)^2)^7(-4) \equiv 16^7(-4) \equiv (-1)^7(-4) \equiv 4 \pmod{17}.$$

■

Problem 8.14.4. Show how to use Euler's Formula and Lemmas 8.14.2 and 8.14.3 in the Appendix to prove that if a connected planar graph has $v \geq 3$ vertices and e edges, then

$$e \leq 3v - 6.$$

Solution. See Corollary 4.4 in Notes 6.5. ■

Appendix

Theorem 8.14.1 (Euler's Formula). *If a connected graph has a planar embedding, then*

$$v - e + f = 2$$

where v is the number of vertices, e is the number of edges, and f is the number of faces.

Lemma 8.14.2. *In a planar embedding of a graph, each edge is traversed a total of two times by the faces of the embedding.*

Lemma 8.14.3. *In a planar embedding of a graph with at least three vertices, each face is of length at least three.*

Lemma 8.14.4. *Any subgraph of a planar graph is planar.*

Definition 8.14.5. a ***divides*** b iff $ak = b$ for some k. This is denoted $a \mid b$.

Theorem 8.14.6 (Division Theorem). *Let n and d be integers such that $d > 0$. Then there exists a unique pair of integers q and r such that $n = qd + r$ and $0 \leq r < d$.*

The remainder r in the Division Theorem is denoted $\text{rem}(n, d)$.

Definition 8.14.7. a ***is congruent to*** b ***modulo*** n iff $n \mid (a - b)$. This is denoted $a \equiv b \pmod{n}$.

Definition 8.14.8. A ***multiplicative inverse*** $\pmod{p}$ of a number x is another number x^{-1} such that:

$$x \cdot x^{-1} \equiv 1 \pmod{p}$$

Theorem (Fermat's (Little) Theorem). *If p is prime and k is not a multiple of p, then*

$$k^{p-1} \equiv 1 \pmod{p}$$

Definition. The value of Euler's totient function, $\phi(n)$, is defined to be the number of positive integers less than n that are relatively prime to n.

Lemma (Euler Function Equations). *If p is prime, then*

$$\phi(p^k) = p^k - p^{k-1} \qquad \text{for prime, } p, \text{ and } k > 0$$
$$\phi(mn) = \phi(m) \cdot \phi(n) \qquad \text{for } \gcd(m,n) = 1.$$

Theorem (Euler's Theorem). *If k and n are relatively prime, then*

$$k^{\phi(n)} \equiv 1 \pmod{n}$$

The Pulverizer

Euclid's algorithm for finding the GCD of two numbers relies on repeated application of the equation:

$$\gcd(a,b) = \gcd(b, \text{rem}(a,b))$$

For example, we can compute the GCD of 259 and 70 as follows:

$$\begin{aligned}
\gcd(259,70) &= \gcd(70,49) && \text{since } \text{rem}(259,70) = 49\\
&= \gcd(49,21) && \text{since } \text{rem}(70,49) = 21\\
&= \gcd(21,7) && \text{since } \text{rem}(49,21) = 7\\
&= \gcd(7,0) && \text{since } \text{rem}(21,7) = 0\\
&= 7.
\end{aligned}$$

The Pulverizer goes through the same steps, but requires some extra bookkeeping along the way: as we compute $\gcd(a,b)$, we keep track of how to write each of the remainders (49, 21, and 7, in the example) as a linear combination of a and b (this is worthwhile, because our objective is to write the last nonzero remainder, which is the GCD, as such a linear combination). For our example, here is this extra bookkeeping:

x	y	$\text{rem}(x,y)$		$x - q \cdot y$
259	70	49	=	$259 - 3 \cdot 70$
70	49	21	=	$70 - 1 \cdot 49$
			=	$70 - 1 \cdot (259 - 3 \cdot 70)$
			=	$-1 \cdot 259 + 4 \cdot 70$
49	21	7	=	$49 - 2 \cdot 21$
			=	$(259 - 3 \cdot 70) - 2 \cdot (-1 \cdot 259 + 4 \cdot 70)$
			=	$\boxed{3 \cdot 259 - 11 \cdot 70}$
21	7	0		

We began by initializing two variables, $x = a$ and $y = b$. In the first two columns above, we carried out Euclid's algorithm. At each step, we computed $\text{rem}(x,y)$, which can be written in the form $x - q \cdot y$. (Remember that the Division Algorithm says $x = q \cdot y + r$, where r is the remainder. We get $r = x - q \cdot y$ by rearranging terms.) Then we replaced x and y in this equation with equivalent linear combinations of a and b, which we already had computed. After simplifying, we were left with a linear combination of a and b that was equal to the remainder as desired. The final solution is boxed.

8.15 In-Class Problems Week 8, Wed.

Problem 8.15.1. Let's try out RSA! There is a complete description of the algorithm at the bottom of the page. You'll probably need extra paper. **Check your work carefully!**

(a) As a team, go through the **beforehand** steps.

- Choose primes p and q to be relatively small, say in the range 10-40. In practice, p and q might contain several hundred digits, but small numbers are easier to handle with pencil and paper.
- Try $e = 3, 5, 7, \ldots$ until you find something that works. Use Euclid's algorithm to compute the gcd.
- Find d (using the Pulverizer —see appendix for a reminder on how the Pulverizer works —or Euler's Theorem).

When you're done, put your public key on the board. This lets another team send you a message.

(b) Now send an encrypted message to another team using their public key. Select your message m from the codebook below:

- 2 = Greetings and salutations!
- 3 = Yo, wassup?
- 4 = You guys are slow!
- 5 = All your base are belong to us.
- 6 = Someone on *our* team thinks someone on *your* team is kinda cute.
- 7 = You *are* the weakest link. Goodbye.

(c) Decrypt the message sent to you and verify that you received what the other team sent!

(d) Explain how you could read messages encrypted with RSA if you could quickly factor large numbers.

Solution. Suppose you see a public key (e, n). If you can factor n to obtain p and q, then you can compute d using the Pulverizer or Euler's Theorem. This gives you the secret key (d, n), and so you can decode messages as well as the intended recipient. ■

RSA Public Key Encryption

Beforehand The receiver creates a public key and a secret key as follows.

1. Generate two distinct primes, p and q.
2. Let $n = pq$.
3. Select an integer e such that $\gcd(e, (p-1)(q-1)) = 1$.
   The *public key* is the pair (e, n). This should be distributed widely.
4. Compute d such that $de \equiv 1 \pmod{(p-1)(q-1)}$.
   The *secret key* is the pair (d, n). This should be kept hidden!

Encoding The sender encrypts message m to produce m' using the public key:

$$m' = \text{rem}(m^e, n).$$

Decoding The receiver decrypts message m' back to message m using the secret key:

$$m = \text{rem}((m')^d, n).$$

Problem 8.15.2. A critical question is whether decrypting an encrypted message always gives back the original message! Mathematically, this amounts to asking whether:

$$m^{de} \equiv m \pmod{pq}. \tag{8.7}$$

It's just as easy to prove something slightly more general:

Lemma 8.15.1. *Let n be a product of distinct primes and $a \equiv 1 \pmod{\phi(n)}$ for some nonnegative integer, a. Then*

$$m^a \equiv m \pmod{n}. \tag{8.8}$$

(a) Verify that equation (8.7) follows as a special case of Lemma 8.15.1.

Solution. Equation (8.7) is an instance of equation (8.8) with $n = pq$ and $a = de$.

So we need only verify that the conditions on n and a required by the Lemma hold when n, d, and e are chosen according to the RSA protocol. The first condition of the Lemma 8.15.1 is that n be a product of primes. In RSA, $n = pq$ so this condition holds. The second condition is that $a \equiv 1 \pmod{\phi(n)}$. But for $n = pq$, the Euler function equations (see the Appendix) imply that $\phi(n) = (p-1)(q-1)$. So when d and e are chosen according to RSA, this condition holds because $a = de \equiv 1 \pmod{(p-1)(q-1)}$. ■

(b) Explain why Lemma 8.15.1 implies that k and k^5 have the same last digit. For example:

$$\underline{2}^5 = 3\underline{2} \qquad\qquad 79\underline{5} = 3077056399\underline{5}$$

Hint: What is $\phi(10)$?

Solution. Two nonnegative integers have the same last digit iff they are $\equiv \pmod{10}$. Now $\phi(10) = \phi(2)\phi(5) = 4$ and $5 \equiv 1 \pmod{4}$, so by Lemma 8.15.1,

$$k^5 \equiv k \pmod{10}.$$

■

(c) Justify each of the steps in the following proof that:

If p is prime, then

$$m^a \equiv m \pmod{p} \tag{8.9}$$

for all nonnegative integers $a \equiv 1 \pmod{\phi(p)}$.

Proof. Equation (8.9) obviously holds if $p \mid m$. (why?)

So assume p does not divide m. Now if $a \equiv 1 \pmod{\phi(p)}$, then for some k,

$$\begin{aligned}
a &= 1 + k\phi(p), && \text{(0. why?)}\\
\text{so}&\\
m^a &= m^{1+k\phi(p)}\\
&= m \cdot \left(m^{\phi(p)}\right)^k && \text{(1. why?)}\\
&= m \cdot \left(m^{p-1}\right)^k && \text{(2. why?)}\\
&\equiv m \cdot (1)^k \pmod{p} && \text{(3. why?)}\\
&\equiv m \pmod{p}. && \text{(4. why?)}
\end{aligned}$$

Solution. If $p \mid m$, then (8.9) holds trivially since both sides are congruent to 0 modulo p.

- 0. $a \equiv 1 \pmod{c}$ iff $c \mid (a-1)$ iff $kc = a - 1$ iff $a = 1 + kc$.
- 1. uses algebraic property of exponentiation that $m^{kj} = m^{jk} = \left(m^j\right)^k$.
- 2. follows since $\phi(p) = p - 1$ by the Euler function equations.
- 3. $m^{p-1} \equiv 1 \pmod{p}$ by Fermat's Little Theorem.
- 4. $m \cdot (1)^k = m$ of course!

■

□

(d) Show that for any positive integers j, k, if $a \equiv b \pmod{k}$ and $j \mid k$, then $a \equiv b \pmod{j}$

Solution. $a \equiv b \pmod{k}$ iff $k \mid (a-b)$. But if $k \mid (a-b)$ and $j \mid k$, then also $j \mid (a-b)$, which implies $a \equiv b \pmod{j}$ ■

(e) Prove that if n is a product of distinct primes, and $a \equiv b \pmod{p}$ for all prime factors, p, of n, then $a \equiv b \pmod{n}$.

Solution. By definition of congruence, $a \equiv b \pmod{k}$ iff $k \mid (a-b)$. So if $a \equiv b \pmod{p}$ for each prime factor, p, of n, then $p \mid (a-b)$ for each prime factor, p, and hence, so does their product (by the Unique Factorization Theorem). That is, $n \mid (a-b)$, which means $a \equiv b \pmod{n}$. ■

(f) Verify that for any $n > 1$ and any prime divisor, p, of n,

$$\phi(p) \mid \phi(n).$$

Solution. Let p be a prime factor of n and factor n as $m \cdot p^k$ where p does not divide m. By the Euler function equations, $\phi(n) = \phi(m)\phi(p^k)$, so $\phi(p^k) \mid \phi(n)$. But $\phi(p) = (p-1)$ which divides $(p-1)p^{k-1} = \phi(p^k)$. ■

(g) Combine the previous parts to complete the proof of Lemma 8.15.1.

Solution. Suppose n is a product of distinct primes and $a \equiv 1 \pmod{\phi(n)}$. Let p be any prime factor of n.

Since $\phi(p) \mid \phi(n)$ by part (f), we conclude from part (d) that

$$a \equiv 1 \pmod{\phi(p)}.$$

Hence, by part (c),

$$m^a \equiv m \pmod{p}$$

for all m. Since this holds for all factors, p, of n, we conclude from part (e) that

$$m^a \equiv m \pmod{n},$$

which proves Lemma 8.15.1. ■

Appendix

Inverses, Fermat, Euler

Lemma (Inverses mod n). *If k and n are relatively prime, then there is integer k' called the* modulo n inverse *of k, such that*

$$k \cdot k' \equiv 1 \pmod{n}.$$

Remark: If $\gcd(k,n) = 1$, then $sk + tn = 1$ for some s, t, so we can choose $k' ::= s$ in the previous Lemma. So given k and n, an inverse k' can be found efficiently using the Pulverizer.

Theorem (Fermat's (Little) Theorem). *If p is prime and k is not a multiple of p, then*

$$k^{p-1} \equiv 1 \pmod{p}$$

Definition. The value of *Euler's totient function*, $\phi(n)$, is defined to be the number of positive integers less than n that are relatively prime to n.

Lemma (Euler Totient Function Equations).

$$\begin{aligned} \phi(p^k) &= p^k - p^{k-1} && \text{for prime, } p, \text{ and } k > 0, \\ \phi(mn) &= \phi(m) \cdot \phi(n) && \text{when } \gcd(m,n) = 1. \end{aligned}$$

Theorem (Euler's Theorem). *If k and n are relatively prime, then*

$$k^{\phi(n)} \equiv 1 \pmod{n}$$

Corollary. *If k and n are relatively prime, then $k^{\phi(n)-1}$ is an inverse modulo n of k.*

Remark: Using fast exponentiation to compute $k^{\phi(n)-1}$ is another efficient way to compute an inverse modulo n of k.

The Pulverizer

Euclid's algorithm for finding the GCD of two numbers relies on repeated application of the equation:

$$\gcd(a, b) = \gcd(b, \text{rem}(a, b))$$

For example, we can compute the GCD of 259 and 70 as follows:

$$\begin{array}{rcll} \gcd(259, 70) & = & \gcd(70, 49) & \text{since } \text{rem}(259, 70) = 49 \\ & = & \gcd(49, 21) & \text{since } \text{rem}(70, 49) = 21 \\ & = & \gcd(21, 7) & \text{since } \text{rem}(49, 21) = 7 \\ & = & \gcd(7, 0) & \text{since } \text{rem}(21, 7) = 0 \\ & = & 7. & \end{array}$$

The Pulverizer goes through the same steps, but requires some extra bookkeeping along the way: as we compute $\gcd(a, b)$, we keep track of how to write each of the remainders (49, 21, and 7, in the example) as a linear combination of a and b (this is worthwhile, because our objective is to write the last nonzero remainder, which is the GCD, as such a linear combination). For our example, here is this extra bookkeeping:

x	y	$\text{rem}(x, y)$	$=$	$x - q \cdot y$
259	70	49	$=$	$259 - 3 \cdot 70$
70	49	21	$=$	$70 - 1 \cdot 49$
			$=$	$70 - 1 \cdot (259 - 3 \cdot 70)$
			$=$	$-1 \cdot 259 + 4 \cdot 70$
49	21	7	$=$	$49 - 2 \cdot 21$
			$=$	$(259 - 3 \cdot 70) - 2 \cdot (-1 \cdot 259 + 4 \cdot 70)$
			$=$	$\boxed{3 \cdot 259 - 11 \cdot 70}$
21	7	0		

We began by initializing two variables, $x = a$ and $y = b$. In the first two columns above, we carried out Euclid's algorithm. At each step, we computed $\text{rem}(x, y)$, which can be written in the form $x - q \cdot y$. (Remember that the Division Algorithm says $x = q \cdot y + r$, where r is the remainder. We get $r = x - q \cdot y$ by rearranging terms.) Then we replaced x and y in this equation with equivalent linear combinations of a and b, which we already had computed. After simplifying, we were left with a linear combination of a and b that was equal to the remainder as desired. The final solution is boxed.

Chapter 9

Sums, Products & Asymptotics

9.1 Closed Forms and Approximations

Sums and products arise regularly in the analysis of algorithms and in other technical areas such as finance and probabilistic systems. We've already seen that

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

Having a simple *closed form* expression such as $n(n+1)/2$ makes the sum a lot easier to understand and evaluate. We proved by induction that this formula is correct, but not where it came from. In Section 9.4, we'll discuss ways to find such closed forms. Even when there are no closed forms exactly equal to a sum, we may still be able to find a closed form that *approximates* a sum with useful accuracy.

The product we focus on in these notes is the familiar factorial:

$$n! ::= 1 \cdot 2 \cdots (n-1) \cdot n = \prod_{i=1}^{n} i.$$

We'll describe a closed form approximation for it called *Stirling's Formula*.

Finally, when there isn't a good closed form approximation for some expression, there may still be a closed form that characterizes its growth rate. We'll introduce *asymptotic notation*, such as "big Oh", to describe growth rates.

9.2 The Value of an Annuity

Would you prefer a million dollars today or $50,000 a year for the rest of your life? On the one hand, instant gratification is nice. On the other hand, the total dollars received at $50K per year is much larger if you live long enough.

Formally, this is a question about the value of an annuity. An *annuity* is a financial instrument that pays out a fixed amount of money at the beginning of every year for some specified number of

years. In particular, an n-year, m-payment annuity pays m dollars at the start of each year for n years. In some cases, n is finite, but not always. Examples include lottery payouts, student loans, and home mortgages. There are even Wall Street people who specialize in trading annuities.

A key question is what an annuity is worth. For example, lotteries often pay out jackpots over many years. Intuitively, $\$50,000$ a year for 20 years ought to be worth less than a million dollars right now. If you had all the cash right away, you could invest it and begin collecting interest. But what if the choice were between $\$50,000$ a year for 20 years and a *half* million dollars today? Now it is not clear which option is better.

In order to answer such questions, we need to know what a dollar paid out in the future is worth today. To model this, let's assume that money can be invested at a fixed annual interest rate p. We'll assume an 8% rate[1] for the rest of the discussion.

Here is why the interest rate p matters. Ten dollars invested today at interest rate p will become $(1+p)\cdot 10 = 10.80$ dollars in a year, $(1+p)^2\cdot 10 \approx 11.66$ dollars in two years, and so forth. Looked at another way, ten dollars paid out a year from now are only really worth $1/(1+p)\cdot 10 \approx 9.26$ dollars today. The reason is that if we had the $\$9.26$ today, we could invest it and would have $\$10.00$ in a year anyway. Therefore, p determines the value of money paid out in the future.

9.2.1 The Future Value of Money

Our goal is to determine the value of an n-year, m-payment annuity. The first payment of m dollars is truly worth m dollars. But the second payment a year later is worth only $m/(1+p)$ dollars. Similarly, the third payment is worth $m/(1+p)^2$, and the n-th payment is worth only $m/(1+p)^{n-1}$. The total value V of the annuity is equal to the sum of the payment values. This gives:

$$V = \sum_{i=1}^{n} \frac{m}{(1+p)^{i-1}}.$$

To compute the real value of the annuity, we need to evaluate this sum. One way is to plug in m, n, and p, compute each term explicitly, and then add them up. However, this sum has a special closed form that makes the job easier. (The phrase "closed form" refers to a mathematical expression without any summation or product notation.) First, lets make the summation prettier with some substitutions.

$$\begin{aligned} V &= \sum_{i=1}^{n} \frac{m}{(1+p)^{i-1}} \\ &= \sum_{j=0}^{n-1} \frac{m}{(1+p)^{j}} \quad (\text{substitute } j = i-1) \\ &= m\sum_{j=0}^{n-1} x^{j} \quad (\text{substitute } x = \frac{1}{1+p}). \end{aligned}$$

[1] U.S. interest rates have dropped steadily for several years, and ordinary bank deposits now earn around 1.5%. But just a few years ago the rate was 8%; this rate makes some of our examples a little more dramatic. The rate has been as high as 17% in the past twenty years.

In Japan, the standard interest rate is near zero%, and on a few occasions in the past few years has even been slightly negative. It's a mystery to U.S. economists why the Japanese populace keeps any money in their banks.

The goal of these substitutions is to put the summation into a special form so that we can bash it with a theorem given in the next section.

9.2.2 Geometric Sums

Theorem 9.2.1. *For all $n \geq 1$ and all $x \neq 1$,*[2]

$$\sum_{i=0}^{n-1} x^i = \frac{1-x^n}{1-x}.$$

The summation in this theorem is a *geometric sum*. The distinguishing feature of a geometric sum is that each of the terms

$$1,\ x,\ x^2,\ x^3,\ \ldots,\ x^{n-1}$$

in the sum is a constant times the one before; in this case, the constant is x. The theorem gives a closed form for a geometric sum that starts with 1.

We already saw one proof of this theorem in our lectures on induction. As is often the case, the proof by induction gives no hint about how the formula was found in the first place. Here is a more insightful derivation. The trick is to let S be the value of the sum and then observe what $-xS$ is:

$$\begin{array}{rcccccccc} S & = & 1 & +x & +x^2 & +x^3 & + & \cdots & +x^{n-1} \\ -xS & = & & -x & -x^2 & -x^3 & - & \cdots & -x^{n-1} - x^n. \end{array}$$

Adding these two equations gives:

$$S - xS = 1 - x^n,$$

so

$$S = \frac{1-x^n}{1-x}.$$

We'll say more about finding (as opposed to just proving) summation formulas later.

9.2.3 Return of the Annuity Problem

Now we can solve the annuity pricing problem. The value of an annuity that pays m dollars at the start of each year for n years is computed as follows:

$$\begin{aligned} V &= m\sum_{j=0}^{n-1} x^j \\ &= m\frac{1-x^n}{1-x} \\ &= m\frac{1-(\frac{1}{1+p})^n}{1-\frac{1}{1+p}} \\ &= m\frac{1+p-(\frac{1}{1+p})^{n-1}}{p}. \end{aligned}$$

[2] For these Notes, we'll adopt the convention that $0^0 ::= 0$.

The first line is a restatement of the summation we obtained earlier for the value of an annuity. The second line uses the closed form formula for a geometric sum. In the third line, we undo the earlier substitution $x = 1/(1+p)$. In the final step, both the numerator and denominator are multiplied by $1+p$ to simplify the expression.

The resulting formula is much easier to use than a summation with dozens of terms. For example, what is the real value of a winning lottery ticket that pays $\$50,000$ per year for 20 years? Plugging in $m = \$50,000$, $n = 20$, and $p = 0.08$ gives $V \approx \$530,180$. Because payments are deferred, the million dollar lottery is really only worth about a half million dollars! This is a good trick for the lottery advertisers!

9.2.4 Infinite Geometric Series

The question at the beginning of this section was whether you would prefer a million dollars today or $\$50,000$ a year for the rest of your life. Of course, this depends on how long you live, so optimistically assume that the second option is to receive $\$50,000$ a year *forever*. This sounds like infinite money!

We can compute the value of an annuity with an infinite number of payments by taking the limit of our geometric sum in Theorem 9.2.1 as n tends to infinity. This one is worth remembering!

Theorem 9.2.2. *If $|x| < 1$, then*

$$\sum_{i=0}^{\infty} x^i = \frac{1}{1-x}.$$

Proof.

$$\begin{aligned}
\sum_{i=0}^{\infty} x^i &= \lim_{n\to\infty} \sum_{i=0}^{n-1} x^i \\
&= \lim_{n\to\infty} \frac{1-x^n}{1-x} \\
&= \frac{1}{1-x}.
\end{aligned}$$

The first equality follows from the definition of an infinite summation. In the second line, we apply the formula for the sum of an n-term geometric sum given in Theorem 9.2.1. The final line follows by evaluating the limit; the x^n term vanishes since we assumed that $|x| < 1$. □

In our annuity problem, $x = 1/(1+p) < 1$, so the theorem applies. Substituting for x, we get an annuity value of

$$\begin{aligned}
V &= m \cdot \frac{1}{1-x} \\
&= m \cdot \frac{1}{1-1/(1+p)} \\
&= m \cdot \frac{1+p}{(1+p)-1} \\
&= m \cdot \frac{1+p}{p}.
\end{aligned}$$

Plugging in $m = \$50,000$ and $p = 0.08$ gives only $\$675,000$. Amazingly, a million dollars today is worth much more than $\$50,000$ paid every year forever! Then again, if we had a million dollars today in the bank earning 8% interest, we could take out and spend $\$80,000$ a year forever. So the answer makes some sense.

9.2.5 Examples

We now have closed form formulas for geometric sums and series. Some examples are given below. In each case, the solution follows immediately from either Theorem 9.2.1 (for finite sums) or Theorem 9.2.2 (for infinite series).

$$1 + 1/2 + 1/4 + 1/8 + \cdots = \sum_{i=0}^{\infty} (1/2)^i \qquad = \frac{1}{1-(1/2)} = 2 \tag{9.1}$$

$$0.999999999\ldots = 0.9\sum_{i=0}^{\infty} (1/10)^i \quad = 0.9\frac{1}{1-1/10} = 0.9\frac{10}{9} = 1 \tag{9.2}$$

$$1 - 1/2 + 1/4 - 1/8 + \cdots = \sum_{i=0}^{\infty} (-1/2)^i \qquad = \frac{1}{1-(-1/2)} = 2/3 \tag{9.3}$$

$$1 + 2 + 4 + 8 + \cdots + 2^{n-1} = \sum_{i=0}^{n-1} 2^i \qquad = \frac{1-2^n}{1-2} = 2^n - 1 \tag{9.4}$$

$$1 + 3 + 9 + 27 + \cdots + 3^{n-1} = \sum_{i=0}^{n-1} 3^i \qquad = \frac{1-3^n}{1-3} = \frac{3^n-1}{2} \tag{9.5}$$

If the terms in a geometric sum or series grow smaller, as in equation (9.1), then the sum is said to be *geometrically decreasing*. If the terms in a geometric sum grow progressively larger, as in (9.4) and (9.5), then the sum is said to be *geometrically increasing*.

Here is a good rule of thumb: *a geometric sum or series is approximately equal to the term with greatest absolute value*. In equations (9.1) and (9.3), the largest term is equal to 1 and the sums are 2 and 2/3, both relatively close to 1. In equation (9.4), the sum is about twice the largest term. In the final equation (9.5), the largest term is 3^{n-1} and the sum is $(3^n - 1)/2$, which is only about a factor of 1.5 greater.

9.2.6 Related Sums

We now know all about geometric sums. But in practice one often encounters sums that cannot be transformed by simple variable substitutions to the form $\sum x^i$.

A non-obvious, but useful way to obtain new summation formulas from old is by differentiating or integrating with respect to x. As an example, consider the following sum:

$$\sum_{i=1}^{n} ix^i = x + 2x^2 + 3x^3 + \cdots + nx^n$$

This is not a geometric sum, since the ratio between successive terms is not constant. Our formula for the sum of a geometric sum cannot be directly applied. But suppose that we differentiate that

formula:

$$\begin{aligned}
\frac{d}{dx}\sum_{i=0}^{n} x^i &= \frac{d}{dx}\frac{1-x^{n+1}}{1-x} \\
\sum_{i=1}^{n} i x^{i-1} &= \frac{-(n+1)x^n(1-x)-(-1)(1-x^{n+1})}{(1-x)^2} \\
&= \frac{-(n+1)x^n+(n+1)x^{n+1}+1-x^{n+1}}{(1-x)^2} \\
&= \frac{1-(n+1)x^n+nx^{n+1}}{(1-x)^2}.
\end{aligned}$$

Often differentiating or integrating messes up the exponent of x in every term. In this case, we now have a formula for a sum of the form $\sum ix^{i-1}$, but we want a formula for the series $\sum ix^i$. The solution is simple: multiply by x. This gives:

$$\sum_{i=1}^{n} i x^i = \frac{x-(n+1)x^{n+1}+nx^{n+2}}{(1-x)^2}$$

Since we could easily have made a mistake, it is a good idea to go back and validate a formula obtained this way with a proof by induction.

Notice that if $|x| < 1$, then this series converges to a finite value even if there are infinitely many terms. Taking the limit as n tends infinity gives the following theorem:

Theorem 9.2.3. *If $|x| < 1$, then*

$$\sum_{i=1}^{\infty} i x^i = \frac{x}{(1-x)^2}.$$

As a consequence, suppose there is an annuity that pays im dollars at the *end* of each year i forever. For example, if $m = \$50,000$, then the payouts are $\$50,000$ and then $\$100,000$ and then $\$150,000$ and so on. It is hard to believe that the value of this annuity is finite! But we can use the preceding theorem to compute the value:

$$\begin{aligned}
V &= \sum_{i=1}^{\infty} \frac{im}{(1+p)^i} \\
&= m\frac{\frac{1}{1+p}}{(1-\frac{1}{1+p})^2} \\
&= m\frac{1+p}{p^2}.
\end{aligned}$$

The second line follows by an application of Theorem 9.2.3. The third line is obtained by multiplying the numerator and denominator by $(1+p)^2$.

For example, if $m = \$50,000$, and $p = 0.08$ as usual, then the value of the annuity is $V = \$8,437,500$. Even though payments increase every year, the increase is only additive with time; by contrast, dollars paid out in the future decrease in value exponentially with time. The geometric

decrease swamps out the additive increase. Payments in the distant future are almost worthless, so the value of the annuity is finite.

The important thing to remember is the trick of taking the derivative (or integral) of a summation formula. Of course, this technique requires one to compute nasty derivatives correctly, but this is at least theoretically possible!

9.3 Book Stacking

Suppose you have a pile of books and you want to stack them on a table in some off-center way so the top book sticks out past books below it. How far past the edge of the table do you think you could get the top book to go without having the stack fall over? Could the top book stick out completely beyond the edge of table?

Most people's first response to this question—sometimes also their second and third responses—is "No, the top book will never get completely past the edge of the table." But in fact, you can get the top book to stick out as far as you want: one booklength, two booklengths, any number of booklengths!

9.3.1 Formalizing the Problem

We'll approach this problem recursively. How far past the end of the table can we get one book to stick out? It won't tip as long as its center of mass is over the table, so we can get it to stick out half its length, as shown in Figure 9.1.

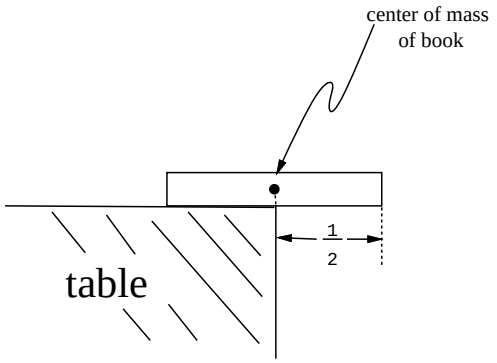


Figure 9.1: One book can overhang half a book length.

Now suppose we have a stack of books that will stick out past the table edge without tipping over—call that a *stable* stack. Let's define the *overhang* of a stable stack to be the largest horizontal distance from the center of mass of the stack to the furthest edge of a book. If we place the center of mass of the stable stack at the edge of the table as in Figure 9.2, that's how far we can get a book in the stack to stick out past the edge.

So we want a formula for the maximum possible overhang, B_n, achievable with a stack of n books.

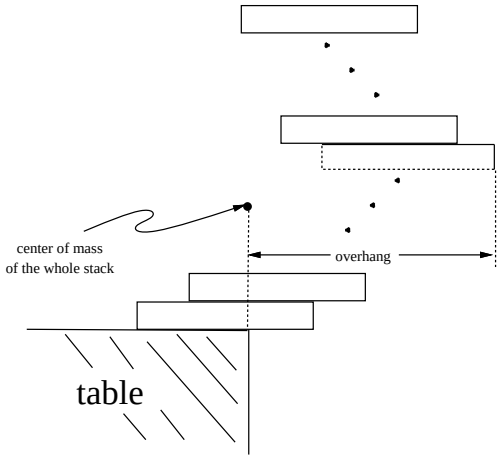


Figure 9.2: Overhanging the edge of the table.

We've already observed that the overhang of one book is $1/2$ a book length. That is,

$$B_1 = \frac{1}{2}.$$

Now suppose we have a stable stack of $n+1$ books with maximum overhang. If the overhang of the n books on top of the bottom book was not maximum, we could get a book to stick out further by replacing the top stack with a stack of n books with larger overhang. So the maximum overhang, B_{n+1}, of a stack of $n+1$ books is obtained by placing a maximum overhang stable stack of n books on top of the bottom book. And we get the biggest overhang for the stack of $n+1$ books by placing the center of mass of the n books right over the edge of the bottom book as in Figure 9.3.

So we know where to place the $n+1$st book to get maximum overhang, and all we have to do is calculate what it is. The simplest way to do that is to let the center of mass of the top n books be the origin. That way the horizontal coordinate of the center of mass of the whole stack of $n+1$ books will equal the increase in the overhang. But now the center of mass of the bottom book has horizontal coordinate $1/2$, so the horizontal coordinate of center of mass of the whole stack of $n+1$ books is

$$\frac{0 \cdot n + (1/2) \cdot 1}{n+1} = \frac{1}{2(n+1)}.$$

In other words,

$$B_{n+1} = B_n + \frac{1}{2(n+1)}, \tag{9.6}$$

as shown in Figure 9.3.

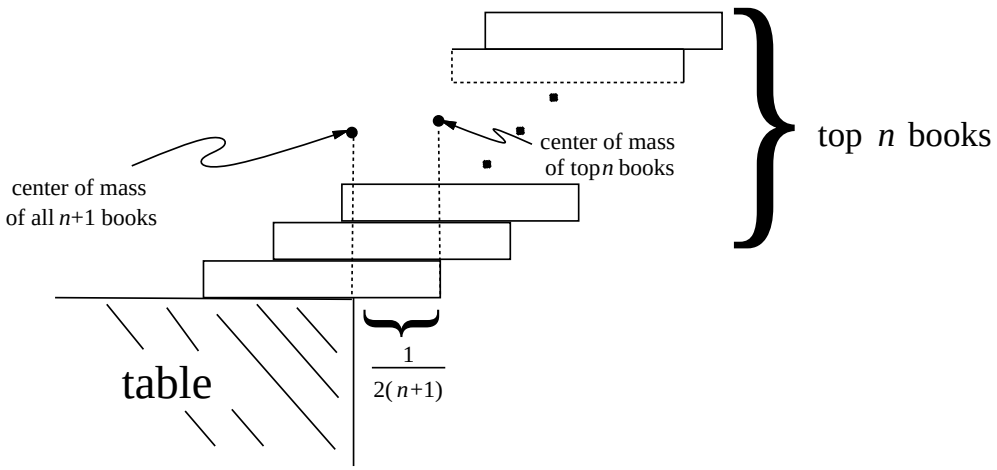


Figure 9.3: Additional overhang with $n+1$ books.

Expanding equation (9.6), we have

$$\begin{aligned} B_{n+1} &= B_{n-1} + \frac{1}{2n} + \frac{1}{2(n+1)} \\ &= B_1 + \frac{1}{2\cdot 2} + \cdots + \frac{1}{2n} + \frac{1}{2(n+1)} \\ &= \frac{1}{2}\sum_{i=1}^{n+1} \frac{1}{i}. \end{aligned}$$

Define

$$H_n ::= \sum_{i=1}^{n} \frac{1}{i}.$$

H_n is called the nth *Harmonic number*, and we have just shown that

$$B_n = \frac{H_n}{2}.$$

The first few Harmonic numbers are easy to compute. For example, $H_1 = 1$, $H_2 = 1 + \frac{1}{2} = \frac{3}{2}$, $H_3 = 1 + \frac{1}{2} + \frac{1}{3} = \frac{11}{6}$, $H_4 = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = \frac{25}{12}$. The fact that H_4 is greater than 2 has special significance; it implies that the total extension of a 4-book stack is greater than one full book! This is the situation shown in Figure 9.4.

In the next section we will prove that H_n grows slowly, but *unboundedly* with n. That means we can get books to overhang *any distance* past the edge of the table by piling them high enough!

9.3.2 Evaluating the Sum—The Integral Method

It would be nice to answer questions like, "How many books are needed to build a stack extending 100 book lengths beyond the table?" One approach to this question would be to keep computing

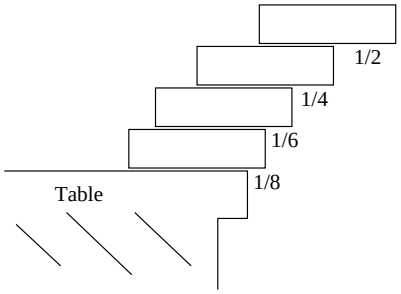


Figure 9.4: Stack of four books with maximum overhang.

Harmonic numbers until we found one exceeding 200. However, as we will see, this is not such a keen idea.

Such questions would be settled if we could express H_n in a closed form. Unfortunately, no closed form is known, and probably none exists. As a second best, however, we can find closed forms for very good approximations to H_n using the Integral Method. The idea of the Integral Method is to bound terms of the sum above and below by simple functions as suggested in Figure 9.5. The integrals of these functions then bound the value of the sum above and below.

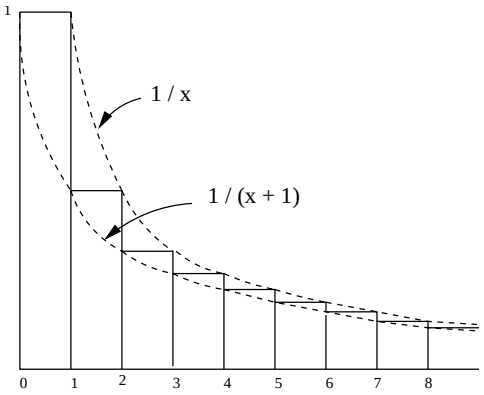


Figure 9.5: *This figure illustrates the Integral Method for bounding a sum. The area under the "stairstep" curve over the interval* $[0, n]$ *is equal to* $H_n = \sum_{i=1}^{n} 1/i$*. The function* $1/x$ *is everywhere greater than or equal to the stairstep and so the integral of* $1/x$ *over this interval is an upper bound on the sum. Similarly,* $1/(x+1)$ *is everywhere less than or equal to the stairstep and so the integral of* $1/(x+1)$ *is a lower bound on the sum.*

The Integral Method gives the following upper and lower bounds on the harmonic number H_n:

$$\begin{aligned} H_n &\leq 1 + \int_1^n \frac{1}{x}\, dx = 1 + \ln n \\ H_n &\geq \int_0^n \frac{1}{x+1}\, dx = \int_1^{n+1} \frac{1}{x}\, dx = \ln(n+1). \end{aligned} \tag{9.7}$$

These bounds imply that the harmonic number H_n is around $\ln n$. Since $\ln n$ grows without bound, albeit slowly, we can make a stack of books that extends arbitrarily far.

For example, to build a stack extending three book lengths beyond the table, we need a number

of books n so that $H_n \geq 6$. Exponentiating the above inequalities gives

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

This implies that we will need somewhere between 149 and 402 books. Actual calculation of H_n shows that 227 books will be the minimum number to overhang three book lengths.

9.3.3 More about Harmonic Numbers

In the preceding section, we showed that H_n is about $\ln n$. An even better approximation is known:

$$H_n = \ln n + \gamma + \frac{1}{2n} + \frac{1}{12n^2} + \frac{\epsilon(n)}{120n^4}$$

Here γ is a value $0.577215664\ldots$ called Euler's constant, and $\epsilon(n)$ is between 0 and 1 for all n. We will not prove this formula.

Asymptotic Equality

The shorthand $H_n \sim \ln n$ is used to indicate that the leading term of H_n is $\ln n$. More precisely:

Definition 9.3.1. For functions $f, g : \mathbb{R} \to \mathbb{R}$, we say f is *asymptotically equal* to g, in symbols,

$$f(x) \sim g(x)$$

iff

$$\lim_{x \to \infty} f(x)/g(x) = 1.$$

We also might write $H_n \sim \ln n + \gamma$ to indicate two leading terms. While this notation is widely used, it is not really right. Referring to the definition of $\sim$, we see that while $H_n \sim \ln n + \gamma$ is a true statement, so is $H_n \sim \ln n + c$ where c is any constant. The correct way to indicate that γ is the second-largest term is $H_n - \ln n \sim \gamma$.

The reason that the $\sim$ notation is useful is that often we do not care about lower order terms. For example, if $n = 100$, then we can compute $H(n)$ to great precision using only the two leading terms:

$$|H_n - \ln n - \gamma| \leq \left| \frac{1}{200} - \frac{1}{120000} + \frac{1}{120 \cdot 100^4} \right| < \frac{1}{200}.$$

9.4 Finding Summation Formulas

The source of the simple formula $\sum_{i=1}^{n} i = n(n+1)/2$ is still a mystery! Sure, we can prove this statement true by induction, but where did the expression on the right come from? Even more inexplicable is the summation formula for consecutive squares:

$$\begin{aligned} \sum_{i=1}^{n} i^2 &= \frac{(2n+1)(n+1)n}{6} \\ &= \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6} \\ &\sim \frac{n^3}{3}. \end{aligned}$$

Here is how we might find the sum-of-squares formula if we forgot it or had never seen it. First, the Integral Method gives a quick estimate of the sum:

$$\int_0^n x^2\,dx \le \sum_{i=1}^{n} i^2 \le \int_0^n (x+1)^2\,dx$$
$$\frac{n^3}{3} \le \sum_{i=1}^{n} i^2 \le \frac{(n+1)^3}{3} - \frac{1}{3}.$$

These upper and lower bounds obtained by the Integral Method show that $\sum_{i=1}^{n} i^2 \sim n^3/3$. To get an exact formula, we then guess the general form of the solution. Where we are uncertain, we can add parameters $a, b, c, \ldots$. For example, we might make the guess:

$$\sum_{i=1}^{n} i^2 = an^3 + bn^2 + cn + d.$$

If the guess is correct, then we can determine the parameters a, b, c, and d by plugging in a few values for n. Each such value gives a linear equation in a, b, c, and d. If we plug in enough values, we may get a linear system with a unique solution. Applying this method to our example gives:

$$\begin{aligned} n = 0 &\rightarrow 0 = d \\ n = 1 &\rightarrow 1 = a + b + c + d \\ n = 2 &\rightarrow 5 = 8a + 4b + 2c + d \\ n = 3 &\rightarrow 14 = 27a + 9b + 3c + d. \end{aligned}$$

Solving this system gives the solution $a = 1/3$, $b = 1/2$, $c = 1/6$, $d = 0$. Therefore, if our initial guess at the form of the solution was correct, then the summation is equal to $n^3/3 + n^2/2 + n/6$. In fact, our initial guess *was* correct, this is the right formula for the sum of squares!

Be careful! After obtaining a formula by this method, always go back and prove it using induction or some other method. This is not merely a check for algebra blunders; if the initial guess at the solution was not of the right form, then the resulting formula will be completely wrong!

9.5 Double Sums

Sometimes we have to evaluate sums of sums, otherwise known as *double summations*. Sometimes it is easy: we can evaluate the inner sum, replace it with a closed form, and then evaluate the outer sum which no longer has a summation inside it.

But there's a special trick that is often extremely useful for sums, which is *exchanging the order of summation.* It's best demonstrated by example. Suppose we want to compute the sum of the harmonic numbers

$$\sum_{k=1}^{n} H_k = \sum_{k=1}^{n} \sum_{j=1}^{k} 1/j$$

For intuition about this sum, we can try the integral method:

$$\sum_{k=1}^{n} H_k \approx \int_1^n \ln x\,dx \approx n \ln n - n.$$

Now let's look for an exact answer. If we think about the pairs (k, j) over which we are summing, they form a triangle:

		j						
		1	2	3	4	5	...	n
k	1	1						
	2	1	1/2					
	3	1	1/2	1/3				
	4	1	1/2	1/3	1/4			
		...						
	n	1	1/2		...			$1/n$

The summation above is summing each row and then adding the row sums. Instead, we can sum the columns and then add the column sums. Inspecting the table we see that this double sum can be written as

$$\begin{aligned}
\sum_{k=1}^{n} H_k &= \sum_{k=1}^{n}\sum_{j=1}^{k} 1/j \\
&= \sum_{j=1}^{n}\sum_{k=j}^{n} 1/j \\
&= \sum_{j=1}^{n} 1/j \sum_{k=j}^{n} 1 \\
&= \sum_{j=1}^{n} \frac{1}{j}(n-j+1) \\
&= \sum_{j=1}^{n} \frac{n-j+1}{j} \\
&= \sum_{j=1}^{n} \frac{n+1}{j} - \sum_{j=1}^{n} \frac{j}{j} \\
&= (n+1)\sum_{j=1}^{n} \frac{1}{j} - \sum_{j=1}^{n} 1 \\
&= (n+1)H_n - n. \qquad (9.8)
\end{aligned}$$

9.6 Stirling's Approximation

The familiar factorial notation, $n!$, is an abbreviation for the product

$$\prod_{i=1}^{n} i.$$

This is by far the most common product in Discrete Mathematics. In this section we describe a good closed-form estimate of $n!$ called *Stirling's Approximation*. Unfortunately, all we can do is estimate: there is no closed form for $n!$ — though proving so would take us beyond the scope of 6.042.

9.6.1 Products to Sums

A good way to handle a product is often to convert it into a sum by taking the logarithm. In the case of factorial, this gives

$$\begin{aligned}\ln(n!) &= \ln(1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n) \\ &= \ln 1 + \ln 2 + \ln 3 + \cdots + \ln(n-1) + \ln n \\ &= \sum_{i=1}^{n} \ln i.\end{aligned}$$

We've not seen a summation containing a logarithm before! Fortunately, one tool that we used in evaluating sums is still applicable: the Integral Method. We can bound the terms of this sum with $\ln x$ and $\ln(x+1)$ as shown in Figure 9.6. This gives bounds on $\ln(n!)$ as follows:

$$\begin{aligned}\int_1^n \ln x \, dx \leq &\quad \textstyle\sum_{i=1}^n \ln i &&\leq \int_0^n \ln(x+1)\, dx \\ n\ln(\frac{n}{e}) + 1 \leq &\quad \textstyle\sum_{i=1}^n \ln i &&\leq (n+1)\ln\left(\frac{n+1}{e}\right) + 1 \\ \left(\frac{n}{e}\right)^n e \leq &\quad n! &&\leq \left(\frac{n+1}{e}\right)^{n+1} e.\end{aligned}$$

The second line follows from the first by completing the integrations. The third line is obtained by exponentiating.

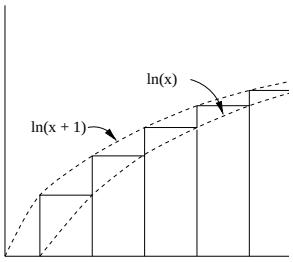


Figure 9.6: *This figure illustrates the Integral Method for bounding the sum* $\sum_{i=1}^{n} ln\ i$.

So $n!$ behaves something like the closed form formula $(n/e)^n$. A more careful analysis yields an unexpected closed form formula that is asymptotically exact:

Lemma (Stirling's Formula).

$$n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n},$$

Stirling's Formula describes how $n!$ behaves in the limit, but to use it effectively, we need to know how close it is to the limit for different values of n. That information is given by the bounding formulas:

Fact (Stirling's Approximation).

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{1/(12n+1)} \leq n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{1/12n}.$$

The Approximation implies the asymptotic Formula, since $e^{1/(12n+1)}$ and $e^{1/12n}$ both approach 1 as n grows large. These inequalities can be verified by induction, but the details are nasty.

The bounds in Stirling's formula are very tight. For example, if $n = 100$, then Stirling's bounds are:

$$\begin{aligned} 100! &\geq \sqrt{200\pi}\left(\frac{100}{e}\right)^{100} e^{1/1201} \\ 100! &\leq \sqrt{200\pi}\left(\frac{100}{e}\right)^{100} e^{1/1200} \end{aligned}$$

The only difference between the upper bound and the lower bound is in the final term. In particular $e^{1/1201} \approx 1.00083299$ and $e^{1/1200} \approx 1.00083368$. As a result, the upper bound is no more than $1 + 10^{-6}$ times the lower bound. This is amazingly tight! Remember Stirling's formula; we will use it often.

9.6.2 Bounds by Double Summing

Another way to derive Stirling's approximation is to remember that $\ln n$ is roughly the same as H_n. This lets us use the result we derived before for $\sum H_k$ via double summation. Our approximation for H_k told us that $\ln(k+1) \leq H_k \leq 1 + \ln k$. Rewriting, we find that $H_k - 1 \leq \ln k \leq H_{k-1}$. It follows that (leaving out the $i = 1$ term in the sum, which contributes 0),

$$\begin{aligned} \sum_{i=2}^{n} \ln i &\leq \sum_{i=2}^{n} H_{i-1} \\ &= \sum_{i=1}^{n-1} H_i \\ &= nH_{n-1} - (n-1) && \text{by (9.8)} \\ &\leq n(1 + \ln(n-1)) - (n-1) && \text{by (9.7)} \\ &= n\ln(n-1) + 1, \end{aligned}$$

roughly the same bound as we proved before via the integral method. We can derive a similar lower bound.

9.7 Asymptotic Notation

Asymptotic notation is a shorthand used to give a quick measure of the behavior of a function $f(n)$ as n grows large.

9.7.1 Little Oh

The asymptotic notation $\sim$ of Definition 9.3.1 is a binary relation indicating that two functions grow at the *same* rate. There is a related strict partial order on functions indicating that one function grows at a significantly *slower* rate. Namely,

Definition 9.7.1. For functions $f, g : \mathbb{R} \to \mathbb{R}$, we say f is *asymptotically smaller* than g, in symbols,

$$f(x) = o(g(x)),$$

iff

$$\lim_{x\to\infty} f(x)/g(x) = 0.$$

For example, $1000x^{1.9} = o(x^2)$, because $1000x^{1.9}/x^2 = 1000/x^{0.1}$ and since $x^{0.1}$ goes to infinity with x and 1000 is constant, we have $\lim_{x\to\infty} 1000x^{1.9}/x^2 = 0$. This argument generalizes directly to yield

Lemma 9.7.2. *$x^a = o(x^b)$ for all nonnegative constants $a < b$.*

Using the familiar fact that $\log x < x$ for all $x > 1$, we can prove

Lemma 9.7.3. *$\log x = o(x^\epsilon)$ for all $\epsilon > 0$ and $x > 1$.*

Proof. Choose $\epsilon > \delta > 0$ and let $x = z^\delta$ in the inequality $\log x < x$. This implies

$$\log z < z^\delta/\delta = o(z^\epsilon) \qquad \text{by Lemma 9.7.2.} \tag{9.9}$$

□

Corollary 9.7.4. *$x^b = o(a^x)$ for any $a, b \in \mathbb{R}$ with $a > 1$.*

Proof. From (9.9),

$$\log z < z^\delta/\delta$$

for all $z > 1$, $\delta > 0$. Hence

$$\begin{aligned}
(e^b)^{\log z} &< (e^b)^{z^\delta/\delta} \\
z^b &< \left(e^{\log a(b/\log a)}\right)^{z^\delta/\delta} \\
&= a^{(b/\delta \log a) z^\delta} \\
&< a^z
\end{aligned}$$

for all z such that

$$(b/\delta \log a) z^\delta < z.$$

But choosing $\delta < 1$, we know $z^\delta = o(z)$, so this last inequality holds for all large enough z. □

Lemma 9.7.3 and Corollary 9.7.4 can also be proved easily in several other ways, for example, using L'Hopital's Rule or the McLaurin Series for $\log x$ and e^x. Proofs can be found in most calculus texts.

Problem 9.7.1. Prove the initial claim that $\log x < x$ for all $x > 1$ (requires elementary calculus).

Problem 9.7.2. Prove that the relation, R, on functions such that fRg iff $f = o(g)$ is a strict partial order, namely, R is transitive and *asymmetric*: if fRg then $\neg gRf$.

Problem 9.7.3. Prove that $f \sim g$ iff $f = g + h$ for some function $h = o(g)$.

9.7.2 Big Oh

Big Oh is the most frequently used asymptotic notation. It is used to give an upper bound on the growth of a function, such as the running time of an algorithm.

Definition 9.7.5. Given functions $f, g : \mathbb{R} \to \mathbb{R}$, with g nonnegative, we say that

$$f = O(g)$$

iff

$$\limsup_{x\to\infty} |f(x)| / g(x) < \infty.$$

This definition[3] makes it clear that

Lemma 9.7.6. *If $f = o(g)$ or $f \sim g$, then $f = O(g)$.*

Proof. $\lim f/g = 0$ or $\lim f/g = 1$ implies $\lim f/g < \infty$. □

It is easy to see that the converse of Lemma 9.7.6 is not true. For example, $2x = O(x)$, but $2x \not\sim x$ and $2x \neq o(x)$.

The usual formulation of Big Oh spells out the definition of $\limsup$ without mentioning it. Namely, here is an equivalent definition:

Definition 9.7.7. Given functions $f, g : \mathbb{R} \to \mathbb{R}$, we say that

$$f = O(g)$$

iff there exists a constant $c \geq 0$ and an x_0 such that for all $x \geq x_0$, $|f(x)| \leq cg(x)$.

This definition is rather complicated, but the idea is simple: $f(x) = O(g(x))$ means $f(x)$ is less than or equal to $g(x)$, except that we're willing to ignore a constant factor, namely, c, and to allow exceptions for small x, namely, $x < x_0$.

We observe,

Lemma 9.7.8. *Assume that g is nonnegative. If $f = o(g)$, then it is* not *true that $g = O(f)$.*

Proof.

$$\lim_{x\to\infty} \frac{g(x)}{f(x)} = \frac{1}{\lim_{x\to\infty} \frac{f(x)}{g(x)}} = \frac{1}{0} = \infty,$$

so $g \neq O(f)$.

□

[3] $$\limsup_{x\to\infty} h(x) ::= \lim_{x\to\infty} \text{lub}_{y\geq x} h(y).$$

We need the $\limsup$ in the definition of $O()$ because if $f(x)/g(x)$ oscillates between, say, 3 and 5 as x grows, then $f = O(g)$ because $f \leq 5g$, but $\lim_{x\to\infty} f(x)/g(x)$ does not exist. However, in this case we would have $\limsup_{x\to\infty} f(x)/g(x) = 5$.

Proposition 9.7.9. $100x^2 = O(x^2)$.

Proof. Choose $c = 100$ and $x_0 = 1$. Then the proposition holds, since for all $x \geq 1$, $\left|100x^2\right| \leq 100x^2$. □

Proposition 9.7.10. $x^2 + 100x + 10 = O(x^2)$.

Proof. $(x^2+100x+10)/x^2 = 1+100/x+10/x^2$ and so its limit as x approaches infinity is $1+0+0 = 1$. So in fact, $x^2 + 100x + 10 \sim x^2$, and therefore $x^2 + 100x + 10 = O(x^2)$. Indeed, it's conversely true that $x^2 = O(x^2 + 100x + 10)$. □

Proposition 9.7.10 generalizes to an arbitrary polynomial:

Proposition 9.7.11. *For $a_k \neq 0$, $a_k x^k + a_{k-1}x^{k-1} + \cdots + a_1 x + a_0 = O(x^k)$.*

The routine proof is left to the reader.

Big Oh notation is especially useful when describing the running time of an algorithm. For example, the usual algorithm for multiplying $n \times n$ matrices requires proportional to n^3 operations in the worst case. This fact can be expressed concisely by saying that the running time is $O(n^3)$. So this asymptotic notation allows the speed of the algorithm to be discussed without reference to constant factors or lower-order terms that might be machine specific. In this case there is another, ingenious matrix multiplication procedure that requires $O(n^{2.55})$ operations. This procedure will therefore be much more efficient on large enough matrices. Unfortunately, the $O(n^{2.55})$-operation multiplication procedure is almost never used because it happens to be less efficient than the usual $O(n^3)$ procedure on matrices of practical size. It is even conceivable that there is an $O(n^2)$ matrix multiplication procedure, but none is known.

9.7.3 Theta

Definition 9.7.12.

$$f = \Theta(g) \quad \text{iff} \quad f = O(g) \text{ and } g = O(f).$$

The statement $f = \Theta(g)$ can be paraphrased intuitively as "f and g are equal to within a constant factor."

The value of these notations is that they highlight growth rates and allow suppression of distracting factors and low-order terms. For example, if the running time of an algorithm is

$$T(n) = 10n^3 - 20n^2 + 1,$$

then

$$T(n) = \Theta(n^3).$$

In this case, we would say that T *is of order* n^3 or that $T(n)$ *grows cubically*.

Another such example is

$$\pi^2 3^{x-7} + \frac{(2.7x^{113} + x^9 - 86)^4}{\sqrt{x}} - 1.08^{3x} = \Theta(3^x).$$

Just knowing that the running time of an algorithm is $\Theta(n^3)$, for example, is useful, because if n doubles we can predict that the running time will *by and large*[4] increase by a factor of at most 8 for large n. In this way, Theta notation preserves information about the scalability of an algorithm or system. Scalability is, of course, a big issue in the design of algorithms and systems.

9.7.4 Pitfalls with Big Oh

There is a long list of ways to make mistakes with Big Oh notation. This section presents some of the ways that Big Oh notation can lead to ruin and despair.

The Exponential Fiasco

Sometimes relationships involving Big Oh are not so obvious. For example, one might guess that $4^x = O(2^x)$ since 4 is only a constant factor larger than 2. This reasoning is incorrect, however; actually 4^x grows much faster than 2^x.

Proposition 9.7.13. $4^x \neq O(2^x)$

Proof. $2^x/4^x = 2^x/(2^x 2^x) = 1/2^x$. Hence, $\lim_{x\to\infty} 2^x/4^x = 0$, so in fact $2^x = o(4^x)$. We observed earlier that this implies that $4^x \neq O(2^x)$. □

Constant Confusion

Every constant is $O(1)$. For example, $17 = O(1)$. This is true because if we let $f(x) = 17$ and $g(x) = 1$, then there exists a $c > 0$ and an x_0 such that $|f(x)| \leq cg(x)$. In particular, we could choose c = 17 and $x_0 = 1$, since $|17| \leq 17 \cdot 1$ for all $x \geq 1$. We can construct a false theorem that exploits this fact.

False Theorem 9.7.14.

$$\sum_{i=1}^{n} i = O(n)$$

False proof. Define $f(n) = \sum_{i=1}^{n} i = 1 + 2 + 3 + \cdots + n$. Since we have shown that every constant i is $O(1)$, $f(n) = O(1) + O(1) + \cdots + O(1) = O(n)$. □

Of course in reality $\sum_{i=1}^{n} i = n(n+1)/2 \neq O(n)$.

The error stems from confusion over what is meant in the statement $i = O(1)$. For any *constant* $i \in \mathbb{N}$ it is true that $i = O(1)$. More precisely, if f is any constant function, then $f = O(1)$. But in this False Theorem, i is not constant but ranges over a set of values 0,1,...,n that depends on n.

And anyway, we should not be adding $O(1)$'s as though they were numbers. We never even defined what $O(g)$ means by itself; it should only be used in the context "$f = O(g)$" to describe a relation between functions f and g.

[4] Since $\Theta(n^3)$ only implies that the running time, $T(n)$, is between cn^3 and dn^3 for constants $0 < c < d$, the time $T(2n)$ could regularly exceed $T(n)$ by a factor as large as $8d/c$. The factor is sure to be close to 8 for all large n only if $T(n) \sim n^3$.

Lower Bound Blunder

Sometimes people incorrectly use Big Oh in the context of a lower bound. For example, they might say, "The running time, $T(n)$, is at least $O(n^2)$," when they probably mean something like "$O(T(n)) = n^2$," or more properly, "$n^2 = O(T(n))$."

Equality Blunder

The notation $f = O(g)$ is too firmly entrenched to avoid, but the use of "=" is really regrettable. For example, if $f = O(g)$, it seems quite reasonable to write $O(g) = f$. But doing so might tempt us to the following blunder: because $2n = O(n)$, we can say $O(n) = 2n$. But $n = O(n)$, so we conclude that $n = O(n) = 2n$, and therefore $n = 2n$. To avoid such nonsense, we will never write "$O(f) = g$."

9.8 In-Class Problems Week 8, Fri.

Problem 9.8.1. We begin with two large glasses. The first glass contains a pint of water, and the second contains a pint of wine. We pour $1/3$ of a pint from the first glass into the second, stir up the wine/water mixture in the second glass, and then pour $1/3$ of a pint of the mix back into the first glass and repeat this pouring back-and-forth process a total of n times.

(a) Describe a closed form formula for the amount of wine in the first glass after n back-and-forth pourings.

Solution. The state of the system of glasses/wine/water at the beginning of a round of pouring and pouring back is determined by the total amount of wine in the first glass. Suppose at the beginning of some round, the first glass contains w pints of wine, $0 \leq w \leq 1$ and $1 - w$ pints of water. The second glass contains the rest of the wine and water.

Pouring $1/3$ pint from the first glass to the second leaves $2/3$ pints of liquid and $(2/3)w$ wine in the first glass, and $4/3$ pints of liquid and $1 - (2/3)w$ wine in the second glass. Pouring $1/3$ pint back from the second into the first transfers a proportion of $(1/3)/(4/3)$ of the wine in the second glass into the first. So the round completes with both glasses containing a pint of liquid, and the first glass containing

$$(2/3)w + (1/4)(1 - (2/3)w) = 1/4 + w/2$$

pints of wine. After one more round, the first glass contains

$$1/4 + (1/4 + w/2)/2 = 1/4 + 1/8 + w/2^2$$

pints of wine, and after n more rounds

$$\begin{aligned} w/2^n + \Sigma_{i=1}^{n}(1/2)^{i+1} &= w/2^n + (1/2)\Sigma_{i=1}^{n}(1/2)^i \\ &= w/2^n + (1/2)(-1 + \Sigma_{i=0}^{n}(1/2)^i) \\ &= w/2^n + (1/2)(-1 + (1 - (1/2)^{n+1})/(1 - 1/2)) \\ &= w/2^n - 1/2 + 1 - (1/2)^{n+1} \\ &= w/2^n + 1/2 - (1/2)^{n+1}. \end{aligned}$$

Since $w = 0$ initially, the pints of wine in the first glass after n rounds is

$$1/2 - (1/2)^{n+1}.$$

■

(b) What is the limit of the amount of wine in each glass as n approaches infinity?

Solution. The limiting amount of wine in the first glass approaches $1/2$ from below as n approaches infinity. In fact, it approaches $1/2$ no matter how the wine was initially distributed. This of course is what you would expect: after a thorough mixing the glasses should contain essentially the same amount of wine. ■

Problem 9.8.2. Suppose you were about to enter college today and a college loan officer offered you the following deal: \$25,000 at the start of each year for four years to pay for your college tuition and an option of choosing one of the following repayment plans:

Plan A: Wait four years, then repay \$20,000 at the start of each year for the next ten years.

Plan B: Wait five years, then repay \$30,000 at the start of each year for the next five years.

Suppose the annual interest rate paid by banks is 7% and does not change in the future.

(a) Assuming that it's no hardship for you to meet the terms of either payback plan, which one is a better deal? (You will need a calculator.)

Solution. \$1 today will be worth $\$1.07$ next year, and $\$1.07^2$ the year after, *etc*. So set $r = \frac{1}{1.07}$. Then:

$$\begin{aligned}
&\text{current value of Plan A}\\
&= \sum_{y=4}^{13} 20000 \cdot r^y\\
&= \sum_{y=0}^{9} 20000 \cdot r^{y+4}\\
&= r^4 \cdot \sum_{y=0}^{9} 20000 \cdot r^y\\
&= 20000r^4 \cdot \sum_{y=0}^{9} r^y\\
&= 20000r^4 \cdot \frac{1-r^{10}}{1-r}\\
&= \$114,666.69
\end{aligned}$$

$$\begin{aligned}
&\text{current value of Plan B}\\
&= \sum_{y=5}^{9} 30000 \cdot r^y\\
&= \sum_{y=0}^{4} 30000 \cdot r^{y+5}\\
&= r^5 \cdot \sum_{y=0}^{4} 30000 \cdot r^y\\
&= 30000r^5 \cdot \sum_{y=0}^{4} r^y\\
&= 30000r^5 \cdot \frac{1-r^5}{1-r}\\
&= \$93,840.63
\end{aligned}$$

You should clearly take Plan B. You will be paying back much less in today's dollars. ■

(b) What is the loan officer's effective profit (in today's dollars) on the loan?

Solution. The value of the money you are given is:

$$\begin{aligned} Loan &= \sum_{y=0}^{3} 25000 \cdot r^y \\ &= 25000 \cdot \sum_{y=0}^{3} r^y \\ &= 25000 \cdot \frac{1-r^4}{1-r} \\ &= \$90,607.90 \end{aligned}$$

Therefore, the loan officer's profit is effectively $\$3,233$. (Or $\$24,059$ if we are not on the ball). ■

Problem 9.8.3. Riemann's Zeta Function $\zeta(k)$ is defined to be the infinite summation:

$$1 + \frac{1}{2^k} + \frac{1}{3^k} \cdots = \sum_{j \geq 1} \frac{1}{j^k}$$

Below is a proof that

$$\sum_{k \geq 2} (\zeta(k) - 1) = 1$$

Justify each line of the proof. (P.S. The purpose of this exercise is to highlight some of the rules for manipulating series. Don't worry about the significance of this identity.)

$$\begin{aligned}
\sum_{k\geq 2}(\zeta(k)-1) &= \sum_{k\geq 2}\left[\left(\sum_{j\geq 1}\frac{1}{j^k}\right)-1\right] && (9.10)\\
&= \sum_{k\geq 2}\sum_{j\geq 2}\frac{1}{j^k} && (9.11)\\
&= \sum_{j\geq 2}\sum_{k\geq 2}\frac{1}{j^k} && (9.12)\\
&= \sum_{j\geq 2}\frac{1}{j^2}\sum_{k\geq 0}\frac{1}{j^k} && (9.13)\\
&= \sum_{j\geq 2}\frac{1}{j^2}\cdot\frac{1}{1-1/j} && (9.14)\\
&= \sum_{j\geq 2}\frac{1}{j(j-1)} && (9.15)\\
&= \lim_{n\to\infty}\sum_{j=2}^{n}\frac{1}{j(j-1)} && (9.16)\\
&= \lim_{n\to\infty}\sum_{j=2}^{n}\frac{1}{j-1}-\frac{1}{j} && (9.17)\\
&= \lim_{n\to\infty}(1-\frac{1}{n}) && (9.18)\\
&= 1 && (9.19)
\end{aligned}$$

Solution. (9.10) Definition of $\zeta(k)$.

(9.11) Because $\sum_{j\geq 1}\frac{1}{j^k} = 1+\sum_{j\geq 2}\frac{1}{j^k}$.

(9.12) Reordering; this is ok because at this point all terms are positive.

(9.13) Because $\sum_{k\geq 2}\frac{1}{j^k} = \sum_{k\geq 0}\frac{1}{j^{k+2}} = \sum_{k\geq 0}\frac{1}{j^k\cdot j^2} = \frac{1}{j^2}\sum_{k\geq 0}\frac{1}{j^k}$.

(9.14) Sum of a geometric series.

(9.15) Algebra inside every summand.

(9.16) Definition of infinite summation.

(9.17) Algebra inside every summand.

(9.18) The sum telescopes: 1 is added once; every one of $\frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n-1}$ is subtracted once and then added once; $\frac{1}{n}$ is subtracted once.

(9.19) Simple limits. ■

9.9 In-Class Problems Week 9, Mon.

Problem 9.9.1. There is a bug on the edge of a 1-meter rug. The bug wants to cross to the other side of the rug. It crawls at 1 cm per second. However, at the end of each second, a malicious first-grader named Mildred Anderson *stretches* the rug by 1 meter. Assume that her action is instantaneous and the rug stretches uniformly. Thus, here's what happens in the first few seconds:

- The bug walks 1 cm in the first second, so 99 cm remain ahead.
- Mildred stretches the rug by 1 meter, which doubles its length. So now there are 2 cm behind the bug and 198 cm ahead.
- The bug walks another 1 cm in the next second, leaving 3 cm behind and 197 cm ahead.
- Then Mildred strikes, stretching the rug from 2 meters to 3 meters. So there are now $3 \cdot (3/2) = 4.5$ cm behind the bug and $197 \cdot (3/2) = 295.5$ cm ahead.
- The bug walks another 1 cm in the third second, and so on.

Your job is to determine this poor bug's fate.

(a) During second i, what *fraction* of the rug does the bug cross?

Solution. During second i, the length of the rug is $100i$ cm and the bug crosses 1 cm. Therefore, the fraction that the bug crosses is $1/100i$. ■

(b) Over the first n seconds, what fraction of the rug does the bug cross altogether? Express your answer in terms of the Harmonic number H_n.

Solution. The bug crosses $1/100$ of the rug in the first second, $1/200$ in the second, $1/300$ in the third, and so forth. Thus, over the first n seconds, the fraction crossed by the bug is:

$$\sum_{k=1}^{n} \frac{1}{100k} = H_n/100$$

(This formula is valid only until the bug reaches the far side of the rug.) ■

(c) Approximately how many seconds does the bug need to cross the entire rug?

Solution. The bug arrives at the far side when the fraction it has crossed reaches 1. This occurs when n, the number of seconds elapsed, is sufficiently large that $H_n/100 \geq 1$. Now H_n is approximately $\ln n$, so the bug arrives about when:

$$\begin{aligned} \frac{\ln n}{100} &\geq 1 \\ \ln n &\geq 100 \\ n &\geq e^{100} \approx 10^{43} \text{ seconds} \end{aligned}$$

■

Problem 9.9.2. Using the method described in lecture, a truck can travel across any size desert if there is a large enough supply of gas at the border of the desert. Show that if there is a large enough supply of gas at the border, a truck can also make a *round trip* across any size desert.

Solution. Given that it can make a one-way trip across any desert, it can make a two-way trip by executing the one-way strategy for twice the desert width, but turning around when it gets to the desert edge instead of continuing.

A considerably more efficient approach uses ideas similar to the one-way crossing strategy: let R_n be the distance a truck can travel into the desert *and return* on n tanks of gas. Clearly, $R_1 = 1/2$.

On $n+1$ tanks, the strategy is to have the truck travel distance x and back n times, leaving $1-2x$ tanks of gas at distance x into the desert on each trip. It then makes one more one-way trip to x. This leaves it with $n(1-2x)+1-x$ tanks of gas at position x. Leaving an xth of a tank so it can get back, if the remaining $(n(1-2x)+1-x)-x = (n+1)(1-2x)$ tanks equal n, it can execute the n-tank round trip strategy from position x and still return to the desert border. So, letting

$$\begin{aligned}(n+1)(1-2x) &= n && (9.20)\\ x &= 1/2(n+1) && (9.21)\\ R_{n+1} &= R_n + x = R_n + 1/2(n+1). && (9.22)\end{aligned}$$

Therefore,

$$R_n = 1/2(1 + 1/2 + 1/3 + \cdots + 1/n) = H_n/2.$$

■

Problem 9.9.3. There is a number a such that $\sum_{i=1}^{\infty} i^p$ converges iff $p < a$. What is the value of a? Prove it.

Solution. $a = -1$.

For $p = -1$, the sum is the harmonic series which we know does not converge. Since the term i^p is increasing in p for $i > 1$, the sum will be larger, and hence also diverge for $p > -1$.

For $p < -1$ there exists an $\epsilon > 0$ such that $p = -(1+\epsilon)$. By the integral method,

$$\begin{aligned}\sum_{i=1}^{\infty} i^{-(1+\epsilon)} &\le 1 + \int_1^{\infty} x^{-(1+\epsilon)}\,dx\\ &= 1 + \epsilon^{-1} - \epsilon^{-1}\lim_{\alpha\to\infty}\alpha^{-\epsilon}\\ &= 1 + \epsilon^{-1}\\ &< \infty\end{aligned}$$

Hence the sum is bounded above, and since it is increasing, it has a finite limit, that is, it converges. ■

9.10 Problem Set 7

Problem 9.10.1. Prove that for any prime, p, and integer, $k \geq 1$,

$$\phi(p^k) = p^k - p^{k-1},$$

where ϕ is Euler's function. *Hint:* Which numbers between 0 and $p^k - 1$ *are* divisible by p? How many are there?

Solution. The numbers in the interval from 0 to $p^k - 1$ that are divisible by p are all those of the form mp. For mp to be in the interval, m can take any value from 0 to $p^{k-1} - 1$ and no others, so there are exactly p^{k-1} numbers in the interval that are divisible by p. Now $\phi(p^k)$ equals the number of remaining elements in the interval, namely, $p^k - p^{k-1}$. ■

Problem 9.10.2. Suppose m, n are relatively prime.

(a) Prove that for any a, b, there is an x such that

$$x \equiv a \pmod{m}, \tag{9.23}$$
$$x \equiv b \pmod{n}. \tag{9.24}$$

Hint: Congruence (9.23) holds iff

$$x = jm + a. \tag{9.25}$$

for some j. So there is such an x only if

$$jm + a \equiv b \pmod{n}. \tag{9.26}$$

Solve (9.26) for j.

Solution. *Proof.* [5] Since m, n are relatively prime, there is an inverse, m', modulo n of m. So $j = m'(b - a)$ satisfies (9.26). Now (9.25) leads to the definition

$$x_1 ::= m'(b - a)m + a.$$

So

$$x_1 = (m'(b - a))m + a \equiv a \pmod{m},$$

and

$$x_1 = m'(b - a)m + a = m'm(b - a) + a \equiv 1 \cdot (b - a) + a \equiv b \pmod{n},$$

proving that x_1 satisfies the congruences (9.23) and (9.24). □

■

[5] Adapted from http://www.cut-the-knot.org/blue/chinese.shtml.

(b) Prove that there is an x satisfying the congruences (9.23) and (9.24) such that $0 \leq x < mn$.

Solution. Let

$$x_0 ::= \text{rem}(x_1, mn),$$

where x_1 satisfies (9.23) and (9.24).

Now $0 \leq x_0 < mn$ by definition of remainder. Further, we know $x_0 \equiv x_1 \pmod{mn}$, which immediately implies that $x_0 \equiv x_1 \pmod{m}$ and $x_0 \equiv x_1 \pmod{n}$. So x_0 also satisifies (9.23) and (9.24), and is therefore the desired solution. ■

(c) Prove that the x satisfying part (b) is unique.

Solution. Assume x_0, y both satisfy congruences (9.23) and (9.24). Taking the differences we see that

$$x_0 - y \equiv 0 \pmod{m} \text{ and } x_0 - y \equiv 0 \pmod{n}.$$

So by definition, both m and n divide $x_0 - y$, and since m and n are relatively prime, this implies $mn \mid (x_0 - y)$ (as shown in a previous problem). But if x_0 and y are both in the range 0 to $mn - 1$, then $mn > |x_0 - y|$, so it must be that $y = x_0$, as required. ■

(d) Conclude from the preceding parts of this problem that

$$\phi(mn) = \phi(m)\phi(n)$$

where ϕ is Euler's function.

Solution. For any positive integer, k, let

$$[0, k) ::= \{0, 1, \ldots, k-1\}.$$

By part (c), the mapping from x to $(\text{rem}(x, m), \text{rem}(x, n))$ is a bijection between $[0, mn)$ and $[0, m) \times [0, n)$. Moreover, since x is relatively prime to mn iff x is relatively prime to m and x is relatively prime to n, this mapping also defines a bijection between the integers in $[0, mn)$ that are relatively prime to mn and the pairs of integers in $[0, m) \times [0, n)$ that are relatively prime to m and n, respectively. In particular the number, $\phi(mn)$ of numbers in $[0, mn)$ that are relatively prime to mn is the same as the number $\phi(m)\phi(n)$ of pairs of integers in $[0, m) \times [0, n)$ whose first coordinate is relatively prime to m and whose second coordinate is relatively prime to n. ■

Problem 9.10.3. Let $S_k = 1^k + 2^k + \ldots + (p-1)^k$, where p is an odd prime and k is a positive multiple of $p - 1$. Use Fermat's theorem to prove that $S_k \equiv -1 \pmod{p}$.

Solution. Fermat's theorem says that $x^{p-1} \equiv 1 \pmod{p}$ when $1 \leq x \leq p - 1$. Since k is a multiple of $p - 1$, raising each side to a suitable power proves that $x^k \equiv 1 \pmod{p}$. Thus:

$$\begin{aligned} 1^k + 2^k + \ldots + (p-1)^k &\equiv \underbrace{1 + 1 + \ldots + 1}_{p-1 \text{ terms}} \pmod{p} \\ &\equiv p - 1 \pmod{p} \\ &\equiv -1 \pmod{p} \end{aligned}$$

■

Problem 9.10.4. Find an integer $k > 1$ such that n and n^k agree in the last *two digits* whenever n is a positive number relatively prime to 100.

Solution. Finding a k such that for all n relatively prime to 100, $n^k \equiv n \pmod{100}$ would satisfy the requirement. From Euler's theorem, we know $k = \phi(100) + 1$ satisfies this equation. Thus $c = \phi(100) + 1 = \phi(4)\phi(25) + 1 = (4-2)(25-5) + 1 = 41$. ■

Problem 9.10.5. Is a Harvard degree really worth more than an MIT degree?! Let us say that a person with a Harvard degree starts with \$40,000 and gets a \$20,000 raise every year after graduation, whereas a person with an MIT degree starts with \$30,000, but gets a 20% raise every year. Assume inflation is a fixed 8% every year. That is, \$1.08 a year from now is worth \$1.00 today. (You'll need a calculator to get final answers; if one is not available, it's ok to express the answer as a closed form numerical expression.)

(a) How much is a Harvard degree worth today if the holder will work for n years following graduation?

(b) How much is an MIT degree worth in this case?

(c) If you plan to retire after twenty years, which degree would be worth more?

Solution. One dollar after year i is worth r^i in today's currency, where

$$r = \frac{1}{1.08} = 0.925\,925\,925\ldots$$

So

$$\begin{aligned}
\text{Hvd}_n &= \sum_{i=0}^{n} (40000 + 20000i) r^i \\
&= 40000 \sum_{i=0}^{n} r^i + 20000 \sum_{i=0}^{n} i r^i, \\
\text{MIT}_n &= 30000 \sum_{i=0}^{n} 1.2^i r^i \\
&= 30000 \sum_{i=0}^{n} (1.2r)^i
\end{aligned}$$

But

$$\sum_{i=0}^{n} i r^i = \frac{r - (n+1)r^{n+1} + n r^{n+2}}{(1-r)^2},$$

so

$$\begin{aligned}
\text{Hvd}_n &= 40000\frac{(1-r^{n+1})}{1-r} + 20000\frac{(r-(n+1)r^{n+1}+nr^{n+2})}{(1-r)^2} \\
&= \frac{20000(2(1-r^{n+1}-r+r^{n+2})+r-(n+1)r^{n+1}+nr^{n+2})}{(1-r)^2} \\
&= \frac{20000(2-r-(n+3)r^{n+1}+(n+2)r^{n+2})}{(1-r)^2} \\
\text{MIT}_n &= \frac{30000(1-(1.2r)^{n+1})}{1-1.2r}
\end{aligned}$$

and for $n = 20$,

$$\begin{aligned}
\text{Hvd}_{20} &= \frac{20000(2-r-23r^{21}+22r^{22})}{(1-r)^2} = 2,010,885 \\
\text{MIT}_{20} &= \frac{30000(1-(1.2r)^{21})}{1-1.2r} = 2,197,579.
\end{aligned}$$

so the MIT degree is more valuable! (But we knew that already.) ■

Problem 9.10.6. Suppose you deposit \$100 into your MIT Credit Union account today, \$99 in one month from now, \$98 in two months from now, and so on. Given that the interest rate is constantly 0.3% per month, how long will it take to save \$5,000?

Solution. First note that you will certainly manage to have saved \$5,000 *some* day, since, even without your earnings from the interest, you will have $100 + 99 + \cdots + 1 = 100 \times 101/2 = 5,050$ dollars after 99 months.

But fewer months will be needed: After the first deposit you will have \$100. After the second deposit, you will have $\$(100 \times 1.003 + 99)$. After your third deposit, your saved money will be $\$\big((100 \times 1.003 + 99) \times 1.003 + 98\big) = \$\big(100 \times (1.003)^2 + 99 \times 1.003 + 98\big)$, and so on. So, after the nth deposit,

$$S_n = \sum_{i=0}^{n-1}(100-i)(1.003)^{n-i-1}$$

dollars will be in your account. Substituting $j = n - i - 1$, we can rewrite this as

$$\sum_{j=0}^{n-1}\big(100-(n-j-1)\big)(1.003)^j;$$

and then as

$$(101-n)\left(\sum_{j=0}^{n-1}(1.003)^j\right) + \left(\sum_{j=0}^{n-1}j(1.003)^j\right).$$

Using the closed forms from the Notes, we can finally write S_n as

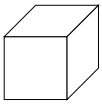
$$(101-n)\left(\frac{1-1.003^n}{1-1.003}\right) + \left(\frac{1.003-n1.003^n+(n-1)1.003^{n+1}}{(1-1.003)^2}\right).$$

Solving

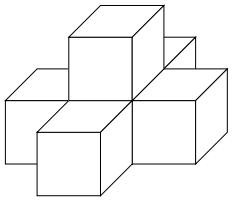
$$S_n \geq 5,000$$

for n, we get $n \geq 67$. That is, you'll need more than 5.5 years to save \$5,000. ■

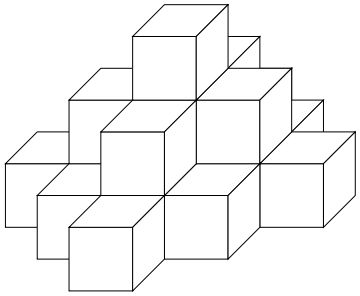
Problem 9.10.7. Pharaoh Aha I decides to build a "pyramid" in his honor consisting of a single block:



His successor, Aha II, trumps him by building a larger pyramid:



Not to be outdone, Aha III, builds a still-larger pyramid:



If this continues, how many blocks will Pharoah Aha n require?

Solution. Vertical cuts divide the nth pyramid into $2n-1$ triangular slabs:

The number of blocks in a triangular slab of height t is:

$$\begin{aligned} 1+2+\ldots+t+\ldots+2+1 &= t+2\sum_{i=1}^{t-1} i \\ &= t+2\cdot\frac{(t-1)t}{2} \\ &= t^2 \end{aligned}$$

Thus, the number of blocks in the nth pyramid is:

$$\begin{aligned} 1^2 + 2^2 + \ldots + n^2 + \ldots + 2^2 + 1^2 &= n^2 + 2\sum_{i=1}^{n-1} i^2 \\ &= n^2 + 2 \cdot \frac{(n-1)n(2n-1)}{6} \\ &= \frac{2n^3 + n}{3} \end{aligned}$$

■

Problem 9.10.8. Use integration to find upper and lower bounds that differ by at most 0.1 for the following sum. (You may need to add the first few terms explicitly and then use integrals to bound the sum of the remaining terms.)

$$\sum_{i=1}^{\infty} \frac{1}{(2i+1)^2}$$

Solution. Let's first try standard bounds:

$$\int_0^\infty \frac{1}{(2x+3)^2}\,dx \quad \leq \quad \sum_{i=1}^{\infty} \frac{1}{(2i+1)^2} \quad \leq \quad \int_0^\infty \frac{1}{(2x+1)^2}\,dx$$

Evaluating the integrals gives:

$$\begin{aligned} -\frac{1}{2(2x+3)}\Big|_0^\infty \quad \leq \quad \sum_{i=1}^{\infty} \frac{1}{(2i+1)^2} \quad \leq \quad -\frac{1}{2(2x+1)}\Big|_0^\infty \\ \frac{1}{6} \quad \leq \quad \sum_{i=1}^{\infty} \frac{1}{(2i+1)^2} \quad \leq \quad \frac{1}{2} \end{aligned}$$

These bounds are too far apart, so let's sum the first couple terms explicitly and bound the rest with integrals.

$$\frac{1}{3^2} + \frac{1}{5^2} + \int_2^\infty \frac{1}{(2x+3)^2}\,dx \quad \leq \quad \sum_{i=1}^{\infty} \frac{1}{(2i+1)^2} \quad \leq \quad \frac{1}{3^2} + \frac{1}{5^2} + \int_2^\infty \frac{1}{(2x+1)^2}\,dx$$

Integration now gives:

$$\begin{aligned} \frac{1}{3^2} + \frac{1}{5^2} + \left(-\frac{1}{2(2x+3)}\Big|_2^\infty\right) \quad \leq \quad \sum_{i=1}^{\infty} \frac{1}{(2i+1)^2} \quad \leq \quad \frac{1}{3^2} + \frac{1}{5^2} + \left(-\frac{1}{2(2x+1)}\Big|_2^\infty\right) \\ \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{14} \quad \leq \quad \sum_{i=1}^{\infty} \frac{1}{(2i+1)^2} \quad \leq \quad \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{10} \end{aligned}$$

Now we have bounds that differ by $1/10 - 1/14 < 1/10 = 0.1$. ■

9.11 Miniquiz Apr. 13

Problem 9.11.1.

(a) Calculate the value of $\phi(100)$.

Solution. $\phi(100) = \phi(25)\phi(4) = \phi(5^2)\phi(4^2) = (5^2 - 5)(2^2 - 2) = 40.$ ■

(b) Assume n is a positive integer greater than 9, and relatively prime to 100. Explain why the last two digits of n and n^{121} are the same.

Solution. Notice that all we have to prove is that n and n^{121} are congruent mod 100, implying they have the same last two digits.

$n^{121} \equiv n^{40 \cdot 3 + 1} \equiv n(n^{40})^3 \pmod{100}$. By Euler's Theorem, since n and 100 are relatively prime, $n^{\phi(100)} \equiv 1 \pmod{100}$. By part (a), we have that $\phi(100) = 40$, implying $n^{40} \equiv 1 \pmod{100}$. Hence, $n(n^{40})^3 \equiv n(1^3) \equiv n \pmod{100}$. ■

Problem 9.11.2. Use the fact that

$$\sum_{i=0}^{\infty} x^i = \frac{1}{1-x}$$

to show that

$$\sum_{i=1}^{\infty} i x^i = \frac{x}{(1-x)^2}.$$

for $|x| < 1$.

Solution. Taking the derivative of both sides of the geometric sum gives

$$\begin{aligned} \frac{d}{dx} \sum_{i=0}^{\infty} x^i &= \frac{d}{dx} \frac{1}{1-x} \\ \sum_{i=1}^{\infty} i x^{i-1} &= \frac{(-1)(-1)}{(1-x)^2} \\ &= \frac{1}{(1-x)^2}. \end{aligned}$$

By multiplying both sides by x, we get

$$\sum_{i=1}^{\infty} i x^i = \frac{x}{(1-x)^2}$$

■

Problem 9.11.3. Let's say you earn \$20,000 immediately and get a \$10,000 raise every year after that. Assume that the interest rate is a fixed 10% every year, that is, \$11 a year from now is worth \$10 today. If you can work forever, how much is the total salary worth in today's dollars?

(a) Write a series summation to express this quantity.

Solution. At year i , you get $20000 + 10000i$ dollars, which is worth $(20000 + 10000i)/1.1^i$ dollars today. Therefore, the total salary is

$$\sum_{i=0}^{\infty} \frac{20000 + 10000i}{1.1^i}$$

■

(b) What is the quantity? You may write a simple arithmetic expression (that is, no indexed sums or products) for its value **or** you may simply give its numerical value. (Do *not* use a calculator.)

Hint: Problem 9.11.2 above.

Solution.

$$\begin{aligned}
\sum_{i=0}^{\infty} \frac{20000 + 10000i}{1.1^i} &= \sum_{i=0}^{\infty} \left(20000 \left(\frac{10}{11}\right)^i + 10000i \left(\frac{10}{11}\right)^i \right) \\
&= 20000 \sum_{i=0}^{\infty} \left(\frac{10}{11}\right)^i + 10000 \sum_{i=0}^{\infty} i \left(\frac{10}{11}\right)^i \\
&= 20000 \frac{1}{1 - \frac{10}{11}} + 10000 \frac{\frac{10}{11}}{(1 - \frac{10}{11})^2} \\
&= 20000 \cdot 11 + 10000 \cdot 110 = 1320000.
\end{aligned}$$

Therefore, it is worth \$1,320,000 in today's dollars. ■

Problem 9.11.4. There is a bug on the edge of a 1-meter rug. The bug wants to cross to the other side of the rug. It crawls at 1 mm per second. However, at the end of each second, a malicious first-grader named Mildred Anderson *stretches* the rug by 1 meter. Assume that her action is instantaneous and the rug stretches uniformly. Thus, here's what happens in the first few seconds:

- The bug walks 1 mm in the first second, so 999 mm remain ahead.
- Mildred stretches the rug by 1 meter, which doubles its length. So now there are 2 mm behind the bug and 1998 mm ahead.
- The bug walks another 1 mm in the next second, leaving 3 mm behind and 1997 mm ahead.
- Then Mildred strikes, stretching the rug from 2 meters to 3 meters. So there are now $3 \cdot (3/2) = 4.5$ mm behind the bug and $1997 \cdot (3/2) = 2995.5$ mm ahead.

- The bug walks another 1 mm in the third second, and so on.

(a) Over the first n seconds, what fraction of the rug does the bug cross altogether? Express your answer in terms of the Harmonic number H_n.

Solution. The bug crosses $1/1000$ of the rug in the first second, $1/2000$ in the second, $1/3000$ in the third, and so forth. Thus, over the first n seconds, the fraction crossed by the bug is:

$$\sum_{k=1}^{n} \frac{1}{1000k} = \frac{1}{1000} H_n$$

(This formula is valid only until the bug reaches the far side of the rug.) ■

(b) Can the bug cross the entire rug? Briefly explain why.

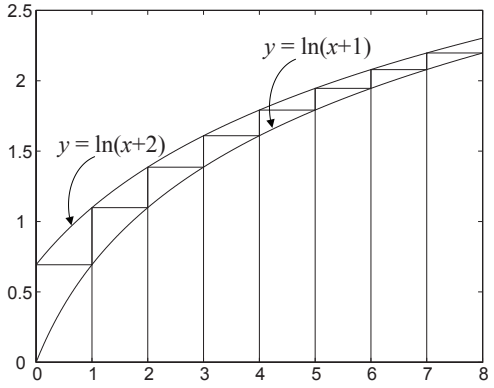
Solution. The bug arrives at the far side when the fraction it has crossed reaches 1. This occurs when n, the number of seconds elapsed, is sufficiently large that $H_n/1000 \geq 1$. Now H_n is approximately $\ln n$, which grows without bound, so H_n will eventually reach 1000.

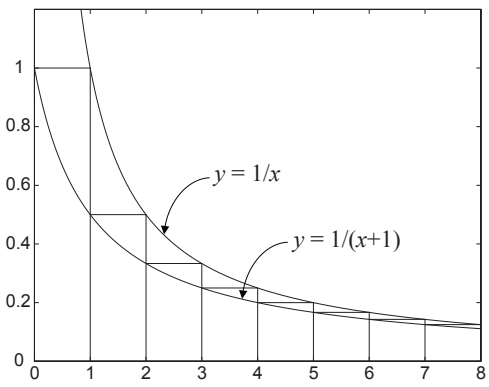
In particular, the bug arrives about when:

$$\begin{aligned} \frac{\ln n}{1000} &\geq 1 \\ \ln n &\geq 1000 \\ n &\geq e^{1000} \approx 10^{430} \text{ seconds} \end{aligned}$$

■

Problem 9.11.5. Circle all the correct inequalities below. Assume n is an integer larger than 1. Do *not* use a calculator. *Hint:* You may find the graphs helpful.





- $\displaystyle\sum_{i=1}^{n} \ln(i+1) \geq \ln 6 + \int_{2}^{n} \ln(x+1)dx$

- $\sum_{i=1}^{n} \ln(i+1) \leq \int_{0}^{n} \ln(x+2)dx$
- $\sum_{i=1}^{n} \ln(i+1) \leq \ln 2 + \int_{1}^{n} \ln(x+1)dx$
- $\sum_{i=1}^{n} \frac{1}{i} \geq \int_{0}^{n} \frac{1}{x+1}dx$
- $\sum_{i=1}^{n} \frac{1}{i} \leq 1.5 + \int_{3}^{n} \frac{1}{x}dx$
- $\sum_{i=1}^{n} \frac{1}{i} \geq 1 + \int_{1}^{n} \frac{1}{x}dx$

Solution. The 1st, 2nd and 4th inequalities hold. ■

Appendix

Definition. The value of *Euler's totient function*, $\phi(n)$, is defined to be the number of positive integers less than n that are relatively prime to n.

Lemma (Euler Totient Function Equations).

$$\begin{aligned} \phi(p^k) &= p^k - p^{k-1} && \text{for prime, } p, \text{ and } k > 0, \\ \phi(mn) &= \phi(m) \cdot \phi(n) && \text{when } \gcd(m,n) = 1. \end{aligned}$$

Theorem (Euler's Theorem). *If k and n are relatively prime, then*

$$k^{\phi(n)} \equiv 1 \pmod{n}$$

Definition. The n-th harmonic number is defined as $H_n ::= \sum_{i=1}^{n} \frac{1}{i}$.

$$\ln(n+1) \leq H_n \leq 1 + \ln n$$

9.12 In-Class Problems Week 9, Wed.

Problem 9.12.1. Prove that asymptotic equality ($\sim$) is a symmetric and transitive relation.

Solution. symmetry: Say $f \sim g$. Then $\lim_{x\to\infty} f(x)/g(x) = 1$. So

$$\begin{aligned}\lim_{x\to\infty} g(x)/f(x) &= \lim_{x\to\infty} 1/(f(x)/g(x)) \\ &= 1/\lim_{x\to\infty} f(x)/g(x) \\ &= 1/1 = 1,\end{aligned}$$

and therefore $g \sim f$.

transitivity: Say $f \sim g$ and $g \sim h$. So

$$\begin{aligned}1 &= 1 \cdot 1 \\ &= [\lim_{x\to\infty} f(x)/g(x)] \cdot [\lim_{x\to\infty} g(x)/h(x)] \\ &= \lim_{x\to\infty} [f(x)/g(x)] \cdot [g(x)/h(x)] \\ &= \lim_{x\to\infty} f(x)/h(x),\end{aligned}$$

so $f \sim h$. ■

Problem 9.12.2. Recall that for functions f, g on the natural numbers, $\mathbb{N}$, $f = O(g)$ iff

$$\exists c \in \mathbb{N}\, \exists n_0 \in \mathbb{N}\, \forall n \geq n_0 \quad c \cdot g(n) \geq |f(n)|\,. \tag{9.27}$$

For each pair of functions below, determine whether $f = O(g)$ and whether $g = O(f)$. In cases where one function is O() of the other, indicate the *smallest natural number*, c, and for that smallest c, the *smallest corresponding natural number* n_0 ensuring that condition (9.27) applies.

(a) $f(n) = n^2, g(n) = 3n$.

$f = O(g)$ YES NO If YES, $c =$ _____, $n_0 =$ _____

Solution. NO. ■

$g = O(f)$ YES NO If YES, $c =$ _____, $n_0 =$ _____

Solution. YES, with $c = 1$, $n_0 = 3$, which works because $3^2 = 9$, $3 \cdot 3 = 9$. ■

(b) $f(n) = (3n - 7)/(n + 4), g(n) = 4$

$f = O(g)$ YES NO If YES, $c =$ _____, $n_0 =$ _____

Solution. YES, with $c = 1, n_0 = 0$ (because $|f(n)| < 3$). ■

$g = O(f)$ YES NO If YES, $c =$ _____, n_0 = _____

Solution. YES, with $c = 2, n_0 = 15$.

Since $\lim_{n\to\infty} f(n) = 3$, the smallest possible c is 2. For $c = 2$, the smallest possible $n_0 = 15$ which follows from the requirement that $2f(n_0) \geq 4$. ■

(c) $f(n) = 1 + (n\sin(n\pi/2))^2, g(n) = 3n$

$f = O(g)$ YES NO If yes, $c =$ _____ n_0 = _____

Solution. NO, because $f(2n) = 1$, which rules out $g = O(f)$ since $g = \Theta(n)$. ■

$g = O(f)$ YES NO If yes, $c =$ _____ n_0 = _____

Solution. NO, because $f(2n+1) = n^2 + 1 \neq O(n)$ which rules out $f = O(g)$. ■

Problem 9.12.3. Indicate which of the following holds for each pair of functions $(f(n), g(n))$ in the table below. Assume $k \geq 1$, $\epsilon > 0$, and $c > 1$ are constants. Be prepared to justify your answers.

$f(n)$	$g(n)$	$f = O(g)$	$f = o(g)$	$g = O(f)$	$g = o(f)$	$f = \Theta(g)$	$f \sim g$
2^n	$2^{n/2}$						
$\sqrt{n}$	$n^{\sin n\pi/2}$						
$\log(n!)$	$\log(n^n)$						
n^k	c^n						
$\log^k n$	n^ϵ						

Solution.

$f(n)$	$g(n)$	$f = O(g)$	$f = o(g)$	$g = O(f)$	$g = o(f)$	$f = \Theta(g)$	$f \sim g$
2^n	$2^{n/2}$	no	no	yes	yes	no	no
$\sqrt{n}$	$n^{\sin n\pi/2}$	no	no	no	no	no	no
$\log(n!)$	$\log(n^n)$	yes	no	yes	no	yes	yes
n^k	c^n	yes	yes	no	no	no	no
$\log^k n$	n^ϵ	yes	yes	no	no	no	no

Following are some hints on deriving the table above:

(a) $\dfrac{2^n}{2^{n/2}} = 2^{n/2}$ grows without bound as n grows—it is not bounded by a constant.

(b) When n is even, then $n^{\sin n\pi/2} = 1$. So, no constant times $n^{\sin n\pi/2}$ will be an upper bound on $\sqrt{n}$ as n ranges over even numbers. When $n \equiv 1 \bmod 4$, then $n^{\sin n\pi/2} = n^1 = n$. So, no constant times $\sqrt{n}$ will be an upper bound on $n^{\sin n\pi/2}$ as n ranges over numbers $\equiv 1 \bmod 4$.

(c)

$$\begin{aligned}\log(n!) &= \log\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \pm c_n && (9.28)\\ &= \log n + n(\log n - 1) \pm d_n && (9.29)\\ &\sim n\log n && (9.30)\\ &= \log n^n.\end{aligned}$$

where $a \le c_n, d_n \le b$ for some constants $a, b \in \mathbb{R}$ and all n. Here equation (9.28) follows by taking logs of Stirling's formula, (9.29) follows from the fact that the log of a product is the sum of the logs, and (9.30) follows because any constant, $\log n$, and n are all $o(n\log n)$ and hence so is their sum.

(d) *Polynomial* growth versus *exponential* growth.

(e) *Polylogarithmic* growth versus *polynomial* growth.

■

Problem 9.12.4. It is a standard fallacy to think that given n quantities each of which is $O(1)$, their sum would have to be $O(n)$.

Namely, let $f_1, f_2, \ldots$ be a sequence of functions from $\mathbb{N}$ to $\mathbb{N}$, and let

$$S(n) ::= \sum_{i=1}^{n} f_i(n).$$

Then given that $f_i = O(1)$ for every f_i in the sequence, we can try to argue as follows:

$$S(n) = \sum_{i=1}^{n} f_i(n) = \sum_{i=1}^{n} O(1) = n \cdot O(1) = O(n).$$

This informal argument may seem plausible, but is fundamentally flawed because it treats O(1) as some kind numerical quantity. In fact, we ask you to show that there is no way to determine how fast the sum, $S(n)$, may grow.

Namely, let g be any function on $\mathbb{N}$. Explain how to define a sequence of functions $f_1, f_2, \ldots$ such that each $f_i = O(1)$, but S is not $O(g)$. *Hint:* Let $f_i(n) ::= i \cdot g(i)$.

Solution. Pick f_i to be the constant function $i \cdot g(i)$. That is,

$$\text{for all } n: \qquad f_i(n) ::= i \cdot g(i),$$

Since f_i is a constant function, it is $O(1)$. But

$$S(n) = \sum_{i=1}^{n} f_i(n) \ge f_n(n) = n \cdot g(n),$$

therefore

$$\lim_{n\to\infty} S(n)/g(n) \ge \lim_{n\to\infty} (n \cdot g(n))/g(n) = \lim_{n\to\infty} n = \infty$$

so, $S \neq O(g)$. ■

Asymptotic Notations

Lemma (Stirling's Formula).

$$n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n},$$

For functions $f, g : \mathbb{R} \to \mathbb{R}$, we say f is *asymptotically equal* to g, in symbols,

$$f(x) \sim g(x)$$

iff

$$\lim_{x\to\infty} f(x)/g(x) = 1.$$

For functions $f, g : \mathbb{R} \to \mathbb{R}$, we say f is *asymptotically smaller* than g, in symbols,

$$f(x) = o(g(x)),$$

iff

$$\lim_{x\to\infty} f(x)/g(x) = 0.$$

Given functions $f, g : \mathbb{R} \mapsto \mathbb{R}$, with g nonnegative, we say that[6]

$$f = O(g)$$

iff

$$\limsup_{x\to\infty} |f(x)|/g(x) < \infty.$$

An alternative, equivalent, definition is

$$f = O(g)$$

iff there exists a constant $c \geq 0$ and an x_0 such that for all $x \geq x_0$, $|f(x)| \leq cg(x)$.

Finally, we say

$$f = \Theta(g) \quad \text{iff} \quad f = O(g) \wedge g = O(f).$$

---

[6] $$\limsup_{x\to\infty} h(x) ::= \lim_{x\to\infty} \text{lub}_{y\geq x} h(y).$$

Chapter 10

Rules for Counting

20480135385502964448038	31710048321735013941130 17	57632573310834796474093 98	82473310000429953116460 21
48944599186691567624099 2	32082344215973686470192 65	58009491235489891226286 63	84962439971234759227663 10
10826620324303796513709 81	34372546563551578648691 13	60429008011992802180260 01	85183991406760026607474 77
11784808947697061789949 93	35748833903058653923711 365	61161717891377378967014 05	85436912834701914523337 63
12531273516832396938513 27	36449099460404801899691 49	61448689730015823697235 12	86753092583741370924613 52
13015051292340778110690 11	37900441327370840944172 46	62473145938511692347461 52	86943211123639968672966 65
13115671111438664338821 94	38703321274379713553228 15	68144289442668749634882 74	87723212036084772458511 54
14700294527212035876862 14	40805058045778014513631 00	68708529455438868491478 81	87914221617225825463410 91
15782710472862574994338 86	41672834610257023481249 20	69149555081209500937323 97	90626280245921262839732 85
16382439218521762431923 54	42359968311237777882112 49	69496324513659871524235 41	91378455669255263498977 94
17635802191319859631023 65	46709394457494390421112 20	71282111436136198284156 50	91537629668031892919344 19
18262277956018422310296 94	48153793518653842796134 27	71739200836518623079253 94	92708801940776364069842 49
18439718626751020372014 20	48370529482129226044421 90	72156548742117556762205 87	93243014807221034903792 04
23969511937221345261772 37	51063894238550185506715 30	72569328471643910402330 50	94360908321466951471405 81
27813945682685998010963 54	51423681920047692180699 10	73328226570752354316203 17	94753081597345382490132 38
27966051967136104054080 19	51812340961301440840418 56	74264418295415734449641 39	94923766239174869749232 02
29310163947619752631903 47	51982673981256179943913 48	76321981265318093271863 21	95119725587798802885229 79
29334580582944051551972 96	53175929403162312197583 72	77121544322119128823105 11	96024134246191871125522 64
30755144104909759203153 48	53843581267717941283569 47	78589186642402623566100 10	96312171149061292194611 11
31114749852527934528600 17	54392117122489019954234 41	78981567867632129631786 79	99081898531027533359813 19
31456215879361201184387 01	56103798260928381927604 58	81475910170375733378486 16	99132747634176429981398 7
31489012556288811031985 49	56323175554652286776760 44	81494367168713711619320 35	
31576931053251112843219 93	56921683746370196174237 12	81760638316825365713067 91	

Two different subsets of the ninety 25-digit numbers shown above have the same sum. For example, maybe the sum of the numbers in the first column is equal to the sum of the numbers in the second column. Can you find two such subsets? This is a challenging computational problem. But we'll prove that such subsets must exist! This is the sort of weird conclusion one can reach by tricky use of counting, the topic of this chapter.

Counting seems easy enough: 1, 2, 3, 4, etc. This explicit approach works well for counting simple things, like your toes, and for extremely complicated things for which there's no identifiable structure. However, subtler methods can help you count many things in the vast middle ground, such as:

- The number of different ways to select a dozen doughnuts when there are five varieties available.
- The number of 16-bit numbers with exactly 4 ones.

Counting is useful in computer science for several reasons:

- Determining the time and storage required to solve a computational problem— a central objective in computer science— often comes down to solving a counting problem.
- Counting is the basis of probability theory, which in turn is perhaps the most important topic this term.
- Two remarkable proof techniques, the "pigeonhole principle" and "combinatorial proof", rely on counting. These lead to a variety of interesting and useful insights.

We're going to present a lot of rules for counting. These rules are actually theorems, but we're generally not going to prove them. Our objective is to teach you counting as a practical skill, like integration. And most of the rules seem "obvious" anyway.

10.1 Counting One Thing by Counting Another

How do you count the number of people in a crowded room? You could count heads, since for each person there is exactly one head. Alternatively, you could count ears and divide by two. Of course, you might have to adjust the calculation if someone lost an ear in a pirate raid or someone was born with three ears. The point here is that you can often *count one thing by counting another*, though some fudge factors may be required. This is the central theme of counting, from the easiest problems to the hardest.

In more formal terms, every counting problem comes down to determining the size of some set. The ***size*** or ***cardinality*** of a set S is the number of elements in S and is denoted $|S|$. In these terms, we're claiming that we can often *find the size of one set S by finding the size of a related set T*. We already have a mathematical tool for relating one set to another: relations. Not surprisingly, a particular kind of relation is at the heart of counting.

10.1.1 The Bijection Rule

If we can pair up all the girls at a dance with all the boys, then there must be an equal number of each. This simple observation generalizes to a powerful counting rule:

Rule 1 (Bijection Rule). *If there exists a bijection $f : A \to B$, then $|A| = |B|$.*

In the example, A is the set of boys, B is the set of girls, and the function f defines how they are paired.

The Bijection Rule acts as a magnifier of counting ability; if you figure out the size of one set, then you can immediately determine the sizes of many other sets via bijections. For example, let's return to two sets mentioned earlier:

$A =$ all ways to select a dozen doughnuts when five varieties are available

$B =$ all 16-bit sequences with exactly 4 ones

Let's consider a particular element of set A:

$$\underbrace{0\,0}_{\text{chocolate}} \quad \underbrace{}_{\text{lemon-filled}} \quad \underbrace{0\,0\,0\,0\,0\,0}_{\text{sugar}} \quad \underbrace{0\,0}_{\text{glazed}} \quad \underbrace{0\,0}_{\text{plain}}$$

We've depicted each doughnut with a 0 and left a gap between the different varieties. Thus, the selection above contains two chocolate doughnuts, no lemon-filled, six sugar, two glazed, and two plain. Now let's put a 1 into each of the four gaps:

$$\underbrace{0\,0}_{\text{chocolate}}\;1\;\underbrace{}_{\text{lemon-filled}}\;1\;\underbrace{0\,0\,0\,0\,0\,0}_{\text{sugar}}\;1\;\underbrace{0\,0}_{\text{glazed}}\;1\;\underbrace{0\,0}_{\text{plain}}$$

We've just formed a 16-bit number with exactly 4 ones— an element of B!

This example suggests a bijection from set A to set B: map a dozen doughnuts consisting of:

$$c \text{ chocolate, } l \text{ lemon-filled, } s \text{ sugar, } g \text{ glazed, and } p \text{ plain}$$

to the sequence:

$$\underbrace{0\ldots0}_{c}\;1\;\underbrace{0\ldots0}_{l}\;1\;\underbrace{0\ldots0}_{s}\;1\;\underbrace{0\ldots0}_{g}\;1\;\underbrace{0\ldots0}_{p}$$

The resulting sequence always has 16 bits and exactly 4 ones, and thus is an element of B. Moreover, the mapping is a bijection; every such bit sequence is mapped to by exactly one order of a dozen doughnuts. Therefore, $|A| = |B|$ by the Bijection Rule!

This demonstrates the magnifying power of the bijection rule. We managed to prove that two very different sets are actually the same size— even though we don't know exactly how big either one is. But as soon as we figure out the size of one set, we'll immediately know the size of the other.

This particular bijection might seem frighteningly ingenious if you've not seen it before. But you'll use essentially this same argument over and over, and soon you'll consider it boringly routine.

10.1.2 Sequences

The Bijection Rule lets us count one thing by counting another. This suggests a general strategy: get really good at counting just a *few* things and then use bijections to count *everything else*. This is the strategy we'll follow. In particular, we'll get really good at counting *sequences*. When we want to determine the size of some other set T, we'll find a bijection from T to a set of sequences S. Then we'll use our super-ninja sequence-counting skills to determine $|S|$, which immediately gives us $|T|$. We'll need to hone this idea somewhat as we go along, but that's pretty much the plan!

10.2 Two Basic Counting Rules

We'll harvest our first crop of counting problems with two basic rules.

10.2.1 The Sum Rule

Linus allocates his big sister Lucy a quota of 20 crabby days, 40 irritable days, and 60 generally surly days. On how many days can Lucy be out-of-sorts one way or another? Let set C be her crabby days, I be her irritable days, and S be the generally surly. In these terms, the answer to the question is $|C \cup I \cup S|$. Now assuming that she is permitted at most one bad quality each day, the size of this union of sets is given by the Sum Rule:

Rule 2 (Sum Rule). *If $A_1, A_2, \ldots, A_n$ are* disjoint *sets, then:*

$$|A_1 \cup A_2 \cup \ldots \cup A_n| = |A_1| + |A_2| + \ldots + |A_n|$$

Thus, according to Linus' budget, Lucy can be out-of-sorts for:

$$\begin{aligned} |C \cup I \cup S| &= |C| + |I| + |S| \\ &= 20 + 40 + 60 \\ &= 120 \text{ days} \end{aligned}$$

Notice that the Sum Rule holds only for a union of *disjoint* sets. Finding the size of a union of intersecting sets is a more complicated problem that we'll take up later.

10.2.2 The Product Rule

The product rule gives the size of a product of sets. Recall that if $P_1, P_2, \ldots, P_n$ are sets, then

$$P_1 \times P_2 \times \ldots \times P_n$$

is the set of all sequences whose first term is drawn from P_1, second term is drawn from P_2 and so forth.

Rule 3 (Product Rule). *If $P_1, P_2, \ldots P_n$ are sets, then:*

$$|P_1 \times P_2 \times \ldots \times P_n| = |P_1| \cdot |P_2| \cdots |P_n|$$

Unlike the sum rule, the product rule does not require the sets $P_1, \ldots, P_n$ to be disjoint. For example, suppose a *daily diet* consists of a breakfast selected from set B, a lunch from set L, and a dinner from set D:

$$\begin{aligned} B &= \{\text{pancakes, bacon and eggs, bagel, Doritos}\} \\ L &= \{\text{burger and fries, garden salad, Doritos}\} \\ D &= \{\text{macaroni, pizza, frozen burrito, pasta, Doritos}\} \end{aligned}$$

Then $B \times L \times D$ is the set of all possible daily diets. Here are some sample elements:

$$\begin{gathered} (\text{pancakes, burger and fries, pizza}) \\ (\text{bacon and eggs, garden salad, pasta}) \\ (\text{Doritos, Doritos, frozen burrito}) \end{gathered}$$

The Product Rule tells us how many different daily diets are possible:

$$\begin{aligned} |B \times L \times D| &= |B| \cdot |L| \cdot |D| \\ &= 4 \cdot 3 \cdot 5 \\ &= 60 \end{aligned}$$

10.2.3 Putting Rules Together

Few counting problems can be solved with a single rule. More often, a solution is a flurry of sums, products, bijections, and other methods. Let's look at some examples that bring more than one rule into play.

Passwords

The sum and product rules together are useful for solving problems involving passwords, telephone numbers, and license plates. For example, on a certain computer system, a valid password is a sequence of between six and eight symbols. The first symbol must be a letter (which can be lowercase or uppercase), and the remaining symbols must be either letters or digits. How many different passwords are possible?

Let's define two sets, corresponding to valid symbols in the first and subsequent positions in the password.

$$F = \{a, b, \ldots, z, A, B, \ldots, Z\}$$
$$S = \{a, b, \ldots, z, A, B, \ldots, Z, 0, 1, \ldots, 9\}$$

In these terms, the set of all possible passwords is:

$$(F \times S^5) \cup (F \times S^6) \cup (F \times S^7)$$

Thus, the length-six passwords are in set $F \times S^5$, the length-seven passwords are in $F \times S^6$, and the length-eight passwords are in $F \times S^7$. Since these sets are disjoint, we can apply the Sum Rule and count the total number of possible passwords as follows:

$$\begin{aligned}
\left|(F \times S^5) \cup (F \times S^6) \cup (F \times S^7)\right| &= \left|F \times S^5\right| + \left|F \times S^6\right| + \left|F \times S^7\right| && \text{Sum Rule} \\
&= |F| \cdot |S|^5 + |F| \cdot |S|^6 + |F| \cdot |S|^7 && \text{Product Rule} \\
&= 52 \cdot 62^5 + 52 \cdot 62^6 + 52 \cdot 62^7 \\
&\approx 1.8 \cdot 10^{14} \text{ different passwords}
\end{aligned}$$

Subsets of an n-element Set

How many different subsets of an n element set X are there? For example, the set $X = \{x_1, x_2, x_3\}$ has eight different subsets:

$$\begin{array}{cccc} \{\} & \{x_1\} & \{x_2\} & \{x_1, x_2\} \\ \{x_3\} & \{x_1, x_3\} & \{x_2, x_3\} & \{x_1, x_2, x_3\} \end{array}$$

There is a natural bijection from subsets of X to n-bit sequences. Let $x_1, x_2, \ldots, x_n$ be the elements of X. Then a particular subset of X maps to the sequence $(b_1, \ldots, b_n)$ where $b_i = 1$ if and only if x_i is in that subset. For example, if $n = 10$, then the subset $\{x_2, x_3, x_5, x_7, x_{10}\}$ maps to a 10-bit sequence as follows:

$$\begin{array}{rccccccccccc} \text{subset:} & \{ & & x_2, & x_3, & & x_5, & & x_7, & & & x_{10} & \} \\ \text{sequence:} & (& 0, & 1, & 1, & 0, & 1, & 0, & 1, & 0, & 0, & 1 &) \end{array}$$

We just used a bijection to transform the original problem into a question about sequences— *exactly according to plan!* Now if we answer the sequence question, then we've solved our original problem as well.

But how many different n-bit sequences are there? For example, there are 8 different 3-bit sequences:

$$\begin{array}{llll}(0,0,0) & (0,0,1) & (0,1,0) & (0,1,1)\\ (1,0,0) & (1,0,1) & (1,1,0) & (1,1,1)\end{array}$$

Well, we can write the set of all n-bit sequences as a product of sets:

$$\underbrace{\{0,1\} \times \{0,1\} \times \ldots \times \{0,1\}}_{n \text{ terms}} = \{0,1\}^n$$

Then Product Rule gives the answer:

$$\begin{aligned}|\{0,1\}^n| &= |\{0,1\}|^n\\ &= 2^n\end{aligned}$$

This means that the number of subsets of an n-element set X is also 2^n. We'll put this answer to use shortly.

10.3 More Functions: Injections and Surjections

Bijections are both injective and surjective, which makes them a powerful tool for exact counting. We've observed in earlier Notes that surjections and injections by themselves imply certain size relationships between sets. For simplicity we'll assume that functions mentioned in these Notes are total; then we can simply state these rules as:

Rule 4 (Mapping Rule).

1. *If $f : X \to Y$ is surjective, then $|X| \geq |Y|$.*
2. *If $f : X \to Y$ is injective, then $|X| \leq |Y|$.*
3. *If $f : X \to Y$ is bijective, then $|X| = |Y|$.*

10.3.1 The Pigeonhole Principle

Here is an old puzzle:

> A drawer in a dark room contains red socks, green socks, and blue socks. How many socks must you withdraw to be sure that you have a matching pair?

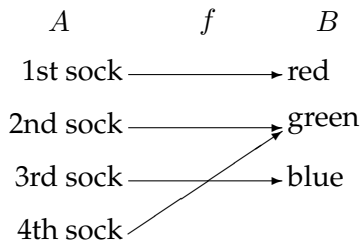
For example, picking out three socks is not enough; you might end up with one red, one green, and one blue. The solution relies on the Pigeonhole Principle, which is a friendly name for the contrapositive of part (2) of the Mapping Rule. Let's write it down:

If $|X| > |Y|$, then no function $f : X \to Y$ is injective.

And now rewrite it again to eliminate the word "injective."

Rule 5 (Pigeonhole Principle). *If $|X| > |Y|$, then for every function $f : X \to Y$, there exist two different elements of X that are mapped to the same element of Y.*

Perhaps the relevance of this abstract mathematical statement to selecting footwear under poor lighting conditions is not obvious. However, let A be the set of socks you pick out, let B be the set of colors available, and let f map each sock to its color. The Pigeonhole Principle says that if $|A| > |B| = 3$, then at least two elements of A (that is, at least two socks) must be mapped to the same element of B (that is, the same color). For example, one possible mapping of four socks to three colors is shown below.



Therefore, four socks are enough to ensure a matched pair.

Not surprisingly, the pigeonhole principle is often described in terms of pigeons:

> *If there are more pigeons than holes they fly into, then at least two pigeons must fly into the same hole.*

In this case, the pigeons form set A, the pigeonholes are set B, and f describes which hole each pigeon flies into.

Mathematicians have come up with many ingenious applications for the pigeonhole principle. If there were a cookbook procedure for generating such arguments, we'd give it to you. Unfortunately, there isn't one. One helpful tip, though: when you try to solve a problem with the pigeonhole principle, the key is to clearly identify three things:

1. The set A (the pigeons).
2. The set B (the pigeonholes).
3. The function f (the rule for assigning pigeons to pigeonholes).

Hairs on Heads

There are a number of generalizations of the pigeonhole principle. For example:

Rule 6 (Generalized Pigeonhole Principle). *If $|X| > k \cdot |Y|$, then every function $f : X \to Y$ maps at least $k + 1$ different elements of X to the same element of Y.*

For example, if you pick two people at random, surely they are extremely unlikely to have *exactly* the same number of hairs on their heads. However, in the remarkable city of Boston, Massachusetts there are actually *three* people who have exactly the same number of hairs! Of course, there are many bald people in Boston, and they all have zero hairs. But we're talking about non-bald people.

Boston has about 500,000 non-bald people, and the number of hairs on a person's head is at most 200,000. Let A be the set of non-bald people in Boston, let $B = \{1, \ldots, 200,000\}$, and let f map a person to the number of hairs on his or her head. Since $|A| > 2\,|B|$, the Generalized Pigeonhole Principle implies that at least three people have exactly the same number of hairs. We don't know who they are, but we know they exist!

Subsets with the Same Sum

We asserted that two different subsets of the ninety 25-digit numbers listed on the first page have the same sum. This actually follows from the Pigeonhole Principle. Let A be the collection of all subsets of the 90 numbers in the list. Now the sum of any subset of numbers is at most $90 \cdot 10^{25}$, since there are only 90 numbers and every 25-digit number is less than 10^{25}. So let B be the set of integers $\left\{0, 1, \ldots, 90 \cdot 10^{25}\right\}$, and let f map each subset of numbers (in A) to its sum (in B).

We proved that an n-element set has 2^n different subsets. Therefore:

$$\begin{aligned} |A| &= 2^{90} \\ &\geq 1.237 \times 10^{27} \end{aligned}$$

On the other hand:

$$\begin{aligned} |B| &= 90 \cdot 10^{25} + 1 \\ &\leq 0.901 \times 10^{27} \end{aligned}$$

Both quantities are enormous, but $|A|$ is a bit greater than $|B|$. This means that f maps at least two elements of A to the same element of B. In other words, by the Pigeonhole Principle, two different subsets must have the same sum!

Notice that this proof gives no indication *which* two sets of numbers have the same sum. This frustrating variety of argument is called a ***nonconstructive proof***.

Sets with Distinct Subset Sums

How can we construct a set of n positive integers such that all its subsets have *distinct* sums? One way is to use powers of two:

$$\{1, 2, 4, 8, 16\}$$

This approach is so natural that one suspects all other such sets must involve larger numbers. (For example, we could safely replace 16 by 17, but not by 15.) Remarkably, there are examples involving *smaller* numbers. Here is one:

$$\{6, 9, 11, 12, 13\}$$

One of the top mathematicans of the century, Paul Erdős, conjectured in 1931 that there are no such sets involving *significantly* smaller numbers. More precisely, he conjectured that the largest number must be $> c2^n$ for some constant $c > 0$. He offered $500 to anyone who could prove or disprove his conjecture, but the problem remains unsolved.

10.4 The Generalized Product Rule

We realize everyone has been working pretty hard this term, and we're considering awarding some prizes for *truly exceptional* coursework. Here are some possible categories:

Best Administrative Critique We asserted that the quiz was closed-book. On the cover page, one strong candidate for this award wrote, "There is no book."

Awkward Question Award "Okay, the left sock, right sock, and pants are in an antichain, but how— even with assistance— could I put on all three at once?"

Best Collaboration Statement Inspired by a student who wrote "I worked alone" on Quiz 1.

In how many ways can, say, three different prizes be awarded to n people? This is easy to answer using our strategy of translating the problem about awards into a problem about sequences. Let P be the set of n people in 6.042. Then there is a bijection from ways of awarding the three prizes to the set $P^3 ::= P \times P \times P$. In particular, the assignment:

"person x wins prize #1, y wins prize #2, and z wins prize #3"

maps to the sequence (x, y, z). By the Product Rule, we have $\left|P^3\right| = |P|^3 = n^3$, so there are n^3 ways to award the prizes to a class of n people.

But what if the three prizes must be awarded to *different* students? As before, we could map the assignment

"person x wins prize #1, y wins prize #2, and z wins prize #3"

to the triple $(x, y, z) \in P^3$. But this function is *no longer a bijection*. For example, no valid assignment maps to the triple (Dave, Dave, Becky) because Dave is not allowed to receive two awards. However, there *is* a bijection from prize assignments to the set:

$$S = \{(x, y, z) \in P^3 \mid x, y, \text{and } z \text{ are different people}\}$$

This reduces the original problem to a problem of counting sequences. Unfortunately, the Product Rule is of no help in counting sequences of this type because the entries depend on one another; in particular, they must all be different. However, a slightly sharper tool does the trick.

Rule 7 (Generalized Product Rule). *Let S be a set of length-k sequences. If there are:*

- *n_1 possible first entries,*
- *n_2 possible second entries for each first entry,*
- *n_3 possible third entries for each combination of first and second entries, etc.*

then:

$$|S| = n_1 \cdot n_2 \cdot n_3 \cdots n_k$$

In the awards example, S consists of sequences (x, y, z). There are n ways to choose x, the recipient of prize #1. For each of these, there are $n - 1$ ways to choose y, the recipient of prize #2, since everyone except for person x is eligible. For each combination of x and y, there are $n - 2$ ways to choose z, the recipient of prize #3, because everyone except for x and y is eligible. Thus, according to the Generalized Product Rule, there are

$$|S| = n \cdot (n - 1) \cdot (n - 2)$$

ways to award the 3 prizes to different people.

10.4.1 Defective Dollars

A dollar is *defective* if some digit appears more than once in the 8-digit serial number. If you check your wallet, you'll be sad to discover that defective dollars are all-too-common. In fact, how common are *nondefective* dollars? Assuming that the digit portions of serial numbers all occur equally often, we could answer this question by computing:

$$\text{fraction dollars that are nondefective} = \frac{\text{\# of serial \#'s with all digits different}}{\text{total \# of serial \#'s}}$$

Let's first consider the denominator. Here there are no restrictions; there are are 10 possible first digits, 10 possible second digits, 10 third digits, and so on. Thus, the total number of 8-digit serial numbers is 10^8 by the Product Rule.

Next, let's turn to the numerator. Now we're not permitted to use any digit twice. So there are still 10 possible first digits, but only 9 possible second digits, 8 possible third digits, and so forth. Thus, by the Generalized Product Rule, there are

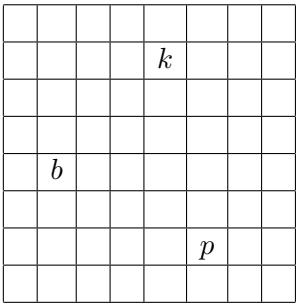
$$\begin{aligned} 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 &= \frac{10!}{2} \\ &= 1,814,400 \end{aligned}$$

serial numbers with all digits different. Plugging these results into the equation above, we find:

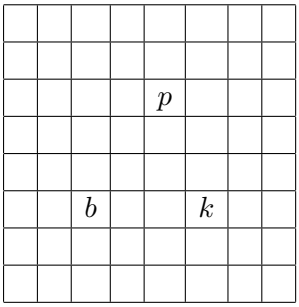
$$\begin{aligned}\text{fraction dollars that are nondefective} &= \frac{1,814,400}{100,000,000}\\ &= 1.8144\%\end{aligned}$$

10.4.2 A Chess Problem

In how many different ways can we place a pawn (p), a knight (k), and a bishop (b) on a chessboard so that no two pieces share a row or a column? A valid configuration is shown below on the left, and an invalid configuration is shown on the right.



valid



invalid

First, we map this problem about chess pieces to a question about sequences. There is a bijection from configurations to sequences

$$(r_p, c_p, r_k, c_k, r_b, c_b)$$

where r_p, r_k, and r_b are distinct rows and c_p, c_k, and c_b are distinct columns. In particular, r_p is the pawn's row, c_p is the pawn's column, r_k is the knight's row, etc. Now we can count the number of such sequences using the Generalized Product Rule:

- r_p is one of 8 rows
- c_p is one of 8 columns
- r_k is one of 7 rows (any one but r_p)
- c_k is one of 7 columns (any one but c_p)
- r_b is one of 6 rows (any one but r_p or r_k)
- c_b is one of 6 columns (any one but c_p or c_k)

Thus, the total number of configurations is $(8 \cdot 7 \cdot 6)^2$.

10.4.3 Permutations

A ***permutation*** of a set S is a sequence that contains every element of S exactly once. For example, here are all the permutations of the set $\{a, b, c\}$:

$$\begin{array}{lll}(a,b,c) & (a,c,b) & (b,a,c)\\ (b,c,a) & (c,a,b) & (c,b,a)\end{array}$$

How many permutations of an n-element set are there? Well, there are n choices for the first element. For each of these, there are $n-1$ remaining choices for the second element. For every combination of the first two elements, there are $n-2$ ways to choose the third element, and so forth. Thus, there are a total of

$$n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1 = n!$$

permutations of an n-element set. In particular, this formula says that there are $3! = 6$ permuations of the 3-element set $\{a, b, c\}$, which is the number we found above.

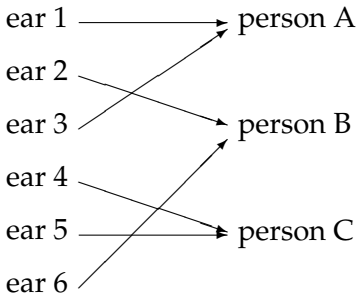
Permutations will come up again in this course approximately 1.6 bazillion times. In fact, permutations are the reason why factorial comes up so often and why we taught you Stirling's approximation:

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

10.5 The Division Rule

Counting ears and dividing by two is a silly way to count the number of people in a room, but this approach is representative of a powerful counting principle.

A ***k-to-1 function*** maps exactly k elements of the domain to every element of the codomain. For example, the function mapping each ear to its owner is 2-to-1:



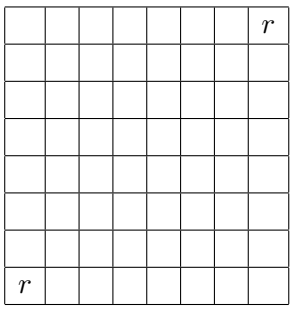
Similarly, the function mapping each finger to its owner is 10-to-1, and the function mapping each finger and toe to its owner is 20-to-1. The general rule is:

Rule 8 (Division Rule). *If $f : A \to B$ is k-to-1, then $|A| = k \cdot |B|$.*

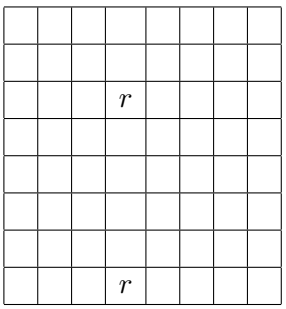
For example, suppose A is the set of ears in the room and B is the set of people. There is a 2-to-1 mapping from ears to people, so by the Division Rule $|A| = 2 \cdot |B|$ or, equivalently, $|B| = |A|/2$, expressing what we knew all along: the number of people is half the number of ears. Unlikely as it may seem, many counting problems are made much easier by initially counting every item multiple times and then correcting the answer using the Division Rule. Let's look at some examples.

10.5.1 Another Chess Problem

In how many different ways can you place two identical rooks on a chessboard so that they do not share a row or column? A valid configuration is shown below on the left, and an invalid configuration is shown on the right.



valid



invalid

Let A be the set of all sequences

$$(r_1, c_1, r_2, c_2)$$

where r_1 and r_2 are distinct rows and c_1 and c_2 are distinct columns. Let B be the set of all valid rook configurations. There is a natural function f from set A to set B; in particular, f maps the sequence (r_1, c_1, r_2, c_2) to a configuration with one rook in row r_1, column c_1 and the other rook in row r_2, column c_2.

But now there's a snag. Consider the sequences:

$$(1, 1, 8, 8) \qquad \text{and} \qquad (8, 8, 1, 1)$$

The first sequence maps to a configuration with a rook in the lower-left corner and a rook in the upper-right corner. The second sequence maps to a configuration with a rook in the upper-right corner and a rook in the lower-left corner. The problem is that those are two different ways of describing the *same* configuration! In fact, this arrangement is shown on the left side in the diagram above.

More generally, the function f maps exactly two sequences to *every* board configuration; that is f is a 2-to-1 function. Thus, by the quotient rule, $|A| = 2 \cdot |B|$. Rearranging terms gives:

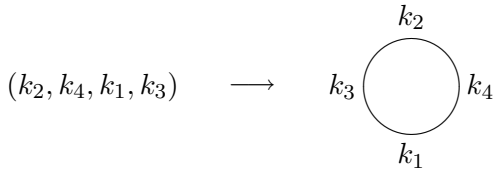
$$\begin{aligned} |B| &= \frac{|A|}{2} \\ &= \frac{(8 \cdot 7)^2}{2} \end{aligned}$$

On the second line, we've computed the size of A using the General Product Rule just as in the earlier chess problem.

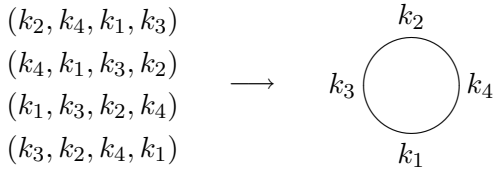
10.5.2 Knights of the Round Table

In how many ways can King Arthur seat n different knights at his round table? Two seatings are considered equivalent if one can be obtained from the other by rotation. For example, the following two arrangements are equivalent:

Let A be all the permutations of the knights, and let B be the set of all possible seating arrangements at the round table. We can map each permutation in set A to a circular seating arrangement in set B by seating the first knight in the permutation anywhere, putting the second knight to his left, the third knight to the left of the second, and so forth all the way around the table. For example:



This mapping is actually an n-to-1 function from A to B, since all n cyclic shifts of the original sequence map to the same seating arrangement. In the example, $n = 4$ different sequences map to the same seating arrangement:



Therefore, by the division rule, the number of circular seating arrangements is:

$$\begin{aligned} |B| &= \frac{|A|}{n} \\ &= \frac{n!}{n} \\ &= (n-1)! \end{aligned}$$

Note that $|A| = n!$ since there are $n!$ permutations of n knights.

10.6 Inclusion-Exclusion

How big is a union of sets? For example, suppose there are 60 Math majors, 200 EECS majors, and 40 Physics majors. How many students are there in these three departments? Let M be the set

of Math majors, E be the set of EECS majors, and P be the set of Physics majors. In these terms, we're asking for $|M \cup E \cup P|$.

The Sum Rule says that the size of union of *disjoint* sets is the sum of their sizes:

$$|M \cup E \cup P| = |M| + |E| + |P| \qquad \text{(if } M, E, \text{ and } P \text{ are disjoint)}$$

However, the sets M, E, and P might *not* be disjoint. For example, there might be a student majoring in both Math and Physics. Such a student would be counted twice on the right sides of this equation, once as an element of M and once as an element of P. Worse, there might be a triple-major counting *three* times on the right side!

Our last counting rule determines the size of a union of sets that are not necessarily disjoint. Before we state the rule, let's build some intuition by considering some easier special cases: unions of just two or three sets.

10.6.1 Union of Two Sets

For two sets, S_1 and S_2, the Inclusion-Exclusion rule is that the size of their union is:

$$|S_1 \cup S_2| = |S_1| + |S_2| - |S_1 \cap S_2| \tag{10.1}$$

Intuitively, each element of S_1 is accounted for in the first term, and each element of S_2 is accounted for in the second term. Elements in *both* S_1 and S_2 are counted *twice*— once in the first term and once in the second. This double-counting is corrected by the final term.

We can capture this double-counting idea in a precise way by decomposing the union of S_1 and S_2 into three disjoint sets, the elements in each set but not the other, and the elements in both:

$$S_1 \cup S_2 = (S_1 - S_2) \cup (S_2 - S_1) \cup (S_1 \cap S_2). \tag{10.2}$$

Similarly, we can decompose each of S_1 and S_2 into the elements exclusively in each set and the elements in both:

$$S_1 = (S_1 - S_2) \cup (S_1 \cap S_2), \tag{10.3}$$
$$S_2 = (S_2 - S_1) \cup (S_1 \cap S_2). \tag{10.4}$$

Now we have from (10.3) and (10.4)

$$\begin{aligned} |S_1| + |S_2| &= (|S_1 - S_2| + |S_1 \cap S_2|) + (|S_2 - S_1| + |S_1 \cap S_2|) \\ &= |S_1 - S_2| + |S_2 - S_1| + 2\,|S_1 \cap S_2|\,, \end{aligned} \tag{10.5}$$

which shows the double-counting of $S_1 \cap S_2$ in the sum. On the other hand, we have from (10.2)

$$|S_1 \cup S_2| = |S_1 - S_2| + |S_2 - S_1| + |S_1 \cap S_2|\,. \tag{10.6}$$

Subtracting (10.6) from (10.5), we get

$$(|S_1| + |S_2|) - |S_1 \cup S_2| = |S_1 \cap S_2|$$

which proves (10.1).

10.6.2 Union of Three Sets

So how many students are there in the Math, EECS, and Physics departments? In other words, what is $|M \cup E \cup P|$ if:

$$\begin{aligned} |M| &= 60 \\ |E| &= 200 \\ |P| &= 40 \end{aligned}$$

The size of a union of three sets is given by a more complicated Inclusion-Exclusion formula:

$$\begin{aligned} |S_1 \cup S_2 \cup S_3| = &|S_1| + |S_2| + |S_3| \\ &- |S_1 \cap S_2| - |S_1 \cap S_3| - |S_2 \cap S_3| \\ &+ |S_1 \cap S_2 \cap S_3| \end{aligned}$$

Remarkably, the expression on the right accounts for each element in the union of S_1, S_2, and S_3 exactly once. For example, suppose that x is an element of all three sets. Then x is counted three times (by the $|S_1|$, $|S_2|$, and $|S_3|$ terms), subtracted off three times (by the $|S_1 \cap S_2|$, $|S_1 \cap S_3|$, and $|S_2 \cap S_3|$ terms), and then counted once more (by the $|S_1 \cap S_2 \cap S_3|$ term). The net effect is that x is counted just once.

So we can't answer the original question without knowing the sizes of the various intersections. Let's suppose that there are:

4	Math - EECS double majors
3	Math - Physics double majors
11	EECS - Physics double majors
2	triple majors

Then $|M \cap E| = 4 + 2$, $|M \cap P| = 3 + 2$, $|E \cap P| = 11 + 2$, and $|M \cap E \cap P| = 2$. Plugging all this into the formula gives:

$$\begin{aligned} |M \cup E \cup P| &= |M| + |E| + |P| - |M \cap E| - |M \cap P| - |E \cap P| + |M \cap E \cap P| \\ &= 60 + 200 + 40 - 6 - 5 - 13 + 2 \\ &= 278 \end{aligned}$$

Sequences with 42, 04, or 60

In how many permutations of the set $\{0, 1, 2, \ldots, 9\}$ do either 4 and 2, 0 and 4, or 6 and 0 appear consecutively? For example, none of these pairs appears in:

$$(7, 2, 9, 5, 4, 1, 3, 8, 0, 6)$$

The 06 at the end doesn't count; we need 60. On the other hand, both 04 and 60 appear consecutively in this permutation:

$$(7, 2, 5, \underline{6}, \underline{0}, \underline{4}, 3, 8, 1, 9)$$

Let P_{42} be the set of all permutations in which 42 appears; define P_{60} and P_{04} similarly. Thus, for example, the permutation above is contained in both P_{60} and P_{04}. In these terms, we're looking for the size of the set $P_{42} \cup P_{04} \cup P_{60}$.

First, we must determine the sizes of the individual sets, such as P_{60}. We can use a trick: group the 6 and 0 together as a single symbol. Then there is a natural bijection between permutations of $\{0, 1, 2, \ldots 9\}$ containing 6 and 0 consecutively and permutations of:

$$\{60, 1, 2, 3, 4, 5, 7, 8, 9\}$$

For example, the following two sequences correspond:

$$(7, 2, 5, \underline{6}, \underline{0}, 4, 3, 8, 1, 9) \qquad \leftrightarrow \qquad (7, 2, 5, \underline{60}, 4, 3, 8, 1, 9)$$

There are $9!$ permutations of the set containing 60, so $|P_{60}| = 9!$ by the Bijection Rule. Similarly, $|P_{04}| = |P_{42}| = 9!$ as well.

Next, we must determine the sizes of the two-way intersections, such as $P_{42} \cap P_{60}$. Using the grouping trick again, there is a bijection with permutations of the set:

$$\{42, 60, 1, 3, 5, 7, 8, 9\}$$

Thus, $|P_{42} \cap P_{60}| = 8!$. Similarly, $|P_{60} \cap P_{04}| = 8!$ by a bijection with the set:

$$\{604, 1, 2, 3, 5, 7, 8, 9\}$$

And $|P_{42} \cap P_{04}| = 8!$ as well by a similar argument. Finally, note that $|P_{60} \cap P_{04} \cap P_{42}| = 7!$ by a bijection with the set:

$$\{6042, 1, 3, 5, 7, 8, 9\}$$

Plugging all this into the formula gives:

$$|P_{42} \cup P_{04} \cup P_{60}| = 9! + 9! + 9! - 8! - 8! - 8! + 7!$$

10.6.3 Union of *n* Sets

The size of a union of n sets is given by the following rule.

Rule 9 (Inclusion-Exclusion).

$$|S_1 \cup S_2 \cup \cdots \cup S_n| =$$

	the sum of the sizes of the individual sets
minus	*the sizes of all two-way intersections*
plus	*the sizes of all three-way intersections*
minus	*the sizes of all four-way intersections*
plus	*the sizes of all five-way intersections, etc.*

There are various ways to write the Inclusion-Exclusion formula in mathematical symbols, but none are particularly clear, so we've just used words. The formulas for unions of two and three sets are special cases of this general rule.

10.6.4 Computing Euler's Function

We will now use Inclusion-Exclusion to calculate Euler's function, $\phi(n)$. By definition, $\phi(n)$ is the number of nonnegative integers less than a positive integer n that are relatively prime to n. But the set, S, of nonnegative integers less than n that are *not* relatively prime to n will be easier to count.

Suppose the prime factorization of n is $p_1^{e_1} \cdots p_m^{e_m}$ for distinct primes p_i. This means that the integers in S are precisely the nonnegative integers less than n that are divisible by at least one of the p_i's. So, letting C_i be the set of nonnegative integers less than n that are divisible by p_i, we have

$$S = \bigcup_{i=1}^{m} C_i.$$

Next, observe that if r is a positive divisor of n, then exactly n/r nonnegative integers less than n are divisible by r, namely, $0, r, 2r, \ldots, ((n/r) - 1)r$.

Now by inclusion-exclusion,

$$\begin{aligned}
|S| &= \left|\bigcup_{i=1}^{m} C_i\right| \\
&= \sum_{i=1}^{m} |C_i| - \sum_{1\le i<j\le m} |C_i \cap C_j| + \sum_{1\le i<j<k\le m} |C_i \cap C_j \cap C_k| - \cdots \pm \left|\bigcap_{i=1}^{m} C_i\right| \\
&= \sum_{i=1}^{m} \frac{n}{p_i} - \sum_{1\le i<j\le m} \frac{n}{p_i p_j} + \sum_{1\le i<j<k\le m} \frac{n}{p_i p_j p_k} - \cdots \pm \frac{n}{\prod_{i=1}^{m} p_i} \\
&= n\left(1 - \prod_{i=1}^{m}\left(1 - \frac{1}{p_i}\right)\right)
\end{aligned}$$

But $\phi(n) = n - |S|$ by definition, so

$$\phi(n) = n - n\left(1 - \prod_{i=1}^{m}\left(1 - \frac{1}{p_i}\right)\right) = n\prod_{i=1}^{m}\left(1 - \frac{1}{p_i}\right). \tag{10.7}$$

Notice that in case $n = p^k$ for some prime, p, then (10.7) simplifies to

$$\phi(p^k) = p^k\left(1 - \frac{1}{p}\right) = p^k - p^{k-1}$$

as claimed in the Notes on Number Theory.

Problem 10.6.1. Use equation (10.7) to prove that, as claimed in the Notes on Number Theory,

$$\phi(ab) = \phi(a)\phi(b)$$

for relatively prime integers $a, b > 1$.

10.7 In-Class Problems Week 9, Fri.

Problem 10.7.1. A license plate consists of either:

- 3 letters followed by 3 digits (standard plate)
- 5 letters (vanity plate)
- 2 characters – letters or numbers (big shot plate)

Let L be the set of all possible license plates.

(a) Express L in terms of

$$\mathcal{A} = \{A, B, C, \ldots, Z\}$$
$$\mathcal{D} = \{0, 1, 2, \ldots, 9\}$$

using unions ($\cup$) and set products ($\times$).

Solution.

$$L = (A^3 \times D^3) \cup A^5 \cup (A \cup D)^2$$

■

(b) Compute $|L|$, the number of different license plates, using the sum and product rules.

Solution.

$$\begin{aligned}
|L| &= \left|(A^3 \times D^3) \cup A^5 \cup (A \cup D)^2\right| \\
&= \left|(A^3 \times D^3)\right| + \left|A^5\right| + \left|(A \cup D)^2\right| && \text{Sum Rule} \\
&= |A|^3 \cdot |D|^3 + |A|^5 + |A \cup D|^2 && \text{Product Rule} \\
&= |A|^3 \cdot |D|^3 + |A|^5 + (|A| + |D|)^2 && \text{Sum Rule} \\
&= 26^3 \cdot 10^3 + 26^5 + 36^2 = 29458672
\end{aligned}$$

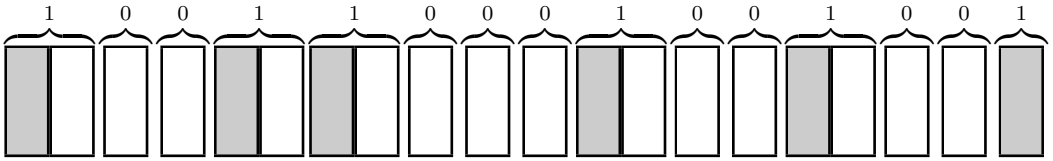
■

Problem 10.7.2. **(a)** How many of the billion numbers in the range from 1 to 10^9 contain the digit 1? (*Hint:* How many don't?)

Solution. We can count up how many *do not* contain the digit 1 and subtract. So (total number) - (number without 1's) $= 10^9 - (9^9 - 1) = 612,579,512$ (the -1 is for 0 which is not in our range).■

(b) There are 20 books arranged in a row on a shelf. Describe a bijection between ways of choosing 6 of these books so that no two adjacent books are selected and 15-bit sequences with exactly 6 ones.

Solution. There is a bijection from 15-bit sequences with exactly six 1's to valid book selections: given such a sequence, map each zero to a non-chosen book, each of the first five 1's to a chosen book followed by a non-chosen book, and the last 1 to a chosen book. For example, here is a configuration of books and the corresponding binary sequence:



Selected books are darkened. Notice that the first fives ones are mapped to a chosen book *and* a non-chosen book in order to ensure that the binary sequence maps to a valid selection of books.■

Problem 10.7.3. A *numbered tree* is a tree whose vertex set is $\{1, 2, \ldots, n\}$ for some $n \geq 2$. We define the *code* of the numbered tree to be a sequence of $n-2$ integers from 1 to n obtained by the following recursive process:

> If $n = 2$, stop—the code is the empty sequence. Otherwise, write down the *father* of the largest leaf[1], delete this *leaf*, and continue the process on the resulting smaller tree.

For example, the codes of a couple of numbered trees are shown in the Figure 10.1.

(a) Describe a procedure for reconstructing a numbered tree from its code.

Solution. The key observation is that, given a code of length $n-2$, the numbers between 1 and n which *do not appear* in the code must be leaves of the tree. Hence, the largest missing number is a leaf attached to the first number of the code. The rest of the tree can now be reconstructed by deleting the first number in the code, henceforth ignoring the largest leaf, and proceeding recursively on the rest of the code. (We're using the obvious fact that what's left after deleting a leaf from a tree is another tree.)

More precisely, the reconstruction procedure applies to any finite tree whose vertex set is totally ordered. The procedure takes *two* parameters: the vertex set, V, and a length $|V| - 2$ "code" sequence, S, of elements in V. If l is the largest element in V which does not appear in S, and f is the first element of S, then the reconstructed tree is obtained by adding edge (l, f) to the tree reconstructed by calling the procedure recursively with first argument $V - \{l\}$ and second argument equal to the code obtained by erasing the initial f from S. The procedure terminates when $|V| = 2$, returning the edge between the two numbers in V.

To justify the key observation, note that any vertex that gets deleted by the process and was not a leaf to begin with, must have been the father of a previously deleted leaf, which means it would

[1]The necessarily unique node adjacent to a leaf is called its *father*.

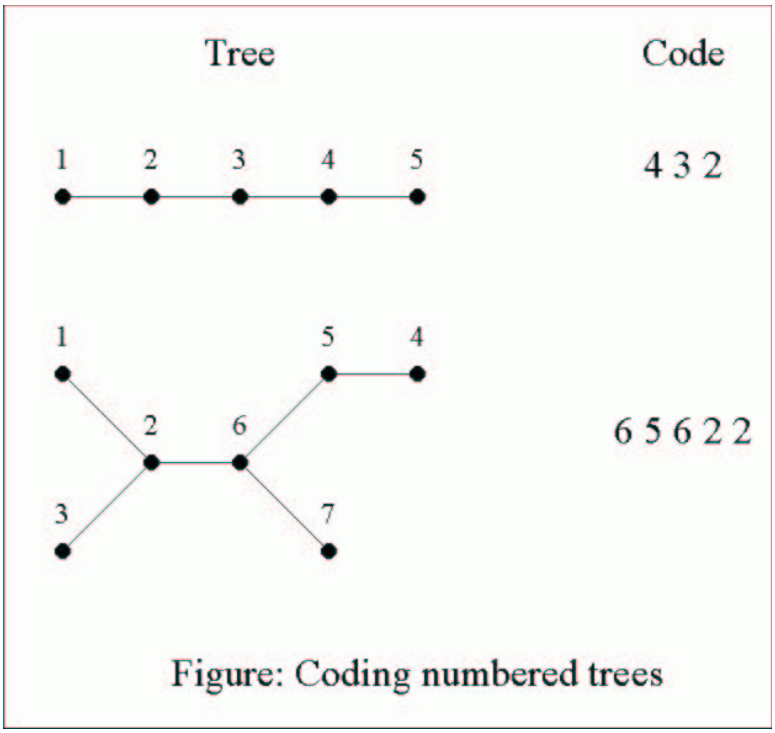


Figure 10.1:

appear in the code. So the missing integers must have been leaves to begin with or must be one of the two undeleted vertices left when the coding process terminates. But by the end of the process the two remaining vertices are leaves, and if they weren't leaves to begin with, they must have become leaves by having their sons deleted, which means they would not have been missing from the code. So the two vertices remaining at the end must also have been leaves of the original tree. ■

(b) How many numbered trees with n vertices are there? Justify your answer assuming the result of the previous problem part.

Solution. There are exactly as many n-vertex numbered trees as the number of possible code words, that is, the number of length $n-2$ sequences of integers between 1 and n. So there are n^{n-2} numbered trees.

The reason is that the map from trees to codes is a bijection. To see this, note that the tree reconstruction procedure finds *the only possible tree* with that code. So there can't be two trees with the same code, i.e., the map from a tree to its code is an injection. But since the reconstruction procedure finds a tree for every possible codeword, the map from trees to codes is also a surjection. ■

10.8 Problem Set 8

Problem 10.8.1. Prove that $\sum_{k=1}^{n} k^6 = \Theta(n^7)$.

Solution. Let $S_n ::= \sum_{k=1}^{n} k^6$.

One approach is to use the Integral Method:

$$\frac{n^7}{7} = \int_0^n x^6\, dx \leq S_n \leq \int_0^n (x+1)^6\, dx = \frac{(n+1)^7}{7} - \frac{1}{7}.$$

So we have $n^7 \leq 7S_n$, and so $n^7 = O(S_n)$. Also $(n+1)^7/7 - 1/7 = O(n^7)$, and so $S_n = O(n^7)$. Hence, $S_n = \Theta(n^7)$.

An alternative approach not using the Integral Method goes as follows. There are n terms in S_n and each term is at most n^6, so $S_n \leq n \cdot n^6 = n^7 = O(n^7)$. So $S_n = O(n^7)$.

On the other hand, at least $(n-1)/2$ of the terms are as large as $[(n-1)/2]^6$, so

$$\begin{aligned} S_n &\geq ((n-1)/2) \cdot [(n-1)/2]^6 \\ &= [(n-1)/2]^7 \\ &\geq (n/3)^7 \end{aligned}$$

for $n > 3$, so $n^7 \leq 3^7 \cdot S_n$. In other words, $n^7 = O(S_n)$. ■

Problem 10.8.2. Determine which of these choices

$$\Theta(n), \quad \Theta(n^2 \log n), \quad \Theta(n^2), \quad \Theta(1), \quad \Theta(2^n), \quad \Theta(2^{n \ln n}), \quad \text{none of these}$$

describes each function's asymptotic behavior. Full proofs are not required, but briefly explain your answers.

(a)

$$n + \ln n + (\ln n)^2$$

Solution. Both $n > \ln n$ and $n > (\ln n)^2$ hold for all sufficiently large n. Thus, for all sufficiently large n:

$$n < n + \ln n + (\ln n)^2 < n + n + n$$

So $n + \ln n + (\ln n)^2 = \Theta(n)$. ■

(b)

$$\frac{n^2 + 2n - 3}{n^2 - 7}$$

Solution. Observe that:

$$\lim_{n\to\infty} \frac{n^2+2n-3}{n^2-7} = 1$$

This means that, for all sufficiently large n, the fraction lies, for example, between 0.99 and 1.01 and is therefore $\Theta(1)$. ■

(c)

$$\sum_{i=0}^{n} 2^{2i+1}$$

Solution. Geometric sums are dominated by their largest term, which is $2^{2n+1} = 2 \cdot 4^n$. This is $\Theta(4^n)$, which does not appear in the list provided. ■

(d)

$$\ln(n^2!)$$

Solution. By Stirling's formula:

$$n^2! \sim \sqrt{2\pi n^2}\left(\frac{n^2}{e}\right)^{n^2}$$

Taking logarithms gives:

$$\begin{aligned}
\ln(n^2!) &\sim \ln(\sqrt{2\pi n^2}\left(\frac{n^2}{e}\right)^{n^2}) \\
&= \ln(\sqrt{2\pi n^2}) + \ln\left(\frac{n^2}{e}\right)^{n^2} \\
&= \frac{1}{2}\ln 2\pi + \ln n + n^2 \ln(\frac{n^2}{e}) \\
&= \frac{1}{2}\ln 2\pi + \ln n + n^2(2\ln n - 1)
\end{aligned}$$

It is then easy to see that this expression and $n^2 \ln n$ are big-O of each other by looking at limits as n goes to ∞, so we conclude that $\ln(n^2!) = \Theta(n^2 \ln n)$. ■

(e)

$$\sum_{k=1}^{n} k\left(1 - \frac{1}{2^k}\right)$$

Solution. The expression in parentheses is always at least $1/2$ and at most 1. Thus, we have the bounds:

$$\frac{1}{2}\sum_{k=1}^{n} k \leq \sum_{k=1}^{n} k\left(1 - \frac{1}{2^k}\right) \leq \sum_{k=1}^{n} k$$

Since the first expression and the last are both $\Theta(n^2)$, so is the one in the middle. ■

Problem 10.8.3. **(a)** Prove that the relation, R, on functions such that fRg iff $f = o(g)$ is a strict partial order.

Solution. We need only show that R is irreflexive and transitive.

Now for any function, f,

$$\lim_{x\to\infty} f(x)/f(x) = 1 \neq 0,$$

so $\neg(fRf)$. This implies R is irreflexive.

To show R is transitive, assume fRg and gRh. This means that $\lim_{x\to\infty} f(x)/g(x) = 0$ and $\lim_{x\to\infty} g(x)/h(x) = 0$. Since these limits exist, there must be an x_0 such that the denominators, $g(x)$ and $h(x)$, are nonzero for all $x \geq x_0$. Therefore,

$$\frac{f(x)}{g(x)} \cdot \frac{g(x)}{h(x)} = \frac{f(x)}{h(x)} \tag{10.8}$$

for all $x \geq x_0$.

So, we have

$$\begin{aligned}
0 &= 0 \cdot 0 \\
&= \left(\lim_{x\to\infty} \frac{f(x)}{g(x)}\right) \cdot \left(\lim_{x\to\infty} \frac{g(x)}{h(x)}\right) \\
&= \lim_{x\to\infty} \frac{f(x)}{g(x)} \cdot \frac{g(x)}{h(x)} && \text{(property of limits)} \\
&= \lim_{x\to\infty} f(x)/h(x) && \text{(by (10.8)),}
\end{aligned}$$

This means that fRh holds, proving that R is transitive. ■

(b) Describe two incomparable elements in this partial order.

Solution. The functions

$$\begin{aligned}
f(n) &::= 1 + (n\sin(n\pi/2))^2, \\
g(n) &::= 3n
\end{aligned}$$

from Class Problems 8W, problem 2(c) are not even O() of each other, and so are certainly not o() of each other. ■

Problem 10.8.4. **(a)** Describe a bijection between the set of paths from $(0, 0)$ to $(10, 20)$ consisting of right-steps (which increment the first coordinate) and up-steps (which increment the second coordinate) and the set of 30-bit sequences with 10 zeros and 20 ones.

Solution. Map a path

$$p ::= (x_0, y_0), (x_1, y_1), \ldots, (x_{30}, y_{30})$$

to a bit string $b_1 b_2 \ldots b_{30}$ defined by the rule that

$$b_i ::= y_i - y_{i-1}$$

for $1 \leq i \leq 30$.

The inverse map is even easier to describe: map the 30-bit sequence $b_1 b_2 \ldots b_{30}$ to a path where the i-th step is right if $b_i = 0$ and up if $b_i = 1$. ■

(b) Mr. and Mrs. Grumperson have collected 13 identical pieces of coal as Christmas presents for their beloved children, Lucy and Spud. Describe a bijection between the set of all ways of distributing the 13 coal pieces to the two children and the set of 14-bit sequences with exactly 1 one.

Solution. Map a distribution in which Lucy get l pieces and Spud gets s pieces to a 14-bit sequence with l zeros, a one, and then s zeros. ■

(c) On Christmas Eve, Mr. and Mrs. Grumperson remember that they have a third child, little Bottlecap, locked in the attic. Describe a bijection between the set of all ways of distributing the 13 coal pieces to the three children and the set of 15-bit sequences with exactly 2 ones.

Solution. Map a distribution in which Lucy gets l pieces, Spud gets s pieces, and Bottlecap gets b pieces to a 15-bit sequence with l zeros, a one, s zeros, a one, and b zeros. ■

(d) On reflection, Mr. and Mrs. Grumperson decide that each of their three children should receive *at least two* pieces of coal for Christmas. Describe a bijection between the set of all ways of distributing the 13 coal pieces to the three Grumperson children given this constraint and the set of 9-bit sequences with exactly 2 ones.

Solution. Map a distribution in which Lucy gets $l \geq 2$ pieces, Spud gets $s \geq 2$ pieces, and Bottlecap gets $b \geq 2$ pieces to a 9-bit sequence with exactly $l-2$ zeros, a one, $s-2$ zeros, a one, and $b-2$ zeros. ■

(e) Describe a bijection between the set of solutions over the natural numbers to the inequality:

$$x_1 + x_2 + \cdots + x_{10} \leq 100, \tag{10.9}$$

and the set of 110-bit sequences with exactly 10 ones.

Solution. Map a solution $(x_1, x_2, \ldots, x_{10})$ to the string

$$0^{x_1}10^{x_2}1 \ldots 10^{x_{10}}10^{100-(x_1+x_2+\cdots+x_{10})}.$$

■

(f) Describe a bijection between solutions to the inequality (10.9) and sequences $(y_1, y_2, \ldots, y_{10})$ such that:

$$0 \leq y_1 \leq y_2 \leq \cdots \leq y_{10} \leq 100.$$

Solution. Let $y_i ::= x_1 + \cdots + x_i$ for $1 \leq i \leq 10$. ■

10.9 Miniquiz Apr. 20

Problem 10.9.1.

Circle each of the true statements below:

- $3n = o(n^2)$
- $\ln n = O(n^k)$, $k > 1$ a constant
- $3^{n/2} = O(3^n)$
- $(3n-7)/(n+4) = \Theta(1)$
- $(3n-7)/(n+4) \sim 1$
- $n^k = O(k^n)$, $k > 1$ a constant
- $k^n = O(n^k)$, $k > 1$ a constant
- $3^n = O(2^n)$
- $2^n = o(3^n)$
- $\sum_{i=1}^{n} i = O(n)$

Solution. Items 1, 2, 3, 4, 6, and 9 are true. ■

Problem 10.9.2. Show that

$$\ln(n^2!) = \Theta(n^2 \ln n)$$

Solution. By Stirling's formula:

$$n^2! \sim \sqrt{2\pi n^2}\left(\frac{n^2}{e}\right)^{n^2}$$

Taking logarithms gives:

$$\begin{aligned}
\ln(n^2!) &\sim \ln(\sqrt{2\pi n^2}\left(\frac{n^2}{e}\right)^{n^2}) \\
&= \ln(\sqrt{2\pi n^2}) + \ln\left(\frac{n^2}{e}\right)^{n^2} \\
&= \frac{1}{2}\ln 2\pi + \ln n + n^2 \ln(\frac{n^2}{e}) \\
&= \frac{1}{2}\ln 2\pi + \ln n + n^2(2\ln n - 1)
\end{aligned}$$

It is then easy to see that this expression and $n^2 \ln n$ are big-O of each other, so we conclude that $\ln(n^2!) = \Theta(n^2 \ln n)$. ■

Problem 10.9.3. A license plate consists of either:

- 3 letters followed by 3 digits (standard plate)
- 5 letters (vanity plate)
- 2 characters – letters or numbers (big shot plate)

Let L be the set of all possible license plates.

(a) Express L in terms of

$$\mathcal{A} = \{A, B, C, \ldots, Z\}$$
$$\mathcal{D} = \{0, 1, 2, \ldots, 9\}$$

using unions ($\cup$) and set products ($\times$ or the compact exponent notation).

Solution.

$$L = (A^3 \times D^3) \cup A^5 \cup (A \cup D)^2$$

■

(b) Compute $|L|$, the number of different license plates, using the sum and product rules.

Solution.

$$\begin{aligned}
|L| &= \left|(A^3 \times D^3) \cup A^5 \cup (A \cup D)^2\right| \\
&= \left|(A^3 \times D^3)\right| + \left|A^5\right| + \left|(A \cup D)^2\right| && \text{Sum Rule} \\
&= |A|^3 \cdot |D|^3 + |A|^5 + |A \cup D|^2 && \text{Product Rule} \\
&= |A|^3 \cdot |D|^3 + |A|^5 + (|A| + |D|)^2 && \text{Sum Rule} \\
&= 26^3 \cdot 10^3 + 26^5 + 36^2 = 29458672
\end{aligned}$$

■

Problem 10.9.4. There are 20 different books arranged in a row on a shelf.

(a) Describe a bijection between ways of choosing x distinct books, $0 \leq x \leq 20$, and 20-bit sequences with exactly x ones.

Solution. There is a bijection from 20-bit sequences with x ones to book selections: map each 1 to a chosen book and each 0 to a non-chosen book. ■

(b) How many ways are there of choosing at least 2 books from the 20? (Hint: How many ways are there of not choosing at least 2 books?)

Solution. If you do not choose at least 2 books, you have chosen 0 or 1 book. There is only 1 way of choosing no books, and 20 ways of choosing 1 book. Since there are 2^{20} 20-bit sequences, there are $2^{20} - 20 - 1$ 20-bit sequences with at least one 1. ■

(c) Describe a bijection between ways of choosing 6 of these books so that no two adjacent books are selected and 15-bit sequences with exactly 6 ones.

Solution. There is a bijection from 15-bit sequences with exactly six 1's to valid book selections: given such a sequence, map each zero to a non-chosen book, each of the first five 1's to a chosen book followed by a non-chosen book, and the last 1 to a chosen book. ■

(d) Using your answer from part (c), give the bit representation corresponding to the selection of the 1st, 3rd, 5th, 10th, 13th, and 19th books.

Solution. 111000101000010 ■

Appendix

Lemma (Stirling's Formula).

$$n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n},$$

For functions $f, g : \mathbb{R} \to \mathbb{R}$, we say f is *asymptotically equal* to g, in symbols,

$$f(x) \sim g(x)$$

iff

$$\lim_{x\to\infty} f(x)/g(x) = 1.$$

For functions $f, g : \mathbb{R} \to \mathbb{R}$, we say f is *asymptotically smaller* than g, in symbols,

$$f(x) = o(g(x)),$$

iff

$$\lim_{x\to\infty} f(x)/g(x) = 0.$$

Given functions $f, g : \mathbb{R} \mapsto \mathbb{R}$, with g nonnegative, we say that[2]

$$f = O(g)$$

iff

$$\limsup_{x\to\infty} |f(x)|/g(x) < \infty.$$

An alternative, equivalent, definition is

$$f = O(g)$$

iff there exists a constant $c \geq 0$ and an x_0 such that for all $x \geq x_0$, $|f(x)| \leq cg(x)$.

Finally, we say

$$f = \Theta(g) \quad \text{iff} \quad f = O(g) \wedge g = O(f).$$

[2]

$$\limsup_{x\to\infty} h(x) ::= \lim_{x\to\infty} \text{lub}_{y\geq x} h(y).$$

10.10 In-Class Problems Week 10, Wed.

Problem 10.10.1. Solve the following problems using the Pigeonhole Principle. For each problem, try to identify the *pigeons*, the *pigeonholes*, and a *rule* assigning each pigeon to a pigeonhole.

(a) In a room of 500 people, there exist two who share a birthday.

Solution. The pigeons are the 500 people. The pigeonholes are 366 possible birthdays. Map each person to his or her own birthday. Since there 500 people and 366 birthdays, some two people must have the same birthday by the Pigeonhole Principle. ■

(b) Every MIT ID number starts with a 9 (we think). Suppose that each of the 101 students in 6.042 sums the nine digits of his or her ID number and doubles the result. Explain why two students' results must be the same.

Solution. The students are the pigeons, the possible results are the pigeonholes, and we map each student to the result calculated from his or her MIT ID number. Every sum is in the range from 9 to $9 + 8 \cdot 9 = 81$, and since results are obtained by doubling these sums, there are also 73 possible pigeonholes. Since there are more pigeons than pigeonholes, there must be two pigeons in the same pigeonhole; in other words, there must be two students with the same result. ■

(c) In every set of 100 integers, there exist two whose difference is a multiple of 37.

Solution. The pigeons are the 100 integers. The pigeonholes are the numbers 0 to 36. Map integer k to $\mathrm{rem}(k, 37)$. Since there are 100 pigeons and only 37 pigeonholes, two pigeons must go in the same pigeonhole. This means $\mathrm{rem}(k_1, 37) = \mathrm{rem}(k_2, 37)$, which implies that $k_1 - k_2$ is a multiple of 37. ■

(d) For any five points inside a unit square, there are two points at distance less than $1/\sqrt{2}$.

Solution. The pigeons are the points. The pigeonholes are the four subsquares of the unit square, each of side length $1/2$. There are five pigeons and four pigeonholes, and a pigeon maps to the subsquare it is in (points on a boundary get assigned to the leftmost lowest possible subsquare) so more than one point must be in the same subsquare. The points in the same subsquare are at distance less than $1/\sqrt{2}$, because the most distant points in the quadrant are at opposite corners at exactly this distance, but only one of these corners can be *inside* the square (not on the boundary), so the distance between the points in the quadrant must actually be less than $1/\sqrt{2}$. ■

(e) For any five points inside an equilateral triangle of side length 2, there are two points at distance less than 1.

Solution. The pigeons are the points. The pigeonholes are the four sub-equilateral triangles of side length 1. There are five pigeons and four pigeonholes, so more than one point must be in the same sub-equilateral triangle. Points inside the same sub-equilateral triangle are at distance less than 1. ■

Problem 10.10.2. Your 6.001 tutorial has 12 students, who are supposed to break up into 4 groups of 3 students each. Your TA has observed that the students waste too much time trying to form balanced groups, so he decided to pre-assign students to groups and email the group assignments to his students.

(a) Your TA has a list of the 12 students in front of him, so he divides the list into consecutive groups of 3. For example, if the list is ABCDEFGHIJKL, the TA would define a sequence of four groups to be $(\{A, B, C\}, \{D, E, F\}, \{G, H, I\}, \{J, K, L\})$. This way of forming groups defines a mapping from a list of twelve students to a sequence of four groups. This is a k-to-1 mapping for what k?

Solution. Two lists map to the same sequence of groups iff the first 3 students are the same on both lists, and likewise for the second, third, and fourth consecutive sublists of 3 students. So for a given sequence of 4 groups, the number of lists which map to it is

$$(3!)^4$$

because there are 3! ways to order the students in each of the 4 consecutive sublists. ■

(b) A group assignment specifies which students are in the same group, but not any order in which the groups should be listed. If we map a sequence of 4 groups,

$$(\{A, B, C\}, \{D, E, F\}, \{G, H, I\}, \{J, K, L\}),$$

into a group assignment

$$\{\{A, B, C\}, \{D, E, F\}, \{G, H, I\}, \{J, K, L\}\},$$

this mapping is j-to-1 for what j?

Solution. $4!$. ■

(c) How many group assignments are possible?

Solution.

$$\frac{12!}{4! \cdot (3!)^4} = 15400$$

different assignments.

There are $12!$ possible lists of students, and we can map each list to an assignment by first mapping the list to a sequence of four groups, and then mapping the sequence to the assignment. Since the first map is $(3!)^4$-to-1 and and the second is $4!$-to-1, the composite map is $(3!)^4 \cdot 4!$-to-1. So by the Division Rule, $12! = \left((3!)^4 \cdot 4!\right) A$ where A is the number of assignments. ■

(d) In how many ways can $3n$ students be broken up into n groups of 3?

Solution.

$$\frac{(3n)!}{(3!)^n n!}.$$

This follows simply by replacing "12" by "3n" in the solution to the previous problem parts. ■

Problem 10.10.3. Answer the following quesions using the Generalized Product Rule.

(a) Next week, I'm going to get really fit! On day 1, I'll exercise for 5 minutes. On each subsequent day, I'll exercise 0, 1, 2, or 3 minutes more than the previous day. For example, the number of minutes that I exercise on the seven days of next week might be 5, 6, 9, 9, 9, 11, 12. How many such sequences are possible?

Solution. The number of minutes on the first day can be selected in 1 way. The number of minutes on each subsequent day can be selected in 4 ways. Therefore, the number of exercise sequences is $1 \cdot 4^6$ by the extended product rule. ■

(b) An ***r-permutation*** of a set is a sequence of r distinct elements of that set. For example, here are all the 2-permutations of $\{a, b, c, d\}$:

$$\begin{array}{lll} (a,b) & (a,c) & (a,d) \\ (b,a) & (b,c) & (b,d) \\ (c,a) & (c,b) & (c,d) \\ (d,a) & (d,b) & (d,c) \end{array}$$

How many r-permutations of an n-element set are there? Express your answer using factorial notation.

Solution. There are n ways to choose the first element, $n-1$ ways to choose the second, $n-2$ ways to choose the third, ..., and $n-r+1$ ways to choose the r-th element. Thus, there are:

$$n \cdot (n-1) \cdot (n-2) \cdots (n-r+1) = \frac{n!}{(n-r)!}$$

r-permutations of an n-element set. ■

(c) How many $n \times n$ matrices are there with *distinct* entries drawn from $\{1, \ldots, p\}$, where $p \geq n^2$?

Solution. There are p ways to choose the first entry, $p-1$ ways to choose the second for each way of choosing the first, $p-2$ ways of choosing the third, and so forth. In all there are

$$p(p-1)(p-2)\cdots(p-n^2+1) = \frac{p!}{(p-n^2)!}$$

such matrices. Alternatively, this is the number of n^2-permutations of a p element set, which is $p!/(p-n^2)!$. ■

Problem 10.10.4. A certain company wants to have security for their computer systems. So they have given everyone a name and password. A length 10 word containing each of the characters:

a, d, e, f, i, l, o, p, r, s,

is called a *cword*. A password will be a cword which does not contain any of the subwords "fails", "failed", or "drop".

For example, the following two words are passwords:

adefiloprs, srpolifeda,

but the following three cwords are not:

a**drop**eflrs, **failedrop**s, **drop**e**fails**.

(a) How many cwords contain the subword "drop"?

Solution. Such cwords are obtainable by taking the word "drop" and the remaining 6 letters in any order. There are 7! permutations of these 7 items. ■

(b) How many cwords contain both "drop" and "fails"?

Solution. Take the words "drop" and "fails" and the remaining letter "e" in any order. So there are 3! such cwords. ■

(c) Use the Inclusion-Exclusion Principle to find a simple formula for the number of passwords.

Solution. There are 7! cwords that contain "drop", 6! that contain "fails", and 5! that contain "failed". There are 3! cwords containing both "drop" and "fails". No cword can contain both "fails" and "failed". The cwords containing both "drop" and "failed" come from taking the subword "failedrop" and the remaining letter "s" in any order, so there are 2! of them. So by Inclusion-exclusion, we have the number of cwords containing at least one of the three forbidden subwords is

$$(7! + 6! + 5!) - (3! + 0 + 2!) + 0 = 5!(49) - 8.$$

Among the 10! cwords, the remaining ones are passwords, so the number of passwords is

$$10! - 5!(48) + 8 = 3,622,928.$$

■

Counting Principles

Rule (Pigeonhole Principle). *If $|A| > |B|$, then for every function $f : A \to B$ there exist two different elements of A that are mapped to the same element of B.*

> "If more than n pigeons are assigned to n holes, then there must exist two pigeons assigned to the same hole."

A ***k-to-1 function*** maps exactly k elements of the domain to every element of the range. For example, the function mapping ears of 6.042 students to their owners is 2-to-1.

Rule (Division Rule). *If $f : A \to B$ is k-to-1, then $|A| = k \cdot |B|$.*

Rule (Generalized Product Rule). *Let S be a set of length-k sequences. If there are:*

- *n_1 possible first entries,*

- n_2 *possible second entries for each first entry,*
- n_3 *possible third entries for each combination of first and second entries,*

$\vdots$

then:

$$|S| = n_1 \cdot n_2 \cdot n_3 \cdots n_k$$

Rule (Inclusion-Exclusion for Three Sets).

$$\begin{aligned} |S_1 \cup S_2 \cup S_3| = &|S_1| + |S_2| + |S_3| \\ &- |S_1 \cap S_2| - |S_1 \cap S_3| - |S_2 \cap S_3| \\ &+ |S_1 \cap S_2 \cap S_3| \end{aligned}$$

Chapter 11

More Counting

11.1 Counting Subsets

How many k-element subsets of an n-element set are there? This question arises all the time in various guises:

- In how many ways can I select 5 books from my collection of 100 to bring on vacation?
- How many different 13-card Bridge hands can be dealt from a 52-card deck?
- In how many ways can I select 5 toppings for my pizza if there are 14 available toppings?

This number comes up so often that there is a special notation for it:

$$\binom{n}{k} ::= \text{ the number of } k\text{-element subsets of an } n\text{-element set.}$$

The expression $\binom{n}{k}$ is read "n choose k." Now we can immediately express the answers to all three questions above:

- I can select 5 books from 100 in $\binom{100}{5}$ ways.
- There are $\binom{52}{13}$ different Bridge hands.
- There are $\binom{14}{5}$ different 5-topping pizzas, if 14 toppings are available.

11.1.1 The Subset Rule

We can derive a simple formula for the n-choose-k number using the Division Rule. We do this by mapping any permutation of an n-element set $\{a_1, \ldots, a_n\}$ into a k-element subset simply by taking the first k elements of the permutation. That is, the permutation $a_1 a_2 \ldots a_n$ will map to the set $\{a_1, a_2, \ldots, a_k\}$.

Notice that any other permutation with the same first k elements $a_1, \ldots, a_k$ *in any order* and the same remaining elements $n-k$ elements *in any order* will also map to this set. What's more, a permutation can only map to $\{a_1, a_2, \ldots, a_k\}$ if its first k elements are the elements $a_1, \ldots, a_k$ in some order. Since there are $k!$ possible permutations of the first k elements and $(n-k)!$ permutations of the remaining elements, we conclude from the Product Rule that exactly $k!(n-k)!$ permutations of the n-element set map to the the particular subset, S. In other words, the mapping from permutations to k-element subsets is $k!(n-k)!$-to-1.

But we know there are $n!$ permutations of an n-element set, so by the Division Rule, we conclude that

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

which proves:

Rule 10 (Subset Rule). *The number,*

$$\binom{n}{k},$$

of k-element subsets of an n-element set is

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

Notice that this works even for 0-element subsets: $n!/0!n! = 1$. Here we use the fact that $0!$ is a *product* of 0 terms, which by convention equals 1. (A *sum* of zero terms equals 0.)

11.1.2 Bit Sequences

How many n-bit sequences contain exactly k ones? We've already seen the straightforward bijection between subsets of an n-element set and n-bit sequences. For example, here is a 3-element subset of $\{x_1, x_2, \ldots, x_8\}$ and the associated 8-bit sequence:

$$\begin{array}{llllllllll} \{ & x_1, & & & x_4, & x_5 & & & & \} \\ (& 1, & 0, & 0, & 1, & 1, & 0, & 0, & 0 &) \end{array}$$

Notice that this sequence has exactly 3 ones, each corresponding to an element of the 3-element subset. More generally, the n-bit sequences corresponding to a k-element subset will have exactly k ones. So by the Bijection Rule,

The number of n-bit sequences with exactly k ones is $\binom{n}{k}$.

11.2 Magic Trick

There is a Magician and an Assistant. The Assistant goes into the audience with a deck of 52 cards while the Magician looks away. Five audience members each select one card from the deck. The Assistant then gathers up the five cards and holds up four of them so the Magician can see them. The Magician concentrates for a short time and then correctly names the secret, fifth card!

Since we don't really believe the Magician can read minds, we know the Assistant has somehow communicated the secret card to the Magician. Since real Magicians and Assistants are not to be trusted, we can expect that the Assistant would illegitimately signal the Magician with coded phrases or body language, but they don't have to cheat in this way. In fact, the Magician and Assistant could be kept out of sight of each other while some audience member holds up the 4 cards designated by the Assistant for the Magician to see.

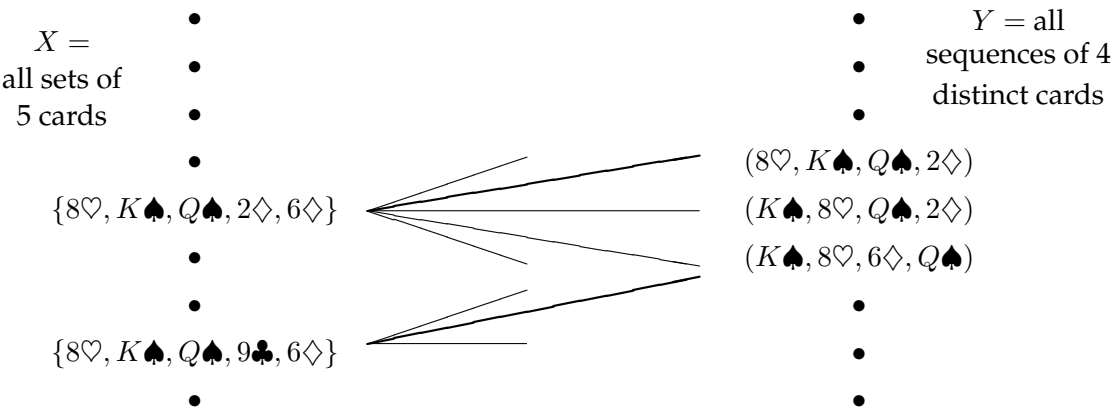
Of course, without cheating, there is still an obvious way the Assistant can communicate to the Magician: he can choose any of the $4! = 24$ permutations of the 4 cards as the order in which to hold up the cards. However, this alone won't quite work: there are 48 cards remaining in the deck, so the Assistant doesn't have enough choices of orders to indicate exactly what the secret card is (though he could narrow it down to two cards).

11.2.1 The Secret

The method the Assistant can use to communicate the fifth card exactly is a nice application of what we know about counting and matching.

The Assistant really has another legitimate ways to communicate: he can choose *which of the five cards to keep hidden*. Of course, it's not clear how the Magician could determine which of these five possibilities the Assistant selected by looking at the four visible cards, but there is a way as we'll now explain.

The problem facing the Magician and Assistant is actually a bipartite matching problem. Put all the *sets* of 5 cards in a collection X on the left. And put all the sequences of 4 distinct cards in a collection Y on the right. These are the two sets of vertices in the bipartite graph. There is an edge between a set of 5 cards and a sequence of 4 if every card in the sequence is also in the set. In other words, if the audience selects a set of cards, then the Assistant must reveal a sequence of cards that is adjacent in the bipartite graph. Some edges are shown in the diagram below.



For example, $\{8\heartsuit, K\spadesuit, Q\spadesuit, 2\diamondsuit, 6\diamondsuit\}$ is an element of X on the left. If the audience selects this set of 5 cards, then there are many different 4-card sequences on the right in set Y that the Assistant could choose to reveal, including $(8\heartsuit, K\spadesuit, Q\spadesuit, 2\diamondsuit)$, $(K\spadesuit, 8\heartsuit, Q\spadesuit, 2\diamondsuit)$, and $(K\spadesuit, 8\heartsuit, 6\diamondsuit, Q\spadesuit)$.

What the Magician and his Assistant need to perform the trick is a *matching* for the X vertices. If they agree in advance on some matching, then when the audience selects a set of 5 cards, the Assistant reveals the matching sequence of 4 cards. The Magician uses the reverse of the matching to find the audience's chosen set of 5 cards, and so he can name the one not already revealed.

For example, suppose the Assistant and Magician agree on a matching containing the two bold edges in the diagram above. If the audience selects the set $\{8\heartsuit, K\spadesuit, Q\spadesuit, 9\clubsuit, 6\diamondsuit\}$, then the Assistant reveals the corresponding sequence $(K\spadesuit, 8\heartsuit, 6\diamondsuit, Q\spadesuit)$. Using the matching, the Magician sees that $(K\spadesuit, 8\heartsuit, 6\diamondsuit, Q\spadesuit)$ is indeed matched to $\{8\heartsuit, K\spadesuit, Q\spadesuit, 9\clubsuit, 6\diamondsuit\}$, so he can name the one card in the corresponding set not already revealed, namely, the $9\clubsuit$. Notice that the fact that the sets are *matched*, that is, that different sets are paired with *distinct* sequences, is essential. For example, the Assistant could have revealed the same sequence, $(K\spadesuit, 8\heartsuit, 6\diamondsuit, Q\spadesuit)$, if the audience picked a different set $\{8\heartsuit, K\spadesuit, Q\spadesuit, 2\diamondsuit, 6\diamondsuit\}$, but if that happened, then the Magician would have no way to tell if remaining card was the $9\clubsuit$ or $2\diamondsuit$.

So how can we be sure the needed matching can be found? The reason is that each vertex on the left has degree $5 \cdot 4! = 120$, since there are five ways to select the card kept secret and there are $4!$ permutations of the remaining 4 cards. In addition, each vertex on the right has degree 48, since there are 48 possibilities for the fifth card. So this graph is *degree-constrained* (see Notes 6), and therefore satisfies Hall's matching condition.

In fact, this reasoning show that the Magician could still pull off the trick if 120 cards were left instead of 48, that is, the trick would work with a deck as large as 124 different cards —without any magic!

11.2.2 The Real Secret

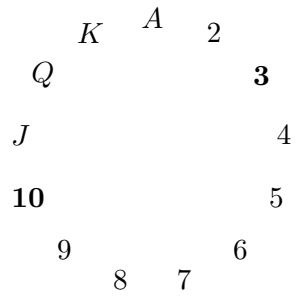
But wait a minute! It's all very well in principle to have the Magician and his Assistant agree on a matching, but how are they supposed to remember a matching with $\binom{52}{5} = 2,598,960$ edges? For the trick to work in practice, there has to be a way to match hands and card sequences mentally and on the fly.

We'll describe one approach. As a running example, suppose that the audience selects:

$$10\heartsuit \quad 9\diamondsuit \quad 3\heartsuit \quad Q\spadesuit \quad J\diamondsuit$$

- The Assistant picks out two cards of the same suit. In the example, the assistant might choose the $3\heartsuit$ and $10\heartsuit$.

- The Assistant locates the values of these two cards on the cycle shown below:



For any two distinct values on this cycle, one is always between 1 and 6 hops clockwise from the other. For example, the $3\heartsuit$ is 6 hops clockwise from the $10\heartsuit$.

- The more counterclockwise of these two cards is revealed first, and the other becomes the secret card. Thus, in our example, the $10\heartsuit$ would be revealed, and the $3\heartsuit$ would be the secret card. Therefore:
  - The suit of the secret card is the same as the suit of the first card revealed.
  - The value of the secret card is between 1 and 6 hops clockwise from the value of the first card revealed.
- All that remains is to communicate a number between 1 and 6. The Magician and Assistant agree beforehand on an ordering of all the cards in the deck from smallest to largest such as:

$$A\clubsuit\ 2\clubsuit\ \ldots\ K\clubsuit\ A\diamondsuit\ 2\diamondsuit\ \ldots\ K\diamondsuit\ A\heartsuit\ 2\heartsuit\ \ldots\ K\heartsuit\ A\spadesuit\ 2\spadesuit\ \ldots\ K\spadesuit$$

The order in which the last three cards are revealed communicates the number according to the following scheme:

$$\begin{array}{lllll}
(\ \text{small}, & \text{medium}, & \text{large}\ ) & = 1\\
(\ \text{small}, & \text{large}, & \text{medium}\ ) & = 2\\
(\ \text{medium}, & \text{small}, & \text{large}\ ) & = 3\\
(\ \text{medium}, & \text{large}, & \text{small}\ ) & = 4\\
(\ \text{large}, & \text{small}, & \text{medium}\ ) & = 5\\
(\ \text{large}, & \text{medium}, & \text{small}\ ) & = 6
\end{array}$$

In the example, the Assistant wants to send 6 and so reveals the remaining three cards in large, medium, small order. Here is the complete sequence that the Magician sees:

$$10\heartsuit \quad Q\spadesuit \quad J\diamondsuit \quad 9\diamondsuit$$

- The Magician starts with the first card, $10\heartsuit$, and hops 6 values clockwise to reach $3\heartsuit$, which is the secret card!

So that's how the trick can work with a standard deck of 52 cards. On the other hand, Hall's Theorem implied that the Magician and Assistant could in principle perform the trick with a deck of up to 124 cards. Until very recently we didn't know how to perform it in practice for decks larger than the standard one, but on March 30, 2007, David Shin, who was 6.042 TA in Spring '06, sent an email describing an ingenious matching for a 124 card deck that a Magician and Assistant could actually figure out in their heads. We'll describe Shin's method another time.

11.2.3 Same Trick with Four Cards?

Suppose that the audience selects only *four* cards and the Assistant reveals a sequence of *three* to the Magician. Can the Magician determine the fourth card?

Let X be all the sets of four cards that the audience might select, and let Y be all the sequences of three cards that the Assistant might reveal. Now, on one hand, we have

$$|X| = \binom{52}{4} = 270,725$$

by the Subset Rule. On the other hand, we have

$$|Y| = 52 \cdot 51 \cdot 50 = 132,600$$

by the Generalized Product Rule. Thus, by the Pigeonhole Principle, the Assistant must reveal the *same* sequence of three cards for some two *different* sets of four. This is bad news for the Magician: if he sees that sequence of three, then there are at least two possibilities for the fourth card which he cannot distinguish. So there is no legitimate way for the Assistant to communicate exactly what the fourth card is!

11.3 The Bookkeeper Rule

11.3.1 Sequences of Subsets

Choosing a k-element subset of an n-element set is the same as splitting the set into a pair of subsets: the first subset of size k and the second subset consisting of the remaining $n-k$ elements. So the Subset Rule can be understood as a rule for counting the number of such splits into pairs of subsets.

We can generalize this to splits into m subsets. Namely, let A be an n-element set and $k_1, k_2, \ldots, k_m$ be nonnegative integers whose sum is n. A $(k_1, k_2, \ldots, k_m)$***-split of*** A is a sequence

$$(A_1, A_2, \ldots, A_m)$$

where the A_i are pairwise disjoint subsets of A and $|A_i| = k_i$ for $i = 1, \ldots, m$.

The same reasoning used to explain the Subset Rule extends directly to a rule for counting the number of splits into subsets of given sizes.

Rule 11 (Subset Split Rule). *The number of $(k_1, k_2, \ldots, k_m)$-splits of an n-element set is*

$$\binom{n}{k_1, \ldots, k_m} ::= \frac{n!}{k_1!\, k_2!\, \cdots k_m!}$$

The proof of this Rule is essentially the same as for the Subset Rule. Namely, we map any permutation $a_1 a_2 \ldots a_n$ of an n-element set, A, into a $(k_1, k_2, \ldots, k_m)$-split by letting the 1st subset in the split be the first k_1 elements of the permutation, the 2nd subset of the split be the next k_2 elements, $\ldots$, and the mth subset of the split be the final k_m elements of the permutation. This map is a $k_1!\, k_2!\, \cdots\, k_m!$-to-1 from the $n!$ permutations to the $(k_1, k_2, \ldots, k_m)$-splits of A, and the Subset Split Rule now follows from the Division Rule.

11.3.2 Sequences over an alphabet

We can also generalize our count of n-bit sequences with k-ones to counting length n sequences of letters over an alphabet with more than two letters. For example, how many sequences can be formed by permuting the letters in the 10-letter word BOOKKEEPER?

Notice that there are 1 B, 2 O's, 2 K's, 3 E's, 1 P, and 1 R in BOOKKEEPER. This leads to a straightforward bijection between permutations of BOOKKEEPER and (1,2,2,3,1,1)-splits of $\{1, \ldots, n\}$. Namely, map a permutation to the sequence of sets of positions where each of the different letters occur.

For example, in the permutation BOOKKEEPER itself, the B is in the 1st position, the O's occur in the 2nd and 3rd positions, K's in 4th and 5th, the E's in the 6th, 7th and 9th, P in the 8th, and R is in the 10th position, so BOOKKEEPER maps to

$$(\{1\}, \{2,3\}, \{4,5\}, \{6,7,9\}, \{8\}, \{10\}).$$

From this bijection and the Subset Split Rule, we conclude that the number of ways to rearrange the letters in the word BOOKKEEPER is:

$$\frac{\overbrace{10!}^{\text{total letters}}}{\underbrace{1!}_{\text{B's}}\ \underbrace{2!}_{\text{O's}}\ \underbrace{2!}_{\text{K's}}\ \underbrace{3!}_{\text{E's}}\ \underbrace{1!}_{\text{P's}}\ \underbrace{1!}_{\text{R's}}}$$

This example generalizes directly to an exceptionally useful counting principle which we will call the

Rule 12 (Bookkeeper Rule). *Let $l_1, \ldots, l_m$ be distinct elements. The number of sequences with k_1 occurrences of l_1, and k_2 occurrences of l_2, ..., and k_m occurrences of l_m is*

$$\frac{(k_1 + k_2 + \ldots + k_m)!}{k_1!\, k_2!\, \ldots\, k_m!}$$

Example. 20-Mile Walks.

I'm planning a 20-mile walk, which should include 5 northward miles, 5 eastward miles, 5 southward miles, and 5 westward miles. How many different walks are possible?

There is a bijection between such walks and sequences with 5 N's, 5 E's, 5 S's, and 5 W's. By the Bookkeeper Rule, the number of such sequences is:

$$\frac{20!}{5!^4}$$

11.3.3 A Word about Words

Someday you might refer to the Subset Split Rule or the Bookkeeper Rule in front of a roomful of colleagues and discover that they're all staring back at you blankly. This is not because they're dumb, but rather because we made up the name "Bookkeeper Rule". However, the rule is excellent and the name is apt, so we suggest that you play through: "You know? The Bookkeeper Rule? Don't you guys know *anything???*"

The Bookkeeper Rule is sometimes called the "formula for permutations with indistinguishable objects." The size k subsets of an n-element set are sometimes called ***k-combinations***. Other similar-sounding descriptions are "combinations with repetition, permutations with repetition, r-permutations, permutations with indistinguishable objects," and so on. However, the counting rules we've taught you are sufficient to solve all these sorts of problems without knowing this jargon, so we'll skip it.

11.4 Poker Hands

There are 52 cards in a deck. Each card has a *suit* and a *value*. There are four suits:

$$\underset{\spadesuit}{\text{spades}} \qquad \underset{\heartsuit}{\text{hearts}} \qquad \underset{\clubsuit}{\text{clubs}} \qquad \underset{\diamondsuit}{\text{diamonds}}$$

And there are 13 values:

$$2\,,\ 3\,,\ 4\,,\ 5\,,\ 6\,,\ 7\,,\ 8\,,\ 9\,,\ \overset{\text{jack}}{J}\,,\ \overset{\text{queen}}{Q}\,,\ \overset{\text{king}}{K}\,,\ \overset{\text{ace}}{A}$$

Thus, for example, $8\heartsuit$ is the 8 of hearts and $A\spadesuit$ is the ace of spades. Values farther to the right in this list are considered "higher" and values to the left are "lower".

Five-Card Draw is a card game in which each player is initially dealt a *hand*, a subset of 5 cards. (Then the game gets complicated, but let's not worry about that.) The number of different hands in Five-Card Draw is the number of 5-element subsets of a 52-element set, which is 52 choose 5:

$$\text{total \# of hands} = \binom{52}{5} = 2,598,960$$

Let's get some counting practice by working out the number of hands with various special properties.

11.4.1 Hands with a Four-of-a-Kind

A *Four-of-a-Kind* is a set of four cards with the same value. How many different hands contain a Four-of-a-Kind? Here are a couple examples:

$$\begin{array}{llllll} \{ & 8\spadesuit, & 8\diamondsuit, & Q\heartsuit, & 8\heartsuit, & 8\clubsuit \ \} \\ \{ & A\clubsuit, & 2\clubsuit, & 2\heartsuit, & 2\diamondsuit, & 2\spadesuit \ \} \end{array}$$

As usual, the first step is to map this question to a sequence-counting problem. A hand with a Four-of-a-Kind is completely described by a sequence specifying:

1. The value of the four cards.
2. The value of the extra card.
3. The suit of the extra card.

Thus, there is a bijection between hands with a Four-of-a-Kind and sequences consisting of two distinct values followed by a suit. For example, the three hands above are associated with the following sequences:

$$\begin{array}{lcl} (8, Q, \heartsuit) & \leftrightarrow & \{\ 8\spadesuit,\ 8\diamondsuit,\ 8\heartsuit,\ 8\clubsuit,\ Q\heartsuit\ \} \\ (2, A, \clubsuit) & \leftrightarrow & \{\ 2\clubsuit,\ 2\heartsuit,\ 2\diamondsuit,\ 2\spadesuit,\ A\clubsuit\ \} \end{array}$$

Now we need only count the sequences. There are 13 ways to choose the first value, 12 ways to choose the second value, and 4 ways to choose the suit. Thus, by the Generalized Product Rule, there are $13 \cdot 12 \cdot 4 = 624$ hands with a Four-of-a-Kind. This means that only 1 hand in about 4165 has a Four-of-a-Kind; not surprisingly, this is considered a very good poker hand!

11.4.2 Hands with a Full House

A *Full House* is a hand with three cards of one value and two cards of another value. Here are some examples:

$$\begin{array}{l} \{\ 2\spadesuit,\ 2\clubsuit,\ 2\diamondsuit,\ J\clubsuit,\ J\diamondsuit\ \} \\ \{\ 5\diamondsuit,\ 5\clubsuit,\ 5\heartsuit,\ 7\heartsuit,\ 7\clubsuit\ \} \end{array}$$

Again, we shift to a problem about sequences. There is a bijection between Full Houses and sequences specifying:

1. The value of the triple, which can be chosen in 13 ways.
2. The suits of the triple, which can be selected in $\binom{4}{3}$ ways.
3. The value of the pair, which can be chosen in 12 ways.
4. The suits of the pair, which can be selected in $\binom{4}{2}$ ways.

The example hands correspond to sequences as shown below:

$$\begin{array}{lcl} (2, \{\spadesuit, \clubsuit, \diamondsuit\}, J, \{\clubsuit, \diamondsuit\}) & \leftrightarrow & \{\ 2\spadesuit,\ 2\clubsuit,\ 2\diamondsuit,\ J\clubsuit,\ J\diamondsuit\ \} \\ (5, \{\diamondsuit, \clubsuit, \heartsuit\}, 7, \{\heartsuit, \clubsuit\}) & \leftrightarrow & \{\ 5\diamondsuit,\ 5\clubsuit,\ 5\heartsuit,\ 7\heartsuit,\ 7\clubsuit\ \} \end{array}$$

By the Generalized Product Rule, the number of Full Houses is:

$$13 \cdot \binom{4}{3} \cdot 12 \cdot \binom{4}{2}$$

We're on a roll— but we're about to hit a speedbump.

11.4.3 Hands with Two Pairs

How many hands have *Two Pairs*; that is, two cards of one value, two cards of another value, and one card of a third value? Here are examples:

$$\begin{array}{l} \{\ 3\diamondsuit,\ 3\spadesuit,\ Q\diamondsuit,\ Q\heartsuit,\ A\clubsuit\ \} \\ \{\ 9\heartsuit,\ 9\diamondsuit,\ 5\heartsuit,\ 5\clubsuit,\ K\spadesuit\ \} \end{array}$$

Each hand with Two Pairs is described by a sequence consisting of:

1. The value of the first pair, which can be chosen in 13 ways.
2. The suits of the first pair, which can be selected $\binom{4}{2}$ ways.
3. The value of the second pair, which can be chosen in 12 ways.
4. The suits of the second pair, which can be selected in $\binom{4}{2}$ ways.
5. The value of the extra card, which can be chosen in 11 ways.
6. The suit of the extra card, which can be selected in $\binom{4}{1} = 4$ ways.

Thus, it might appear that the number of hands with Two Pairs is:

$$13 \cdot \binom{4}{2} \cdot 12 \cdot \binom{4}{2} \cdot 11 \cdot 4$$

Wrong answer! The problem is that there is *not* a bijection from such sequences to hands with Two Pairs. This is actually a 2-to-1 mapping. For example, here are the pairs of sequences that map to the hands given above:

$$\begin{array}{ll} (3, \{\diamondsuit, \spadesuit\}, Q, \{\diamondsuit, \heartsuit\}, A, \clubsuit) & \searrow \\ & \{\ 3\diamondsuit,\ \ 3\spadesuit,\ \ Q\diamondsuit,\ \ Q\heartsuit,\ \ A\clubsuit\ \ \} \\ (Q, \{\diamondsuit, \heartsuit\}, 3, \{\diamondsuit, \spadesuit\}, A, \clubsuit) & \nearrow \\ \\ (9, \{\heartsuit, \diamondsuit\}, 5, \{\heartsuit, \clubsuit\}, K, \spadesuit) & \searrow \\ & \{\ 9\heartsuit,\ \ 9\diamondsuit,\ \ 5\heartsuit,\ \ 5\clubsuit,\ \ K\spadesuit\ \ \} \\ (5, \{\heartsuit, \clubsuit\}, 9, \{\heartsuit, \diamondsuit\}, K, \spadesuit) & \nearrow \end{array}$$

The problem is that nothing distinguishes the first pair from the second. A pair of 5's and a pair of 9's is the same as a pair of 9's and a pair of 5's. We avoided this difficulty in counting Full Houses because, for example, a pair of 6's and a triple of kings is different from a pair of kings and a triple of 6's.

We ran into precisely this difficulty last time, when we went from counting arrangements of *different* pieces on a chessboard to counting arrangements of two *identical* rooks. The solution then was to apply the Division Rule, and we can do the same here. In this case, the Division rule says there are twice as many sequences as hands, so the number of hands with Two Pairs is actually:

$$\frac{13 \cdot \binom{4}{2} \cdot 12 \cdot \binom{4}{2} \cdot 11 \cdot 4}{2}$$

Another Approach

The preceding example was disturbing! One could easily overlook the fact that the mapping was 2-to-1 on an exam, fail the course, and turn to a life of crime. You can make the world a safer place in two ways:

1. Whenever you use a mapping $f : A \to B$ to translate one counting problem to another, check the number of elements in A that are mapped to each element in B. This determines the size of A relative to B. You can then apply the Division Rule with the appropriate correction factor.

2. As an extra check, try solving the same problem in a different way. Multiple approaches are often available— and all had better give the same answer! (Sometimes different approaches give answers that *look* different, but turn out to be the same after some algebra.)

We already used the first method; let's try the second. There is a bijection between hands with two pairs and sequences that specify:

1. The values of the two pairs, which can be chosen in $\binom{13}{2}$ ways.
2. The suits of the lower-value pair, which can be selected in $\binom{4}{2}$ ways.
3. The suits of the higher-value pair, which can be selected in $\binom{4}{2}$ ways.
4. The value of the extra card, which can be chosen in 11 ways.
5. The suit of the extra card, which can be selected in $\binom{4}{1} = 4$ ways.

For example, the following sequences and hands correspond:

$$\begin{array}{rcl} (\{3, Q\}, \{\diamondsuit, \spadesuit\}, \{\diamondsuit, \heartsuit\}, A, \clubsuit) & \leftrightarrow & \{\ 3\diamondsuit,\ \ 3\spadesuit,\ \ Q\diamondsuit,\ \ Q\heartsuit,\ \ A\clubsuit\ \} \\ (\{9, 5\}, \{\heartsuit, \clubsuit\}, \{\heartsuit, \diamondsuit\}, K, \spadesuit) & \leftrightarrow & \{\ 9\heartsuit,\ \ 9\diamondsuit,\ \ 5\heartsuit,\ \ 5\clubsuit,\ \ K\spadesuit\ \} \end{array}$$

Thus, the number of hands with two pairs is:

$$\binom{13}{2} \cdot \binom{4}{2} \cdot \binom{4}{2} \cdot 11 \cdot 4$$

This is the same answer we got before, though in a slightly different form.

11.4.4 Hands with Every Suit

How many hands contain at least one card from every suit? Here is an example of such a hand:

$$\{\ \ 7\diamondsuit,\ \ K\clubsuit,\ \ 3\diamondsuit,\ \ A\heartsuit,\ \ 2\spadesuit\ \ \}$$

Each such hand is described by a sequence that specifies:

1. The values of the diamond, the club, the heart, and the spade, which can be selected in $13 \cdot 13 \cdot 13 \cdot 13 = 13^4$ ways.
2. The suit of the extra card, which can be selected in 4 ways.
3. The value of the extra card, which can be selected in 12 ways.

For example, the hand above is described by the sequence:

$$(7, K, A, 2, \diamondsuit, 3) \quad \leftrightarrow \quad \{\ \ 7\diamondsuit,\ \ K\clubsuit,\ \ A\heartsuit,\ \ 2\spadesuit,\ \ 3\diamondsuit\ \ \}$$

Are there other sequences that correspond to the same hand? There is one more! We could equally well regard either the $3\diamondsuit$ or the $7\diamondsuit$ as the extra card, so this is actually a 2-to-1 mapping. Here are the two sequences corresponding to the example hand:

$$\begin{array}{ll} (7, K, A, 2, \diamondsuit, 3) & \searrow \\ & \{\ 7\diamondsuit,\ \ K\clubsuit,\ \ A\heartsuit,\ \ 2\spadesuit,\ \ 3\diamondsuit\ \ \} \\ (3, K, A, 2, \diamondsuit, 7) & \nearrow \end{array}$$

Therefore, the number of hands with every suit is:

$$\frac{13^4 \cdot 4 \cdot 12}{2}$$

11.5 Binomial Theorem

Counting gives insight into one of the basic theorems of algebra. A ***binomial*** is a sum of two terms, such as $a + b$. Now consider its 4th power, $(a + b)^4$.

If we multiply out this 4th power expression completely, we get

$$\begin{array}{rcccccccc} (a+b)^4 & = & aaaa & + & aaab & + & aaba & + & aabb \\ & + & abaa & + & abab & + & abba & + & abbb \\ & + & baaa & + & baab & + & baba & + & babb \\ & + & bbaa & + & bbab & + & bbba & + & bbbb \end{array}$$

Notice that there is one term for every sequence of a's and b's. So there are 2^4 terms, and the number of terms with k copies of b and $n - k$ copies of a is:

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

by the Bookkeeper Rule. Now let's group equivalent terms, such as $aaab = aaba = abaa = baaa$. Then the coefficient of $a^{n-k}b^k$ is $\binom{n}{k}$. So for $n = 4$, this means:

$$(a+b)^4 = \binom{4}{0} \cdot a^4b^0 + \binom{4}{1} \cdot a^3b^1 + \binom{4}{2} \cdot a^2b^2 + \binom{4}{3} \cdot a^1b^3 + \binom{4}{4} \cdot a^0b^4$$

In general, this reasoning gives the Binomial Theorem:

Theorem 11.5.1 (Binomial Theorem). *For all $n \in \mathbb{N}$ and $a, b \in \mathbb{R}$:*

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k}b^k$$

The expression $\binom{n}{k}$ is often called a "binomial coefficient" in honor of its appearance here.

This reasoning about binomials extends nicely to ***multinomials***, which are sums of two or more terms. For example, suppose we wanted the coefficient of

$$bo^2k^2e^3pr$$

in the expansion of $(b+o+k+e+p+r)^{10}$. Each term in this expansion is a product of 10 variables where each variable is one of b, o, k, e, p, or r. Now, the coefficient of $bo^2k^2e^3pr$ is the number of those terms with exactly 1 b, 2 o's, 2 k's, 3 e's, 1 p, and 1 r. And the number of such terms is precisely the number of rearrangments of the word BOOKKEEPER:

$$\binom{10}{1,2,2,3,1,1} = \frac{10!}{1!\ 2!\ 2!\ 3!\ 1!\ 1!}.$$

The expression on the left is called a "multinomial coefficient." This reasoning extends to a general theorem:

Theorem 11.5.2 (Multinomial Theorem). *For all $n \in \mathbb{N}$ and $z_1, \ldots z_m \in \mathbb{R}$:*

$$(z_1 + z_2 + \cdots + z_m)^n = \sum_{\substack{k_1,\ldots,k_m \in \mathbb{N} \\ k_1+\cdots+k_m=n}} \binom{n}{k_1, k_2, \ldots, k_m} z_1^{k_1} z_2^{k_2} \cdots z_m^{k_m}$$

You'll be better off remembering the reasoning behind the Multinomial Theorem rather than this ugly formal statement.

11.6 Combinatorial Proof

Suppose you have n different T-shirts, but only want to keep k. You could equally well select the k shirts you want to keep or select the complementary set of $n-k$ shirts you want to throw out. Thus, the number of ways to select k shirts from among n must be equal to the number of ways to select $n-k$ shirts from among n. Therefore:

$$\binom{n}{k} = \binom{n}{n-k}$$

This is easy to prove algebraically, since both sides are equal to:

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

But we didn't really have to resort to algebra; we just used counting principles.

Hmm.

11.6.1 Boxing

Jay, famed 6.042 TA, has decided to try out for the US Olympic boxing team. After all, he's watched all of the *Rocky* movies and spent hours in front of a mirror sneering, "Yo, you wanna piece a' *me*?!" Jay figures that n people (including himself) are competing for spots on the team and only k will be selected. As part of maneuvering for a spot on the team, he needs to work out how many different teams are possible. There are two cases to consider:

- Jay *is* selected for the team, and his $k-1$ teammates are selected from among the other $n-1$ competitors. The number of different teams that can be formed in this way is:

$$\binom{n-1}{k-1}$$

- Jay is *not* selected for the team, and all k team members are selected from among the other $n-1$ competitors. The number of teams that can be formed this way is:

$$\binom{n-1}{k}$$

All teams of the first type contain Jay, and no team of the second type does; therefore, the two sets of teams are disjoint. Thus, by the Sum Rule, the total number of possible Olympic boxing teams is:

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

Chiyoun, equally-famed 6.042 TA, thinks Jay isn't so tough and so he might as well also try out. He reasons that n people (including himself) are trying out for k spots. Thus, the number of ways to select the team is simply:

$$\binom{n}{k}$$

Chiyoun and Jay each correctly counted the number of possible boxing teams; thus, their answers must be equal. So we know:

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

This is called ***Pascal's Identity***. And we proved it *without any algebra!* Instead, we relied purely on counting techniques.

11.6.2 Finding a Combinatorial Proof

A ***combinatorial proof*** is an argument that establishes an algebraic fact by relying on counting principles. Many such proofs follow the same basic outline:

1. Define a set S.
2. Show that $|S| = n$ by counting one way.
3. Show that $|S| = m$ by counting another way.
4. Conclude that $n = m$.

In the preceding example, S was the set of all possible Olympic boxing teams. Jay computed

$$|S| = \binom{n-1}{k-1} + \binom{n-1}{k}$$

by counting one way, and Chiyoun computed

$$|S| = \binom{n}{k}$$

by counting another. Equating these two expressions gave Pascal's Identity.

More typically, the set S is defined in terms of simple sequences or sets rather than an elaborate story. Here is less colorful example of a combinatorial argument.

Theorem 11.6.1.

$$\sum_{r=0}^{n} \binom{n}{r}\binom{2n}{n-r} = \binom{3n}{n}$$

Proof. We give a combinatorial proof. Let S be all n-card hands that can be dealt from a deck containing n red cards (numbered $1, \ldots, n$) and $2n$ black cards (numbered $1, \ldots, 2n$). First, note that every $3n$-element set has

$$|S| = \binom{3n}{n}$$

n-element subsets.

From another perspective, the number of hands with exactly r red cards is

$$\binom{n}{r}\binom{2n}{n-r}$$

since there are $\binom{n}{r}$ ways to choose the r red cards and $\binom{2n}{n-r}$ ways to choose the $n-r$ black cards. Since the number of red cards can be anywhere from 0 to n, the total number of n-card hands is:

$$|S| = \sum_{r=0}^{n} \binom{n}{r}\binom{2n}{n-r}$$

Equating these two expressions for $|S|$ proves the theorem. □

Combinatorial proofs are almost magical. Theorem 11.6.1 looks pretty scary, but we proved it without any algebraic manipulations at all. The key to constructing a combinatorial proof is choosing the set S properly, which can be tricky. Generally, the simpler side of the equation should provide some guidance. For example, the right side of Theorem 11.6.1 is $\binom{3n}{n}$, which suggests choosing S to be all n-element subsets of some $3n$-element set.

11.7 In-Class Problems Week 10, Fri.

Problem 11.7.1. **(a)** Show that the Magician could not pull off the trick with a deck larger than 124 cards.

Hint: Compare the number of 5-card hands in an n-card deck with the number of 4-card sequences.

Solution. For a match to be possible with a n-card deck, the number, $\binom{n}{5}$, of 5-card hands must be at most as large as the number, $(n)_4$, of 4-card sequences. So

$$(n)_4(n-4)/5! = \binom{n}{5} \leq (n)_4,$$

which implies

$$n - 4 \leq 5!$$

and hence $n \leq 124$. ■

(b) Show that, in principle, the Magician could pull off the Card Trick with a deck of 124 cards.

Hint: Hall's Theorem and degree-constrained graphs.

Solution. In principle the trick is possible iff the bipartite graph between 5-card hands and 4-card sequences has a matching for the hands. In this graph, the degree of each hand is $5! = 120$, whatever the size of deck. The degree of each sequence of 4 will be the number of cards remaining in the deck. With a deck of 124, there will be 120 cards remaining, so the degree of each sequence of 4 will also be 120. Hence, the graph is degree-constrained, and so satisfies Hall's condition for a matching. ■

Problem 11.7.2. The Tao of BOOKKEEPER: we seek enlightenment through contemplation of the word $BOOKKEEPER$.

(a) In how many ways can you arrange the letters in the word $POKE$?

Solution. There are $4!$ arrangements corresponding to the $4!$ permutations of the set $\{P, O, K, E\}$. ■

(b) In how many ways can you arrange the letters in the word BO_1O_2K? Observe that we have subscripted the O's to make them distinct symbols.

Solution. There are $4!$ arrangements corresponding to the $4!$ permutations of the set $\{B, O_1, O_2, K\}$. ■

(c) Suppose we map arrangements of the letters in BO_1O_2K to arrangements of the letters in $BOOK$ by erasing the subscripts. Indicate with arrows how the arrangements on the left are mapped to the arrangements on the right.

O_2BO_1K
KO_2BO_1
O_1BO_2K
KO_1BO_2
BO_1O_2K
BO_2O_1K
...

$BOOK$
$OBOK$
$KOBO$
...

(d) What kind of mapping is this, young grasshopper?

Solution. 2-to-1 ■

(e) In light of the Division Rule, how many arrangements are there of $BOOK$?

Solution. $4!/2$ ■

(f) Very good, young master! How many arrangements are there of the letters in $KE_1E_2PE_3R$?

Solution. $6!$ ■

(g) Suppose we map each arrangement of $KE_1E_2PE_3R$ to an arrangement of $KEEPER$ by erasing subscripts. List all the different arrangements of $KE_1E_2PE_3R$ that are mapped to $REPEEK$ in this way.

Solution. $RE_1PE_2E_3K$, $RE_1PE_3E_2K$, $RE_2PE_1E_3K$, $RE_2PE_3E_1K$, $RE_3PE_1E_2K$, $RE_3PE_2E_1K$ ■

(h) What kind of mapping is this?

Solution. 3!-to-1 ■

(i) So how many arrangements are there of the letters in $KEEPER$?

Solution. $6!/3!$ ■

(j) *Now you are ready to face the BOOKKEEPER!*
How many arrangements of $BO_1O_2K_1K_2E_1E_2PE_3R$ are there?

Solution. $10!$ ■

(k) How many arrangements of $BOOK_1K_2E_1E_2PE_3R$ are there?

Solution. $10!/2!$ ■

(l) How many arrangements of $BOOKKE_1E_2PE_3R$ are there?

Solution. $10!/(2! \cdot 2!)$ ■

(m) How many arrangements of $BOOKKEEPER$ are there?

Solution. $10!/(2! \cdot 2! \cdot 3!)$ ■

(n) How many arrangements of $VOODOODOLL$ are there?

Solution. $10!/(2! \cdot 2! \cdot 5!)$ ■

(o) (IMPORTANT) How many n-bit sequences contain k zeros and $(n-k)$ ones?

Solution. $\binom{n}{k}$ ■

Remember well what you have learned: subscripts on, subscripts off.

This is the Tao of Bookkeeper.

Problem 11.7.3. Solve the following counting problems. Define an appropriate mapping (bijective or k-to-1) between a set whose size you know and the set in question.

(a) How many different ways are there to select a dozen donuts if four varieties are available?

Solution. There is a bijection from selections of a dozen donuts to 15-bit sequences with exactly 3 ones. In particular, suppose that the varieties are glazed, chocolate, lemon, and Boston creme. Then a selection of g glazed, c chocolate, l lemon, and b Boston creme maps to the sequence:

$$(g\ 0's)\ 1\ (c\ 0's)\ 1\ (l\ 0's)\ 1\ (b\ 0's)$$

Therefore, the number of selections is equal to the number of 15-bit sequences with exactly 3 ones, which is:

$$\frac{15!}{3!\ 12!} = \binom{15}{3}$$

■

(b) How many paths are there from $(0,0)$ to $(10,20)$ consisting of right-steps (which increment the first coordinate) and up-steps (which increment the second coordinate)?

Solution. There is a bijection from 30-bit sequences with 10 zeros and 20 ones. The sequence $(b_1, \ldots, b_{30})$ maps to a path where the i-th step is right if $b_i = 0$ and up if $b_i = 1$. Therefore, the number of paths is equal to $\binom{30}{10}$. ■

(c) An independent living group is hosting nine new candidates for membership. Each candidate must be assigned a task: 1 must wash pots, 2 must clean the kitchen, 3 must clean the bathrooms, 1 must clean the common area, and 2 must serve dinner. In how many ways can this be done?

Solution. There is a bijection from sequences containing one P, two K's, three B's, a C, and two D's. In any such sequence, the letter in the ith position specifies the task assigned to the ith candidate. Therefore, the number of possible assignments is:

$$\frac{9!}{1!\ 2!\ 3!\ 1!\ 2!}$$

■

(d) In how many ways can Mr. and Mrs. Grumperson distribute 13 identical pieces of coal to their two— no, three!— children for Christmas?

Solution. There is a bijection from 15-bit strings with two ones. In particular, the bit string $0^a10^b10^c$ maps to the assignment of a coals to the first child, b coals to the second, and c coals to the third. Therefore, there are $\binom{15}{2}$ assignments. ■

(e) How many solutions over the natural numbers are there to the equation:

$$x_1 + x_2 + \ldots + x_{10} \leq 100$$

Solution. There is a bijection from 110-bit sequences with 10 ones to solutions to this equation. In particular, x_i is the number of zeros before the i-th one but after the $(i-1)$-st one (or the beginning of the sequence). Therefore, there are $\binom{110}{10}$ solutions. ■

(f) Suppose that two identical 52-card decks are mixed together. In how many ways can the cards in this double-size deck be arranged?

Solution. The number of sequences of the 104 cards containing 2 of each card is $104!/(2!)^{52}$. ■

11.8 In-Class Problems Week 11, Mon.

Problem 11.8.1. Find the coefficients of

(a) x^5 in $(1+x)^{11}$

Solution.

$$\binom{11}{5} = 462$$

■

(b) x^8y^9 in $(3x+2y)^{17}$

Solution.

$$\binom{17}{8}3^8 2^9$$

■

(c) a^6b^6 in $(a^2+b^3)^5$

Solution. $a^6b^6 = (a^2)^3(b^3)^2$, so the coefficient is

$$\binom{5}{3} = 10$$

■

Problem 11.8.2. According to the Multinomial theorem, $(w+x+y+z)^n$ can be expressed as a sum of terms of the form

$$\binom{n}{r_1, r_2, r_3, r_4} w^{r_1}x^{r_2}y^{r_3}z^{r_4}.$$

How many terms are there in the sum?

Solution. The sum is over all 4-tuples of nonnegative integers (r_1, r_2, r_3, r_4) such that

$$r_1 + r_2 + r_3 + r_4 = n.$$

We know this is the same as the number of binary words with n zeroes and 3 ones, namely

$$\binom{n+3}{3}.$$

■

Combinatorial proof

Combinatorial proofs of identities

Recall the basic plan for a combinatorial proof of an identity $x = y$:

1. Define a set S.
2. Show that $|S| = x$ by counting one way.
3. Show that $|S| = y$ by counting another way.
4. Conclude that $x = y$.

Problem 11.8.3. You want to choose a team of m people from a pool of n people for your startup company, and from these m people you want to choose k to be the team managers. You took 6.042, so you know you can do this in

$$\binom{n}{m}\binom{m}{k}$$

ways. But your CFO, who went to Harvard Business School, comes up with the formula

$$\binom{n}{k}\binom{n-k}{m-k}.$$

Before doing the reasonable thing —dump on your CFO or Harvard Business School— you decide to check his answer against yours.

(a) Start by giving an *algebraic proof* that your CFO's formula agrees with yours.

Solution.

$$\begin{aligned}
\binom{n}{m}\binom{m}{k} &= \frac{n!}{m!(n-m)!}\frac{m!}{k!(m-k)!} \\
&= \frac{n!}{(n-m)!k!(m-k)!} \\
&= \frac{n!(n-k)!}{(n-m)!k!(m-k)!(n-k)!} \\
&= \frac{n!}{k!(n-k)!}\frac{(n-k)!}{(n-m)!(m-k)!} \\
&= \frac{n!}{k!(n-k)!}\frac{(n-k)!}{((n-k)-(m-k))!(m-k)!} \\
&= \binom{n}{k}\binom{n-k}{m-k}.
\end{aligned}$$

■

(b) Now give a *combinatorial argument* proving this same fact.

Solution. Instead of choosing first m from n and then k from m, you could alternately choose the k managers from the n people and then choose $m-k$ people to fill out the team from the remaining $n-k$ people. This gives you $\binom{n}{k}\binom{n-k}{m-k}$ ways of picking your team. Since you must have the same number of options regardless of the order in which you choose to pick team members and managers,

$$\binom{n}{m}\binom{m}{k} = \binom{n}{k}\binom{n-k}{m-k}.$$

■

Problem 11.8.4. Now give a combinatorial proof of the following, more interesting theorem:

$$n2^{n-1} = \sum_{k=1}^{n} k\binom{n}{k}$$

Hint: Let S be the set of all length-n sequences of 0's, 1's and a single *.

Solution. Let $P ::= \{0,\ldots,n-1\} \times \{0,1\}^{n-1}$. On the one hand, there is a bijection from P to S by mapping (k,x) to the word obtained by inserting a * just after the kth bit in the length-$n-1$ binary word, x. So

$$|S| = |P| = n2^{n-1} \tag{11.1}$$

by the Product Rule.

On the other hand, every sequence in S contains between 1 and n nonzero entries since the $*$, at least, is nonzero. The mapping from a sequence in S with exactly k nonzero entries to a pair consisting of the set of positions of the nonzero entries and the position of the * among these entries is a bijection, and the number of such pairs is $\binom{n}{k}k$ by the Generalized Product Rule. Thus, by the Sum Rule:

$$|S| = \sum_{k=1}^{n} k\binom{n}{k}$$

Equating this expression and the expression (11.1) for $|S|$ proves the theorem. ■

Problem 11.8.5. What do the following expressions equal? Give both algebraic and combinatorial proofs for your answers.

(a)

$$\sum_{i=0}^{n} \binom{n}{i}$$

Solution. 2^n.

Algebraic proof: This is the Binomial theorem with $x = y = 1$.

Combinatorial proof: There are 2^n length n bit strings. The number of such sequences is also equal to the number of length n bit strings with 0 ones, plus the number with 1 one, plus the number with 2 ones, etc., which is precisely $\sum_{i=0}^{n} \binom{n}{i}$. ■

(b)

$$\sum_{i=0}^{n} \binom{n}{i} (-1)^i$$

Hint: Consider the bit strings with an even number of ones and an odd number of ones.

Solution. 0.

Algebraic proof: This is just the Binomial theorem with $x = 1$ and $y = -1$.

Combinatorial proof: Consider the $n-$bit sequences, and divide them into two sets, those with an even number of ones (even i terms) and those with an odd number of ones (odd i terms). The sum is then equal to the number of strings with an even number of ones, minus the number of strings with an odd number of ones.

Next, we note that the number of strings with an even number of ones is equal to the number with an odd number of ones. This can be seen by establishing a bijection between the two sets: any string in one set can be made into a string in the other set by complementing the first bit in the string. Since the number of strings with an even number of ones is equal to the number with an odd number, the entire expression must be equal to 0. ■

11.9 Problem Set 9

Problem 11.9.1. **(a)** Jellybeans of 6 different flavors are stored in 5 jars. There are 11 jellybeans of each flavor. Prove that some jar contains at least three jellybeans of one flavor and also at least three jellybeans of some other flavor.

Solution. We use the pigeonhole principle twice. Since there are 11 beans of a given flavor and there are only 5 jars, some jar must get at least $\lceil 11/5 \rceil = 3$ beans of that flavor. Since there are 6 flavors and only 5 jars, some jar must get two triples of same-flavored beans. ■

(b) Prove that every finite undirected graph has two vertices of the same degree.

Solution. If G is a graph with n vertices, the possible degrees of its vertices are $0, 1, ..., n-1$. If the graph contains a vertex of degree $n-1$, it is connected to all other vertices, meaning that a vertex of degree 0 cannot exist. So there are $n-1$ possible values for the degrees. If the graph does not have a vertex of degree $n-1$, there are $n-1$ possible values for the degrees. In either case, we have n vertices and $n-1$ possible degree values. By the Pigeonhole Principle, two of the vertices must have the same degree. ■

Problem 11.9.2. There is a robot that steps between integer positions in 3-dimensional space. Each step of the robot increments one coordinate and leaves the other two unchanged.

(a) Let m, n be nonnegative integers. Describe a bijection between the length-$(m+n)$ binary strings with exactly m ones and the paths the robot can take to go from $(0, 0, 0)$ to $(m, n, 0)$. How many such paths are there?

Solution. The bijection is defined by letting the ith bit of the binary string corresponding to a path be 1 iff the ith step in the path increments the 1st coordinate. The number of such strings is

$$\binom{m+n}{m}$$

so this is also the number of paths. ■

(b) How many different paths are there from point $(0, 0, 0)$ to $(12, 24, 36)$?

Solution. There is a bijection between the set of all such paths and the set of strings containing 12 X's, 24 Y's, and 36 Z's. In particular, we obtain a path by working through a string from left to right. An X corresponds to a step that increments the first coordinate, a Y increments the second coordinate, and a Z increments the third. The number of such strings is:

$$\frac{72!}{12!\ 24!\ 36!}$$

Therefore, this is also the number of paths. ■

Problem 11.9.3. Suppose you have seven dice— each a different color of the rainbow; otherwise the dice are standard, with six faces numbered 1 to 6. A *roll* is a sequence specifying a value for each die in rainbow (ROYGBIV) order. For example, one roll is $(3, 1, 6, 1, 4, 5, 2)$ indicating that the red die showed a 3, the orange die showed 1, the yellow 6, the green 1, the blue 4, the indigo 5, and the violet 2.

For the problems below, describe a bijection between the specified set of rolls and another set that is easily counted using the Product, Generalized Product, and similar rules. Then write a simple numerical expression for the size of the set of rolls. You do not need to prove that the correspondence between sets you describe is a bijection, and you do not need to simplify the expression you come up with.

For example, let A be the set of rolls where 4 dice come up showing the same number, and the other 3 dice also come up the same, but with a different number. Let R be the set of seven rainbow colors and S be the set $\{1, \ldots, 6\}$ of dice values.

Define $B ::= S_2 \times \{3, 4\} \times R_3$, where S_2 is the set of size 2 subsets of S, and R_3 is the set of size 3 subsets of R. Then define a bijection from A to B by mapping a roll in A to the sequence in B whose first element is the set of two numbers that came up, whose second element is the number of times the smaller of the two numbers came up in the roll, and whose third element is the set of colors of the three matching dice.

For example, the roll

$$(4, 4, 2, 2, 4, 2, 4) \in A$$

maps to the triple

$$(\{2, 4\}, 3, \{\text{yellow,green,indigo}\}) \in B.$$

Now by the Bijection rule $|A| = |B|$, and by the Product rule,

$$|B| = \binom{6}{2} \cdot 2 \cdot \binom{7}{3}.$$

(a) For how many rolls is the value on every die different?

Solution. None, by the Pigeonhole Principle. ■

(b) For how many rolls do two dice have the value 6 and the remaining five dice all have different values?

Example: $(6, 2, 6, 1, 3, 4, 5)$ is a roll of this type, but $(1, 1, 2, 6, 3, 4, 5)$ and $(6, 6, 1, 2, 4, 3, 4)$ are not.

Solution. As in the example, map a roll into an element of $B ::= R_2 \times P_5$ where P_5 is the set of permutations of $\{1, \ldots, 5\}$. A roll maps to the pair whose first element is the set of colors of the two dice with value 6, and whose second element is the sequence of values of the remaining dice (in rainbow order). So $(6, 2, 6, 1, 3, 4, 5)$ above maps to $(\{\text{red,yellow}\}, (2, 1, 3, 4, 5))$. By the Product rule,

$$|B| = \binom{7}{2} \cdot 5!.$$

■

(c) For how many rolls do two dice have the same value and the remaining five dice all have different values?

Example: $(4, 2, 4, 1, 3, 6, 5)$ is a roll of this type, but $(1, 1, 2, 6, 1, 4, 5)$ and $(6, 6, 1, 2, 4, 3, 4)$ are not.

Solution. Map a roll into a triple whose first element is in S, indicating the value of the pair of matching dice, whose second element is the set of colors of the two matching dice, and whose third element is the sequence of the remaining five dice values (in rainbow order).

So $(4, 2, 4, 1, 3, 6, 5)$ above maps to $(4, \{\text{red,yellow}\}, (2, 1, 3, 6, 5))$. Notice that the number of choices for the third element of a triple is the number of permutations of the remaining five values, namely $5!$. This mapping is a bijection, so the number of such rolls equals the number of such triples. By the Generalized Product rule, the number of such triples is

$$6 \cdot \binom{7}{2} \cdot 5!.$$

Alternatively, we can define a map from rolls in this part to the rolls in part (b), by replacing the value of the duplicated values with 6's and replacing any 6 in the remaining values by the value of the duplicated pair. So the roll $(4, 2, 4, 1, 3, 6, 5)$ would map to the roll $(6, 2, 6, 1, 3, 4, 5)$. Now a type b roll, r, is mapped to by exactly the rolls obtainable from r by exchanging occurrences of 6's and i's, for $i = 1, \ldots, 6$. So this map is 6-to-1, and by the Division rule, the number of rolls here is 6 times the number of rolls in part (b).

■

(d) For how many rolls do two dice have one value, two different dice have a second value, and the remaining three dice a third value?

Example: $(6, 1, 2, 1, 2, 6, 6)$ is a roll of this type, but $(4, 4, 4, 4, 1, 3, 5)$ and $(5, 5, 5, 6, 6, 1, 2)$ are not.

Solution. Map a roll of this kind into a 4-tuple whose first element is the set of two numbers of the two pairs of matching dice, whose second element is the set of two colors of the pair of matching dice with the smaller number, whose third element is the set of two colors of the larger of the matching pairs, and whose fourth element is the value of the remaining three dice. For example, the roll $(6, 1, 2, 1, 2, 6, 6)$ maps to the triple $(\{1, 2\}, \{\text{orange,green}\}, \{\text{yellow,blue}\}, 6)$.

There are $\binom{6}{2}$ possible first elements of a triple, $\binom{7}{2}$ second elements, $\binom{5}{2}$ third elements since the second set of two colors must be different from the first two, and 4 ways to choose the value of the three dice since their value must differ from the values of the two pairs. So by the Generalized Product rule, there are

$$\binom{6}{2} \cdot \binom{7}{2} \cdot \binom{5}{2} \cdot 4$$

possible rolls of this kind. ■

Problem 11.9.4. Answer the following questions with a number or a simple formula involving factorials and binomial coefficients. Briefly explain your answers.

(a) How many ways are there to order the 26 letters of the alphabet so that no two of the vowels `a`, `e`, `i`, `o`, `u` appear consecutively and the last letter in the ordering is not a vowel?

Hint: Every vowel appears to the left of a consonant.

Solution. The constraint on where vowels can appear is equivalent to the requirement that every vowel appears to the left of a consonant. So given a sequence of the 21 consonants, there are $\binom{21}{5}$ positions where the 5 vowels can be placed. After determining such a placement, we can reorder the consonants and vowels in any order. Thus, the number is:

$$\binom{21}{5} \cdot 21! \cdot 5!.$$

■

(b) In how many different ways can the letters in the name of the popular 1980's band $BANANARAMA$ be arranged?

Solution. There are 5 A's, 2 N's, 1 B, 1 R, and 1 M. Therefore, the number of arrangements is

$$\frac{10!}{5!\ 2!\ 1!\ 1!\ 1!}$$

by the Bookkeeper Rule. ■

(c) In how many different ways can $2n$ students be paired up?

Solution. Pair up students by the following procedure. Line up the students and pair the first and second, the third and fourth, the fifth and sixth, etc. The students can be lined up in $(2n)!$ ways. However, this overcounts by a factor of 2^n, because we would get the same pairing if the first and second students were swapped, the third and fourth were swapped, etc. Furthermore, we are still overcounting by a factor of $n!$, because we would get the same pairing even if pairs of students were permuted, e.g. the first and second were swapped with the ninth and tenth. Therefore, the number of pairings is:

$$\frac{(2n)!}{2^n \cdot n!}$$

■

(d) How many different solutions over the natural numbers are there to the following equation?

$$x_1 + x_2 + x_3 + \ldots + x_8 = 100$$

A solution is a specification of the value of each variable x_i. Two solutions are different if different values are specified for some variable x_i.

Solution. There is a bijection between sequences containing 100 zeros and 7 ones. Specifically, the 7 ones divide the zeros into 8 segments. Let x_i be the number of zeros in the i-th segment. Therefore, the number of solutions is:

$$\binom{100+7}{7}$$

■

(e) How many simple graphs are there with n vertices numbered $1, \ldots n$?

Solution. There are $\binom{n}{2}$ potential edges, each of which may or may not appear in a given graph. Therefore, the number of graphs is:

$$2^{\binom{n}{2}}$$

■

Problem 11.9.5. How many of the numbers $2, \ldots, n$ are prime? One way to answer this question is to test each number up to n for primality and keep a count. A somewhat more efficient method is to use the "Sieve of Erastosthenes" procedure which you may have learned about in 6.001 (but, don't worry, you needn't know about this). In this problem, we will use the Inclusion-Exclusion Principle to get the count; this approach turns out to be much more efficient when n is large.

Actually, we will use Inclusion-Exclusion to count the number of *composite* (nonprime) integers from 2 to n. Subtracting this from $n - 1$ gives the number of primes.

Let C_n be the set of composites from 2 to n, and let A_m be the set of numbers in the range $m + 1, \ldots, n$ that are divisible by m. Notice that by definition, $A_m = \emptyset$ for $m \geq n$. So

$$C_n = \bigcup_{i=2}^{n-1} A_i.$$

(a) Write C_n in terms of a union of A_p's, where p is prime. Explain why $\sqrt{n}$ is an upper bound on the largest p needed.

Solution. If $p \mid k$, then $A_p \supseteq A_k$, so there is no need to include A_k as long as A_p is included in the union. Also, any composite $\leq n$ must be divisible by a prime $\leq \sqrt{n}$ (because it is a product of at least two primes, and they can't both be bigger than $\sqrt{n}$). So we have

$$C_n = \bigcup_{p \leq \sqrt{n}} A_p.$$

■

(b) What is the cardinality of A_p?

Solution. $|A_p| = \lfloor n/p \rfloor - 1$. ■

(c) Let P be a set of primes. Give a simple formula for

$$\left| \bigcap_{p \in P} A_p \right|.$$

Solution. If $|P| = 1$, just look at the previous part.

Otherwise, if $|P| > 1$, let $m ::= \prod_{p \in P} p$. Then

$$\left| \bigcap_{p \in P} A_p \right| = \lfloor n/m \rfloor .$$

■

(d) Use the Inclusion-Exclusion principle to obtain a formula for $|C_{150}|$ in terms of nonempty intersections among the sets $A_2, A_3, A_5, A_7, A_{11}$.

Solution.

$$\begin{aligned} |C| \quad = \quad & |A_2| + |A_3| + |A_5| + |A_7| + |A_{11}| \\ & -|A_2 \cap A_3| - |A_2 \cap A_5| - |A_2 \cap A_7| - |A_2 \cap A_{11}| \\ & -|A_3 \cap A_5| - |A_3 \cap A_7| - |A_3 \cap A_{11}| \\ & -|A_5 \cap A_7| - |A_5 \cap A_{11}| \\ & -|A_7 \cap A_{11}| \\ & +|A_2 \cap A_3 \cap A_5| + |A_2 \cap A_3 \cap A_7| + |A_2 \cap A_3 \cap A_{11}| \\ & +|A_2 \cap A_5 \cap A_7| + |A_2 \cap A_5 \cap A_{11}| \\ & +|A_3 \cap A_5 \cap A_7| \end{aligned}$$

■

(e) Use this formula to find the number of primes up to 150.

Solution. We have:

$$\begin{aligned} |C_{150}| \quad = \quad & 74 + 49 + 29 + 20 + 12 \\ & -25 - 15 - 10 - 6 \\ & -10 - 7 - 4 \\ & -4 - 2 \\ & -1 \\ & +5 + 3 + 2 \\ & +2 + 1 \\ & +1 \\ = \quad & 114 \end{aligned}$$

The number of primes from 2 to 150 is $(150 - 1) - C_{150} = 149 - 114 = 35$. ■

Problem 11.9.6. Find the coefficients of

(a) x^{10} in $(x + (1/x))^{100}$

Solution. $x^{55}(1/x)^{45} = x^{10}$ so the coefficient is

$$\binom{100}{55}$$

■

(b) x^k in $(x^2 - (1/x))^n$.

Solution. Take $a = -1/x$ and $b = x^2$ in the Binomial theorem of the notes. We then have that:

$$(a+b)^n = \sum_{j=0}^{n} \binom{n}{j} a^{n-j} b^j = \sum_{j=0}^{n} \binom{n}{j} (\frac{-1}{x})^{n-j} x^{2j} = \sum_{j=0}^{n} \binom{n}{j} (-1)^{n-j} x^{-(n-j)+2j}$$

We know that x^k is $x^{-(n-j)+2j}$ for some j where $0 \leq j \leq n$, namely $j = (n+k)/3$. By the algebra above, we know that the coefficient for the term $x^{-(n-j)+2j}$ is $\binom{n}{j}(-1)^{n-j}$. Substituting $(n+k)/3$ for j, we get that the coefficient for x^k is:

$$\binom{n}{(n+k)/3}(-1)^{n-(n+k)/3} = \binom{n}{(n+k)/3}(-1)^{(2n-k)/3}$$

■

Problem 11.9.7. Suppose a generalized World Series between the Sox and the Cardinals involved $2n+1$ games. As usual, the generalized Series will stop as soon as one team has won more than half the possible games.

(a) Suppose that when the Sox finally win the GSeries, the Cards have managed to win *exactly* r games (so $r \leq n$). How many possible win-loss patterns are possible for the Sox to win the GSeries in this way? Express your answer as a binomial coefficient.

Solution.

$$\binom{n+r}{r} \tag{11.2}$$

Stars and bars, or better "S"'s and "C"'s: we can represent a win-loss pattern as a sequence of r C's and $n+1$ S's, where an S in the ith position indicates that the Sox won the ith game. However, the sequence must end with an S, so the number of such sequences is the same as the number of sequences of r C's and n S's, namely (11.2). ■

(b) How many possible win-loss patterns are possible for the Sox to win the GSeries when the Cards win *at most* r games? Express your answer as a binomial coefficient.

Solution.

$$\binom{n+r+1}{r} \tag{11.3}$$

We can represent a win-loss pattern as a sequence of r C's and $n+1$ S's, as in part (a). The number of C's which occur before the $n+1$st (last) S is the number of games the Cards won when the GSeries ends. ■

(c) Give a combinatorial proof that

$$\sum_{i=0}^{r} \binom{n+i}{i} = \binom{n+r+1}{r}. \tag{11.4}$$

Solution. The righthand side of (11.4) is the number of patterns where the Cards win at most r games. But they can win at most r by winning exactly i games, where $0 \leq i \leq r$. So by part (a), the number of win-loss patterns is given by the expression on the lefthand side of (11.4). ■

(d) Verify equation (11.4) by induction using algebra.

Solution. By induction on r, taking (11.4) as $P(r)$.

Proof. **Base case ($r = 0$):**

$$\binom{n}{0} = 1 = \binom{n+1}{0}.$$

Inductive step:

$$\begin{aligned}
\sum_{i=0}^{r+1} \binom{n+i}{i} &= \binom{n+r+1}{r+1} + \sum_{i=0}^{r} \binom{n+i}{i} \\
&= \binom{n+r+1}{r+1} + \binom{n+r+1}{r} && \text{(by Ind. Hyp.)} \\
&= \binom{n+(r+1)+1}{r+1} && \text{(Pascal's identity),}
\end{aligned}$$

Which proves $P(r+1)$. □

■

Problem 11.9.8. Below is a combinatorial proof of an equation. What is the equation?

Proof. Stinky Peterson owns n newts, t toads, and s slugs. Conveniently, he lives in a dorm with $n + t + s$ other students. (The students are distinguishable, but creatures of the same variety are not distinguishable.) Stinky wants to put one creature in each neighbor's bed. Let W be the set of all ways in which this can be done.

On one hand, he could first determine who gets the slugs. Then, he could decide who among his remaining neighbors has earned a toad. Therefore, $|W|$ is equal to the expression on the left.

On the other hand, Stinky could first decide which people deserve newts and slugs and then, from among those, determine who truly merits a newt. This shows that $|W|$ is equal to the expression on the right.

Since both expressions are equal to $|W|$, they must be equal to each other. □

(Combinatorial proofs are real proofs. They are not only rigorous, but also convey an intuitive understanding that a purely algebraic argument might not reveal. However, combinatorial proofs are usually less colorful than this one.)

Solution.

$$\binom{n+t+s}{s} \cdot \binom{n+t}{t} = \binom{n+t+s}{n+s} \cdot \binom{n+s}{n}$$

■

11.10 Miniquiz Apr. 27

Problem 11.10.1. Albert needs to pick a 10 digit numeric password containing *each* of the digits $0, 1, \ldots, 9$. He recognizes that it clearly should not contain either of the subwords "6042" or "18062". Just to be safe he decides that it should also not contain the subword "624" (his office number in the Gates Tower).

Use the Inclusion-Exclusion Principle to find a simple formula for the number of such passwords that do not contain "6042", "18062" or "624".

Solution. There are $7!$ passwords that contain "6042", $6!$ that contain "18062", and $8!$ that contain "624". No password can contain both "6042" and "18062". No password can contain both "6042" and "624". The only way a password can contain both "18062" and "624" is if it contains the subword "180624". There are $5!$ passwords that contain "180624".

By Inclusion-Exclusion, we have that the number of passwords containing at least one of the three forbidden subwords is

$$(7! + 6! + 8!) - (0 + 0 + 5!) + 0 = (7 + 1 + 8 \cdot 7)6! - 5! = 64 \cdot 6! - 5!.$$

Among the $10!$ passwords, the number of remaining ones that do not contain any of the forbidden subwords is

$$10! - (64)6! + 5! = 3,582,840.$$

■

Problem 11.10.2. For given positive integers i and m, how many different solutions[1] over the nonnegative integers are there to the following inequality?

$$x_1 + x_2 + x_3 + \ldots + x_m \leq i.$$

Your answer should be a simple formula (no indexed sums or three dots) which may involve factorials and binomial coefficients.

Solution.

$$\binom{m+i}{i}.$$

This follows because there is a bijection between sequences containing i zeros and m ones and solutions. Specifically, the m ones divide the zeros into $m + 1$ segments. In the corresponding solution, the value specified for x_k is the number of zeros in the kth segment, for $1 \leq k \leq m$. ■

[1] A ***solution*** is a specification of m nonnegative integer values for the successive variables $x_1, x_2, \ldots, x_m$. Two solutions are different if they specify different values for some variable.

Problem 11.10.3. Provide the combinatorial identity implied by the following:

Suppose it's 4:30pm Friday and you must choose a team of m people with 1 leader among them from a pool of n people.

Your first inclination is to select m people and let them pick the leader. Then you remember how much time it will take to do your hair for your hot date at 7pm, so instead you choose the leader from the pool and let that person choose the rest of the team after you leave.

You know the same teams could potentially result from either method and so the numbers of ways to choose teams either way are equal. What is the identity?

Solution.

$$\binom{n}{m}\binom{m}{1} = \binom{n}{1}\binom{n-1}{m-1}.$$

■

Problem 11.10.4. Find the coefficient of x^{15} in $(2x^2 - x)^{11}$.

Solution. Since $(2x^2 - x)^{11} = x^{11}(2x - 1)^{11}$, the coefficient of x^{15} in $(2x^2 - x)^{11}$ is equal to the coefficient of x^{15-11} in $(2x - 1)^{11}$, which is $\binom{11}{4}2^4(-1)^{11-4} = -5,280$. ■

Appendix

Rule (Inclusion-Exclusion for Three Sets).

$$\begin{aligned}|S_1 \cup S_2 \cup S_3| = {} & |S_1| + |S_2| + |S_3| \\ & - |S_1 \cap S_2| - |S_1 \cap S_3| - |S_2 \cap S_3| \\ & + |S_1 \cap S_2 \cap S_3|\end{aligned}$$

Theorem 11.10.1 (Binomial Theorem). *For all $n \in \mathbb{N}$ and $a, b \in \mathbb{R}$:*

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$$

Chapter 12

Generating Functions

Generating functions are one of the most surprising and useful inventions in Discrete Math. Roughly speaking, generating functions transform problems about *sequences* into problems about *functions*. This is great because we've got piles of mathematical machinery for manipulating functions. Thanks to generating functions, we can apply all that machinery to problems about sequences. In this way, we can use generating functions to solve all sorts of counting problems. There is a huge chunk of mathematics concerning generating functions, so we will only get a taste of the subject.

In these notes, we'll put sequences in angle brackets to more clearly distinguish them from the many other mathematical expressions floating around.

12.1 Generating Functions

The ***ordinary generating function*** for the infinite sequence $\langle g_0, g_1, g_2, g_3 \dots \rangle$ is the power series:

$$G(x) = g_0 + g_1 x + g_2 x^2 + g_3 x^3 + \cdots .$$

Not all generating functions are ordinary, but those are the only kind we'll consider here.

A generating function is a "formal" power series in the sense that we usually regard x as a placeholder rather than a number. Only in rare cases will we actually evaluate a generating function by letting x take a real number value, so we generally ignore the issue of convergence.

Throughout these notes, we'll indicate the correspondence between a sequence and its generating function with a double-sided arrow as follows:

$$\langle g_0, g_1, g_2, g_3, \dots \rangle \longleftrightarrow g_0 + g_1 x + g_2 x^2 + g_3 x^3 + \cdots$$

For example, here are some sequences and their generating functions:

$$\begin{aligned}
\langle 0, 0, 0, 0, \dots \rangle &\longleftrightarrow 0 + 0x + 0x^2 + 0x^3 + \cdots = 0 \\
\langle 1, 0, 0, 0, \dots \rangle &\longleftrightarrow 1 + 0x + 0x^2 + 0x^3 + \cdots = 1 \\
\langle 3, 2, 1, 0, \dots \rangle &\longleftrightarrow 3 + 2x + 1x^2 + 0x^3 + \cdots = 3 + 2x + x^2
\end{aligned}$$

The pattern here is simple: the ith term in the sequence (indexing from 0) is the coefficient of x^i in the generating function.

Recall that the sum of an infinite geometric series is:

$$1 + z + z^2 + z^3 + \cdots = \frac{1}{1-z}$$

This equation does not hold when $|z| \geq 1$, but as remarked, we don't worry about convergence issues. This formula gives closed-form generating functions for a whole range of sequences. For example:

$$\langle 1, 1, 1, 1, \ldots \rangle \quad \longleftrightarrow \quad 1 + x + x^2 + x^3 + \cdots \qquad = \frac{1}{1-x}$$

$$\langle 1, -1, 1, -1, \ldots \rangle \quad \longleftrightarrow \quad 1 - x + x^2 - x^3 + x^4 - \cdots \qquad = \frac{1}{1+x}$$

$$\langle 1, a, a^2, a^3, \ldots \rangle \quad \longleftrightarrow \quad 1 + ax + a^2x^2 + a^3x^3 + \cdots \qquad = \frac{1}{1-ax}$$

$$\langle 1, 0, 1, 0, 1, 0, \ldots \rangle \quad \longleftrightarrow \quad 1 + x^2 + x^4 + x^6 + \cdots \qquad = \frac{1}{1-x^2}$$

12.2 Operations on Generating Functions

The magic of generating functions is that we can carry out all sorts of manipulations on sequences by performing mathematical operations on their associated generating functions. Let's experiment with various operations and characterize their effects in terms of sequences.

12.2.1 Scaling

Multiplying a generating function by a constant scales every term in the associated sequence by the same constant. For example, we noted above that:

$$\langle 1, 0, 1, 0, 1, 0, \ldots \rangle \longleftrightarrow 1 + x^2 + x^4 + x^6 + \cdots = \frac{1}{1-x^2}$$

Multiplying the generating function by 2 gives

$$\frac{2}{1-x^2} = 2 + 2x^2 + 2x^4 + 2x^6 + \cdots$$

which generates the sequence:

$$\langle 2, 0, 2, 0, 2, 0, \ldots \rangle$$

Rule 13 (Scaling Rule). *If*

$$\langle f_0, f_1, f_2, \ldots \rangle \longleftrightarrow F(x),$$

then

$$\langle cf_0,\ cf_1,\ cf_2,\ \ldots \rangle \longleftrightarrow c \cdot F(x).$$

The idea behind this rule is that:

$$\begin{aligned}\langle cf_0, cf_1, cf_2, \ldots\rangle &\longleftrightarrow cf_0 + cf_1x + cf_2x^2 + \cdots \\ &= c\cdot(f_0 + f_1x + f_2x^2 + \cdots) \\ &= cF(x)\end{aligned}$$

12.2.2 Addition

Adding generating functions corresponds to adding the two sequences term by term. For example, adding two of our earlier examples gives:

$$\begin{array}{rccccccccc} & \langle & 1, & 1, & 1, & 1, & 1, & 1, & \ldots & \rangle \\ + & \langle & 1, & -1, & 1, & -1, & 1, & -1, & \ldots & \rangle \\ \hline & \langle & 2, & 0, & 2, & 0, & 2, & 0, & \ldots & \rangle \end{array} \begin{array}{c} \longleftrightarrow \\ \longleftrightarrow \\ \hline \longleftrightarrow \end{array} \begin{array}{c} \dfrac{1}{1-x} \\ \dfrac{1}{1+x} \\ \hline \dfrac{1}{1-x} + \dfrac{1}{1+x} \end{array}$$

We've now derived two different expressions that both generate the sequence $\langle 2, 0, 2, 0, \ldots\rangle$. They are, of course, equal:

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

Rule 14 (Addition Rule). *If*

$$\begin{aligned}\langle f_0, f_1, f_2, \ldots\rangle &\longleftrightarrow F(x), \qquad\qquad \textit{and} \\ \langle g_0, g_1, g_2, \ldots\rangle &\longleftrightarrow G(x),\end{aligned}$$

then

$$\langle f_0 + g_0,\ f_1 + g_1,\ f_2 + g_2,\ \ldots\rangle \longleftrightarrow F(x) + G(x).$$

The idea behind this rule is that:

$$\begin{aligned}\langle f_0 + g_0,\ f_1 + g_1,\ f_2 + g_2,\ \ldots\rangle &\longleftrightarrow \sum_{n=0}^{\infty}(f_n + g_n)x^n \\ &= \left(\sum_{n=0}^{\infty} f_n x^n\right) + \left(\sum_{n=0}^{\infty} g_n x^n\right) \\ &= F(x) + G(x)\end{aligned}$$

12.2.3 Right Shifting

Let's start over again with a simple sequence and its generating function:

$$\langle 1, 1, 1, 1, \ldots\rangle \longleftrightarrow \frac{1}{1-x}$$

Now let's *right-shift* the sequence by adding k leading zeros:

$$\begin{aligned}\langle \underbrace{0,0,\ldots,0}_{k \text{ zeroes}},1,1,1,\ldots\rangle &\longleftrightarrow x^k + x^{k+1} + x^{k+2} + x^{k+3} + \cdots \\ &= x^k \cdot (1 + x + x^2 + x^3 + \cdots) \\ &= \frac{x^k}{1-x}\end{aligned}$$

Evidently, adding k leading zeros to the sequence corresponds to multiplying the generating function by x^k. This holds true in general.

Rule 15 (Right-Shift Rule). *If* $\langle f_0, f_1, f_2, \ldots\rangle \longleftrightarrow F(x)$*, then:*

$$\langle \underbrace{0,0,\ldots,0}_{k \text{ zeroes}}, f_0, f_1, f_2, \ldots\rangle \longleftrightarrow x^k \cdot F(x)$$

The idea behind this rule is that:

$$\begin{aligned}\langle \overbrace{0,0,\ldots,0}^{k \text{ zeroes}}, f_0, f_1, f_2, \ldots\rangle &\longleftrightarrow f_0 x^k + f_1 x^{k+1} + f_2 x^{k+2} + \cdots \\ &= x^k \cdot (f_0 + f_1 x + f_2 x^2 + f_3 x^3 + \cdots) \\ &= x^k \cdot F(x)\end{aligned}$$

12.2.4 Differentiation

What happens if we take the *derivative* of a generating function? As an example, let's differentiate the now-familiar generating function for an infinite sequence of 1's.

$$\begin{aligned}\frac{d}{dx}(1 + x + x^2 + x^3 + x^4 + \cdots) &= \frac{d}{dx}\left(\frac{1}{1-x}\right) \\ 1 + 2x + 3x^2 + 4x^3 + \cdots &= \frac{1}{(1-x)^2} \\ \langle 1,2,3,4,\ldots\rangle &\longleftrightarrow \frac{1}{(1-x)^2}\end{aligned}$$

We found a generating function for the sequence $\langle 1,2,3,4,\ldots\rangle$ of positive integers!

In general, differentiating a generating function has two effects on the corresponding sequence: each term is multiplied by its index and the entire sequence is shifted left one place.

Rule 16 (Derivative Rule). *If*

$$\langle f_0, f_1, f_2, f_3, \ldots\rangle \longleftrightarrow F(x),$$

then

$$\langle f_1, 2f_2, 3f_3, \ldots\rangle \longleftrightarrow F'(x).$$

The idea behind this rule is that:

$$\begin{aligned}\langle f_1, 2f_2, 3f_3, \dots \rangle &\longleftrightarrow f_1 + 2f_2 x + 3f_3 x^2 + \cdots \\ &= \frac{d}{dx}\left(f_0 + f_1 x + f_2 x^2 + f_3 x^3 + \cdots\right) \\ &= \frac{d}{dx} F(x)\end{aligned}$$

The Derivative Rule is very useful. In fact, there is frequent, independent need for each of differentiation's two effects, multiplying terms by their index and left-shifting one place. Typically, we want just one effect and must somehow cancel out the other. For example, let's try to find the generating function for the sequence of squares, $\langle 0, 1, 4, 9, 16, \dots \rangle$. If we could start with the sequence $\langle 1, 1, 1, 1, \dots \rangle$ and multiply each term by its index two times, then we'd have the desired result:

$$\langle 0 \cdot 0,\ 1 \cdot 1,\ 2 \cdot 2,\ 3 \cdot 3,\ \dots \rangle = \langle 0, 1, 4, 9, \dots \rangle$$

A challenge is that differentiation not only multiplies each term by its index, but also shifts the whole sequence left one place. However, the Right-Shift Rule 15 tells how to cancel out this unwanted left-shift: multiply the generating function by x.

Our procedure, therefore, is to begin with the generating function for $\langle 1, 1, 1, 1, \dots \rangle$, differentiate, multiply by x, and then differentiate and multiply by x once more.

$$\begin{aligned}\langle 1, 1, 1, 1, \dots \rangle &\longleftrightarrow \frac{1}{1-x} \\ \langle 1, 2, 3, 4, \dots \rangle &\longleftrightarrow \frac{d}{dx}\frac{1}{1-x} = \frac{1}{(1-x)^2} \\ \langle 0, 1, 2, 3, \dots \rangle &\longleftrightarrow x \cdot \frac{1}{(1-x)^2} = \frac{x}{(1-x)^2} \\ \langle 1, 4, 9, 16, \dots \rangle &\longleftrightarrow \frac{d}{dx}\frac{x}{(1-x)^2} = \frac{1+x}{(1-x)^3} \\ \langle 0, 1, 4, 9, \dots \rangle &\longleftrightarrow x \cdot \frac{1+x}{(1-x)^3} = \frac{x(1+x)}{(1-x)^3}\end{aligned}$$

Thus, the generating function for squares is:

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

12.2.5 Products

Rule 17 (Product Rule). *If*

$$\begin{aligned}\langle a_0, a_1, a_2, \dots \rangle &\longleftrightarrow A(x), \qquad \textit{and} \\ \langle b_0, b_1, b_2, \dots \rangle &\longleftrightarrow B(x),\end{aligned}$$

then

$$\langle c_0,\ c_1,\ c_2,\ \dots \rangle \longleftrightarrow A(x) \cdot B(x),$$

where

$$c_n ::= a_0 b_n + a_1 b_{n-1} + a_2 b_{n-2} + \cdots + a_n b_0.$$

To understand this rule, let

$$C(x) ::= A(x) \cdot B(x) = \sum_{n=0}^{\infty} c_n x^n.$$

We can evaluate the product $A(x) \cdot B(x)$ by using a table to identify all the cross-terms from the product of the sums:

	b_0x^0	b_1x^1	b_2x^2	b_3x^3	$\ldots$
a_0x^0	$a_0b_0x^0$	$a_0b_1x^1$	$a_0b_2x^2$	$a_0b_3x^3$	$\ldots$
a_1x^1	$a_1b_0x^1$	$a_1b_1x^2$	$a_1b_2x^3$	$\ldots$	
a_2x^2	$a_2b_0x^2$	$a_2b_1x^3$	$\ldots$		
a_3x^3	$a_3b_0x^3$	$\ldots$			
$\vdots$	$\ldots$				

Notice that all terms involving the same power of x lie on a /-sloped diagonal. Collecting these terms together, we find that the coefficient of x^n in the product is the sum of all the terms on the $(n+1)$st diagonal, namely,

$$a_0 b_n + a_1 b_{n-1} + a_2 b_{n-2} + \cdots + a_n b_0. \tag{12.1}$$

This expression (12.1) may be familiar from a signal processing course; the sequence $\langle c_0, c_1, c_2, \ldots \rangle$ is called the ***convolution*** of sequences $\langle a_0, a_1, a_2, \ldots \rangle$ and $\langle b_0, b_1, b_2, \ldots \rangle$.

12.3 The Fibonacci Sequence

Sometimes we can find nice generating functions for more complicated sequences. For example, here is a generating function for the Fibonacci numbers:

$$\langle 0, 1, 1, 2, 3, 5, 8, 13, 21, \ldots \rangle \longleftrightarrow \frac{x}{1 - x - x^2}$$

The Fibonacci numbers may seem like a fairly nasty bunch, but the generating function is simple!

We're going to derive this generating function and then use it to find a closed form for the nth Fibonacci number. The techniques we'll use are applicable to a large class of recurrence equations.

12.3.1 Finding a Generating Function

Let's begin by recalling the definition of the Fibonacci numbers:

$$\begin{aligned} f_0 &= 0 \\ f_1 &= 1 \\ f_n &= f_{n-1} + f_{n-2} \qquad (\text{for } n \geq 2) \end{aligned}$$

We can expand the final clause into an infinite sequence of equations. Thus, the Fibonacci numbers are defined by:

$$\begin{aligned} f_0 &= 0 \\ f_1 &= 1 \\ f_2 &= f_1 + f_0 \\ f_3 &= f_2 + f_1 \\ f_4 &= f_3 + f_2 \\ &\vdots \end{aligned}$$

Now the overall plan is to *define* a function $F(x)$ that generates the sequence on the left side of the equality symbols, which are the Fibonacci numbers. Then we *derive* a function that generates the sequence on the right side. Finally, we equate the two and solve for $F(x)$. Let's try this. First, we define:

$$F(x) = f_0 + f_1 x + f_2 x^2 + f_3 x^3 + f_4 x^4 + \cdots$$

Now we need to derive a generating function for the sequence:

$$\langle 0,\ 1,\ f_1 + f_0,\ f_2 + f_1,\ f_3 + f_2,\ \ldots \rangle$$

One approach is to break this into a sum of three sequences for which we know generating functions and then apply the Addition Rule:

$$\begin{array}{rccccccccc} & \langle & 0, & 1, & 0, & 0, & 0, & \ldots & \rangle & \longleftrightarrow & x \\ & \langle & 0, & f_0, & f_1, & f_2, & f_3, & \ldots & \rangle & \longleftrightarrow & xF(x) \\ + & \langle & 0, & 0, & f_0, & f_1, & f_2, & \ldots & \rangle & \longleftrightarrow & x^2F(x) \\ \hline & \langle & 0, & 1+f_0, & f_1+f_0, & f_2+f_1, & f_3+f_2, & \ldots & \rangle & \longleftrightarrow & x + xF(x) + x^2F(x) \end{array}$$

This sequence is almost identical to the right sides of the Fibonacci equations. The one blemish is that the second term is $1 + f_0$ instead of simply 1. However, this amounts to nothing, since $f_0 = 0$ anyway.

Now if we equate $F(x)$ with the new function $x + xF(x) + x^2F(x)$, then we're implicitly writing down *all* of the equations that define the Fibonacci numbers in one fell swoop:

$$\begin{array}{rcccccccccc} F(x) & = & f_0 + & f_1 & x + & f_2 & x^2 + & f_3 & x^3 + & f_4 & x^4 + \ldots \\ \| & & \| & \| & & \| & & \| & & \| & \\ x + xF(x) + x^2F(x) & = & 0 + & (1+f_0)\, x + & & (f_1+f_0)\, x^2 + & & (f_2+f_1)\, x^3 + & & (f_3+f_2)\, x^4 + \cdots & \end{array}$$

Solving for $F(x)$ gives the generating function for the Fibonacci sequence:

$$F(x) = x + xF(x) + x^2F(x)$$

so

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

Sure enough, this is the simple generating function we claimed at the outset.

12.3.2 Finding a Closed Form

Why should one care about the generating function for a sequence? There are several answers, but here is one: if we can find a generating function for a sequence, then we can often find a closed form for the nth coefficient— which can be pretty useful! For example, a closed form for the coefficient of x^n in the power series for $x/(1-x-x^2)$ would be an explicit formula for the nth Fibonacci number.

So our next task is to extract coefficients from a generating function. There are several approaches. For a generating function that is a ratio of polynomials, we can use the method of partial fractions, which you learned in calculus. Just as the terms in a partial fraction expansion are easier to integrate, the coefficients of those terms are easy to compute.

Let's try this approach with the generating function for Fibonacci numbers. First, we factor the denominator:

$$1-x-x^2 = (1-\alpha_1 x)(1-\alpha_2 x)$$

where $\alpha_1 = \frac{1}{2}(1+\sqrt{5})$ and $\alpha_2 = \frac{1}{2}(1-\sqrt{5})$. Next, we find A_1 and A_2 which satisfy:

$$\frac{x}{1-x-x^2} = \frac{A_1}{1-\alpha_1 x} + \frac{A_2}{1-\alpha_2 x}$$

We do this by plugging in various values of x to generate linear equations in A_1 and A_2. We can then find A_1 and A_2 by solving a linear system. This gives:

$$A_1 = \frac{1}{\alpha_1 - \alpha_2} = \frac{1}{\sqrt{5}}$$
$$A_2 = \frac{-1}{\alpha_1 - \alpha_2} = -\frac{1}{\sqrt{5}}$$

Substituting into the equation above gives the partial fractions expansion of $F(x)$:

$$\frac{x}{1-x-x^2} = \frac{1}{\sqrt{5}}\left(\frac{1}{1-\alpha_1 x} - \frac{1}{1-\alpha_2 x}\right)$$

Each term in the partial fractions expansion has a simple power series given by the geometric sum formula:

$$\frac{1}{1-\alpha_1 x} = 1 + \alpha_1 x + \alpha_1^2 x^2 + \cdots$$
$$\frac{1}{1-\alpha_2 x} = 1 + \alpha_2 x + \alpha_2^2 x^2 + \cdots$$

Substituting in these series gives a power series for the generating function:

$$\begin{aligned} F(x) &= \frac{1}{\sqrt{5}}\left(\frac{1}{1-\alpha_1 x} - \frac{1}{1-\alpha_2 x}\right) \\ &= \frac{1}{\sqrt{5}}\left((1+\alpha_1 x + \alpha_1^2 x^2 + \cdots) - (1+\alpha_2 x + \alpha_2^2 x^2 + \cdots)\right), \end{aligned}$$

so

$$\begin{aligned} f_n &= \frac{\alpha_1^n - \alpha_2^n}{\sqrt{5}} \\ &= \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right) \end{aligned}$$

This formula may be scary and astonishing —it's not even obvious that its value is an integer —but it's very useful. For example, it provides (via the repeated squaring method) a much more efficient way to compute Fibonacci numbers than crunching through the recurrence, and it also clearly reveals the exponential growth of these numbers.

12.4 Counting with Generating Functions

Generating functions are particularly useful for solving counting problems. In particular, problems involving choosing items from a set often lead to nice generating functions by letting the coefficient of x^n be the number of ways to choose n items.

12.4.1 Choosing Distinct Items from a Set

The generating function for binomial coefficients follows directly from the Binomial Theorem:

$$\begin{aligned} \left\langle \binom{k}{0}, \binom{k}{1}, \binom{k}{2}, \ldots, \binom{k}{k}, 0, 0, 0, \ldots \right\rangle &\longleftrightarrow \binom{k}{0} + \binom{k}{1}x + \binom{k}{2}x^2 + \cdots + \binom{k}{k}x^k \\ &= (1+x)^k \end{aligned}$$

Thus, the coefficient of x^n in $(1+x)^k$ is the number of ways to choose n distinct items from a set of size k. For example, the coefficient of x^2 is $\binom{k}{2}$, the number of ways to choose 2 items from a set with k elements. Similarly, the coefficient of x^{k+1} is the number of ways to choose $k+1$ items from a size k set, which is zero.

12.4.2 Building Generating Functions that Count

Often we can translate the description of a counting problem directly into a generating function for the solution. For example, we could figure out that $(1+x)^k$ generates the number of ways to select n distinct items from a k-element set without resorting to the Binomial Theorem or even fussing with binomial coefficients!

Here is how. First, consider a single-element set $\{a_1\}$. The generating function for the number of ways to choose n elements from this set is simply $1+x$: we have 1 way to choose zero elements, 1 way to choose one element, and 0 ways to choose more than one element. Similarly, the number of ways to choose n elements from the set $\{a_2\}$ is also given by the generating function $1+x$. The fact that the elements differ in the two cases is irrelevant.

Now here is the main trick: *the generating function for choosing elements from a union of disjoint sets is the product of the generating functions for choosing from each set.* We'll justify this in a moment, but

let's first look at an example. According to this principle, the generating function for the number of ways to choose n elements from the $\{a_1, a_2\}$ is:

$$\underbrace{(1+x)}_{\substack{\text{OGF for}\\ \{a_1\}}} \cdot \underbrace{(1+x)}_{\substack{\text{OGF for}\\ \{a_2\}}} = \underbrace{(1+x)^2}_{\substack{\text{OGF for}\\ \{a_1, a_2\}}} = 1 + 2x + x^2$$

Sure enough, for the set $\{a_1, a_2\}$, we have 1 way to choose zero elements, 2 ways to choose one element, 1 way to choose two elements, and 0 ways to choose more than two elements.

Repeated application of this rule gives the generating function for choosing n items from a k-element set $\{a_1, a_2, \ldots, a_k\}$:

$$\underbrace{(1+x)}_{\substack{\text{OGF for}\\ \{a_1\}}} \cdot \underbrace{(1+x)}_{\substack{\text{OGF for}\\ \{a_2\}}} \cdots \underbrace{(1+x)}_{\substack{\text{OGF for}\\ \{a_k\}}} = \underbrace{(1+x)^k}_{\substack{\text{OGF for}\\ \{a_1, a_2, \ldots, a_k\}}}$$

This is the same generating function that we obtained by using the Binomial Theorem. But this time around we translated directly from the counting problem to the generating function.

We can extend these ideas to a general principle:

Rule 18 (Convolution Rule). *Let $A(x)$ be the generating function for selecting items from set $\mathcal{A}$, and let $B(x)$ be the generating function for selecting items from set $\mathcal{B}$. If $\mathcal{A}$ and $\mathcal{B}$ are disjoint, then the generating function for selecting items from the union $\mathcal{A} \cup \mathcal{B}$ is the product $A(x) \cdot B(x)$.*

This rule is rather ambiguous: what exactly are the rules governing the selection of items from a set? Remarkably, the Convolution Rule remains valid under *many* interpretations of selection. For example, we could insist that distinct items be selected or we might allow the same item to be picked a limited number of times or any number of times. Informally, the only restrictions are that (1) the order in which items are selected is disregarded and (2) restrictions on the selection of items from sets $\mathcal{A}$ and $\mathcal{B}$ also apply in selecting items from $\mathcal{A} \cup \mathcal{B}$. (Formally, there must be a bijection between n-element selections from $\mathcal{A} \cup \mathcal{B}$ and ordered pairs of selections from $\mathcal{A}$ and $\mathcal{B}$ containing a total of n elements.)

To count the number of ways to select n items from $\mathcal{A} \cup \mathcal{B}$, we observe that we can select n items by choosing j items from $\mathcal{A}$ and $n - j$ items from $\mathcal{B}$, where j is any number from 0 to n. This can be done in $a_j b_{n-j}$ ways. Summing over all the possible values of j gives a total of

$$a_0 b_n + a_1 b_{n-1} + a_2 b_{n-2} + \cdots + a_n b_0$$

ways to select n items from $\mathcal{A} \cup \mathcal{B}$. By the Product Rule, this is precisely the coefficient of x^n in the series for $A(x)B(x)$.

12.4.3 Choosing Items with Repetition

The first counting problem we considered was the number of ways to select a dozen doughnuts when five flavors were available. We can generalize this question as follows: in how many ways can we select n items from a k-element set if we're allowed to pick the same item multiple times?

In these terms, the doughnut problem asks in how many ways we can select $n = 12$ doughnuts from the set of $k = 5$ flavors

$$\{\text{chocolate}, \text{lemon-filled}, \text{sugar}, \text{glazed}, \text{plain}\}$$

where, of course, we're allowed to pick several doughnuts of the same flavor. Let's approach this question from a generating functions perspective.

Suppose we choose n items (with repetition allowed) from a set containing a single item. Then there is one way to choose zero items, one way to choose one item, one way to choose two items, etc. Thus, the generating function for choosing n elements with repetition from a 1-element set is:

$$\begin{aligned} \langle 1, 1, 1, 1, \ldots \rangle &\longleftrightarrow 1 + x + x^2 + x^3 + \cdots \\ &= \frac{1}{1-x} \end{aligned}$$

The Convolution Rule says that the generating function for selecting items from a union of disjoint sets is the product of the generating functions for selecting items from each set:

$$\underbrace{\frac{1}{1-x}}_{\substack{\text{OGF for} \\ \{a_1\}}} \cdot \underbrace{\frac{1}{1-x}}_{\substack{\text{OGF for} \\ \{a_2\}}} \cdots \underbrace{\frac{1}{1-x}}_{\substack{\text{OGF for} \\ \{a_k\}}} = \underbrace{\frac{1}{(1-x)^k}}_{\substack{\text{OGF for} \\ \{a_1, a_2, \ldots, a_k\}}}$$

Therefore, the generating function for selecting items from a k-element set with repetition allowed is $1/(1-x)^k$.

Now the Bookkeeper Rule tells us that the number of ways to select n items with repetition from an k element set is

$$\binom{n+k-1}{n},$$

so this is the coefficient of x^n in the series expansion of $1/(1-x)^k$.

On the other hand, it's instructive to derive this coefficient algebraically, which we can do using Taylor's Theorem:

Theorem 12.4.1 (Taylor's Theorem).

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + \cdots.$$

This theorem says that the nth coefficient of $1/(1-x)^k$ is equal to its nth derivative evaluated at 0 and divided by $n!$. Computing the nth derivative turns out not to be very difficult. Let

$$G(x) ::= \frac{1}{(1-x)^k} = (1-x)^{-k}.$$

Then we have:

$$\begin{aligned} G'(x) &= k(1-x)^{-(k+1)} \\ G''(x) &= k(k+1)(1-x)^{-(k+2)} \\ G'''(x) &= k(k+1)(k+2)(1-x)^{-(k+3)} \\ G^{(n)}(x) &= k(k+1)\cdots(k+n-1)(1-x)^{-(k+n)} \end{aligned}$$

Thus, the coefficient of x^n in the generating function is:

$$\begin{aligned} G^{(n)}(0)/n! &= \frac{k(k+1)\cdots(k+n-1)}{n!} \\ &= \frac{(k+n-1)!}{(k-1)!\,n!} \\ &= \binom{n+k-1}{n}. \end{aligned}$$

So if we didn't already know the Bookkeeper Rule, we could have proved it using generating functions.

12.5 An "Impossible" Counting Problem

So far everything we've done with generating functions we could have done another way. But here is an absurd counting problem— really over the top! In how many ways can we fill a bag with n fruits subject to the following constraints?

- The number of apples must be even.
- The number of bananas must be a multiple of 5.
- There can be at most four oranges.
- There can be at most one pear.

For example, there are 7 ways to form a bag with 6 fruits:

Apples	6	4	4	2	2	0	0
Bananas	0	0	0	0	0	5	5
Oranges	0	2	1	4	3	1	0
Pears	0	0	1	0	1	0	1

These constraints are so complicated that the problem seems hopeless! But let's see what generating functions reveal.

Let's first construct a generating function for selecting apples. We can select a set of 0 apples in one way, a set of 1 apple in zero ways (since the number of apples must be even), a set of 2 apples in one way, a set of 3 apples in zero ways, and so forth. So we have:

$$A(x) = 1 + x^2 + x^4 + x^6 + \cdots = \frac{1}{1-x^2}$$

Similarly, the generating function for selecting bananas is:

$$B(x) = 1 + x^5 + x^{10} + x^{15} + \cdots = \frac{1}{1-x^5}$$

Now, we can select a set of 0 oranges in one way, a set of 1 orange in one way, and so on. However, we can not select more than four oranges, so we have the generating function:

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

Here we're using the geometric sum formula. Finally, we can select only zero or one pear, so we have:

$$P(x) = 1 + x$$

The Convolution Rule says that the generating function for selecting from among all four kinds of fruit is:

$$\begin{aligned} A(x)B(x)O(x)P(x) &= \frac{1}{1 - x^2}\frac{1}{1 - x^5}\frac{1 - x^5}{1 - x}(1 + x) \\ &= \frac{1}{(1 - x)^2} \\ &= 1 + 2x + 3x^2 + 4x^3 + \cdots \end{aligned}$$

Almost everything cancels! We're left with $1/(1 - x)^2$, which we found a power series for earlier: the coefficient of x^n is simply $n + 1$. Thus, the number of ways to form a bag of n fruits is just $n + 1$. This is consistent with the example we worked out, since there were 7 different fruit bags containing 6 fruits. *Amazing!*

12.6 In-Class Problems Week 11, Wed.

Problem 12.6.1. We are interested in generating functions for the number of different ways to compose a bag of n donuts subject to various restrictions. For each of the restrictions in (a)-(e) below, find a closed form for the corresponding generating function.

(a) All the donuts are chocolate and there are at least 3.

Solution.

$$\frac{x^3}{1-x}$$

■

(b) All the donuts are glazed and there are at most 2.

Solution.

$$1 + x + x^2$$

■

(c) All the donuts are coconut and there are exactly 2 or there are none.

Solution.

$$1 + x^2$$

■

(d) All the donuts are plain and their number is a multiple of 4.

Solution.

$$\frac{1}{1-x^4} = \frac{1}{(1-x)(1+x)(1+x^2)}$$

■

(e) The donuts must be chocolate, glazed, coconut, or plain and:

- there must be at least 3 chocolate donuts, and
- there must be at most 2 glazed, and
- there must be exactly 0 or 2 coconut, and
- there must be a multiple of 4 plain.

Solution.

$$\begin{aligned}\frac{x^3}{1-x}(1+x+x^2)(1+x^2)\frac{1}{1-x^4} &= \frac{x^3(1+x+x^2)(1+x^2)}{(1-x)^2(1+x)(1+x^2)} \\ &= x^3\frac{1+x+x^2}{(1-x)^2(1+x)}\end{aligned}$$

■

(f) Find a closed form for the number of ways to select n donuts subject to the constraints of the previous part.

Solution. Let

$$G(x) ::= \frac{1+x+x^2}{(1-x)^2(1+x)},$$

so the generating function for donut selections is $x^3G(x)$. By partial fractions

$$\frac{1+x+x^2}{(1-x)^2(1+x)} = \frac{A}{1-x} + \frac{B}{(1-x)^2} + \frac{C}{1+x} \tag{12.2}$$

for some constants, A, B, C. Using the fact the Convolution Counting Property from lecture (also see Problem 12.6.2.(d)) that the coefficient of x^n in the series for $(1-x)^2$ is $\binom{n+1}{1}$, we conclude that the nth coefficient in the series for $G(x)$ is

$$A + B\binom{n+1}{1} + C(-1)^n. \tag{12.3}$$

To find A, B, C, we multiply both sides of (12.2) by the denominator $(1-x)^2(1+x)$ to obtain

$$1+x+x^2 = A(1-x)(1+x) + B(1+x) + C(1-x)^2. \tag{12.4}$$

Letting $x = 1$ in (12.4), we conclude that $3 = 2B$, so $B = 3/2$. Then, letting $x = -1$, we conclude $(-1)^2 = C2^2$, so $C = 1/4$. Finally, letting $x = 0$, we have

$$1 = A + B + C = A + \frac{3}{2} + \frac{1}{4},$$

so $A = -3/4$. Then from (12.3), we conclude that the nth coefficient in the series for $G(x)$ is

$$-\frac{3}{4} + \frac{3(n+1)}{2} + \frac{(-1)^n}{4} = \frac{6n+3+(-1)^n}{4}.$$

So the nth coefficient in the series for the generating function, $x^3G(x)$, for donut selections is zero for $n < 3$, and, for $n \geq 3$, is the $(n-3)$rd coefficient of G, namely,

$$\frac{6(n-3)+3+(-1)^{n-3}}{4} = \frac{6n-15+(-1)^{n-1}}{4}.$$

■

Problem 12.6.2. **(a)** Let

$$S(x) ::= \frac{x^2+x}{(1-x)^3}.$$

What is the coefficient of x^n in the generating function series for $S(x)$?

Hint: A formula for the coefficient of x^n in $1/(1-x)^k$ follows from the Convolution Counting Principle and is given in the Appendix (and in part (d) below).

Solution. n^2. That is, $S(x) = \sum_{n=1}^{\infty} n^2 x^n$.

To see why, note that the coefficient of x^n in $1/(1-x)^3$ is

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

by the formula in the Appendix.

Now the coefficient of x^n in $x^2/(1-x)^3$ is the same as the coefficient of x^{n-2} in $1/(1-x)^3$, namely, $((n-2)+2)((n-2)+1)/2 = n(n-1)/2$. Similarly, the coefficient of x^n in $x/(1-x)^3$ is the same as the coefficient of x^{n-1} in $1/(1-x)^3$, namely, $((n-1)+2)((n-1)+1)/2 = (n+1)n/2$. The coefficient of x^n in $S(x)$ is the sum of these two coefficients, namely,

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

■

(b) Explain why $S(x)/(1-x)$ is the generating function for the sums of squares. That is, the coefficient of x^n in the series for $S(x)/(1-x)$ is $\sum_{k=1}^{n} k^2$.

Solution.

$$\left(\sum_{n=0}^{\infty} a_n x^n\right)\left(\sum_{n=0}^{\infty} x^n\right) = \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} a_k \cdot 1\right) x^n = \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} a_k\right) x^n \tag{12.5}$$

by the convolution formula for the product of series. For $S(x)$, the coefficient of x^k is $a_k = k^2$, and

$$S(x)/(1-x) = S(x)\left(\sum_{n=0}^{\infty} x^n\right),$$

so (12.5) implies that the coefficient of x^n in $S(x)/(1-x)$ is the sum of the first n squares. ■

(c) Use the previous parts to prove that

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

Solution. We have

$$\frac{S(x)}{1-x} = \frac{\frac{x(1+x)}{(1-x)^3}}{1-x} = \frac{x+x^2}{(1-x)^4}. \tag{12.6}$$

From part (d), the coefficient of x^n in the series expansion of $1/(1-x)^4$ is

$$\binom{n+3}{3} = \frac{(n+1)(n+2)(n+3)}{3!}.$$

But by (12.6),

$$\frac{S(x)}{1-x} = \frac{x}{(1-x)^4} + \frac{x^2}{(1-x)^4},$$

so the coefficient of x^n is the sum of the $(n-1)$st and $(n-2)$nd coefficients of $(1-x)^4$, namely,

$$\frac{n(n+1)(n+2)}{3!} + \frac{(n-1)n(n+1)}{3!} = \frac{n(n+1)(2n+1)}{6}.$$

■

(d) Let $A(x) = \sum_{n=0}^{\infty} a_n x^n$. Then it's easy to check that

$$a_n = \frac{A^{(n)}(0)}{n!},$$

where $A^{(n)}$ is the nth derivative of A. Use this fact (which you may assume) instead of the Convolution Counting Principle, to prove that

$$\frac{1}{(1-x)^k} = \sum_{n=0}^{\infty} \binom{n+k-1}{k-1} x^n.$$

Solution.

$$\begin{aligned}
\frac{d\,(1-x)^{-k}}{dx} &= k\,(1-x)^{-(k+1)}\,.\\
\frac{d^2\,(1-x)^{-k}}{(dx)^2} &= \frac{d\,k\,(1-x)^{-(k+1)}}{dx} = (k+1)k\,(1-x)^{-(k+2)}\\
\frac{d^3\,(1-x)^{-k}}{(dx)^3} &= \frac{d\,(k+1)k\,(1-x)^{-(k+2)}}{dx} = (k+2)(k+1)k\,(1-x)^{-(k+3)}\\
&\vdots\\
\frac{d^n\,(1-x)^{-k}}{(dx)^n} &= (k+n-1)\cdots(k+2)(k+1)k\,(1-x)^{-(k+n)}\,.
\end{aligned}$$

Now suppose $(1-x)^{-k} = A(x)$. Then we have

$$\begin{aligned}
a_n &= \frac{A^{(n)}(0)}{n!}\\
&= \frac{(k+n-1)\cdots(k+2)(k+1)k\,(1-0)^{-(k+n)}}{n!}\\
&= \frac{\dfrac{(n+k-1)!}{(k-1)!}\cdot 1}{n!}\\
&= \frac{(n+k-1)!}{(k-1)!n!}\\
&= \binom{n+k-1}{k-1}
\end{aligned}$$

■

Appendix

Products of Series

Let

$$A(x) = \sum_{n=0}^{\infty} a_n x^n, \qquad B(x) = \sum_{n=0}^{\infty} b_n x^n, \qquad C(x) = A(x)\cdot B(x) = \sum_{n=0}^{\infty} c_n x^n.$$

Then

$$c_n = a_0 b_n + a_1 b_{n-1} + a_2 b_{n-2} + \cdots + a_n b_0.$$

By the Convolution Counting Property, or by Problem 12.6.2.(d),

$$\frac{1}{(1-x)^k} = \sum_{n=0}^{\infty} \binom{n+k-1}{k-1} x^n.$$

12.7 In-Class Problems Week 11, Fri.

Problem 12.7.1. Define the function $f : \mathbb{N} \to \mathbb{N}$ recursively by the rules

$$\begin{aligned} f(0) &= 1, \\ f(1) &= 6, \\ f(n) &= 2f(n-1) + 3f(n-2) + 4 \qquad \text{for } n \geq 2. \end{aligned}$$

(a) Find a closed form for the generating function

$$G(x) ::= f(0) + f(1)x + f(2)x^2 + \cdots + f(n)x^n + \cdots .$$

Solution.

$$\begin{array}{rcccccccc} G(x) & = & f(0) & + & f(1)x & + & f(2)x^2 & +\cdots & + & f(n)x^n + \cdots \\ 2xG(x) & = & & & 2f(0)x & + & 2f(1)x^2 & +\cdots & + & 2f(n-1)x^n + \cdots \\ 3x^2G(x) & = & & & & & 3f(0)x^2 & +\cdots & + & 3f(n-2)x^n + \cdots \\ 4/(1-x) & = & 4 & + & 4x & & 4x^2 & +\cdots & + & 4x^n + \cdots \end{array}$$

Therefore,

$$\begin{aligned} G(x) &= 2xG(x) + 3x^2G(x) + \frac{4}{1-x} + (f(0) - 4) + (f(1) - 2f(0) - 4)x \\ &= 2xG(x) + 3x^2G(x) + \frac{4}{1-x} + (1-4) + (6-2-4)x \\ &= 2xG(x) + 3x^2G(x) + \frac{4}{1-x} - 3. \end{aligned}$$

It follows that

$$G(x)(1 - 2x - 3x^2) = \frac{4}{1-x} - 3,$$

and hence

$$\begin{aligned} G(x) &= \frac{\dfrac{4}{1-x} - 3}{(1+x)(1-3x)} \\ &= \frac{4}{(1-x)(1+x)(1-3x)} - \frac{3}{(1+x)(1-3x)} \\ &= \frac{4 - 3(1-x)}{(1-x)(1+x)(1-3x)} \\ &= \frac{3x+1}{(1-x)(1+x)(1-3x)}. \end{aligned} \tag{12.7}$$

■

(b) Find a closed form for $f(n)$. *Hint:* Find numbers a, b, c, d, e, g such that

$$G(x) = \frac{a}{1+dx} + \frac{b}{1+ex} + \frac{c}{1+gx}.$$

Solution. From (12.7) and the method of partial fractions, we conclude that $d, e, g = -1, 1, -3$, respectively. So we want a, b, c such that

$$\frac{3x+1}{(1-x)(1+x)(1-3x)} = \frac{a}{1-x} + \frac{b}{1+x} + \frac{c}{1-3x} \tag{12.8}$$

$$3x+1 = a(1+x)(1-3x) + b(1-x)(1-3x) + c(1-x)(1+x). \tag{12.9}$$

Setting $x = 1$ in (12.9), we conclude that $4 = a \cdot 2 \cdot (-2)$, so

$$a = -1.$$

Setting $x = -1$ in (12.9), we conclude that $4 - 3 \cdot 2 = b \cdot 2 \cdot 4$, so

$$b = -\frac{1}{4}.$$

Setting $x = 1/3$ in (12.9), we conclude that $4 - 3(2/3) = c \cdot (2/3)(4/3)$, so

$$c = \frac{9}{4}.$$

So from (12.7) and (12.8), we have

$$G(x) = \frac{-1}{1-x} + \frac{1/4}{1+x} + \frac{9/4}{1-3x}.$$

Now the coefficient of x^n in $a/(1-x)$ is a, the coefficient in $b/(1+x)$ is $b(-1)^n$ and the coefficient in $c/(1-3x)$ is $c3^n$. For $n \geq 2$, the coefficient in $G(x)$ is the sum of these coefficients. So

$$f(n) = -1 + \frac{(-1)^n}{4} + \frac{9}{4}3^n = \frac{3^{n+2} + (-1)^n}{4} - 1.$$

■

Problem 12.7.2. (Carried over from Wednesday, April 25)

(a) Let

$$S(x) ::= \frac{x^2 + x}{(1-x)^3}.$$

What is the coefficient of x^n in the generating function series for $S(x)$?

Hint: A formula for the coefficient of x^n in $1/(1-x)^k$ follows from the Convolution Counting Principle and is given in the Appendix (and in part (d) below).

Solution. n^2. That is, $S(x) = \sum_{n=1}^{\infty} n^2 x^n$.

To see why, note that the coefficient of x^n in $1/(1-x)^3$ is

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

by the formula in the Appendix.

Now the coefficient of x^n in $x^2/(1-x)^3$ is the same as the coefficient of x^{n-2} in $1/(1-x)^3$, namely, $((n-2)+2)((n-2)+1)/2 = n(n-1)/2$. Similarly, the coefficient of x^n in $x/(1-x)^3$ is the same as the coefficient of x^{n-1} in $1/(1-x)^3$, namely, $((n-1)+2)((n-1)+1)/2 = (n+1)n/2$. The coefficient of x^n in $S(x)$ is the sum of these two coefficients, namely,

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

■

(b) Explain why $S(x)/(1-x)$ is the generating function for the sums of squares. That is, the coefficient of x^n in the series for $S(x)/(1-x)$ is $\sum_{k=1}^{n} k^2$.

Solution.

$$\left(\sum_{n=0}^{\infty} a_n x^n\right)\left(\sum_{n=0}^{\infty} x^n\right) = \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} a_k \cdot 1\right)x^n = \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} a_k\right)x^n \tag{12.10}$$

by the convolution formula for the product of series. For $S(x)$, the coefficient of x^k is $a_k = k^2$, and

$$S(x)/(1-x) = S(x)\left(\sum_{n=0}^{\infty} x^n\right),$$

so (12.10) implies that the coefficient of x^n in $S(x)/(1-x)$ is the sum of the first n squares. ■

(c) Use the previous parts to prove that

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

Solution. We have

$$\frac{S(x)}{1-x} = \frac{\dfrac{x(1+x)}{(1-x)^3}}{1-x} = \frac{x+x^2}{(1-x)^4}. \tag{12.11}$$

From part (d), the coefficient of x^n in the series expansion of $1/(1-x)^4$ is

$$\binom{n+3}{3} = \frac{(n+1)(n+2)(n+3)}{3!}.$$

But by (12.11),

$$\frac{S(x)}{1-x} = \frac{x}{(1-x)^4} + \frac{x^2}{(1-x)^4},$$

so the coefficient of x^n is the sum of the $(n-1)$st and $(n-2)$nd coefficients of $(1-x)^4$, namely,

$$\frac{n(n+1)(n+2)}{3!} + \frac{(n-1)n(n+1)}{3!} = \frac{n(n+1)(2n+1)}{6}.$$

■

(d) Let $A(x) = \sum_{n=0}^{\infty} a_n x^n$. Then it's easy to check that

$$a_n = \frac{A^{(n)}(0)}{n!},$$

where $A^{(n)}$ is the nth derivative of A. Use this fact (which you may assume) instead of the Convolution Counting Principle, to prove that

$$\frac{1}{(1-x)^k} = \sum_{n=0}^{\infty} \binom{n+k-1}{k-1} x^n.$$

Solution.

$$\begin{aligned}
\frac{d\,(1-x)^{-k}}{dx} &= k\,(1-x)^{-(k+1)}\,. \\
\frac{d^2\,(1-x)^{-k}}{(dx)^2} &= \frac{d\,k\,(1-x)^{-(k+1)}}{dx} = (k+1)k\,(1-x)^{-(k+2)} \\
\frac{d^3\,(1-x)^{-k}}{(dx)^3} &= \frac{d\,(k+1)k\,(1-x)^{-(k+2)}}{dx} = (k+2)(k+1)k\,(1-x)^{-(k+3)} \\
&\vdots \\
\frac{d^n\,(1-x)^{-k}}{(dx)^n} &= (k+n-1)\cdots(k+2)(k+1)k\,(1-x)^{-(k+n)}\,.
\end{aligned}$$

Now suppose $(1-x)^{-k} = A(x)$. Then we have

$$\begin{aligned}
a_n &= \frac{A^{(n)}(0)}{n!} \\
&= \frac{(k+n-1)\cdots(k+2)(k+1)k\,(1-0)^{-(k+n)}}{n!} \\
&= \frac{\dfrac{(n+k-1)!}{(k-1)!}\cdot 1}{n!} \\
&= \frac{(n+k-1)!}{(k-1)!n!} \\
&= \binom{n+k-1}{k-1}
\end{aligned}$$

■

Appendix

Products of Series

Let

$$A(x) = \sum_{n=0}^{\infty} a_n x^n, \qquad B(x) = \sum_{n=0}^{\infty} b_n x^n, \qquad C(x) = A(x)\cdot B(x) = \sum_{n=0}^{\infty} c_n x^n.$$

Then

$$c_n = a_0 b_n + a_1 b_{n-1} + a_2 b_{n-2} + \cdots + a_n b_0.$$

By the Convolution Counting Property, or by Problem 12.7.2.(d),

$$\frac{1}{(1-x)^k} = \sum_{n=0}^{\infty} \binom{n+k-1}{k-1} x^n.$$

Finding a Generating Function for Fibonacci Numbers

The Fibonacci numbers are defined by:

$$\begin{aligned} f_0 &::= 0 \\ f_1 &::= 1 \\ f_n &::= f_{n-1} + f_{n-2} \qquad (\text{for } n \geq 2) \end{aligned}$$

Let F be the generating function for the Fibonacci numbers, that is,

$$F(x) ::= f_0 + f_1 x + f_2 x^2 + f_3 x^3 + f_4 x^4 + \cdots$$

So we need to derive a generating function whose series has coefficients:

$$\langle 0,\ 1,\ f_1 + f_0,\ f_2 + f_1,\ f_3 + f_2,\ \ldots \rangle$$

Now we observe that

$$\begin{array}{rccccccccc} & \langle & 0, & 1, & 0, & 0, & 0, & \ldots & \rangle & \longleftrightarrow & x \\ & \langle & 0, & f_0, & f_1, & f_2, & f_3, & \ldots & \rangle & \longleftrightarrow & xF(x) \\ + & \langle & 0, & 0, & f_0, & f_1, & f_2, & \ldots & \rangle & \longleftrightarrow & x^2F(x) \\ \hline & \langle & 0, & 1+f_0, & f_1+f_0, & f_2+f_1, & f_3+f_2, & \ldots & \rangle & \longleftrightarrow & x + xF(x) + x^2F(x) \end{array}$$

This sequence is almost identical to the right sides of the Fibonacci equations. The one blemish is that the second term is $1 + f_0$ instead of simply 1. But since $f_0 = 0$, the second term is ok.

So we have

$$\begin{aligned} F(x) &= x + xF(x) + x^2F(x). \\ F(x) &= \frac{x}{1-x-x^2}. \end{aligned} \tag{12.12}$$

Finding a Closed Form for the Coefficients

Now we expand the righthand side of (12.12) into partial fractions. To do this, we first factor the denominator

$$1 - x - x^2 = (1 - \alpha_1 x)(1 - \alpha_2 x)$$

where $\alpha_1 = \frac{1}{2}(1 + \sqrt{5})$ and $\alpha_2 = \frac{1}{2}(1 - \sqrt{5})$ by the quadratic formula. Next, we find A_1 and A_2 which satisfy:

$$F(x) = \frac{x}{1-x-x^2} = \frac{A_1}{1-\alpha_1 x} + \frac{A_2}{1-\alpha_2 x} \tag{12.13}$$

Now the coefficient of x^n in $F(x)$ will be A_1 times the coefficient of x^n in $1/(1-\alpha_1 x)$ plus A_2 times the coefficient of x^n in $1/(1-\alpha_2 x)$. The coefficients of these fractions will simply be the terms α_1^n and α_2^n because

$$\frac{1}{1-\alpha_1 x} = 1 + \alpha_1 x + \alpha_1^2 x^2 + \cdots$$
$$\frac{1}{1-\alpha_2 x} = 1 + \alpha_2 x + \alpha_2^2 x^2 + \cdots$$

by the formula for geometric series.

So we just need to find find A_1 and A_2. We do this by plugging values of x into (12.13) to generate linear equations in A_1 and A_2. It helps to note that from (12.13), we have

$$x = A_1(1-\alpha_2 x) + A_2(1-\alpha_1 x),$$

so simple values to use are $x = 0$ and $x = 1/\alpha_2$. We can then find A_1 and A_2 by solving the linear equations. This gives:

$$A_1 = \frac{1}{\alpha_1 - \alpha_2} = \frac{1}{\sqrt{5}}$$
$$A_2 = -A_1 = -\frac{1}{\sqrt{5}}$$

Substituting into (12.13) gives the partial fractions expansion of $F(x)$:

$$F(x) = \frac{1}{\sqrt{5}}\left(\frac{1}{1-\alpha_1 x} - \frac{1}{1-\alpha_2 x}\right).$$

So we conclude that the coefficient, f_n, of x^n in the series for $F(x)$ is

$$f_n = \frac{\alpha_1^n - \alpha_2^n}{\sqrt{5}}$$
$$= \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right)$$

12.8 Problem Set 10

Problem 12.8.1. Miss McGillicuddy never goes outside without a collection of pets. In particular:

- She brings a positive number of songbirds, which always come in pairs.
- She may or may not bring her alligator, Freddy.
- She brings at least 2 cats.
- She brings two or more chihuahuas and labradors leashed together in a line.

Let P_n denote the number of different collections of n pets that can accompany her, where we regard chihuahuas and labradors leashed up in different orders as different collections, even if there are the same number chihuahuas and labradors leashed in the line.

For example, $P_6 = 4$ since there are 4 possible collections of 6 pets:

- 2 songbirds, 2 cats, 2 chihuahuas leashed in line
- 2 songbirds, 2 cats, 2 labradors leashed in line
- 2 songbirds, 2 cats, a labrador leashed behind a chihuahua
- 2 songbirds, 2 cats, a chihuahua leashed behind a labrador

And $P_7 = 16$ since there are 16 possible collections of 7 pets:

- 2 songbirds, 3 cats, 2 chihuahuas leashed in line
- 2 songbirds, 3 cats, 2 labradors leashed in line
- 2 songbirds, 3 cats, a labrador leashed behind a chihuahua
- 2 songbirds, 3 cats, a chihuahua leashed behind a labrador
- 4 collections consisting of 2 songbirds, 2 cats, 1 alligator, and a line of 2 dogs
- 8 collections consisting of 2 songbirds, 2 cats, and a line of 3 dogs.

(a) Let

$$P(x) ::= P_0 + P_1 x + P_2 x^2 + P_3 x^3 + \cdots$$

be the generating function for the number of Miss McGillicuddy's pet collections. Verify that

$$P(x) = \frac{4x^6}{(1-x)^2(1-2x)}.$$

Solution.

$$P(x) = \underbrace{(x^2+x^4+x^6+x^8+\cdots)}_{\text{collections of songbirds}} \cdot \underbrace{(1+x)}_{\text{collections of gators}} \cdot \underbrace{(x^2+x^3+x^4+\cdots)}_{\text{collections of cats}} \cdot \underbrace{(2^2x^2+2^3x^3+2^4x^4+\cdots)}_{\text{lines of dogs}}$$

$$= \frac{x^2}{1-x^2}\cdot(1+x)\cdot\frac{x^2}{1-x}\cdot\frac{4x^2}{1-2x}$$

$$= \frac{x^2}{(1-x)(1+x)}\cdot(1+x)\cdot\frac{x^2}{1-x}\cdot\frac{4x^2}{1-2x}$$

$$= \frac{4x^6}{(1-x)^2(1-2x)}.$$

■

(b) Find a simple formula for P_n.

Solution. P_n is the coefficient of x^n in the power series for $4x^6/(1-x)^2(1-2x)$, which means it is 4 times the coefficient of x^{n-6} in the series for $1/(1-x)^2(1-2x)$ when $n \geq 6$, and $P_n = 0$ for $n < 6$.

But we can express $1/(1-x)^2(1-2x)$ using partial fractions as

$$\frac{1}{(1-x)^2(1-2x)} = \frac{A}{1-x} + \frac{B}{(1-x)^2} + \frac{C}{1-2x} \tag{12.14}$$

for some constants A, B, C, so P_n will be 4 times the sum of the coefficients of x^{n-6} in each of $A/(1-x)$, $B/(1-x)^2$, and $C/(1-2x)$, namely

$$P_n = 4(A + B\binom{n-5}{1} + C2^{n-6}) = 4A + 4B(n-5) + C2^{n-4}. \tag{12.15}$$

So we need only find the values of A, B, C. But multiplying both sides of (12.14) by the lefthand denominator $(1-x)^2(1-2x)$ yields

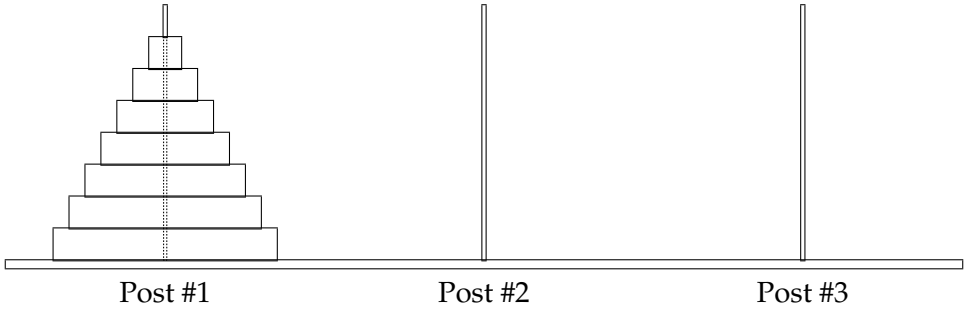
$$1 = A(1-x)(1-2x) + B(1-2x) + C(1-x)^2. \tag{12.16}$$

Now letting $x = 1$ in (12.16) gives $B = -1$. Similarly, letting $x = 1/2$ gives $C = 4$. Finally, letting $x = 0$ gives $A + B + C = 1$ and so $A = -2$. Substituting these values into (12.15) finally gives

$$P_n = 4(-2) - 4(n-5) + 4(2^{n-4}) = 2^{n-2} - 4n + 12.$$

■

Problem 12.8.2. Less well-known than the Towers of Hanoi— but no less fascinating— are Towers of Sheboygan, WI. As in Hanoi, the puzzle in Sheboygan involves 3 posts and n disks of different sizes. Initially, all the disks are on post #1:



The objective is to transfer all n disks to post #2 via a sequence of moves. A move consists of removing the top disk from one post and dropping it onto another post with the restriction that a larger disk can never lie above a smaller disk. Furthermore, a local ordinance requires that *a disk can be moved only from a post to the next post on its right —or from post #3 to post #1.* Thus, for example, moving a disk directly from post #1 to post #3 is not permitted.

(a) One procedure that solves the Sheboygan puzzle is defined recursively: to move an initial stack of n disks to the next post, move the top stack of $n-1$ disks to the furthest post by moving it to the next post two times, then move the big, nth disk to the next post, and finally move the top stack another two times to land on top of the big disk. Let S_n be the number of moves that this procedure uses. Write a simple linear recurrence for S_n.

Solution.

$$\begin{aligned} &S_1 = 1, \\ &S_n = 2S_{n-1} + 1 + 2S_{n-1} = 4S_{n-1} + 1 \qquad \text{for } n > 1. \end{aligned} \tag{12.17}$$

■

(b) Let $S(x)$ be the generating function for the sequence $\langle S_0, S_1, S_2, \ldots \rangle$. Show that $S(x)$ is a quotient of polynomials.

Solution.

$$S(x) - 4xS(x) - \frac{1}{1-x} = S_1 x - 1 - x = -1$$

so

$$S(x) = \frac{\frac{1}{1-x} - 1}{1 - 4x} \tag{12.18}$$

$$= \frac{x}{(1-x)(1-4x)} \tag{12.19}$$

■

(c) Give a simple formula for S_n.

Solution. We can express $x/(1-x)(1-4x)$ using partial fractions as

$$\frac{x}{(1-x)(1-4x)} = \frac{A}{1-x} + \frac{B}{1-4x} \tag{12.20}$$

for some constants A, B. Multiplying both sides of (12.20) by the left hand denominator yields

$$x = A(1-4x) + B(1-x). \tag{12.21}$$

Letting $x = 1$ yields $A = -1/3$ and letting $x = 1/4$ yields $B = 1/3$. Now from (12.20), we have

$$S(x) = \frac{-1/3}{1-x} + \frac{1/3}{1-4x}$$

so

$$S_n = -\frac{1}{3} + \frac{1}{3}4^n = \frac{4^n - 1}{3}.$$

■

(d) A better (indeed optimal, but we won't prove this) procedure can be defined in terms of two mutually recursive procedures, procedure $P_1(n)$ for moving a stack of n disks 1 pole forward, and $P_2(n)$ for moving a stack of n disks 2 poles forward. It's obvious how to do this for $n = 1$. For $n > 1$, define:

$P_1(n)$: Apply $P_2(n-1)$ to move the top $n-1$ disks two poles forward to the third pole. Then move the remaining big disk once to land on the second pole. Then apply $P_2(n-1)$ again to move the stack of $n-1$ disks two poles forward from the third pole to land on top of the big disk.

$P_2(n)$: Apply $P_2(n-1)$ to move the top $n-1$ disks two poles forward to land on the third pole. Then move the remaining big disk to the second pole. Then apply $P_1(n-1)$ to move the stack of $n-1$ disks one pole forward to land on the first pole. Now move the big disk 1 pole forward again to land on the third pole. Finally, apply $P_2(n-1)$ again to move the stack of $n-1$ disks two poles forward to land on the big disk.

Let T_n be the number of moves needed to solve the Sheboygan puzzle using procedure $P_1(n)$. Write a simple linear recurrence for T_n.

Hint: Let R_n be the number of moves used by procedure $P_2(n)$. Express each of T_n and R_n as linear combinations of T_{n-1} and R_{n-1} and solve for T_n.

Solution. From the definitions of procedures P_1 and P_2 we have

$$\begin{aligned}
T_1 &= 1, \\
R_1 &= 2, \\
T_n &= R_{n-1} + 1 + R_{n-1} && \text{for } n > 1, && (12.22)\\
R_n &= R_{n-1} + 1 + T_{n-1} + 1 + R_{n-1} && \text{for } n > 1. && (12.23)
\end{aligned}$$

Using (12.22) to get $R_{n-1} = (T_n - 1)/2$ and then substituting for R in (12.23) with this expression in T, we conclude that for $n \geq 2$,

$$\frac{T_{n+1} - 1}{2} = (T_n - 1) + T_{n-1} + 2$$

so

$$T_{n+1} = 2T_n + 2T_{n-1} + 3,$$

and hence

$$T_n = 2T_{n-1} + 2T_{n-2} + 3, \tag{12.24}$$

for $n > 2$. ■

(e) Find a simple expression for T_n and conclude that $T_n = o(S_n)$.

Solution.

$$T_n = \frac{1}{3-\sqrt{3}}(1+\sqrt{3})^n + \frac{1}{3+\sqrt{3}}(1-\sqrt{3})^n - 1. \tag{12.25}$$

In particular, we conclude that $T_n = \Theta((1+\sqrt{3})^n)$. Since $S_n = \Theta(4^n)$, we conclude that $T_n = o(S_n)$, so the second procedure for moving a stack of n disks is significantly more efficient than the first one for large n.

To derive (12.25), we let $T(x)$ be the generating function for $\langle T_0, T_1, T_2, \ldots \rangle$ and conclude from (12.24) that

$$T(x) - 2xT(x) - 2x^2T(x) - \frac{3}{1-x} = T_1x - 3 - 3x = -(2x+3)$$

and so

$$T(x) = \frac{2x^2+x}{2x^3-3x+1} \tag{12.26}$$

The roots of the denominator of (12.26) are $x = \alpha_1^{-1}$, $x = \alpha_2^{-1}$ and $x = \alpha_3^{-1}$ where $\alpha_1 = 1+\sqrt{3}$, $\alpha_2 = 1-\sqrt{3}$ and $\alpha_3 = 1$. Therefore, $T(x)$ can be expressed using partial fractions as

$$\frac{2x^2+x}{2x^3-3x+1} = \frac{A_1}{1-\alpha_1 x} + \frac{A_2}{1-\alpha_2 x} + \frac{A_3}{1-\alpha_3 x} \tag{12.27}$$

Multiplying both sides of (12.27) by $(1-\alpha_1 x)(1-\alpha_2 x)(1-\alpha_3 x)$ gives

$$2x^2 + x = A_1(1-\alpha_2 x)(1-\alpha_3 x) + A_2(1-\alpha_1 x)(1-\alpha_3 x) + A_3(1-\alpha_1 x)(1-\alpha_2 x) \tag{12.28}$$

and for each $i \in \{1,2,3\}$ we can solve for A_i by substituting $x = \alpha_i^{-1}$ into (12.28):

$$\begin{aligned} A_1 &= \frac{1}{3-\sqrt{3}} \\ A_2 &= \frac{1}{3+\sqrt{3}} \\ A_3 &= -1 \end{aligned}$$

Since the nth-coefficient of $\frac{A_i}{1-\alpha_i x}$ is $A_i\alpha_i^n$,

$$\begin{aligned} T_n &= A_1\alpha_1^n + A_2\alpha_2^n + A_3\alpha_3^n \\ &= \frac{1}{3-\sqrt{3}}(1+\sqrt{3})^n + \frac{1}{3+\sqrt{3}}(1-\sqrt{3})^n - 1 \end{aligned}$$

■

12.9 Miniquiz May. 4

Problem 12.9.1. You would like to give a bouquet for Mother's Day, but you know nothing about flower arrangement. You google and find an online service where you just enter the number of each type of flower you want, and they make a nice bouquet and send it to your home. You decide to buy a bouquet with some number of lilies and red and white roses, but with the following restrictions:

- there must be at most 3 lilies,
- there must be at least 4 red roses,
- there must be an odd number of white roses.

Let f_n be the number of ways to compose a bouquet of n flowers satisfying the restrictions.

Find a simple closed form for $F(x)$, the generating function for the sequence $f_0, f_1, f_2, \ldots$.

Solution. Generating function for the number of ways to choose lilies:

$$F_L(x) = 1 + x + x^2 + x^3$$

Generating function for the number of ways to choose red roses:

$$F_R(x) = x^4 + x^5 + x^6 + \cdots = \frac{x^4}{1-x}$$

Generating function for the number of ways to choose white roses:

$$F_W(x) = x + x^3 + x^5 + \cdots = \frac{x}{1-x^2}$$

Therefore the generating function for f_n is

$$\begin{aligned} F(x) &= F_L(x)F_R(x)F_W(x) \\ &= \frac{x^5(1+x+x^2+x^3)}{(1-x)(1-x^2)} \\ &= \frac{x^5(1-x^4)}{(1-x)^3(1+x)} \end{aligned}$$

■

Problem 12.9.2. A sequence a_n is defined recursively by the following rules

$$\begin{aligned} a_0 &= 0 \\ a_1 &= 2 \\ a_n &= 4a_{n-1} - 3a_{n-2} + 2, \text{for } n > 1 \end{aligned}$$

Find a simple closed form for $A(x)$, the generating function for the sequence $a_0, a_1, a_2, \ldots$.

Solution.

$$\begin{array}{rcccccccccl} A(x) & = & a_0 & + & a_1x & + & a_2x^2 & + & a_3x^3 & +\cdots & \\ -4xA(x) & = & & - & 4a_0x & - & 4a_1x^2 & - & 4a_2x^3 & -\cdots & \\ 3x^2A(x) & = & & & & & 3a_0x^2 & + & 3a_1x^3 & +\cdots & \\ \hline (1-4x+3x^2)A(x) & = & & & 2x & + & 2x^2 & + & 2x^3 & +\cdots & = 2x/(1-x) \end{array}$$

Therefore

$$A(x) = \frac{2x}{(1-x)(1-4x+3x^2)} = \frac{2x}{(1-x)^2(1-3x)}$$

■

Problem 12.9.3. The following is the generating function of an infinite sequence $\langle g_0, g_1, g_2, g_3, \ldots \rangle$. Find a simple closed form for the value of g_n.

$$G(x) = \frac{e}{(1-ex)^2} + \frac{2}{(1-x)^3} - \frac{2}{1-x}$$

Solution.

$$\begin{aligned} g_n &= e(n+1)e^n + 2\binom{n+2}{2} - 2 \\ &= (n+1)e^{n+1} + n^2 + 3n \end{aligned}$$

■

Appendix

Definition 12.9.1. The *generating function* for the infinite sequence $\langle g_0, g_1, g_2, g_3, \cdots \rangle$ is the power series:

$$G(x) = g_0 + g_1x + g_2x^2 + g_3x^3 + \cdots .$$

Useful series expansion

$$\frac{1}{(1-x)^k} = \sum_{n=0}^{\infty} \binom{n+k-1}{k-1} x^n, \text{for } |x| < 1.$$

Chapter 13

Introduction to Probability

Probability is the last topic in this course and perhaps the most important. Many algorithms rely on randomization. Investigating their correctness and performance requires probability theory. Moreover, many aspects of computer systems, such as memory management, branch prediction, packet routing, and load balancing are designed around probabilistic assumptions and analyses. Probability also comes up in information theory, cryptography, artificial intelligence, and game theory. Beyond these engineering applications, an understanding of probability gives insight into many everyday issues, such as polling, DNA testing, risk assessment, investing, and gambling.

So probability is good stuff.

13.1 Monty Hall

In the September 9, 1990 issue of *Parade* magazine, the columnist Marilyn vos Savant responded to this letter:

> *Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say number 1, and the host, who knows what's behind the doors, opens another door, say number 3, which has a goat. He says to you, "Do you want to pick door number 2?" Is it to your advantage to switch your choice of doors?*
>
> Craig. F. Whitaker
> Columbia, MD

The letter roughly describes a situation faced by contestants on the 1970's game show *Let's Make a Deal*, hosted by Monty Hall and Carol Merrill. Marilyn replied that the contestant should indeed switch. But she soon received a torrent of letters— many from mathematicians— telling her that she was wrong. The problem generated thousands of hours of heated debate.

Yet this is is an elementary problem with an elementary solution. Why was there so much dispute? Apparently, most people *believe* they have an intuitive grasp of probability. (This is in stark contrast to other branches of mathematics; few people believe they have an intuitive ability to compute integrals or factor large integers!) Unfortunately, approximately 100% of those people

are *wrong*. In fact, everyone who has studied probability at length can name a half-dozen problems in which their intuition led them astray— often embarrassingly so.

The way to avoid errors is to distrust informal arguments and rely instead on a rigorous, systematic approach. If you insist on relying on intuition, then there are lots of compelling financial deals we'd love to offer you! In short: intuition *bad*, rigor *good* —at least until you've gained some solid experience with probabilities.

13.1.1 The Four-Step Method

Every probability problem involves some sort of randomized experiment, process, or game. And each such problem involves two distinct challenges:

1. How do we model the situation mathematically?

2. How do we solve the resulting mathematical problem?

In this section, we introduce a four-step approach to questions of the form, "What is the probability that —— ?" In this approach, we build a probabilistic model step-by-step, formalizing the original question in terms of that model. Remarkably, the structured thinking that this approach imposes reduces many famously-confusing problems to near triviality. For example, as you'll see, the four-step method cuts through the confusion surrounding the Monty Hall problem like a Ginsu knife. However, more complex probability questions may spin off challenging counting, summing, and approximation problems— which, fortunately, you've already spent weeks learning how to solve!

13.1.2 Clarifying the Problem

Craig's original letter to Marilyn vos Savant is a bit vague, so we must make some assumptions in order to have any hope of modeling the game formally:

1. The car is equally likely to be hidden behind each of the three doors.

2. The player is equally likely to pick each of the three doors, regardless of the car's location.

3. After the player picks a door, the host *must* open a different door with a goat behind it and offer the player the choice of staying with the original door or switching.

4. If the host has a choice of which door to open, then he is equally likely to select each of them.

In making these assumptions, we're reading a lot into Craig Whitaker's letter. Other interpretations are at least as defensible, and some actually lead to different answers. But let's accept these assumptions for now and address the question, "What is the probability that a player who switches wins the car?"

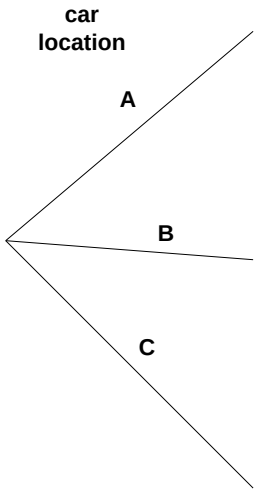
13.1.3 Step 1: Find the Sample Space

Our first objective is to identify all the possible outcomes of the experiment. A typical experiment involves several randomly-determined quantities. For example, the Monty Hall game involves three such quantities:

1. The door concealing the car.
2. The door initially chosen by the player.
3. The door that the host opens to reveal a goat.

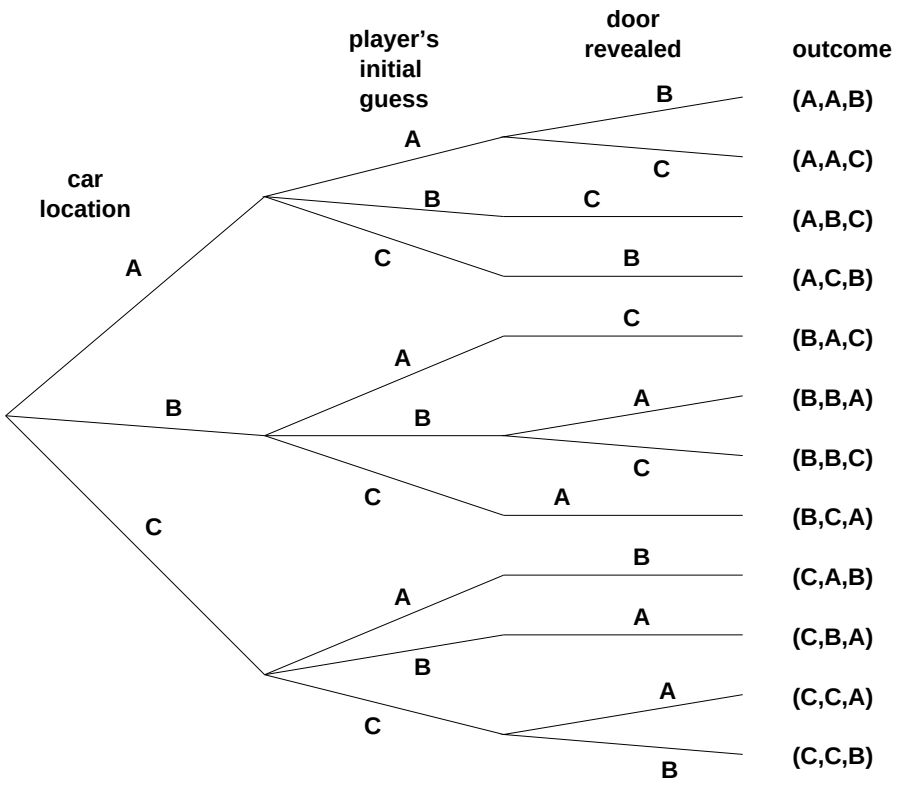
Every possible combination of these randomly-determined quantities is called an ***outcome***. The set of all possible outcomes is called the ***sample space*** for the experiment.

A ***tree diagram*** is a graphical tool that can help us work through the four-step approach when the number of outcomes is not too large or the problem is nicely structured. In particular, we can use a tree diagram to help understand the sample space of an experiment. The first randomly-determined quantity in our experiment is the door concealing the prize. We represent this as a tree with three branches:



In this diagram, the doors are called A, B, and C instead of 1, 2, and 3 because we'll be adding a lot of other numbers to the picture later.

Now, for each possible location of the prize, the player could initially choose any of the three doors. We represent this in a second layer added to the tree. Then a third layer represents the possibilities of the final step when the host opens a door to reveal a goat:



Notice that the third layer reflects the fact that the host has either one choice or two, depending on the position of the car and the door initially selected by the player. For example, if the prize is behind door A and the player picks door B, then the host must open door C. However, if the prize is behind door A and the player picks door A, then the host could open either door B or door C.

Now let's relate this picture to the terms we introduced earlier: the leaves of the tree represent *outcomes* of the experiment, and the set of all leaves represents the *sample space*. Thus, for this experiment, the sample space consists of 12 outcomes. For reference, we've labeled each outcome with a triple of doors indicating:

$$(\text{door concealing prize},\ \text{door initially chosen},\ \text{door opened to reveal a goat})$$

In these terms, the sample space is the set:

$$\mathcal{S} = \left\{ \begin{array}{llllll} (A,A,B), & (A,A,C), & (A,B,C), & (A,C,B), & (B,A,C), & (B,B,A), \\ (B,B,C), & (B,C,A), & (C,A,B), & (C,B,A), & (C,C,A), & (C,C,B) \end{array} \right\}$$

The tree diagram has a broader interpretation as well: we can regard the whole experiment as "walk" from the root down to a leaf, where the branch taken at each stage is randomly determined. Keep this interpretation in mind; we'll use it again later.

13.1.4 Step 2: Define Events of Interest

Our objective is to answer questions of the form "What is the probability that —— ?", where the horizontal line stands for some phrase such as "the player wins by switching", "the player initially

picked the door concealing the prize", or "the prize is behind door C". Almost any such phrase can be modeled mathematically as an ***event***, which is defined to be a subset of the sample space.

For example, the event that the prize is behind door C is the set of outcomes:

$$\{(C,A,B),(C,B,A),(C,C,A),(C,C,B)\}$$

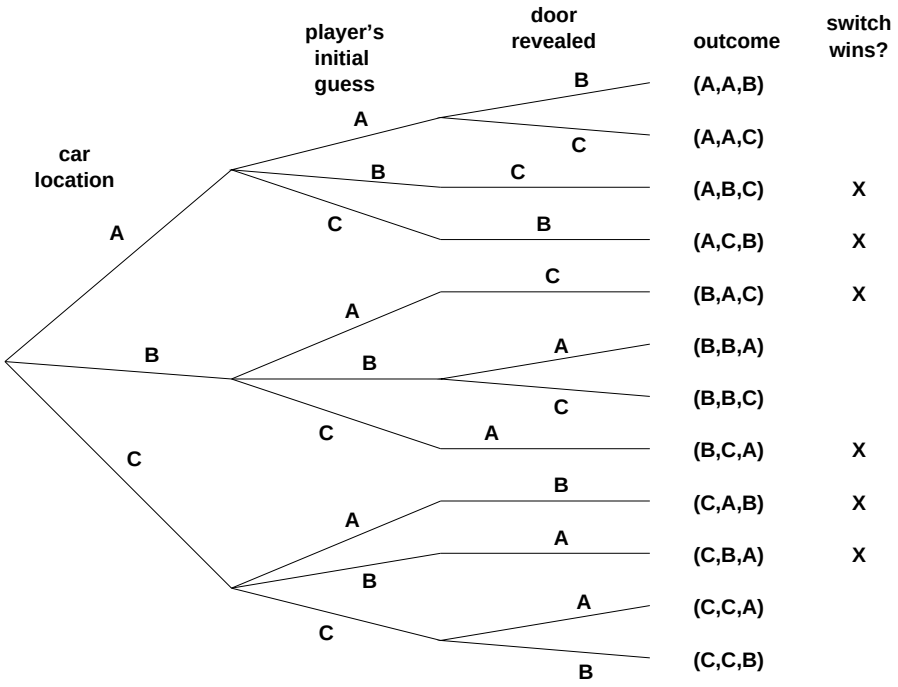
The event that the player initially picked the door concealing the prize is the set of outcomes:

$$\{(A,A,B),(A,A,C),(B,B,A),(B,B,C),(C,C,A),(C,C,B)\}$$

And what we're really after, the event that the player wins by switching, is the set of outcomes:

$$\{(A,B,C),(A,C,B),(B,A,C),(B,C,A),(C,A,B),(C,B,A)\}$$

Let's annotate our tree diagram to indicate the outcomes in this event.



Notice that exactly half of the outcomes are marked, meaning that the player wins by switching in half of all outcomes. You might be tempted to conclude that a player who switches wins with probability $1/2$. *This is wrong.* The reason is that these outcomes are not all equally likely, as we'll see shortly.

13.1.5 Step 3: Determine Outcome Probabilities

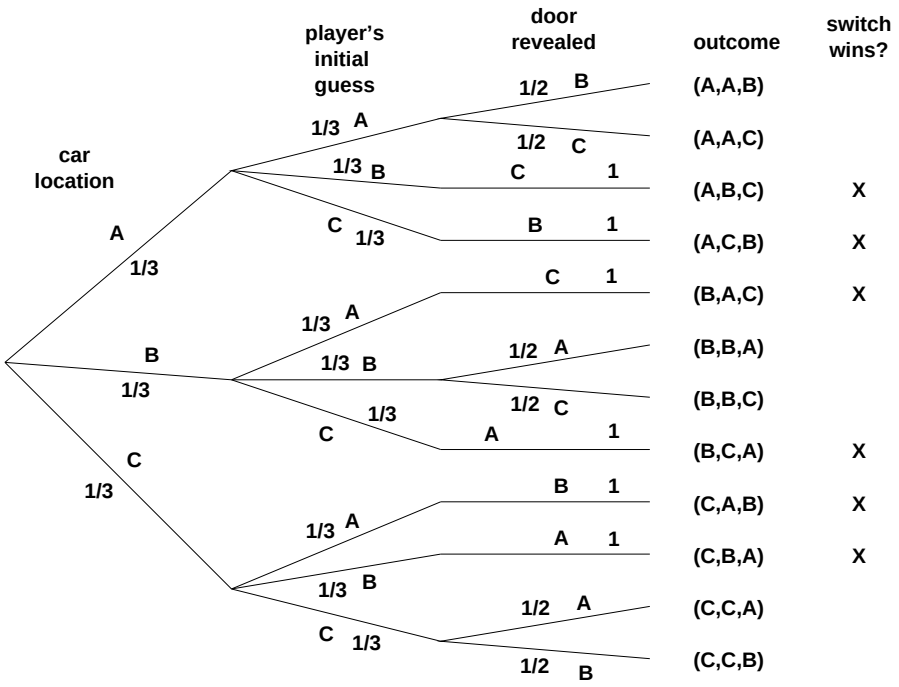
So far we've enumerated all the possible outcomes of the experiment. Now we must start assessing the likelihood of those outcomes. In particular, the goal of this step is to assign each outcome

a probability, which is a real number between 0 and 1. The sum of all outcome probabilities must be 1, reflecting the fact that exactly one outcome must occur.

Ultimately, outcome probabilities are determined by the phenomenon we're modeling and thus are not quantities that we can derive mathematically. However, mathematics can help us compute the probability of every outcome *based on fewer and more elementary modeling decisions.* In particular, we'll break the task of determining outcome probabilities into two stages.

Step 3a: Assign Edge Probabilities

First, we record a probability on each *edge* of the tree diagram. These edge-probabilities are determined by the assumptions we made at the outset: that the prize is equally likely to be behind each door, that the player is equally likely to pick each door, and that the host is equally likely to reveal each goat, if he has a choice. Notice that when the host has no choice regarding which door to open, the single branch is assigned probability 1.



Step 3b: Compute Outcome Probabilities

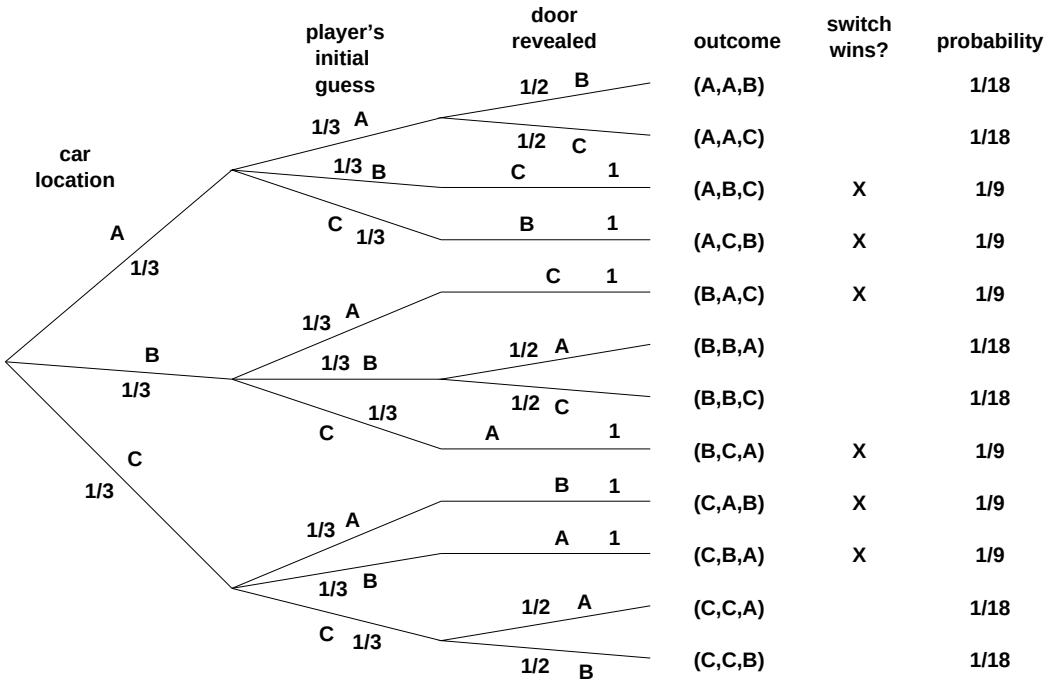
Our next job is to convert edge probabilities into outcome probabilities. This is a purely mechanical process: *the probability of an outcome is equal to the product of the edge-probabilities on the path from the root to that outcome*. For example, the probability of the topmost outcome, (A, A, B) is

$$\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{18}.$$

We'll justify this process formally later on. In the meantime, here is a nice informal justification to tide you over. Remember that the whole experiment can be regarded as a walk from the root of the tree diagram down to a leaf, where the branch taken at each step is randomly determined. In particular, the probabilities on the edges indicate how likely the walk is to proceed along each path. For example, a walk starting at the root in our example is equally likely to go down each of the three top-level branches.

Now, how likely is such a walk to arrive at the topmost outcome, (A, A, B)? Well, there is a 1-in-3 chance that a walk would follow the A-branch at the top level, a 1-in-3 chance it would continue along the A-branch at the second level, and 1-in-2 chance it would follow the B-branch at the third level. Thus, it seems that about 1 walk in 18 should arrive at the (A, A, B) leaf, which is precisely the probability we assign it.

Anyway, let's record all the outcome probabilities in our tree diagram.



Specifying the probability of each outcome amounts to defining a function that maps each outcome to a probability. This function is usually called **Pr**. In these terms, we've just determined that:

$$\Pr\{(A, A, B)\} = \frac{1}{18}$$
$$\Pr\{(A, A, C)\} = \frac{1}{18}$$
$$\Pr\{(A, B, C)\} = \frac{1}{9}$$
$$\text{etc.}$$

Earlier, we noted that the sum of all outcome probabilties must be 1 since exactly one outcome must occur. We can now express this symbolically:

$$\sum_{x \in \mathcal{S}} \Pr\{x\} = 1$$

In this equation, $\mathcal{S}$ denotes the sample space.

Though $\Pr$ is an ordinary function, just like your old friends f and g from calculus, we will subject it to all sorts of horrible notational abuses that f and g were mercifully spared. Just for starters, all of the following are common notations for the probability of an outcome x:

$$\Pr\{x\} \qquad \Pr(x) \qquad \Pr[x] \qquad \Pr x \qquad p(x)$$

A sample space $\mathcal{S}$ and a probability function $\Pr : \mathcal{S} \to [0,1]$ together form a ***probability space***. Thus, a probability space describes all possible outcomes of an experiment *and* the probability of each outcome. A probability space is a complete mathematical model of an experiment.

13.1.6 Step 4: Compute Event Probabilities

We now have a probability for each *outcome*, but we want to determine the probability of an *event*. We can bridge this gap with a definition:

> The *probability of an event* is the sum of the probabilities of the outcomes it contains.

As a notational matter, the probability of an event $E \subseteq \mathcal{S}$ is written $\Pr\{E\}$. Thus, our definition of the probability of an event can be written:

$$\Pr\{E\} ::= \sum_{x \in E} \Pr\{x\}.$$

For example, the probability of the event that the player wins by switching is:

$$\begin{aligned}\Pr\{\text{switching wins}\} &= \Pr\{A,B,C\} + \Pr\{A,C,B\} + \Pr\{B,A,C\} + \\ &\quad \Pr\{B,C,A\} + \Pr\{C,A,B\} + \Pr\{C,B,A\} \\ &= \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} \\ &= \frac{2}{3}\end{aligned}$$

It seems Marilyn's answer is correct; a player who switches doors wins the car with probability $2/3$! In contrast, a player who stays with his or her original door wins with probability $1/3$, since staying wins if and only if switching loses.

We're done with the problem! We didn't need any appeals to intuition or ingenious analogies. In fact, no mathematics more difficult than adding and multiplying fractions was required. The only hard part was resisting the temptation to leap to an "intuitively obvious" answer.

13.1.7 An Alternative Interpretation of the Monty Hall Problem

Was Marilyn really right? Our analysis suggests she was. But a more accurate conclusion is that her answer is correct *provided we accept her interpretation of the question*. There is an equally plausible interpretation in which Marilyn's answer is wrong. Notice that Craig Whitaker's original letter does not say that the host is *required* to reveal a goat and offer the player the option to switch, merely that he *did* these things. In fact, on the *Let's Make a Deal* show, Monty Hall sometimes simply opened the door that the contestant picked initially. Therefore, if he wanted to, Monty could give the option of switching only to contestants who picked the correct door initially. In this case, switching never works!

13.1.8 Probability Identities

The definitions we've introduced lead to some useful identities involving probabilities. Many probability problems can be solved quickly with such identities, once you're used to them. If E is an event, then the ***complement of*** E consists of all outcomes not in E and is denoted $\overline{E}$. The probabilities of complementary events sum to 1:

$$\Pr\{E\} + \Pr\left\{\overline{E}\right\} = 1.$$

About half of the time, the easiest way to compute the probability of an event is to compute the probability of its complement and then apply this formula.

Suppose that events $E_1, \ldots, E_n$ are disjoint; that is, every outcome is in at most one event E_i. The ***sum rule*** says that the probability of the union of these events is equal to the sum of their probabilities:

$$\Pr\{E_1 \cup \cdots \cup E_n\} = \Pr\{E_1\} + \cdots + \Pr\{E_n\}.$$

The probability of the union of events that are not necessarily disjoint is given by an ***inclusion-exclusion formula*** analogous to the one for set sizes:

$$\Pr\{E_1 \cup \cdots \cup E_n\} = \sum_i \Pr\{E_i\} - \sum_{i,j} \Pr\{E_i \cap E_j\} + \sum_{i,j,k} \Pr\{E_i \cap E_j \cap E_k\} - \cdots.$$

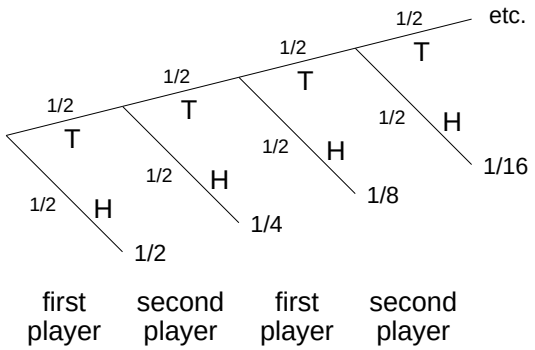
The following inequality, called the ***Union Bound***, also holds even if events $E_1, \ldots, E_n$ are not disjoint:

$$\Pr\{E_1 \cup \cdots \cup E_n\} \leq \Pr\{E_1\} + \cdots + \Pr\{E_n\}.$$

The Union Bound is simple and "good enough" for many probability calculations. For example, suppose that E_i is the event that the i-th critical component in a spacecraft fails. Then $E_1 \cup \cdots \cup E_n$ is the event that *some* critical component fails. The Union Bound gives an upper bound on this vital probability and does not require engineers to estimate all the terms in the gigantic inclusion-exclusion formula.

13.2 Infinite Sample Spaces

Suppose two players take turns flipping a fair coin. Whoever flips heads first is declared the winner. What is the probability that the first player wins? A tree diagram for this problem is shown below:



The event that the first player wins contains an infinite number of outcomes, but we can still sum their probabilities:

$$\begin{aligned}\Pr\{\text{first player wins}\} &= \frac{1}{2} + \frac{1}{8} + \frac{1}{32} + \frac{1}{128} + \cdots \\ &= \frac{2}{3}.\end{aligned}$$

The second step uses the formula for the sum of a geometric series. Similarly, we can compute the probability that the second player wins:

$$\begin{aligned}\Pr\{\text{second player wins}\} &= \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \frac{1}{256} + \cdots \\ &= \frac{1}{3}.\end{aligned}$$

In principle, the game could go on forever if both players kept flipping tails. In our tree diagram, this situation does not correspond to any leaf —rather, it corresponds to the infinite path. So, this is not an outcome. If we wanted to consider it as such, we could add an extra leaf as a child of the root. The probability on the edge to this leaf should then be the probability of the infinite path

$$\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdots = \lim_{n\to\infty} \left(\frac{1}{2}\right)^n = 0.$$

Since this probability is 0, there would be no impact on our calculations.

The mathematical machinery we've developed is adequate to model and analyze many interesting probability problems with infinite sample spaces. However, some intricate infinite processes require more powerful (and more complex) measure-theoretic notions of probability. For example, if we generate an infinite sequence of random bits $b_1, b_2, b_3, \ldots$, then what is the probability that

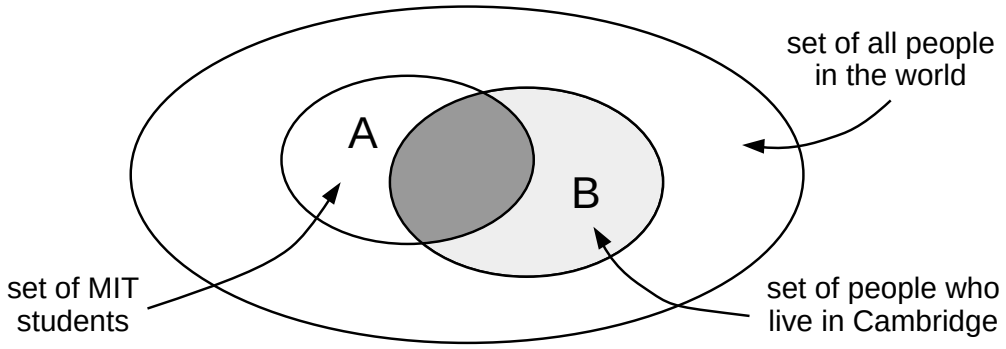
$$\frac{b_1}{2^1} + \frac{b_2}{2^2} + \frac{b_3}{2^3} + \cdots$$

is a rational number? We won't take up such problems in this course.

13.3 Conditional Probability

Suppose that we pick a random person in the world. Everyone has an equal chance of being selected. Let A be the event that the person is an MIT student, and let B be the event that the

person lives in Cambridge. What are the probabilities of these events? Intuitively, we're picking a random point in the big ellipse shown below and asking how likely that point is to fall into region A or B:



The vast majority of people in the world neither live in Cambridge nor are MIT students, so events A and B both have low probability. But what is the probability that a person is an MIT student, *given* that the person lives in Cambridge? This should be much greater— but what is it exactly?

What we're asking for is called a ***conditional probability***; that is, the probability that one event happens, given that some other event definitely happens. Questions about conditional probabilities come up all the time:

- What is the probability that it will rain this afternoon, given that it is cloudy this morning?
- What is the probability that two rolled dice sum to 10, given that both are odd?
- What is the probability that I'll get four-of-a-kind in Texas No Limit Hold 'Em Poker, given that I'm initially dealt two queens?

There is a special notation for conditional probabilities. In general, $\Pr\{A \mid B\}$ denotes the probability of event A, given that event B happens. So, in our example, $\Pr\{A \mid B\}$ is the probability that a random person is an MIT student, given that he or she is a Cambridge resident.

How do we compute $\Pr\{A \mid B\}$? Since we are *given* that the person lives in Cambridge, we can forget about everyone in the world who does not. Thus, all outcomes outside event B are irrelevant. So, intuitively, $\Pr\{A \mid B\}$ should be the fraction of Cambridge residents that are also MIT students; that is, the answer should be the probability that the person is in set $A \cap B$ (darkly shaded) divided by the probability that the person is in set B (lightly shaded). This motivates the definition of conditional probability:

$$\Pr\{A \mid B\} ::= \frac{\Pr\{A \cap B\}}{\Pr\{B\}}$$

If $\Pr\{B\} = 0$, then the conditional probability $\Pr\{A \mid B\}$ is undefined.

Pure probability is often counterintuitive, but conditional probability is worse! Conditioning can subtly alter probabilities and produce unexpected results in randomized algorithms and computer systems as well as in betting games. Yet, the mathematical definition of conditional probability given above is very simple and should give you no trouble— provided you rely on formal reasoning and not intuition.

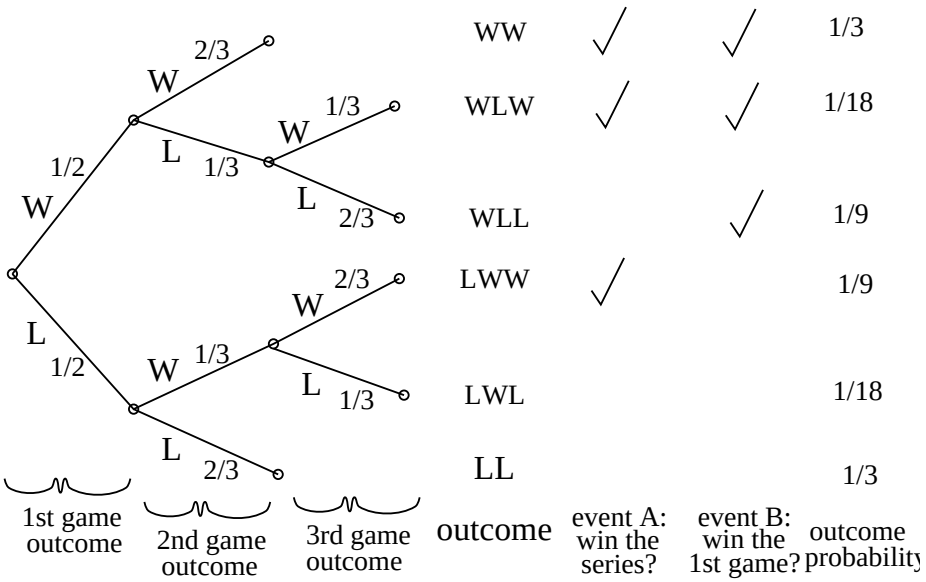
13.3.1 The Halting Problem

The *Halting Problem* is the canonical undecidable problem in computation theory that was first introduced by Alan Turing in his seminal 1936 paper. The problem is to determine whether a Turing machine halts on a given . . . blah, blah, blah. But what's *much more important*, it is the name of the MIT EECS department's famed C-league hockey team.

In a best-of-three tournament, the Halting Problem wins the first game with probability $1/2$. In subsequent games, their probability of winning is determined by the outcome of the previous game. If the Halting Problem won the previous game, then they are invigorated by victory and win the current game with probability $2/3$. If they lost the previous game, then they are demoralized by defeat and win the current game with probablity only $1/3$. What is the probability that the Halting Problem wins the tournament, given that they win the first game?

This is a question about a conditional probability. Let A be the event that the Halting Problem wins the tournament, and let B be the event that they win the first game. Our goal is then to determine the conditional probability $\Pr\{A \mid B\}$.

We can tackle conditional probability questions just like ordinary probability problems: using a tree diagram and the four-step method. A complete tree diagram is shown below, followed by an explanation of its construction and use.



Step 1: Find the Sample Space

Each internal vertex in the tree diagram has two children, one corresponding to a win for the Halting Problem (labeled W) and one corresponding to a loss (labeled L). The complete sample space is:

$$\mathcal{S} = \{WW,\ WLW,\ WLL,\ LWW,\ LWL,\ LL\}$$

Step 2: Define Events of Interest

The event that the Halting Problem wins the whole tournament is:

$$T = \{WW,\ WLW,\ LWW\}$$

And the event that the Halting Problem wins the first game is:

$$F = \{WW, WLW, WLL\}$$

The outcomes in these events are indicated with checkmarks in the tree diagram.

Step 3: Determine Outcome Probabilities

Next, we must assign a probability to each outcome. We begin by labeling edges as specified in the problem statement. Specifically, The Halting Problem has a $1/2$ chance of winning the first game, so the two edges leaving the root are each assigned probability $1/2$. Other edges are labeled $1/3$ or $2/3$ based on the outcome of the preceding game. We then find the probability of each outcome by multiplying all probabilities along the corresponding root-to-leaf path. For example, the probability of outcome WLL is:

$$\frac{1}{2} \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{1}{9}$$

Step 4: Compute Event Probabilities

We can now compute the probability that The Halting Problem wins the tournament, given that they win the first game:

$$\begin{aligned}
\Pr\{A \mid B\} &= \frac{\Pr\{A \cap B\}}{\Pr\{B\}} \\
&= \frac{\Pr\{\{WW, WLW\}\}}{\Pr\{\{WW, WLW, WLL\}\}} \\
&= \frac{1/3 + 1/18}{1/3 + 1/18 + 1/9} \\
&= \frac{7}{9}
\end{aligned}$$

We're done! If the Halting Problem wins the first game, then they win the whole tournament with probability $7/9$.

13.3.2 Why Tree Diagrams Work

We've now settled into a routine of solving probability problems using tree diagrams. But we've left a big question unaddressed: what is the mathematical justification behind those funny little pictures? Why do they work?

The answer involves conditional probabilities. In fact, the probabilities that we've been recording on the edges of tree diagrams *are* conditional probabilities. For example, consider the uppermost

path in the tree diagram for the Halting Problem, which corresponds to the outcome WW. The first edge is labeled $1/2$, which is the probability that the Halting Problem wins the first game. The second edge is labeled $2/3$, which is the probability that the Halting Problem wins the second game, *given* that they won the first— that's a conditional probability! More generally, on each edge of a tree diagram, we record the probability that the experiment proceeds along that path, given that it reaches the parent vertex.

So we've been using conditional probabilities all along. But why can we multiply edge probabilities to get outcome probabilities? For example, we concluded that:

$$\begin{aligned}\Pr\{WW\} &= \frac{1}{2}\cdot\frac{2}{3}\\ &= \frac{1}{3}\end{aligned}$$

Why is this correct?

The answer goes back to the definition of conditional probability. Rewriting this in a slightly different form gives the ***Product Rule*** for probabilities:

Definition 13.3.1 (Product Rule for 2 Events). If $\Pr\{A_2\} \neq 0$, then:

$$\Pr\{A_1 \cap A_2\} = \Pr\{A_1\} \cdot \Pr\{A_2 \mid A_1\}$$

Multiplying edge probabilities in a tree diagram amounts to evaluating the right side of this equation. For example:

$$\begin{aligned}&\Pr\{\text{win first game} \cap \text{win second game}\}\\ &\quad= \Pr\{\text{win first game}\} \cdot \Pr\{\text{win second game} \mid \text{win first game}\}\\ &\quad= \frac{1}{2}\cdot\frac{2}{3}\end{aligned}$$

So the Product Rule is the formal justification for multiplying edge probabilities to get outcome probabilities!

To justify multiplying edge probabilities along longer paths, we need a more general form of the Product Rule:

Definition 13.3.2 (Product Rule for n Events). If $\Pr\{A_1 \cap \ldots \cap A_{n-1}\} \neq 0$, then:

$$\Pr\{A_1 \cap \ldots \cap A_n\} = \Pr\{A_1\} \cdot \Pr\{A_2 \mid A_1\} \cdot \Pr\{A_3 \mid A_1 \cap A_2\} \cdots \Pr\{A_n \mid A_1 \cap \ldots \cap A_{n-1}\}$$

Let's interpret this big formula in terms of tree diagrams. Suppose we want to compute the probability that an experiment traverses a particular root-to-leaf path of length n. Let A_i be the event that the experiment traverses the i-th edge of the path. Then $A_1 \cap \ldots \cap A_n$ is the event that the experiment traverses the whole path. The Product Rule says that the probability of this is the probability that the experiment takes the first edge times the probability that it takes the second, *given* it takes the first edge, times the probability it takes the third, *given* it takes the first two edges, and so forth. In other words, the probability of an outcome is the product of the edge probabilities along the corresponding root-to-leaf path.

13.3.3 The Law of Total Probability

The following identity

$$\Pr\{A\} = \Pr\{A \mid E\} \cdot \Pr\{E\} + \Pr\left\{A \mid \overline{E}\right\} \cdot \Pr\left\{\overline{E}\right\}.$$

is called the Law of Total Probability and lets you compute the probability of an event A using case analysis based on whether or not event E occurs. For example, suppose we conduct the following experiment. First, we flip a coin. If heads comes up, then we roll one die and take the result. If tails comes up, then we roll two dice and take the sum of the two results. What is the probability that this process yields a 2? Let E be the event that the coin comes up heads, and let A be the event that we get a 2 overall. Assuming that the coin is fair, $\Pr\{E\} = \Pr\left\{\overline{E}\right\} = 1/2$. There are now two cases. If we flip heads, then we roll a 2 on a single die with probabilty $\Pr\{A \mid E\} = 1/6$. On the other hand, if we flip tails, then we get a sum of 2 on two dice with probability $\Pr\left\{A \mid \overline{E}\right\} = 1/36$. Therefore, the probability that the whole process yields a 2 is

$$\Pr\{A\} = \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{1}{36} = \frac{7}{72}.$$

More generally, if $E_1, \ldots, E_n$ are disjoint events whose union is the whole sample space, then:

$$\Pr\{A\} = \sum_{i=1}^{n} \Pr\{A \mid E_i\} \cdot \Pr\{E_i\}.$$

13.3.4 *A Posteriori* Probabilities

Suppose that we turn the hockey question around: what is the probability that the Halting Problem won their first game, given that they won the series?

This seems like an absurd question! After all, if the Halting Problem won the series, then the winner of the first game has already been determined. Therefore, who won the first game is a question of fact, not a question of probability. However, our mathematical theory of probability contains no notion of one event preceding another— there is no notion of time at all. Therefore, from a mathematical perspective, this is a perfectly valid question. And this is also a meaningful question from a practical perspective. Suppose that you're told that the Halting Problem won the series, but not told the results of individual games. Then, from your perspective, it makes perfect sense to wonder how likely it is that The Halting Problem won the first game.

A conditional probability $\Pr\{B \mid A\}$ is called ***a posteriori*** if event B precedes event A in time. Here are some other examples of a posteriori probabilities:

- The probability it was cloudy this morning, given that it rained in the afternoon.
- The probability that I was initially dealt two queens in Texas No Limit Hold 'Em poker, given that I eventually got four-of-a-kind.

Mathematically, a posteriori probabilities are *no different* from ordinary probabilities; the distinction is only at a higher, philosophical level. Our only reason for drawing attention to them is to say, "Don't let them rattle you."

Let's return to the original problem. The probability that the Halting Problem won their first game, given that they won the series is $\Pr\{B \mid A\}$. We can compute this using the definition of conditional probability and our earlier tree diagram:

$$\begin{aligned}\Pr\{B \mid A\} &= \frac{\Pr\{B \cap A\}}{\Pr\{A\}} \\ &= \frac{1/3 + 1/18}{1/3 + 1/18 + 1/9} \\ &= \frac{7}{9}\end{aligned}$$

This answer is suspicious! In the preceding section, we showed that $\Pr\{A \mid B\}$ was also $7/9$. Could it be true that $\Pr\{A \mid B\} = \Pr\{B \mid A\}$ in general? Some reflection suggests this is unlikely. For example, the probability that I feel uneasy, given that I was abducted by aliens, is pretty large. But the probability that I was abducted by aliens, given that I feel uneasy, is rather small.

Let's work out the general conditions under which $\Pr\{A \mid B\} = \Pr\{B \mid A\}$. By the definition of conditional probability, this equation holds if an only if:

$$\frac{\Pr\{A \cap B\}}{\Pr\{B\}} = \frac{\Pr\{A \cap B\}}{\Pr\{A\}}$$

This equation, in turn, holds only if the denominators are equal or the numerator is 0:

$$\Pr\{B\} = \Pr\{A\} \qquad \text{or} \qquad \Pr\{A \cap B\} = 0$$

The former condition holds in the hockey example; the probability that the Halting Problem wins the series (event A) is equal to the probability that it wins the first game (event B). In fact, both probabilities are $1/2$.

13.3.5 Medical Testing

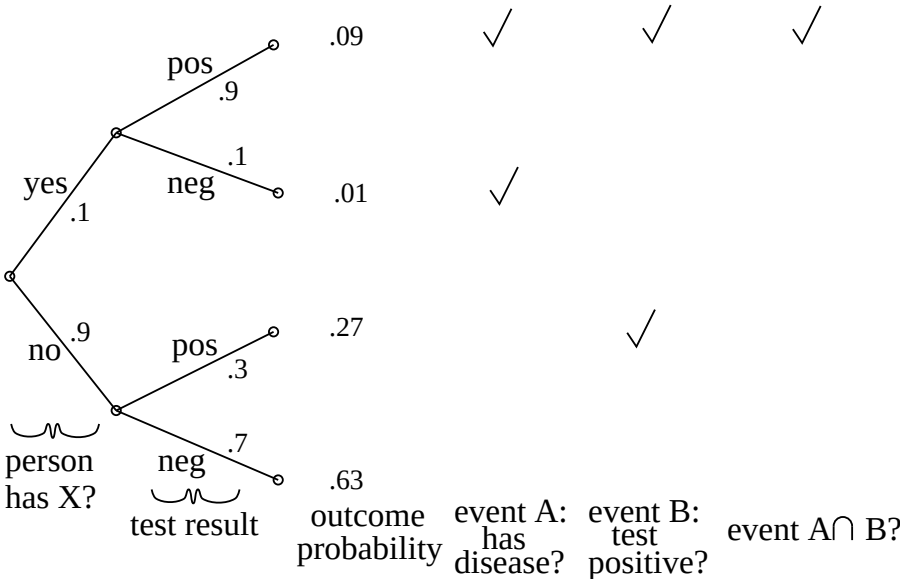
There is a deadly disease called X that has infected 10% of the population. There are no symptoms; victims just drop dead one day. Fortunately, there is a test for the disease. The test is not perfect, however:

- If you have the disease, there is a 10% chance that the test will say you do not. (These are called "false negatives".)
- If you do not have the disease, there is a 30% chance that the test will say you do. (These are "false positives".)

A random person is tested for the disease. If the test is positive, then what is the probability that the person has the disease?

Step 1: Find the Sample Space

The sample space is found with the tree diagram below.



Step 2: Define Events of Interest

Let A be the event that the person has the disease. Let B be the event that the test was positive. The outcomes in each event are marked in the tree diagram. We want to find $\Pr\{A \mid B\}$, the probability that a person has disease X, given that the test was positive.

Step 3: Find Outcome Probabilities

First, we assign probabilities to edges. These probabilities are drawn directly from the problem statement. By the Product Rule, the probability of an outcome is the product of the probabilities on the corresponding root-to-leaf path. All probabilities are shown in the figure.

Step 4: Compute Event Probabilities

$$\begin{aligned}\Pr\{A \mid B\} &= \frac{\Pr\{A \cap B\}}{\Pr\{B\}} \\ &= \frac{0.09}{0.09 + 0.27} \\ &= \frac{1}{4}\end{aligned}$$

If you test positive, then there is only a 25% chance that you have the disease!

This answer is initially surprising, but makes sense on reflection. There are two ways you could test positive. First, it could be that you are sick and the test is correct. Second, it could be that you

are healthy and the test is incorrect. The problem is that almost everyone is healthy; therefore, most of the positive results arise from incorrect tests of healthy people!

We can also compute the probability that the test is correct for a random person. This event consists of two outcomes. The person could be sick and the test positive (probability 0.09), or the person could be healthy and the test negative (probability 0.63). Therefore, the test is correct with probability $0.09 + 0.63 = 0.72$. This is a relief; the test is correct almost three-quarters of the time.

But wait! There is a simple way to make the test correct 90% of the time: always return a negative result! This "test" gives the right answer for all healthy people and the wrong answer only for the 10% that actually have the disease. The best strategy is to completely ignore the test result!

There is a similar paradox in weather forecasting. During winter, almost all days in Boston are wet and overcast. Predicting miserable weather every day may be more accurate than really trying to get it right!

13.3.6 Other Identities

There is a close relationship between computing the size of a set and computing the probability of an event. The inclusion-exclusion formula is one such example; the probability of a union of events and the cardinality of a union of sets are computed using similar formulas.

In fact, all of the methods we developed for computing sizes of sets carry over to computing probabilities. This is because a probability space is just a weighted set; the sample space is the set and the probability function assigns a weight to each element. Earlier, we were counting the number of items in a set. Now, when we compute the probability of an event, we are just summing the weights of items. We'll see many examples of the close relationship between probability and counting over the next few weeks.

Many general probability identities still hold when all probabilities are conditioned on the same event. For example, the following identity is analogous to the Inclusion-Exclusion formula for two sets, except that all probabilities are conditioned on an event C.

$$\Pr\{A \cup B \mid C\} = \Pr\{A \mid C\} + \Pr\{B \mid C\} - \Pr\{A \cap B \mid C\}.$$

As a special case we have

$$\Pr\{A \cup B \mid C\} = \Pr\{A \mid C\} + \Pr\{B \mid C\} \qquad \text{when } A \cap B = \emptyset.$$

Be careful not to mix up events before and after the conditioning bar! For example, the following is *not* a valid identity:

False Claim.

$$\Pr\{A \mid B \cup C\} = \Pr\{A \mid B\} + \Pr\{A \mid C\} \qquad \textit{when } B \cap C = \emptyset. \tag{13.1}$$

13.4 Independence

Suppose that we flip two fair coins simultaneously on opposite sides of a room. Intuitively, the way one coin lands does not affect the way the other coin lands. The mathematical concept that captures this intuition is called ***independence***:

Definition. Events A and B are independent if and only if:

$$\Pr\{A \cap B\} = \Pr\{A\} \cdot \Pr\{B\}$$

Generally, independence is something you *assume* in modeling a phenomenon— or wish you could realistically assume. Many useful probability formulas only hold if certain events are independent, so a dash of independence can greatly simplify the analysis of a system.

13.4.1 Examples

Let's return to the experiment of flipping two fair coins. Let A be the event that the first coin comes up heads, and let B be the event that the second coin is heads. If we assume that A and B are independent, then the probability that both coins come up heads is:

$$\begin{aligned}\Pr\{A \cap B\} &= \Pr\{A\} \cdot \Pr\{B\} \\ &= \frac{1}{2} \cdot \frac{1}{2} \\ &= \frac{1}{4}\end{aligned}$$

On the other hand, let C be the event that tomorrow is cloudy and R be the event that tomorrow is rainy. Perhaps $\Pr\{C\} = 1/5$ and $\Pr\{R\} = 1/10$ around here. If these events were independent, then we could conclude that the probability of a rainy, cloudy day was quite small:

$$\begin{aligned}\Pr\{R \cap C\} &= \Pr\{R\} \cdot \Pr\{C\} \\ &= \frac{1}{5} \cdot \frac{1}{10} \\ &= \frac{1}{50}\end{aligned}$$

Unfortunately, these events are definitely not independent; in particular, every rainy day is cloudy. Thus, the probability of a rainy, cloudy day is actually $1/10$.

13.4.2 Working with Independence

There is another way to think about independence that you may find more intuitive. According to the definition, events A and B are independent if and only if $\Pr\{A \cap B\} = \Pr\{A\} \cdot \Pr\{B\}$. This equation holds even if $\Pr\{B\} = 0$, but assuming it is not, we can divide both sides by $\Pr\{B\}$ and use the definition of conditional probability to obtain an alternative formulation of independence:

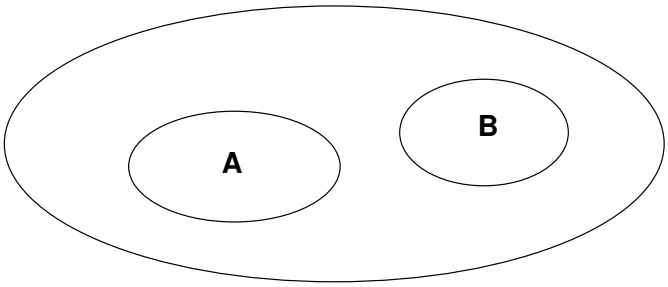
Proposition. *If* $\Pr\{B\} \neq 0$*, then events* A *and* B *are independent if and only if*

$$\Pr\{A \mid B\} = \Pr\{A\}. \tag{13.2}$$

Equation (13.2) says that events A and B are independent if the probability of A is unaffected by the fact that B happens. In these terms, the two coin tosses of the previous section were independent, because the probability that one coin comes up heads is unaffected by the fact that the other came up heads. Turning to our other example, the probability of clouds in the sky is strongly affected by the fact that it is raining. So, as we noted before, these events are not independent.

13.4.3 Some Intuition

Suppose that A and B are disjoint events, as shown in the figure below.



Are these events independent? Let's check. On one hand, we know

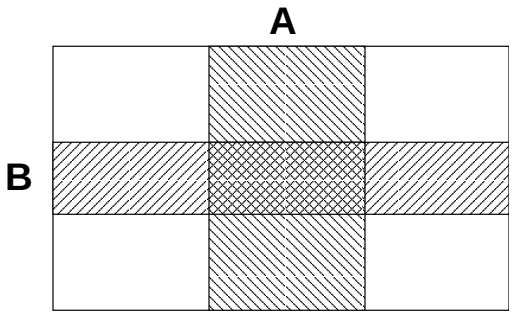
$$\Pr\{A \cap B\} = 0$$

because $A \cap B$ contains no outcomes. On the other hand, we have

$$\Pr\{A\} \cdot \Pr\{B\} > 0$$

except in degenerate cases where A or B has zero probability. Thus, *disjointness and independence are very different ideas*.

Here's a better mental picture of what independent events look like.



The sample space is the whole rectangle. Event A is a vertical stripe, and event B is a horizontal stripe. Assume that the probability of each event is proportional to its area in the diagram. Now if A covers an α-fraction of the sample space, and B covers a β-fraction, then the area of the intersection region is $\alpha \cdot \beta$. In terms of probability:

$$\Pr\{A \cap B\} = \Pr\{A\} \cdot \Pr\{B\}$$

13.5 Mutual Independence

We have defined what it means for two events to be independent. But how can we talk about independence when there are more than two events? For example, how can we say that the orientations of n coins are all independent of one another?

Events $E_1, \ldots, E_n$ are ***mutually independent*** if and only if *for every subset* of the events, the probability of the intersection is the product of the probabilities. In other words, all of the following equations must hold:

$$\begin{aligned}
\Pr\{E_i \cap E_j\} &= \Pr\{E_i\} \cdot \Pr\{E_j\} && \text{for all distinct } i, j \\
\Pr\{E_i \cap E_j \cap E_k\} &= \Pr\{E_i\} \cdot \Pr\{E_j\} \cdot \Pr\{E_k\} && \text{for all distinct } i, j, k \\
\Pr\{E_i \cap E_j \cap E_k \cap E_l\} &= \Pr\{E_i\} \cdot \Pr\{E_j\} \cdot \Pr\{E_k\} \cdot \Pr\{E_l\} && \text{for all distinct } i, j, k, l \\
&\cdots \\
\Pr\{E_1 \cap \cdots \cap E_n\} &= \Pr\{E_1\} \cdots \Pr\{E_n\}
\end{aligned}$$

As an example, if we toss 100 fair coins and let E_i be the event that the ith coin lands heads, then we might reasonably assume that $E_1, \ldots, E_{100}$ are mutually independent.

13.5.1 DNA Testing

This is testimony from the O. J. Simpson murder trial on May 15, 1995:

MR. CLARKE: When you make these estimations of frequency— and I believe you touched a little bit on a concept called independence?

DR. COTTON: Yes, I did.

MR. CLARKE: And what is that again?

DR. COTTON: It means whether or not you inherit one allele that you have is not— does not affect the second allele that you might get. That is, if you inherit a band at 5,000 base pairs, that doesn't mean you'll automatically or with some probability inherit one at 6,000. What you inherit from one parent is what you inherit from the other. *(Got that? – EAL)*

MR. CLARKE: Why is that important?

DR. COTTON: Mathematically that's important because if that were not the case, it would be improper to multiply the frequencies between the different genetic locations.

MR. CLARKE: How do you— well, first of all, are these markers independent that you've described in your testing in this case?

The jury was told that genetic markers in blood found at the crime scene matched Simpson's. Furthermore, the probability that the markers would be found in a randomly-selected person was at most 1 in 170 million. This astronomical figure was derived from statistics such as:

- 1 person in 100 has marker A.
- 1 person in 50 marker B.
- 1 person in 40 has marker C.
- 1 person in 5 has marker D.
- 1 person in 170 has marker E.

Then these numbers were multiplied to give the probability that a randomly-selected person would have all five markers:

$$\begin{aligned}\Pr\{A \cap B \cap C \cap D \cap E\} &= \Pr\{A\} \cdot \Pr\{B\} \cdot \Pr\{C\} \cdot \Pr\{D\} \cdot \Pr\{E\} \\ &= \frac{1}{100} \cdot \frac{1}{50} \cdot \frac{1}{40} \cdot \frac{1}{5} \cdot \frac{1}{170} \\ &= \frac{1}{170,000,000}\end{aligned}$$

The defense pointed out that this assumes that the markers appear mutually independently. Furthermore, all the statistics were based on just a few hundred blood samples. The jury was widely mocked for failing to "understand" the DNA evidence. If you were a juror, would *you* accept the 1 in 170 million calculation?

13.5.2 Pairwise Independence

The definition of mutual independence seems awfully complicated— there are so many conditions! Here's an example that illustrates the subtlety of independence when more than two events are involved and the need for all those conditions. Suppose that we flip three fair, mutually-independent coins. Define the following events:

- A_1 is the event that coin 1 matches coin 2.
- A_2 is the event that coin 2 matches coin 3.
- A_3 is the event that coin 3 matches coin 1.

Are A_1, A_2, A_3 mutually independent?

The sample space for this experiment is:

$$\{HHH,\ HHT,\ HTH,\ HTT,\ THH,\ THT,\ TTH,\ TTT\}$$

Every outcome has probability $(1/2)^3 = 1/8$ by our assumption that the coins are mutually independent.

To see if events A_1, A_2, and A_3 are mutually independent, we must check a sequence of equalities. It will be helpful first to compute the probability of each event A_i:

$$\begin{aligned}\Pr\{A_1\} &= \Pr\{HHH\} + \Pr\{HHT\} + \Pr\{TTH\} + \Pr\{TTT\} \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \\ &= \frac{1}{2}\end{aligned}$$

By symmetry, $\Pr\{A_2\} = \Pr\{A_3\} = 1/2$ as well. Now we can begin checking all the equalities required for mutual independence.

$$\begin{aligned}\Pr\{A_1 \cap A_2\} &= \Pr\{HHH\} + \Pr\{TTT\} \\ &= \frac{1}{8} + \frac{1}{8} \\ &= \frac{1}{4} \\ &= \frac{1}{2} \cdot \frac{1}{2} \\ &= \Pr\{A_1\} \Pr\{A_2\}\end{aligned}$$

By symmetry, $\Pr\{A_1 \cap A_3\} = \Pr\{A_1\} \cdot \Pr\{A_3\}$ and $\Pr\{A_2 \cap A_3\} = \Pr\{A_2\} \cdot \Pr\{A_3\}$ must hold also. Finally, we must check one last condition:

$$\begin{aligned}\Pr\{A_1 \cap A_2 \cap A_3\} &= \Pr\{HHH\} + \Pr\{TTT\} \\ &= \frac{1}{8} + \frac{1}{8} \\ &= \frac{1}{4} \\ &\neq \Pr\{A_1\} \Pr\{A_2\} \Pr\{A_3\} = \frac{1}{8}\end{aligned}$$

The three events A_1, A_2, and A_3 are not mutually independent, even though all *pairs* of events are independent!

A set of events is ***pairwise independent*** if every pair is independent. Pairwise independence is a much weaker property than mutual independence. For example, suppose that the prosecutors in the O. J. Simpson trial were wrong and markers A, B, C, D, and E appear only *pairwise* independently. Then the probability that a randomly-selected person has all five markers is no more than:

$$\begin{aligned}\Pr\{A \cap B \cap C \cap D \cap E\} &\leq \Pr\{A \cap E\} \\ &= \Pr\{A\} \cdot \Pr\{E\} \\ &= \frac{1}{100} \cdot \frac{1}{170} \\ &= \frac{1}{17,000}\end{aligned}$$

The first line uses the fact that $A \cap B \cap C \cap D \cap E$ is a subset of $A \cap E$. (We picked out the A and E markers because they're the rarest.) We use pairwise independence on the second line. Now the probability of a random match is 1 in 17,000— a far cry from 1 in 170 million! And this is the strongest conclusion we can reach assuming only pairwise independence.

13.6 In-Class Problems Week 12, Mon.

Problem 13.6.1. [A Baseball Series] The New York Yankees and the Boston Red Sox are playing a two-out-of-three series. (In other words, they play until one team has won two games. Then that team is declared the overall winner and the series ends.) Assume that the Red Sox win each game with probability $3/5$, regardless of the outcomes of previous games.

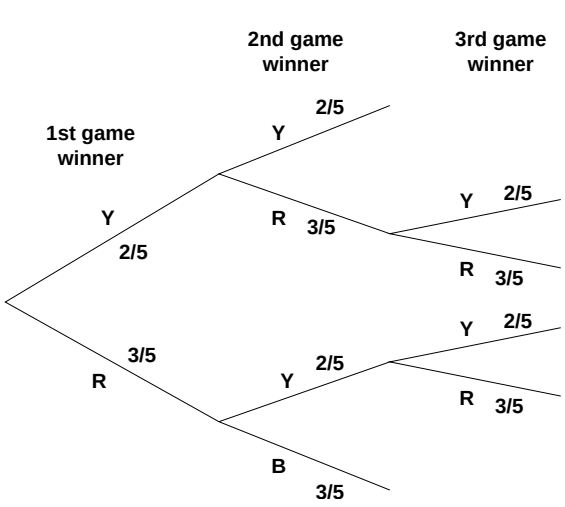
Answer the questions below using the four-step method. You can use the same tree diagram for all three problems.

(a) What is the probability that a total of 3 games are played?

(b) What is the probability that the winner of the series loses the first game?

(c) What is the probability that the *correct* team wins the series?

Solution. A tree diagram is worked out below.



outcome	3 games played?	winner lost first game?	correct team wins?	Probability
YY				4/25
YRY	X			12/125
YRR	X	X	X	18/125
RYY	X	X		12/125
RYR	X		X	18/125
RR			X	9/25

From the tree diagram, we get:

$$\Pr\{\text{3 games played}\} = \frac{12}{125} + \frac{18}{125} + \frac{12}{125} + \frac{18}{125} = \frac{12}{25}$$

$$\Pr\{\text{winner lost first game}\} = \frac{18}{125} + \frac{12}{125} = \frac{6}{25}$$

$$\Pr\{\text{correct team wins}\} = \frac{18}{125} + \frac{18}{125} + \frac{9}{25} = \frac{81}{125}$$

■

Problem 13.6.2. [The Four-Door Deal] Suppose that *Let's Make a Deal* is played according to different rules. Now there are **four** doors, with a prize hidden behind one of them. The contestant is allowed to pick a door. The host must then reveal a different door that has no prize behind it. The contestant is allowed to stay with his or her original door or to pick one of the other two that are still closed. If the contestant chooses the door concealing the prize in this second stage, then he or she wins.

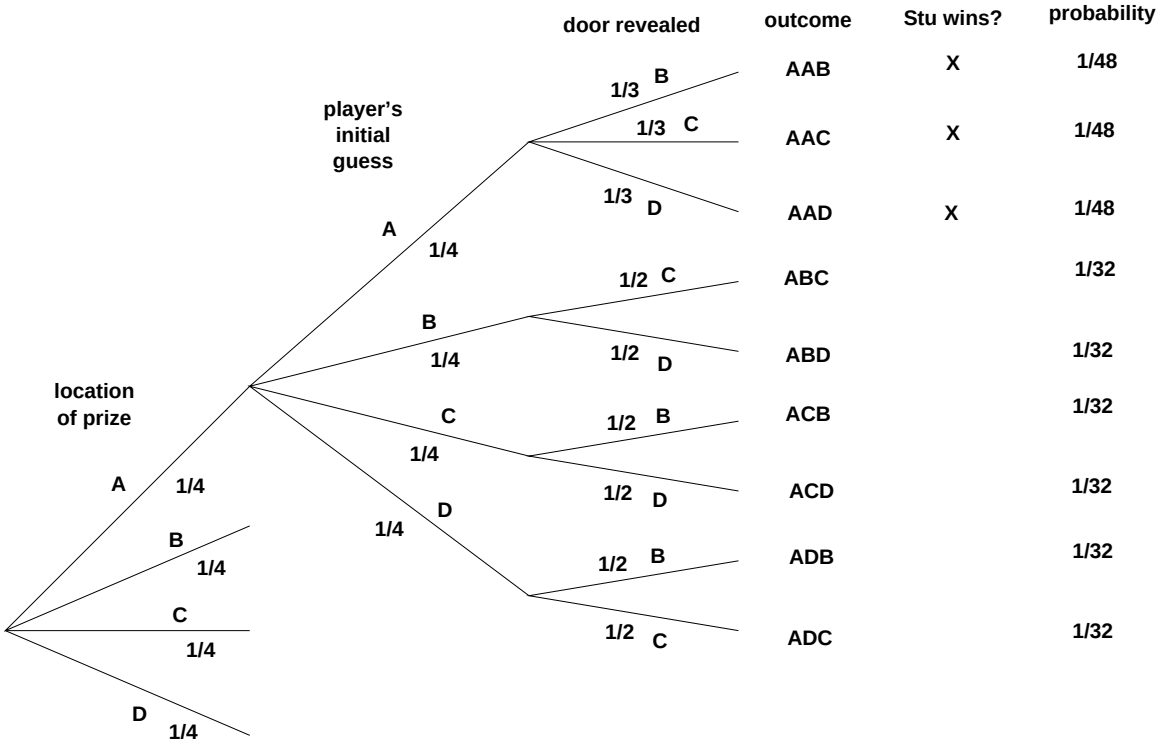
(a) Contestant Stu, a sanitation engineer from Trenton, New Jersey, stays with his original door. What is the probability that he wins the prize?

The tree diagram is awkwardly large. This often happens; in fact, sometimes you'll encounter *infinite* tree diagrams! Try to draw enough of the diagram so that you understand the structure of the remainder.

Solution. Let's make the same assumptions as in the original problem:

1. The prize is equally likely to be behind each door.
2. The contestant is equally likely to pick each door initially, regardless of the prize's location.
3. The host is equally likely to reveal each door that does not conceal the prize and was not selected by the player.

A partial tree diagram is shown below. The remaining subtrees are symmetric to the fully-expanded subtree.



The probability that Stu wins the prize is:

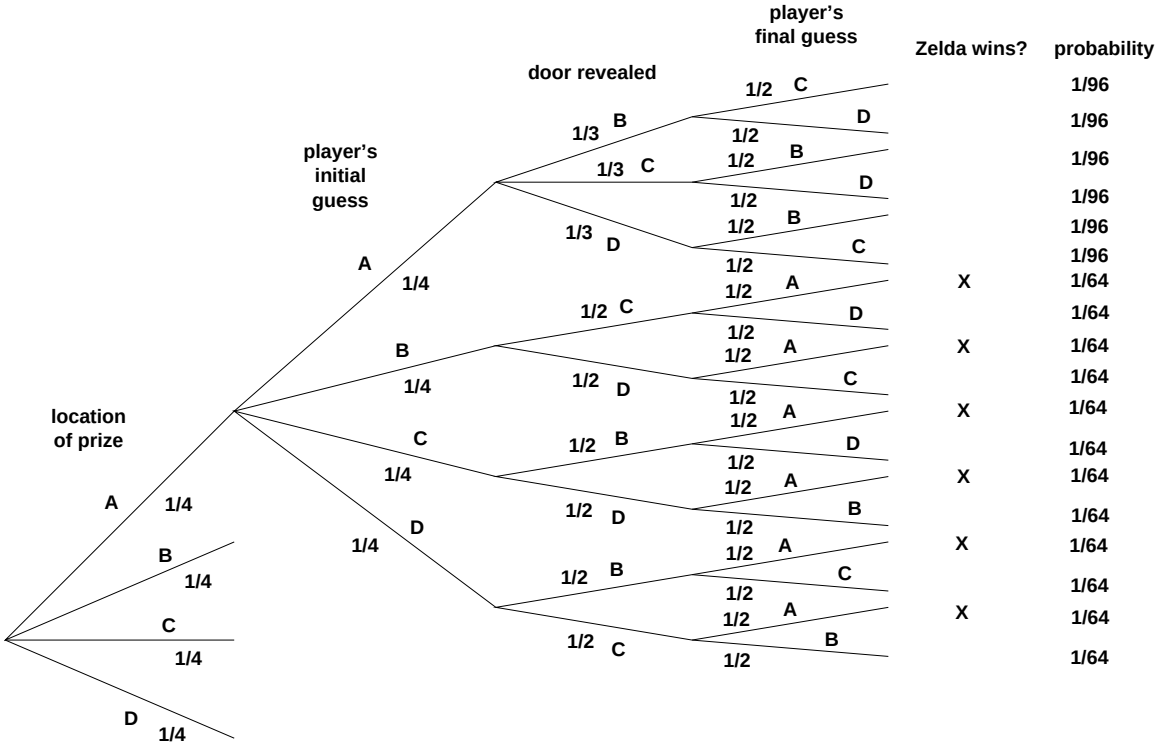
$$\Pr\{\text{Stu wins}\} = 4 \cdot \left(\frac{1}{48} + \frac{1}{48} + \frac{1}{48}\right) = \frac{1}{4}$$

We multiply by 4 to account for the four subtrees, of which we've only drawn one.

Notice that we expanded the tree out to the third ("door revealed") level to spell out the outcomes, but in this case we could, in fact, have stopped at the second level ("player's initial guess). This follows because the win/lose outcome is determined by the prize location and Stu's selected door, regardless of what happens after that. ■

(b) Contestant Zelda, an alien abduction researcher from Helena, Montana, switches to one of the remaining two doors with equal probability. What is the probability that she wins the prize?

Solution. A partial tree diagram is worked out below.



The probability that Zelda wins the prize is:

$$\Pr\{\text{Zelda wins}\} = 4 \cdot \left(\frac{1}{64} + \frac{1}{64} + \frac{1}{64} + \frac{1}{64} + \frac{1}{64} + \frac{1}{64}\right) = \frac{3}{8}$$

■

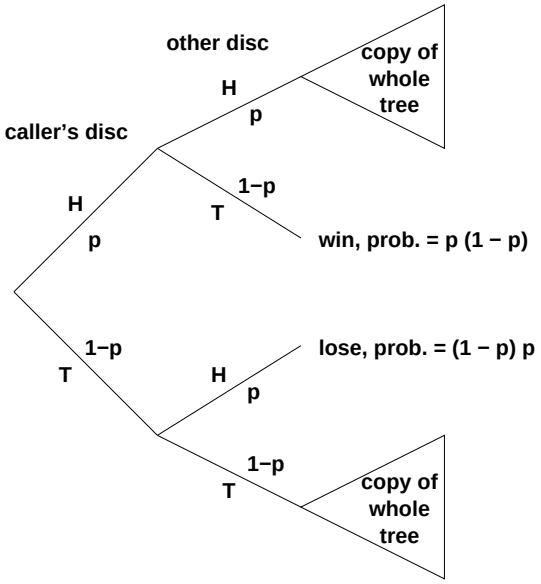
Problem 13.6.3. Suppose that the rules of Ultimate Frisbee were revised as follows:

> *Representatives of the two teams each flip a disc. The representative of one team predicts the orientation of his or her own disc by calling "up" or "down" while the discs are in the air. However, if both discs land the same way, then the call does not count and the process starts over.*

Assume that a disc lands face-up with probability p. What is the probability that the caller wins by always saying "up"?

Suggestions: The tree diagram is infinite, so draw only enough to see a pattern. Summing all the winning outcome probabilities directly is difficult. However, a neat trick solves this problem and many others. Let s be the sum of all winning outcome probabilities in the whole tree. Notice that *you can write the sum of all the winning probabilities in certain subtrees as a function of* s. Use this observation to write an equation in s and then solve.

Solution. In the tree diagram below, the small triangles represent subtrees that are themselves complete copies of the whole tree.



Let s equal the sum of all winning probabilities in the whole tree. There are two extra edges with probability p on the path to each outcome in the top subtree. Therefore, the sum of winning probabilities in the upper tree is $p^2 s$. Similarly, the sum of winning probabilities in the lower subtree is $(1-p)^2 s$. This gives the equation:

$$s = p^2 s + (1-p)^2 s + p(1-p)$$

The solution to this equation is $s = 1/2$, for all p between 0 and 1. ■

Problem 13.6.4. Here are some handy rules for reasoning about probabilities that all follow directly from the Sum Rule for the probabilities of disjoint events. Prove them:

(a) The *Difference Rule*:

$$\Pr\{A - B\} = \Pr\{A\} - \Pr\{A \cap B\}$$

Solution. Any set A is the disjoint union of $A - B$ and $A \cap B$, so

$$\Pr\{A\} = \Pr\{A - B\} + \Pr\{A \cap B\}$$

by the Disjoint Sum Rule. ■

(b) The *Complement Rule*:

$$\Pr\left\{\overline{A}\right\} = 1 - \Pr\{A\}$$

Solution. $\overline{A} ::= \mathcal{S} - A$, so by part (a),

$$\Pr\left\{\overline{A}\right\} = \Pr\{\mathcal{S}\} - \Pr\{A\} = 1 - \Pr\{A\}.$$

■

(c) *Inclusion-Exclusion*:

$$\Pr\{A \cup B\} = \Pr\{A\} + \Pr\{B\} - \Pr\{A \cap B\}$$

Solution. $A \cup B$ is the disjoint union of A and $B - A$ so

$$\begin{aligned}
\Pr\{A \cup B\} &= \Pr\{A\} + \Pr\{B - A\} && \text{(Disjoint Sum Rule)} \\
&= \Pr\{A\} + (\Pr\{B\} - \Pr\{A \cap B\}) && \text{(Difference Rule)}
\end{aligned}$$

■

(d) The *Union Bound*:

$$\Pr\{A \cup B\} \leq \Pr\{A\} + \Pr\{B\}.$$

Solution. This follows immediately from Inclusion-Exclusion and the fact that $\Pr\{A \cap B\} \geq 0$. ■

(e) *Monotonicity*:

$$\text{If } A \subseteq B, \text{ then } \Pr\{A\} \leq \Pr\{B\}.$$

Solution.

$$\begin{aligned}
\Pr\{A\} &= \Pr\{B\} - (\Pr\{B\} - \Pr\{A\}) \\
&= \Pr\{B\} - (\Pr\{B\} - \Pr\{A \cap B\}) && (\text{since } A = A \cap B) \\
&= \Pr\{B\} - \Pr\{B - A\} && (\text{difference rule}) \\
&\leq \Pr\{B\} && (\text{since } \Pr\{B - A\} \geq 0).
\end{aligned}$$

■

The Four-Step Method

This is a good approach to questions of the form, "What is the probability that ——-?" Intuition *will* mislead you, but this formal approach gives the right answer every time.

1. Find the sample space. (Use a tree diagram.)

2. Define events of interest. (Mark leaves corresponding to these events.)
3. Determine outcome probabilities:
   (a) Assign edge probabilities.
   (b) Compute outcome probabilities. (Multiply along root-to-leaf paths.)
4. Compute event probabilities. (Sum the probabilities of all outcomes in the event.)

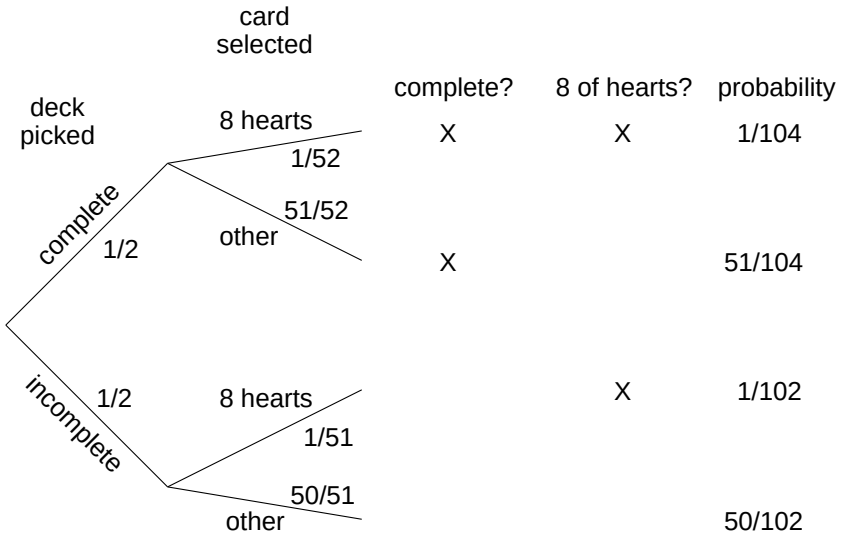
13.7 In-Class Problems Week 12, Wed.

Problem 13.7.1. There are two decks of cards. One is complete, but the other is missing the ace of spades. Suppose you pick one of the two decks with equal probability and then select a card from that deck uniformly at random. What is the probability that you picked the complete deck, given that you selected the eight of hearts? Use the four-step method and a tree diagram.

Solution. Let C be the event that you pick the complete deck, and let H be the event that you select the eight of hearts. In these terms, our aim is to compute:

$$\Pr\{C \mid H\} = \frac{\Pr\{C \cap H\}}{\Pr\{H\}}$$

A tree diagram is worked out below:



Now we can compute the desired conditional probability as follows:

$$\begin{aligned}
\Pr\{C \mid H\} &= \frac{\Pr\{C \cap H\}}{\Pr\{H\}} \\
&= \frac{\frac{1}{2} \cdot \frac{1}{52}}{\frac{1}{2} \cdot \frac{1}{52} + \frac{1}{2} \cdot \frac{1}{51}} \\
&= \frac{51}{103} \\
&= 0.495146\ldots
\end{aligned}$$

Thus, if you selected the eight of hearts, then the deck you picked is more likely to be incomplete. ■

Problem 13.7.2. There is a rare and deadly disease called *Nerditosis* which afflicts about 1 person in 1000. One symptom is a compulsion to refer to everything— fields of study, classes, buildings, etc.— using numbers. It's horrible. As victims enter their final, downward spiral, they're awarded a degree from MIT. Two doctors claim that they can diagnose Nerditosis.

(a) Doctor X received his degree from Harvard Medical School. He practices at Massachusetts General Hospital and has access to the latest scanners, lab tests, and research. Suppose you ask Doctor X whether you have the disease.

- If you have Nerditosis, he says "yes" with probability 0.99.
- If you don't have it, he says "no" with probability 0.97.

Let D be the event that you have the disease, and let E be the event that the diagnosis is erroneous. Use the Total Probability Law to compute $\Pr\{E\}$, the probability that Doctor X makes a mistake.

Solution. By the Total Probability Law:

$$\begin{aligned}\Pr\{E\} &= \Pr\{E \mid D\} \cdot \Pr\{D\} + \Pr\left\{E \mid \overline{D}\right\} \cdot \Pr\left\{\overline{D}\right\} \\ &= 0.01 \cdot 0.001 + 0.03 \cdot 0.999 \\ &= 0.02998\end{aligned}$$

■

(b) "Doctor" Y received his genuine degree from a fully-accredited university for \$49.95 via a special internet offer. He knows that Nerditosis stikes 1 person in 1000, but is a little shaky on how to interpret this. So if you ask him whether you have the disease, he'll randomly say "yes" with probability 1 in 1000 without even examining you.

Let D be the event that you have the disease, and let F be the event that the diagnosis is faulty. Use the Total Probability Law to compute $\Pr\{F\}$, the probability that Doctor Y made a mistake.

Solution. By the Total Probability Law:

$$\begin{aligned}\Pr\{F\} &= \Pr\{F \mid D\} \cdot \Pr\{D\} + \Pr\left\{F \mid \overline{D}\right\} \cdot \Pr\left\{\overline{D}\right\} \\ &= 0.999 \cdot 0.001 + 0.001 \cdot 0.999 \\ &= 0.001998\end{aligned}$$

■

(c) Which doctor is more reliable?

Solution. Doctor X makes more than 15 times as many errors as Doctor Y. ■

(d) Doctor Z, who went to MIT and took 6.042, observes that he can do even better than Doctor Y. How?

Solution. *Always* say "No". ■

Problem 13.7.3. There were n Immortal Warriors born into our world, but in the end *there can be only one*. The Immortals' original plan was to stalk the world for centuries, dueling one another with ancient swords in dramatic landscapes until only one survivor remained. However, after a thought-provoking discussion of probabilistic independence, they opt to give the following protocol a try:

1. The Immortals forge a coin that comes up heads with probability p.
2. Each Immortal flips the coin once.
3. If *exactly one* Immortal flips heads, then he or she is declared The One. Otherwise, the protocol is declared a failure, and they all go back to hacking each other up with swords.

(a) One of the Immortals (Kurgan from the Russian steppe) argues that as n grows large, the probability that this protocol succeeds must tend to zero. Another (McLeod from the Scottish highlands) argues that this need not be the case, provided p is chosen *very carefully*. What does your intuition tell you?

Solution. Your intuition tells you that a short nap would be nice right now. As would a couple cookies to dunk in a cold glass of milk. ■

(b) What is the probability that the experiment succeeds as a function of p and n?

Solution. The sample space consists of all possible results of n coin flips, which we can represent by the length n strings of H's and T's. Since we intend that the flips of the different Immortals are independent of each other, we *define* the probability of an outcome to be the product of the probability of each H or T in the outcome. That is, the probability of any outcome with k H's is defined to be

$$p^k(1-p)^{n-k}.$$

So the probability that Kurgan flips heads and all the other Immortals flip tails is

$$p(1-p)^{n-1}.$$

The same probability applies to each of the n Immortals. The probability of the event, E, that the experiment successfully selects The One is the sum of these n probabilities, namely,

$$\Pr\{E\} = np(1-p)^{n-1} \tag{13.3}$$

■

(c) How should p, the bias of the coin, be chosen in order to maximize the probability that the experiment succeeds? (You're going to have to compute a derivative!)

Solution. We compute the derivative of the success probability:

$$\frac{d}{dp}\, np(1-p)^{n-1} \quad = \quad n(1-p)^{n-1} - np(n-1)(1-p)^{n-2}$$

Now we set the right side equal to zero to find the best probability p:

$$\begin{aligned} n(1-p)^{n-1} &= np(n-1)(1-p)^{n-2} \\ (1-p) &= p(n-1) \\ p &= 1/n \end{aligned}$$

This answer makes sense, since we want the coin to come up heads exactly 1 time in n. ■

(d) What is the probability of success if p is chosen in this way? What quantity does this approach when n, the number of Immortal Warriors, grows large?

Solution. Setting $p = 1/n$ in the formula (13.3) for the probability that the experiment succeeds gives:

$$\Pr\{E\} = \left(1 - \frac{1}{n}\right)^{n-1} = \left(1 + \frac{-1}{n}\right)^{n} \left(1 - \frac{1}{n}\right)^{-1}. \tag{13.4}$$

But using the familiar fact that

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^{n} = e^{x},$$

for all x, we conclude that the limit as n goes to infinity of the righthand term of (13.4) is $e^{-1} \cdot (1-0)^{-1}$. That is,

$$\lim_{n\to\infty} \Pr\{E\} = \frac{1}{e}$$

McLeod is right. ■

Problem 13.7.4. Suppose there is a system with n components, and we know from past experience that any particular component will fail in a given year with probability p. That is, letting F_i be the event that the ith component fails within one year, we have

$$\Pr\{F_i\} = p$$

for $1 \le i \le n$. The *system* will fail if *any one* of its components fails. What can we say about the probability that the system will fail within one year?

Let F be the event that the system fails within one year. Without any additional assumptions, we can't get an exact answer for $\Pr\{F\}$. However, we can give useful upper and lower bounds, namely,

$$p \le \Pr\{F\} \le np. \tag{13.5}$$

We may as well assume $p < 1/n$, since the upper bound is trivial otherwise. For example, if $n = 100$ and $p = 10^{-5}$, we conclude that there is at most one chance in 1000 of system failure within a year and at least one chance in 100,000.

Let's model this situation with the sample space $\mathcal{S} ::= \mathcal{P}(\{1, \ldots, n\})$ of subsets of positive integers $\le n$, where $s \in \mathcal{S}$ corresponds to the indices of the components which fail within one year. For example, $\{2, 5\}$ is the outcome that the second and fifth components failed within a year and none of the other components failed. So the outcome that the system did not fail corresponds to the emptyset, $\emptyset$.

(a) Show that the probability that the system fails could be as small as p by describing appropriate probabilities for the sample points.

Solution. There could be a probability p of system failure if all the individual failures occur together. That is, let $\Pr\{\{1,\ldots,n\}\} ::= p$, $\Pr\{\emptyset\} ::= 1-p$, and let the probability of all other outcomes be zero. So $F_i = \{s \in \mathcal{S} \mid i \in s\}$ and $\Pr\{F_i\} = 0+0+\cdots+0+\Pr\{\{1,\ldots,n\}\} = \Pr\{\{1,\ldots,n\}\} = p$. Also, the only outcome with positive probability in F is $\{1,\ldots,n\}$, so $\Pr\{F\} = p$, as required. ■

(b) Show that the probability that the system fails could actually be as large as np by describing appropriate probabilities for the sample points.

Solution. Suppose at most one component ever fails at a time. That is, $\Pr\{\{i\}\} = p$ for $1 \le i \le n$, $\Pr\{\emptyset\} = 1 - np$, and probability of all other points is zero. The sum of the probabilities of all the points is one, so this is a well-defined probability space. Also, the only sample point in F_i with positive probability is $\{i\}$, so $\Pr\{F_i\} = \Pr\{\{i\}\} = p$ as required. Finally, $\Pr\{F\} = np$ because $F = \{A \subseteq \{1,\ldots,n\} \mid A \neq \emptyset\}$, so F in particular contains all the n outcomes of the form $\{i\}$. ■

(c) Prove the inequality (13.5).

Solution. $F = \bigcup_{i=1}^{n} F_i$ so

$$
\begin{aligned}
p &= \Pr\{F_1\} && \text{(given)} && (13.6)\\
&\le \Pr\{F\} && \text{(since } F_1 \subseteq F\text{)} && (13.7)\\
&= \Pr\left\{\bigcup F_i\right\} && \text{(def. of } F\text{)} && (13.8)\\
&\le \sum_{i=1}^{n} \Pr\{F_i\} && \text{(Union Bound)} && (13.9)\\
&= np. && \text{(since the } F_i\text{'s are disjoint)} && (13.10)
\end{aligned}
$$

■

Appendix

The Total Probability Law is

$$\Pr\{A\} = \Pr\{A \mid E\} \cdot \Pr\{E\} + \Pr\left\{A \mid \overline{E}\right\} \cdot \Pr\left\{\overline{E}\right\}.$$

Chapter 14

Random Variables, Distributions, Sampling

14.1 Random Variables

Last week we focused on probabilities of ***events***: what is the probability of the event that you win the Monty Hall game? ... that you have a rare disease, given that you tested positive. ... that the Red Sox won the pennant, given that they lost the second game?

This week we focus on quantitative questions: *How many* contestants must play the Monty Hall game until one of them finally wins? ... *How long* will this illness last? *How much* will I lose playing 6.042 games all day? A *random variable* is the the mathematical tool for addressing such questions.

Definition 14.1.1. A random variable, R, is a total function whose domain is $\mathcal{S}$, the sample space of outcomes.

The codomain of R can be anything, but will usually be a subset of the real numbers. Notice that the name "random variable" is a misnomer; random variables are actually functions!

14.1.1 Examples

Consider the experiment of tossing three independent, unbiased coins. Let C be the number of heads that appear. Let $M = 1$ if the three coins come up all heads or all tails, and let $M = 0$ otherwise. Now every outcome of the three coin flips uniquely determines the values of C and M. For example, if we flip heads, tails, heads, then $C = 2$ and $M = 0$. If we flip tails, tails, tails, then $C = 0$ and $M = 1$. In effect, C counts the number of heads, and M indicates whether all the coins match.

Since each outcome uniquely determines C and M, we can regard them as functions mapping outcomes to numbers. For this experiment, the sample space is:

$$\mathcal{S} \quad = \quad \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Now C is a function that maps each outcome in the sample space to a number as follows:

$$\begin{array}{rclrcl} C(HHH) & = & 3 & C(THH) & = & 2 \\ C(HHT) & = & 2 & C(THT) & = & 1 \\ C(HTH) & = & 2 & C(TTH) & = & 1 \\ C(HTT) & = & 1 & C(TTT) & = & 0. \end{array}$$

Similarly, M is a function mapping each outcome another way:

$$\begin{array}{rclrcl} M(HHH) & = & 1 & M(THH) & = & 0 \\ M(HHT) & = & 0 & M(THT) & = & 0 \\ M(HTH) & = & 0 & M(TTH) & = & 0 \\ M(HTT) & = & 0 & M(TTT) & = & 1. \end{array}$$

So C and M are ***random variables***.

14.1.2 Indicator Random Variables

An ***indicator random variable*** (or simply an ***indicator***, or a ***Bernoulli random variable***) is a random variable that maps every outcome to either 0 or 1. The random variable M is an example. If all three coins match, then $M = 1$; otherwise, $M = 0$.

Indicator random variables are closely related to events. In particular, an indicator partitions the sample space into those outcomes mapped to 1 and those outcomes mapped to 0. For example, the indicator M partitions the sample space into two blocks as follows:

$$\underbrace{HHH \quad TTT}_{M=1} \quad \underbrace{HHT \quad HTH \quad HTT \quad THH \quad THT \quad TTH}_{M=0}.$$

In the same way, an event, E, partitions the sample space into those outcomes in E and those not in E. So E is naturally associated with an indicator random variable, I_E, for the event, where $I_E(p) = 1$ for outcomes $p \in E$ and $I_E(p) = 0$ for outcomes $p \notin E$. Thus, $M = I_F$ where F is the event that all three coins match.

14.1.3 Random Variables and Events

There is a strong relationship between events and more general random variables as well. A random variable that takes on several values partitions the sample space into several blocks. For example, C partitions the sample space as follows:

$$\underbrace{TTT}_{C=0} \quad \underbrace{TTH \quad THT \quad HTT}_{C=1} \quad \underbrace{THH \quad HTH \quad HHT}_{C=2} \quad \underbrace{HHH}_{C=3}.$$

Each block is a subset of the sample space and is therefore an event. Thus, we can regard an equation or inequality involving a random variable as an event. For example, the event that $C = 2$ consists of the outcomes THH, HTH, and HHT. The event $C \leq 1$ consists of the outcomes TTT, TTH, THT, and HTT.

Naturally enough, we can talk about the probability of events defined by equations involving random variables. For example:

$$\begin{aligned}\Pr\{C = 2\} &= \Pr\{THH\} + \Pr\{HTH\} + \Pr\{HHT\} \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}.\end{aligned}$$

As another example:

$$\begin{aligned}\Pr\{M = 1\} &= \Pr\{TTT\} + \Pr\{HHH\} \\ &= \frac{1}{8} + \frac{1}{8} = \frac{1}{4}.\end{aligned}$$

14.1.4 Conditional Probability

Mixing conditional probabilities and events involving random variables creates no new difficulties. For example, $\Pr\{C \geq 2 \mid M = 0\}$ is the probability that at least two coins are heads ($C \geq 2$), given that not all three coins are the same ($M = 0$). We can compute this probability using the definition of conditional probability:

$$\begin{aligned}\Pr\{C \geq 2 \mid M = 0\} &= \frac{\Pr\{[C \geq 2] \cap [M = 0]\}}{\Pr\{M = 0\}} \\ &= \frac{\Pr\{\{THH, HTH, HHT\}\}}{\Pr\{\{THH, HTH, HHT, HTT, THT, TTH\}\}} \\ &= \frac{3/8}{6/8} = \frac{1}{2}.\end{aligned}$$

The expression $[C \geq 2] \cap [M = 0]$ on the first line may look odd; what is the set operation $\cap$ doing between an inequality and an equality? But recall that, in this context, $[C \geq 2]$ and $[M = 0]$ are *events*, namely, *sets* of outcomes.

14.1.5 Independence

The notion of independence carries over from events to random variables as well. Random variables R_1 and R_2 are ***independent*** iff for all x_1 in the codomain of R_1, and x_2 in the codomain of R_2, we have:

$$\Pr\{[R_1 = x_1] \cap [R_2 = x_2]\} = \Pr\{R_1 = x_1\} \cdot \Pr\{R_2 = x_2\}.$$

As with events, we can formulate independence for random variables in an equivalent and perhaps more intuitive way: random variables R_1 and R_2 are independent if for all x_1 and x_2

$$\Pr\{R_1 = x_1 \mid R_2 = x_2\} = \Pr\{R_1 = x_1\}.$$

whenever the lefthand conditional probability is defined, that is, whenever $\Pr\{R_2 = x_2\} > 0$.

As an example, are C and M independent? Intuitively, the answer should be "no". The number of heads, C, completely determines whether all three coins match; that is, whether $M = 1$. But, to verify this intuition, we must find some $x_1, x_2 \in \mathbb{R}$ such that:

$$\Pr\{[C = x_1] \cap [M = x_2]\} \neq \Pr\{C = x_1\} \cdot \Pr\{M = x_2\}.$$

One appropriate choice of values is $x_1 = 2$ and $x_2 = 1$. In this case, we have:

$$\Pr\{[C = 2] \cap [M = 1]\} = 0 \neq \frac{1}{4} \cdot \frac{3}{8} = \Pr\{M = 1\} \cdot \Pr\{C = 2\}.$$

The first probability is zero because we never have exactly two heads ($C = 2$) when all three coins match ($M = 1$). The other two probabilities were computed earlier.

On the other hand, let H_1 be the indicator variable for event that the first flip is a Head, so

$$[H_1 = 1] = \{HHH, HTH, HHT, HTT\}.$$

Then H_1 is independent of M, since

$$\begin{aligned}\Pr\{M = 1\} = 1/4 = \Pr\{M = 1 \mid H_1 = 1\} = \Pr\{M = 1 \mid H_1 = 0\}\\ \Pr\{M = 0\} = 3/4 = \Pr\{M = 0 \mid H_1 = 1\} = \Pr\{M = 0 \mid H_1 = 0\}\end{aligned}$$

In fact, this example is an instance of the important observation that two events are independent iff their indicator variables are independent.

As with events, the notion of independence generalizes to more than two random variables.

Definition 14.1.2. Random variables $R_1, R_2, \ldots, R_n$ are ***mutually independent*** iff

$$\begin{aligned}&\Pr\{[R_1 = x_1] \cap [R_2 = x_2] \cap \cdots \cap [R_n = x_n]\}\\ &= \Pr\{R_1 = x_1\} \cdot \Pr\{R_2 = x_2\} \cdots \Pr\{R_n = x_n\}.\end{aligned}$$

for all $x_1, x_2, \ldots, x_n$.

It is a simple exercise to show that the probability that any *subset* of the variables takes a particular set of values is equal to the product of the probabilites that the individual variables take their values. Thus, for example, if $R_1, R_2, \ldots, R_{100}$ are mutually independent random variables, then it follows that:

$$\begin{aligned}&\Pr\{[R_1 = 7] \cap [R_7 = 9.1] \cap [R_{23} = \pi] \cap [R_{87} = -1]\}\\ &= \Pr\{R_1 = 7\} \cdot \Pr\{R_7 = 9.1\} \cdot \Pr\{R_{23} = \pi\} \cdot \Pr\{R_{87} = -1\}.\end{aligned}$$

14.2 The Birthday Principle

There are 100 students in a lecture hall. What is the probability that some two people share a birthday? Maybe about $1/3$? Let's check! We'll use the following two variables throughout our analysis:

- Let n be the number of people in the group.
- Let d be the number of days in the year.
- Let D be the number of pairs of students with the same birthday.

We'll assume that a random choice of a student is made independently of the any other choices of students. For simplicity, we'll assume that the probability that a randomly chosen student has a given birthday is $1/365$. This assumption is not really true, since more babies are born at certain times of year. However, our analysis of this problem applies to many situations in Computer Science that are unaffected by leap days, snow days or Spring fever, so we won't dwell on those complications.

A sensible sample space to model this experiment consists of all ways of assigning birthdays to the people of the group. There are d^n such assignments, since the first person can have d different birthdays, the second person can have d different birthdays, and so forth. Furthermore, every such assignment is equally probable by our assumption that birthdays are equally likely and independent of each other.

Now the event that some two students have the same birthday can be expressed as $[D > 0]$. It turns out to be be easier to calculate the complement event $[D = 0]$ that everyone has a distinct birthday, which is good enough since $\Pr\{D > 0\} = 1 - \Pr\{D = 0\}$.

Anyway, the event $[D = 0]$ consists of $d(d-1)(d-2)\cdots(d-n+1)$ outcomes, since we can select the birthday of the first person in d days, the birthday of the second person in $d-1$ ways, and so forth. Therefore, the probability that everyone has a different birthday is:

$$\Pr\{D = 0\} = \frac{d(d-1)(d-2)\cdots(d-n+1)}{d^n}.$$

For $n = 100$, this probability is actually fantastically small —less than one in a million! If there are 100 people in a room, two are almost certain to share a birthday.

Let's rewrite the right side of the preceding equation in a more insightful form that allows us to use the fact that $e^x > 1 + x$ for all x.[1]

$$\begin{aligned}
\Pr\{D = 0\} &= \left(1 - \frac{0}{d}\right)\left(1 - \frac{1}{d}\right)\left(1 - \frac{2}{d}\right)\cdots\left(1 - \frac{n-1}{d}\right) \\
&< e^0 \cdot e^{-1/d} \cdot e^{-2/d} \cdots e^{-(n-1)/d} \\
&= e^{-\frac{\sum_{i=1}^{n-1} i}{d}} \\
&= e^{\frac{-n(n-1)}{2d}} = e^{-\frac{\binom{n}{2}}{d}}.
\end{aligned}$$

The exponent $\binom{n}{2}/d$ in the final expression above is close to 1 when $n \approx \sqrt{2d}$. In this case, the probability that two people share a birthday is close to $1/e$ which is roughly the break-even point, where it's equally likely whether or not a couple of people will share a birthday. This leads to a rule called the ***Birthday Principle***, which is useful in many contexts in Computer Science:

> If there are d days in a year and $\sqrt{2d}$ people in a room, then the probability that two share a birthday is about $1 - 1/e \approx 0.632$.

For example, this principle says that if you have $\sqrt{2 \cdot 365} \approx 27$ people in a room, then the probability that two share a birthday is about 0.632. The actual probability is about 0.626, so the approximation is quite good.

[1] This approximation is obtained by truncating the Taylor series $e^{-x} = 1 - x + x^2/2! - x^3/3! + \cdots$. The approximation $e^{-x} \approx 1 - x$ is pretty accurate when x is small.

The Birthday Principle is a great rule of thumb with surprisingly many applications. For example, cryptographic systems and digital signature schemes must be hardened against "birthday attacks". The principle also tells us how many items can be inserted into a hash table before one starts to experience collisions.

14.3 Probability Distributions

A random variable is defined to be a function whose domain is the sample space of an experiment. Often, however, random variables with essentially the same properties show up in completely different experiments. For example, some random variables that come up in polling, in primality testing, and in coin flipping all share some common properties. If we could study such random variables in the abstract, divorced from the details of any particular experiment, then our conclusions would apply to *all* the experiments where that sort of random variable turned up. Such general conclusions could be very useful. There are a couple tools that capture the essential properties of a random variable, but leave other details of the associated experiment behind.

The ***probability density function (pdf)*** for a random variable R with codomain V is a function $\text{PDF}_R : V \to [0,1]$ defined by:

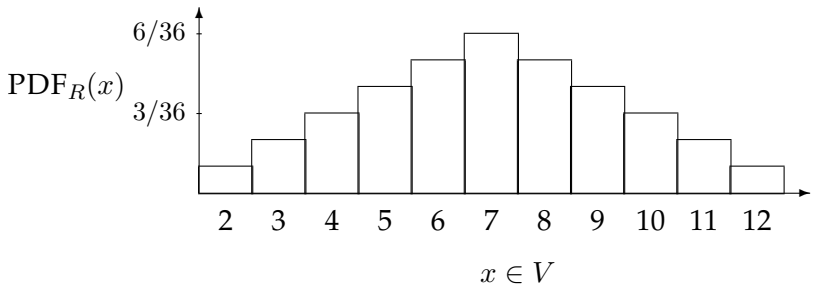
$$\text{PDF}_R(x) = \Pr\{R = x\}$$

A consequence of this definition is that

$$\sum_{x \in V} \text{PDF}_R(x) = 1$$

since the random variable always takes on exactly one value in the set V.

As an example, let's return to the experiment of rolling two fair, independent dice. As before, let T be the total of the two rolls. This random variable takes on values in the set $V = \{2, 3, \ldots, 12\}$. A plot of the probability density function is shown below:

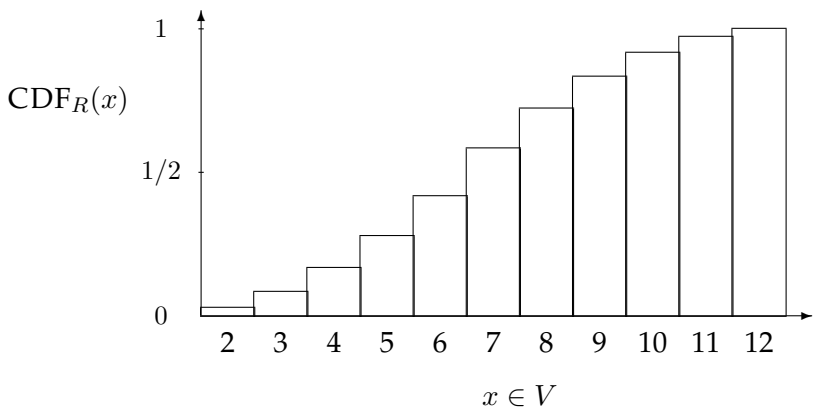


The lump in the middle indicates that sums close to 7 are the most likely. The total area of all the rectangles is 1 since the dice must take on exactly one of the sums in $V = \{2, 3, \ldots, 12\}$.

A closely-related idea is the ***cumulative distribution function (cdf)*** for a random variable R whose range is real numbers. This is a function $\text{CDF}_R : \mathbb{R} \to [0,1]$ defined by:

$$\text{CDF}_R(x) = \Pr\{R \leq x\}$$

As an example, the cumulative distribution function for the random variable T is shown below:



The height of the i-th bar in the cumulative distribution function is equal to the *sum* of the heights of the leftmost i bars in the probability density function. This follows from the definitions of pdf and cdf:

$$\begin{aligned}\text{CDF}_R(x) &= \Pr\{R \leq x\} \\ &= \sum_{y \leq x} \Pr\{R = y\} \\ &= \sum_{y \leq x} \text{PDF}_R(y)\end{aligned}$$

In summary, $\text{PDF}_R(x)$ measures the probability that $R = x$ and $\text{CDF}_R(x)$ measures the probability that $R \leq x$. Both the PDF_R and CDF_R capture the same information about the random variable R— you can derive one from the other— but sometimes one is more convenient. The key point here is that neither the probability density function nor the cumulative distribution function involves the sample space of an experiment. Thus, through these functions, we can study random variables without reference to a particular experiment.

We'll now look at three important distributions and some applications.

14.3.1 Bernoulli Distribution

Indicator random variables are perhaps the most common type because of their close association with events. The probability density function of an indicator random variable B is always

$$\begin{aligned}\text{PDF}_B(0) &= p \\ \text{PDF}_B(1) &= 1 - p\end{aligned}$$

where $0 \leq p \leq 1$. The corresponding cumulative distribution function is:

$$\begin{aligned}\text{CDF}_B(0) &= p \\ \text{CDF}_B(1) &= 1\end{aligned}$$

14.3.2 Uniform Distribution

A random variable that takes on each possible value with the same probability is called ***uniform***. For example, the probability density function of a random variable U that is uniform on the set $\{1, 2, \ldots, N\}$ is:

$$\text{PDF}_U(k) = \frac{1}{N}$$

And the cumulative distribution function is:

$$\text{CDF}_U(k) = \frac{k}{N}$$

Uniform distributions come up all the time. For example, the number rolled on a fair die is uniform on the set $\{1, 2, \ldots, 6\}$.

14.3.3 The Numbers Game

Let's play a game! I have two envelopes. Each contains an integer in the range $0, 1, \ldots, 100$, and the numbers are distinct. To win the game, you must determine which envelope contains the larger number. To give you a fighting chance, I'll let you peek at the number in one envelope selected at random. Can you devise a strategy that gives you a better than 50% chance of winning?

For example, you could just pick an evelope at random and guess that it contains the larger number. But this strategy wins only 50% of the time. Your challenge is to do better.

So you might try to be more clever. Suppose you peek in the left envelope and see the number 12. Since 12 is a small number, you might guess that that other number is larger. But perhaps I'm sort of tricky and put small numbers in *both* envelopes. Then your guess might not be so good!

An important point here is that the numbers in the envelopes may *not* be random. I'm picking the numbers and I'm choosing them in a way that I think will defeat your guessing strategy. I'll only use randomization to choose the numbers if that serves *my* end: making you lose!

Intuition Behind the Winning Strategy

Amazingly, there is a strategy that wins more than 50% of the time, regardless of what numbers I put in the envelopes!

Suppose that you somehow knew a number x *between* my lower number and higher numbers. Now you peek in an envelope and see one or the other. If it is bigger than x, then you know you're peeking at the higher number. If it is smaller than x, then you're peeking at the lower number. In other words, if you know a number x between my lower and higher numbers, then you are certain to win the game.

The only flaw with this brilliant strategy is that you do *not* know x. Oh well.

But what if you try to *guess* x? There is some probability that you guess correctly. In this case, you win 100% of the time. On the other hand, if you guess incorrectly, then you're no worse off than before; your chance of winning is still 50%. Combining these two cases, your overall chance of winning is better than 50%!

Informal arguments about probability, like this one, often sound plausible, but do not hold up under close scrutiny. In contrast, this argument sounds completely implausible— but is actually correct!

Analysis of the Winning Strategy

For generality, suppose that I can choose numbers from the set $\{0, 1, \ldots, n\}$. Call the lower number L and the higher number H.

Your goal is to guess a number x between L and H. To avoid confusing equality cases, you select x at random from among the half-integers:

$$\left\{\frac{1}{2},\ 1\frac{1}{2},\ 2\frac{1}{2},\ \ldots,\ n-\frac{1}{2}\right\}$$

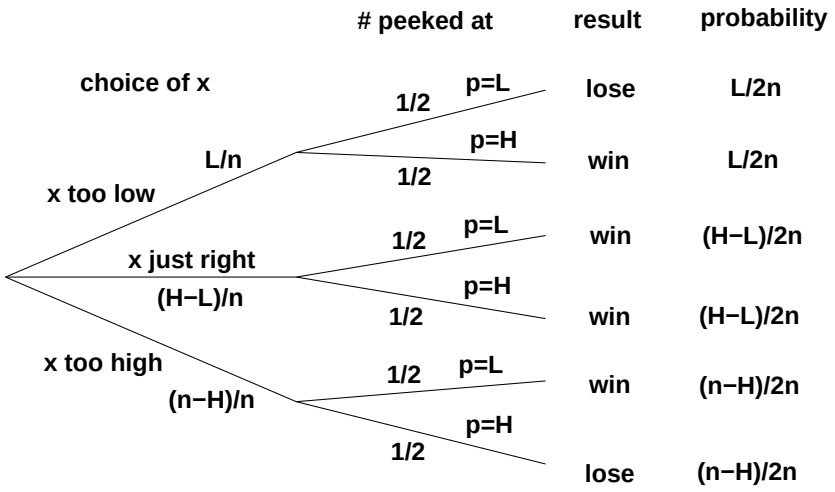
But what probability distribution should you use?

The uniform distribution turns out to be your best bet. An informal justification is that if I figured out that you were unlikely to pick some number— say $50\frac{1}{2}$— then I'd always put 50 and 51 in the evelopes. Then you'd be unlikely to pick an x between L and H and would have less chance of winning.

After you've selected the number x, you peek into an envelope and see some number p. If $p > x$, then you guess that you're looking at the larger number. If $p < x$, then you guess that the other number is larger.

All that remains is to determine the probability that this strategy succeeds. We can do this with the usual four-step method and a tree diagram.

Step 1: Find the sample space. You either choose x too low ($< L$), too high ($> H$), or just right ($L < x < H$). Then you either peek at the lower number ($p = L$) or the higher number ($p = H$). This gives a total of six possible outcomes.



Step 2: Define events of interest. The four outcomes in the event that you win are marked in the tree diagram.

Step 3: Assign outcome probabilities. First, we assign edge probabilities. Your guess x is too low with probability L/n, too high with probability $(n-H)/n$, and just right with probability $(H-L)/n$. Next, you peek at either the lower or higher number with equal probability. Multiplying along root-to-leaf paths gives the outcome probabilities.

Step 4: Compute event probabilities. The probability of the event that you win is the sum of the probabilities of the four outcomes in that event:

$$\begin{aligned}\Pr\{\text{win}\} &= \frac{L}{2n} + \frac{H-L}{2n} + \frac{H-L}{2n} + \frac{n-H}{2n} \\ &= \frac{1}{2} + \frac{H-L}{2n} \\ &\geq \frac{1}{2} + \frac{1}{2n}\end{aligned}$$

The final inequality relies on the fact that the higher number H is at least 1 greater than the lower number L since they are required to be distinct.

Sure enough, you win with this strategy more than half the time, regardless of the numbers in the envelopes! For example, if I choose numbers in the range $0, 1, \ldots, 100$, then you win with probability at least $\frac{1}{2} + \frac{1}{200} = 50.5\%$. Even better, if I'm allowed only numbers in the range $0, \ldots, 10$, then your probability of winning rises to 55%! By Las Vegas standards, those are great odds!

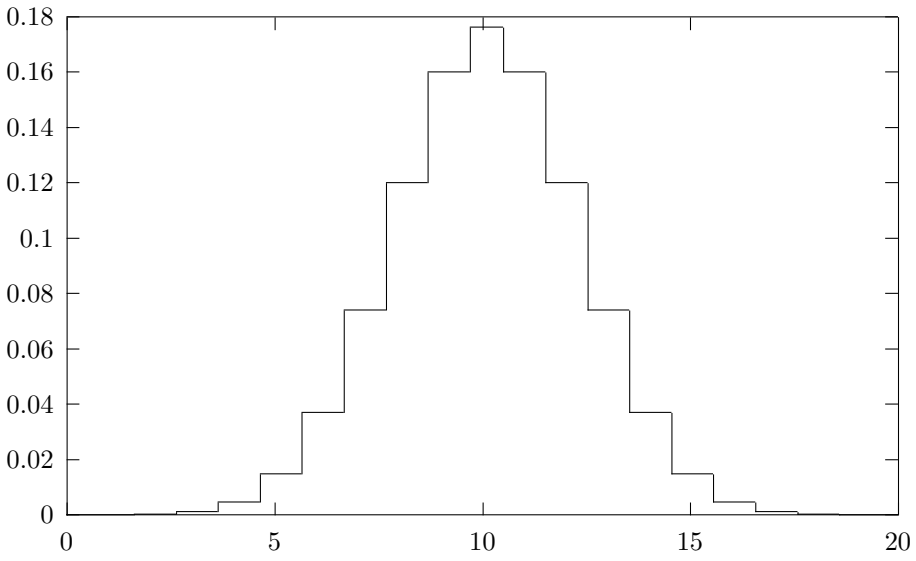
14.3.4 Binomial Distribution

Of the more complex distributions, the ***binomial distribution*** is surely the most important in Computer Science. The standard example of a random variable with a binomial distribution is the number of heads that come up in n independent flips of a coin; call this random variable H_n. If the coin is fair, then H_n has an *unbiased binomial density function*:

$$\text{PDF}_{H_n}(k) = \binom{n}{k} 2^{-n}.$$

This follows because there are $\binom{n}{k}$ sequences of n coin tosses with exactly k heads, and each such sequence has probability 2^{-n}.

Here is a plot of the unbiased probability density function $\text{PDF}_{H_n}(k)$ corresponding to $n = 20$ coins flips. The most likely outcome is $k = 10$ heads, and the probability falls off rapidly for larger and smaller values of k. These falloff regions to the left and right of the main hump are usually called the ***tails of the distribution***.



An enormous number of analyses in Computer Science come down to proving that the tails of the binomial and similar distributions are very small. In the context of a problem, this typically means that there is very small probability that something *bad* happens, which could be a server or communication link overloading or a randomized algorithm running for an exceptionally long time or producing the wrong result.

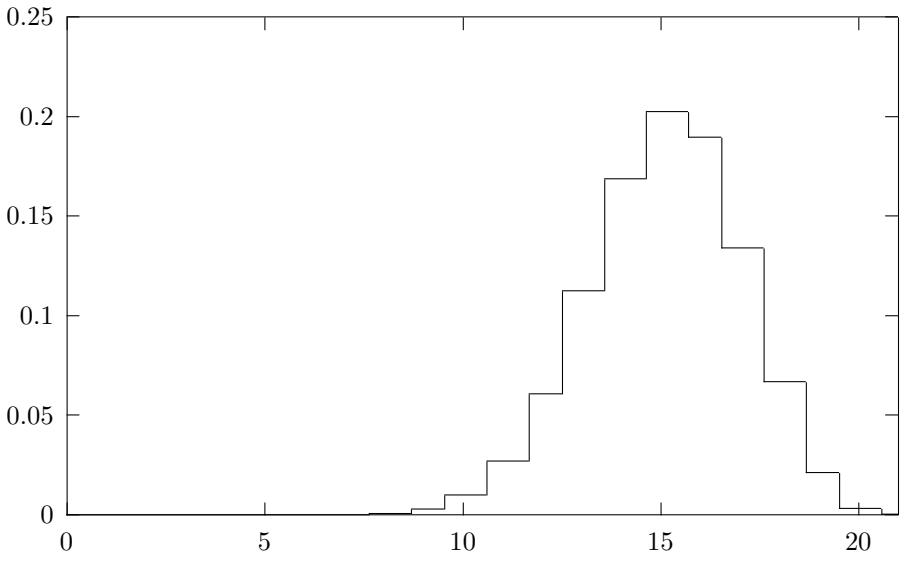
The General Binomial Distribution

Now let J be the number of heads that come up on n independent coins, each of which is heads with probability p. Then J has a *general binomial density function*:

$$\text{PDF}_J(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

As before, there are $\binom{n}{k}$ sequences with k heads and $n-k$ tails, but now the probability of each such sequence is $p^k(1-p)^{n-k}$.

As an example, the plot below shows the probability density function $\text{PDF}_J(k)$ corresponding to flipping $n = 20$ independent coins that are heads with probabilty $p = 0.75$. The graph shows that we are most likely to get around $k = 15$ heads, as you might expect. Once again, the probability falls off quickly for larger and smaller values of k.



Approximating the Binomial Density Function

Computing the general binomial density function is daunting if not impossible when n is large –say $n \geq 2000$. Fortunately, there is an approximate closed-form formula for this function, which, though a bit unwieldy, is easy to calculate. First, we need an approximation for the binomial coefficient in the exact formula. For convenience, let's replace k by αn where α is a number between 0 and 1. Then, from Stirling's formula, we find that:

$$\binom{n}{\alpha n} \leq \frac{2^{nH(\alpha)}}{\sqrt{2\pi\alpha(1-\alpha)n}}$$

where $H(\alpha)$ is the famous ***entropy function***:

$$H(\alpha) ::= \alpha \log_2 \frac{1}{\alpha} + (1-\alpha)\log_2 \frac{1}{1-\alpha}$$

Its graph is shown in Figure 14.1.

This upper bound on $\binom{n}{\alpha n}$ is very tight and serves as an excellent approximation.

Now let's plug this formula into the general binomial density function. The probability of flipping αn heads in n tosses of a coin that comes up heads with probability p is:

$$\text{PDF}_J(\alpha n) \leq \frac{2^{nH(\alpha)}}{\sqrt{2\pi\alpha(1-\alpha)n}} \cdot p^{\alpha n}(1-p)^{(1-\alpha)n} \tag{14.1}$$

This formula is ugly as a bowling shoe, but quite useful. For example, suppose we flip a fair coin n times. What is the probability of getting *exactly* $\frac{1}{2}n$ heads? Plugging $\alpha = 1/2$ and $p = 1/2$ into this formula gives:

$$\begin{aligned}\text{PDF}_J(\alpha n) &\leq \frac{2^{nH(1/2)}}{\sqrt{2\pi(1/2)(1-(1/2))n}} \cdot 2^{-n} \\ &= \sqrt{\frac{2}{\pi n}}\end{aligned}$$

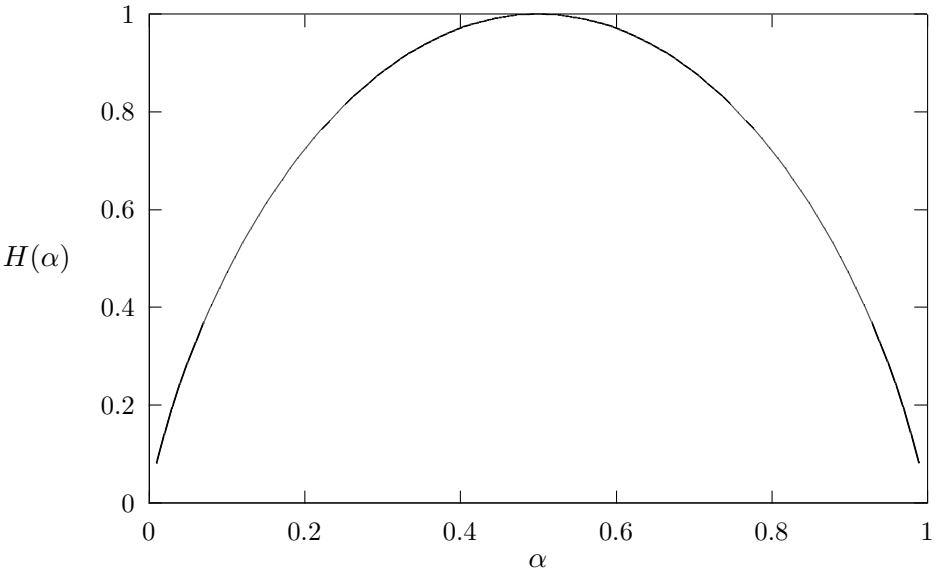


Figure 14.1: The Entropy Function

Thus, for example, if we flip a fair coin 100 times, the probability of getting exactly 50 heads is about $1/\sqrt{50\pi} \approx 0.079$ or around 8%.

14.3.5 Approximating the Cumulative Binomial Distribution Function

Suppose a coin comes up heads with probability p. As before, let the random variable J be the number of heads that come up on n independent flips. Then the probability of getting *at most* k heads is given by the cumulative binomial distribution function:

$$\begin{aligned} \text{CDF}_J(k) &= \Pr\{J \leq k\} \\ &= \sum_{i=0}^{k} \text{PDF}_J(i) \\ &= \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \end{aligned} \tag{14.2}$$

Evaluating this expression directly would be a lot of work for large k and n, so now an approximation would be really helpful. Once again, we can let $k = \alpha n$; that is, instead of thinking of the absolute number of heads (k), we consider the fraction of flips that are heads (α). The following approximation holds provided $\alpha < p$:

$$\begin{aligned} \text{CDF}_J(\alpha n) &\leq \frac{1-\alpha}{1-\alpha/p} \cdot \text{PDF}_J(\alpha n) \\ &\leq \frac{1-\alpha}{1-\alpha/p} \cdot \frac{2^{nH(\alpha)}}{\sqrt{2\pi\alpha(1-\alpha)n}} \cdot p^{\alpha n}(1-p)^{(1-\alpha)n} \qquad \text{(by (14.1))} \end{aligned} \tag{14.3}$$

The first inequality above can be derived by bounding the summation (14.2) with a geometric sum and applying the formula for the sum of a geometric series. (The details are dull and omitted.)

It would be awkward to evaluate (14.3) with a calculator, but it's easy to write a program to do it. So don't look gift blessings in the mouth before they hatch. Or something.

As an example, the probability of flipping at most 25 heads in 100 tosses of a fair coin is obtained by setting $\alpha = 1/4$, $p = 1/2$ and $n = 100$:

$$\text{CDF}_J\left(\frac{n}{4}\right) \leq \frac{1-(1/4)}{1-(1/4)/(1/2)} \cdot \text{PDF}_J\left(\frac{n}{4}\right) \leq \frac{3}{2} \cdot 1.913 \cdot 10^{-7}.$$

This says that flipping 25 or fewer heads is extremely unlikely, which is consistent with our earlier claim that the tails of the binomial distribution are very small. In fact, notice that the probability of flipping *25 or fewer* heads is only 50% more than the probability of flipping *exactly 25* heads. Thus, flipping exactly 25 heads is twice as likely as flipping any number between 0 and 24!

Caveat: The upper bound on $\text{CDF}_J(\alpha n)$ holds only if $\alpha < p$. If this is not the case in your problem, then try thinking in complementary terms; that is, look at the number of tails flipped instead of the number of heads. In our example, the probability of flipping 75 or more heads is the same as the probability of flipping 25 or fewer tails. By the above analysis, this is also extremely small.

14.4 Polling

Suppose we want to estimate the fraction of the U.S. voting population who would favor Hillary Clinton over Rudy Giuliani in the year 2008 presidential election.

Let p be this unknown fraction, and let's suppose we have some random process —say throwing darts at voter registration lists— which will select each voter with equal probability. We can define a Bernoulli variable, K, by the rule that $K = 1$ if the random voter most prefers Clinton, and $K = 0$ otherwise.

Now to estimate p, we take a large number, n, of random choices of voters[2] and count the fraction who favor Clinton. That is, we define variables $K_1, K_2, \ldots$, where K_i is interpreted to be the indicator variable for the event that the ith chosen voter prefers Clinton. Since our choices are made independently, the K_i's are independent. So formally, we model our estimation process by simply assuming we have mutually independent Bernoulli variables $K_1, K_2, \ldots$, each with the same probability, p, of being equal to 1. Now let S_n be their sum, that is,

$$S_n ::= \sum_{i=1}^{n} K_i. \tag{14.4}$$

So S_n has the binomial distribution with parameter n, which we can choose, and unknown parameter p.

The variable S_n/n describes the fraction of voters *in our sample* who favor Clinton. We would expect that S_n/n should be something like p. We will use the sample value, S_n/n, as our *statistical estimate* of p.

[2]We're choosing a random voter n times *with replacement*. That is, we don't remove a chosen voter from the set of voters eligible to be chosen later; so we might choose the same voter more than once in n tries! We would get a slightly better estimate if we required n *different* people to be chosen, but doing so complicates both the selection process and its analysis with little gain in accuracy.

14.4.1 Sampling

Suppose we want our estimate of p to be within 0.04 of p at least 95% of the time. Namely, we want

$$\Pr\left\{\left|\frac{S_n}{n} - p\right| \leq 0.04\right\} \geq 0.95\,.$$

We let ϵ be the margin of error we can tolerate, and let δ be the probability that our result lies outside this margin, so in this case we'd have $\epsilon = 0.04$ and $\delta \leq 0.05$.

We want to determine the number, n, of times we must poll voters so that the value, S_n/n, of our estimate will, with probability at least $1-\delta$, be within ϵ of the actual fraction in the nation favoring Clinton.

We can define δ, the probability that our poll is off by more than the margin of error ϵ, as follows:

$$\begin{aligned}\delta &= \underbrace{\Pr\left\{\frac{S_n}{n} \leq p-\epsilon\right\}}_{\substack{\text{too many in sample}\\ \text{prefer ``Giuliani''}}} + \underbrace{\Pr\left\{\frac{S_n}{n} \geq p+\epsilon\right\}}_{\substack{\text{too many in sample}\\ \text{prefer ``Clinton''}}}\\ &= \Pr\{S_n \leq (p-\epsilon)n\} + \Pr\{S_n \geq (p+\epsilon)n\}\,.\end{aligned}$$

Now

$$\mathrm{CDF}_{S_n}((p-\epsilon)n) ::= \Pr\{S_n \leq (p-\epsilon)n\}$$

Also,

$$\Pr\{S_n \geq (p+\epsilon)n\} = \Pr\{n - S_n \leq ((1-p)-\epsilon)n\}\,.$$

But $T_n ::= n - S_n$ is simply the number of voters in the sample who prefer Giuliani, which is a sum of Bernoulli random variables with parameter $1-p$, and therefore

$$\Pr\{T_n \leq ((1-p)-\epsilon)n\} = \mathrm{CDF}_{T_n}(((1-p)-\epsilon)n).$$

Hence

$$\delta = \mathrm{CDF}_{S_n}((p-\epsilon)n) + \mathrm{CDF}_{T_n}(((1-p)-\epsilon)n). \tag{14.5}$$

So we have reduced getting a good estimate of the required sample size to finding good bounds on two cumulative binomial distributions with parameters p and $1-p$ respectively.

Using the bound on the cumulative binomial distribution function allows us to calculate an expression bounding (14.5) in terms of n, ϵ and p. The problem is that this bound would contain the fraction, p, of Americans that prefer Clinton which is the unknown number we are trying to determine by polling in the first place! Fortunately, there is a simple way out of this circularity. Since (14.5) is symmetric in p, it has an inflection point when $p = 1/2$, and this inflection point is, in fact, its maximum:

Fact. For all ϵ, n, the maximum value of δ in equation (14.5) occurs when $p = 1/2$.

In other words, the binomial tails fall off most slowly when $p = 1/2$. Using this fact, and plugging into the inequality (14.3) bounding $\mathrm{CDF}_{S_n}((p-\epsilon)n)$ and $\mathrm{CDF}_{T_n}(((1-p)-\epsilon)n)$, we get the following theorem:

Theorem 14.4.1 (Binomial Sampling). *Let $K_1, K_2, \ldots,$ be a sequence of mutually independent 0-1-valued random variables with the same probability, p, that $K_i = 1$, and let*

$$S_n ::= \sum_{i=1}^{n} K_i.$$

Then, for $1/2 > \epsilon > 0$,

$$\Pr\left\{\left|\frac{S_n}{n} - p\right| \geq \epsilon\right\} \leq \frac{1+2\epsilon}{2\epsilon} \cdot \frac{2^{-n(1-H((1/2)-\epsilon))}}{\sqrt{2\pi(1/4-\epsilon^2)n}}. \tag{14.6}$$

We want $\epsilon = 0.04$, so plugging into (14.6) gives

$$\delta \leq 13.5 \cdot \frac{2^{-n(0.00462)}}{1.2492\sqrt{n}} \tag{14.7}$$

where δ is the probability that our estimate is not within ϵ of p. We want to poll enough people so that $\delta \leq 0.05$. The easiest way to find the necessary sample size n is to plug in values for n to find the smallest one where the righthand side of (14.7) is ≤ 0.05:

n = people polled	upper bound on probability poll is wrong	
500	9.7%	
600	6.4%	
623	5.9%	
650	5.3%	
664	5.0%	← our poll size
700	4.3%	

So 95% of the time, polling 664[3] people will yield a fraction that is within 0.04 of the actual fraction of voters preferring Clinton. This method of estimation by sampling a quantity —voting preference in this example— is a technique that can obviously be used to estimate many other unknown quantities.

We just showed that sampling merely 664 voters will yield a fraction that, 95% of the time, is within 0.04 of the actual fraction of the voting population who prefer Clinton. Notice that the actual size of the voting population was never considered because *it did not matter*. Polling only a few hundred is always sufficient, whether there are a thousand, a million, or a billion voters in the country —which people often find remarkable.

14.4.2 Confidence Levels

Suppose a pollster uses a sample of 664 random voters to estimate the fraction of voters who prefer Clinton, and the pollster finds that 364 of them prefer Clinton. It's tempting, **but sloppy**, to say that this means:

False Claim. *With probability 0.95, the fraction, p, of voters who prefer Clinton is $364/664 \pm 0.04$. Since $364/664 - 0.04 > 0.50$, there is a 95% chance that more than half the voters prefer Clinton.*

[3] An exact calculation of the binomial CDF shows that a somewhat smaller poll size of 589 would be sufficient.

What's objectionable about this statement is that it talks about the probability or "chance" that a real world fact is true, namely that the actual fraction, p, of voters favoring Clinton is more than 0.50. But p is what it is, and it simply makes no sense to talk about the probability that it is something else. For example, suppose p is actually 0.49; then it's nonsense to ask about the probability that it is within 0.04 of 364/664 —it simply isn't.

This example of voter preference is typical: we want to estimate a fixed, unknown real-world quantity. But being unknown does not make this quantity a random variable, so it makes no sense to talk about the probability that it has some property.

A more careful summary of what we have accomplished goes this way:

> We have described a probabilistic procedure for estimating the value of the actual fraction, p. The probability that *our estimation procedure* will yield a value within 0.04 of p is 0.95.

This is a bit of a mouthful, so special phrasing closer to the sloppy language is commonly used. The pollster would describe his conclusion by saying that

> At the 95% *confidence level*, the fraction of voters who prefer Clinton is $364/664 \pm 0.04$.

So confidence levels refer to the results of estimation procedures for real-world quantities. The phrase "confidence level" should be heard as a reminder that some statistical procedure was used to obtain an estimate, and in judging the credibility of the estimate, it may be important to learn just what this procedure was.

14.5 In-Class Problems Week 12, Fri.

Problem 14.5.1. For the game of "guess the bigger integer from 0 to 7" described in lecture, give a strategy for Team 1 (the team that picks the integers that go on each piece of paper) which guarantees that $\Pr\{\text{Team 1 loses}\} \leq 4/7$.

For this purpose, assume Team 2 (the team that decides to stick or switch) plays by a probabilistic strategy: if the integer they see is k, they switch with probability p_k, for $k = 0, \ldots 7$. Use the 4-step tree approach to define the sample space and the probability of each outcome.

Solution. The strategy for Team 1 is to choose each integer j from 0 to 6 with probability $1/7$, then choose between the two pieces of paper with equal probability and put j on the chosen paper and $j+1$ on the other paper.

Let k be the number on the paper that Team 2 turns over. Now the probability that $k = 0$ is $1/14$: the $1/7$ probability that $j = 0$ times the $1/2$ probability that Team 2 picks the paper with j on it. Similarly, $\Pr\{k = 7\} = 1/14$. For each of the other integers, $n = 1, 2, \ldots, 6$, there is a $1/7$ chance that $k = n$, and in this case it is equally likely to be j or $j+1$. So the probability that Team 1 loses is exactly $1/2$ when k is any of $1, 2, \ldots, 6$.

So $\Pr\{0 < k < 7\} = 6/7$, and now the Total Probability Law implies

$$\begin{aligned}
&\Pr\{\text{Team 1 loses}\} \\
&= \Pr\{\text{Team 1 loses} \mid 0 < k < 7\} \cdot \Pr\{0 < k < 7\} + \Pr\{\text{Team 1 loses} \mid k = 0\} \cdot \Pr\{k = 0\} \\
&\qquad + \Pr\{\text{Team 1 loses} \mid k = 7\} \cdot \Pr\{k = 7\} \\
&= \frac{1}{2} \cdot \frac{6}{7} + p_0 \cdot \frac{1}{14} + (1 - p_7) \cdot \frac{1}{14} \\
&\leq \frac{3}{7} + \frac{1}{14} + \frac{1}{14} = \frac{4}{7}.
\end{aligned}$$

The tree diagram that verifies this reasoning is

TBA

■

Problem 14.5.2. **(a)** Prove that if events E and F are independent then so are $\overline{E}$ and F.

Solution.

$$\begin{aligned}
\Pr\{F\} &= \Pr\{E \cap F\} + \Pr\left\{\overline{E} \cap F\right\} && \text{(total probability)} \\
&= \Pr\{E\} \Pr\{F\} + \Pr\left\{\overline{E} \cap F\right\} && \text{(indep. of } E, F) \qquad (14.8)
\end{aligned}$$

which implies

$$\begin{aligned}\Pr\left\{\overline{E}\right\}\Pr\{F\} &= (1-\Pr\{E\})\Pr\{F\} && \text{(complement rule)}\\ &= \Pr\{F\}-\Pr\{E\}\Pr\{F\}\\ &= \Pr\left\{\overline{E}\cap F\right\} && \text{(by (14.8))}\end{aligned}$$

■

(b) For any event A, its *indicator variable*, I_A, is the 0-1 valued variable such that the event $[I_A = 1]$ is the same as the event A. It follows that $[I_A = 0]$ is the same as $\overline{A}$. Prove that two events E, F are independent iff their indicator random variables I_E, I_F are independent. (The definitions of independence for events and for random variables are not the same, so there is some proving to do.)

Solution. If I_E and I_F are independent, then

$$\begin{aligned}\Pr\{E\cap F\} &= \Pr\{[I_E=1]\cap[F=1]\} && \text{(def of } I_E, I_F)\\ &= \Pr\{I_E=1\}\cdot\Pr\{I_F=1\} && \text{(indep. of } I_E, I_F)\\ &= \Pr\{E\}\cdot\Pr\{F\} && \text{(def of } I_E, I_F)\end{aligned}$$

which proves that E and F are independent.

Conversely, suppose E and F are independent. Then

$$\begin{aligned}\Pr\{[I_E=1]\cap[I_F=1]\} &= \Pr\{E\cap F\} && \text{(def of } I_E, I_F)\\ &= \Pr\{E\}\cdot\Pr\{F\} && \text{(indep. of } E, F)\\ &= \Pr\{I_E=1\}\cdot\Pr\{I_F=1\} && \text{(def of } I_E, I_F)\end{aligned}$$

Now we must similarly prove that

$$\Pr\{[I_E=a]\cap[I_F=b]\} = \Pr\{I_E=a\}\cdot\Pr\{I_F=b\}$$

for the three remaining binary pairs (a, b). Now $[I_E = 0]$ is the same event as $[I_{\overline{E}} = 1]$, so to show, for example, that this equality holds when $a = 0$ and $b = 1$, we need only show that $\overline{E}$ and F are independent. But this follows from part (a). ■

(c) How about for three events E, F, G?

Solution. If E, F and G are mutually independent, then it follows immediately as in the previous part that

$$\Pr\{[I_E=1]\cap[I_F=1]\cap[I_G=1]\} = \Pr\{I_E=1\}\cdot\Pr\{I_F=1\}\cdot\Pr\{I_G=1\}$$

Now as in the previous part, this equality will imply the corresponding equalities with the values (1,1,1) replaced by any of the other seven possible values of (I_E, I_F, I_G) if we can generalize part (a) to prove that if E, F and G are mutually independent, then so are $\overline{E}, F$ and G. The proof is similar to part (a):

$$\begin{aligned}\Pr\{F\}\Pr\{G\} &= \Pr\{F\cap G\} && \text{(indep. of } F, G) && (14.9)\\ &= \Pr\{E\cap F\cap G\}+\Pr\left\{\overline{E}\cap F\cap G\right\} && \text{(total probability)}\\ &= \Pr\{E\}\Pr\{F\}\Pr\{G\}+\Pr\left\{\overline{E}\cap F\cap G\right\} && \text{(indep. of } E, F, G) && (14.10)\end{aligned}$$

which implies

$$\begin{aligned}
\Pr\left\{\overline{E}\right\}\Pr\{F\}\Pr\{G\} &= (1-\Pr\{E\})\Pr\{F\}\Pr\{G\} && \text{(complement rule)}\\
&= \Pr\{F\}\Pr\{G\} - \Pr\{E\}\Pr\{F\}\Pr\{G\}\\
&= \Pr\left\{\overline{E}\cap F\cap G\right\} && \text{(by (14.10))}
\end{aligned}$$

Repeated application of this Lemma implies that E, F, G are independent iff E', F' and G' are independent, where R' is either R or $\overline{R}$. ■

Problem 14.5.3. Independently flip three fair coins ("fair" means equally likely to come up with a head or a tail), and let H_i be the indicator variable for a head occurring on the ith flip, for $i = 1, 2, 3$. Define $C ::= H_1 + H_2 + H_3$ to be the number of heads flipped, M to be the indicator variable for the event $[H_1 = H_2 = H_3]$ that all three coins match, and S be the indicator variable for the event $[C \equiv 1 \mod 2]$ that an odd number of heads are flipped.

(a) Verify that none of these six variables is independent of C.

Solution. If $C = 3$, then the values of all the other variables are determined. That is, $\Pr\{V = 0 \mid C = 3\}$ is 0 or 1 for all six variables, V. So

$$\Pr\{V = 0\} \neq \Pr\{V = 0 \mid C = 3\}$$

for all six variables, V, which shows that none of them is independent of C. ■

(b) Verify that H_1, H_2, H_3, and S are 3-wise independent, but not mutually independent.

Solution. Since H_1, H_2, H_3 determine all the other variables, then by the same kind of argument used in part (a), no set of four or more variables including these three can be mutually independent.

The variables H_1, H_2, H_3 are 3-wise independent by definition of "flipping independently." Now consider the three variables H_1, H_2, S.

$$\begin{aligned}
&\Pr\{S = a \text{ and } H_1 = b \text{ and } H_2 = c\}\\
&= \Pr\{(b + c + H_3 \equiv a \mod 2) \text{ and } H_1 = b \text{ and } H_2 = c\} \quad ([S = a] ::= [H_1 + H_2 + H_3 \equiv a \mod 2])\\
&= \Pr\{H_3 = \text{rem}(a - b - c, 2) \text{ and } H_1 = b \text{ and } H_2 = c\}\\
&= \Pr\{H_3 = \text{rem}(a - b - c, 2)\} \cdot \Pr\{H_1 = b\} \cdot \Pr\{H_2 = c\} && \text{(independence of the flips).}
\end{aligned}$$

Likewise for S and any other two H_i's. So H_1, H_2, H_3, and S are 3-wise independent because any three of them are mutually independent. ■

(c) Verify that H_1, S and M are not mutually independent.

Solution. If $H_1 = 1$ and $S = 0$, then $M = 0$. Hence,

$$\begin{aligned}
\Pr\{H_1 = 1 \text{ and } S = 0 \text{ and } M = 0\} &= \Pr\{H_1 = 1 \text{ and } S = 0\} \\
&= \Pr\{H_1 = 1\} \cdot \Pr\{S = 0\} \qquad \text{(part (b))} \\
&= \frac{1}{2} \cdot \frac{1}{2} \\
&\neq \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{3}{4} \\
&= \Pr\{H_1 = 1\} \cdot \Pr\{S = 0\} \cdot \Pr\{M = 0\}.
\end{aligned}$$

■

Appendix

Random Variables

A (real-valued) *random variable* over a given sample space, $\mathcal{S}$, is a total function from $\mathcal{S} \to \mathbb{R}$.

Random variables $R_1, R_2, \ldots$ are ***mutually independent*** iff

$$\Pr\left\{\bigcap_i [R_i = x_i]\right\} = \prod_i \Pr\{R_i = x_i\},$$

for all $x_1, x_2, \cdots \in \mathbb{R}$. They are ***$k$-wise independent*** iff $\{R_i \mid i \in J\}$ are mutually independent for all subsets $J \subset \mathbb{N}$ with $|J| = k$. Variables that are 2-wise independent are also called ***pairwise independent***.

The Four-Step Method

This is a good approach to questions of the form, "What is the probability that ——-?" Intuition *will* mislead you, but this formal approach gives the right answer every time.

1. Find the sample space. (Use a tree diagram.)
2. Define events of interest. (Mark leaves corresponding to these events.)
3. Determine outcome probabilities:
   (a) Assign edge probabilities.
   (b) Compute outcome probabilities. (Multiply along root-to-leaf paths.)
4. Compute event probabilities. (Sum the probabilities of all outcomes in the event.)

14.6 Problem Set 11

Do any four out of five problems for full credit on this problem set.

Problem 14.6.1. Outside of their hum-drum duties as 6.042 LAs, Jeff is trying to learn to levitate using only intense concentration and Jessica is trying to become the world champion flaming torch juggler. Suppose that Jeff's probability of success is $1/6$, Jessica's chance of success is $1/4$, and these two events are independent.

(a) If at least one of them succeeds, what is the probability that Jeff learns to levitate?

Solution. Let L be the event that Jeff learns to levitate, and let F be the event that Jessica becomes the flaming torch juggler champion. We can work out the desired probability as follows:

$$\begin{aligned}
\Pr\{L \mid (L \cup F)\} &= \frac{\Pr\{L \cap (L \cup F)\}}{\Pr\{L \cup F\}} \\
&= \frac{\Pr\{L\}}{1 - \Pr\{\overline{L} \cap \overline{F}\}} \\
&= \frac{1/6}{1 - (1 - 1/6)(1 - 1/4)} \\
&= \frac{4}{9}
\end{aligned}$$

The first step uses the definition of conditional probability. In the second step, we rewrite both the top and bottom of the fraction using set identities. Then we substitute in the given probability and simplify. ■

(b) If at most one of them succeeds, what is the probability that Jessica becomes the world flaming torch juggler champion?

Solution. Define events L and F as before.

$$\begin{aligned}
\Pr\{F \mid (\overline{L} \cup \overline{F})\} &= \frac{\Pr\{F \cap (\overline{L} \cup \overline{F})\}}{\Pr\{\overline{L} \cup \overline{F}\}} \\
&= \frac{\Pr\{F \cap \overline{L}\}}{1 - \Pr\{L \cap F\}} \\
&= \frac{(1/4) \cdot (5/6)}{1 - (1/6) \cdot (1/4)} \\
&= \frac{5}{23}
\end{aligned}$$

■

(c) If exactly one of them succeeds, what is the probability that it is Jeff?

Solution.

$$\begin{aligned}\Pr\left\{L \mid ((L \cap \overline{F}) \cup (\overline{L} \cap F))\right\} &= \frac{\Pr\left\{L \cap \overline{F}\right\}}{\Pr\left\{((L \cap \overline{F}) \cup (\overline{L} \cap F))\right\}} \\ &= \frac{(1/6)\cdot(3/4)}{(1/6)\cdot(3/4)+(5/6)\cdot(1/4)} \\ &= \frac{3}{8}\end{aligned}$$

■

Problem 14.6.2. Independently flip three fair coins ("fair" means equally likely to come up with a head or a tail), and let H_i be the indicator variable for a head occurring on the ith flip, for $i = 1, 2, 3$. Define $C ::= H_1 + H_2 + H_3$ to be the number of heads flipped, M to be the indicator variable for the event $[H_1 = H_2 = H_3]$ that all three coins match, and S be the indicator variable for the event $[C \equiv 1 \mod 2]$ that an odd number of heads are flipped.

Class Problem 12F, Prob 3 shows that none of these variables is independent of C, and that the variables H_1, H_2, H_3, S are 3-wise independent, but not mutually independent.

(a) Verify that H_1, S and M are not mutually independent.

Solution. If $H_1 = 1$ and $S = 0$, then $M = 0$. Hence,

$$\begin{aligned}\Pr\{H_1 = 1 \text{ and } S = 0 \text{ and } M = 0\} &= \Pr\{H_1 = 1 \text{ and } S = 0\} \\ &= \Pr\{H_1 = 1\}\cdot\Pr\{S = 0\} \qquad \text{(indep. of } H_1 \text{ and } S\text{)} \\ &= \frac{1}{2}\cdot\frac{1}{2} \\ &\neq \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{3}{4} \\ &= \Pr\{H_1 = 1\}\cdot\Pr\{S = 0\}\cdot\Pr\{M = 0\}.\end{aligned}$$

■

(b) Verify that the five variables other than C are pairwise independent.

Since 3-wise independence by definition implies pairwise independence, any two of H_1, H_2, H_3, S are pairwise independent, so we need only verify that H_i and M are pairwise independent and that S and M are pairwise independent.

Solution. To see that H_i and M are pairwise independent, we just check the cases where $i = 1$ and $H_1 = 1$. The other cases are symmetric.

$$\begin{aligned}\Pr\{H_1 = 1 \text{ and } M = 0\} &= \Pr\{HTT, HHT, HTH\} \\ &= \frac{3}{8} = \frac{1}{2}\cdot\frac{3}{4} = \Pr\{H_1 = 1\}\cdot\Pr\{M = 0\} \\ \Pr\{H_1 = 1 \text{ and } M = 1\} &= \Pr\{HHH\} \\ &= \frac{1}{8} = \frac{1}{2}\cdot\frac{1}{4} = \Pr\{H_1 = 1\}\cdot\Pr\{M = 1\}.\end{aligned}$$

To see that S and M are pairwise independent, we check each of the cases.

$$\begin{aligned}
\Pr\{S=0 \text{ and } M=0\} &= \Pr\{HHT, HTH, THH\} \\
&= \frac{3}{8} = \frac{1}{2}\cdot\frac{3}{4} = \Pr\{S=0\}\cdot\Pr\{M=0\} \\
\Pr\{S=0 \text{ and } M=1\} &= \Pr\{TTT\} \\
&= \frac{1}{8} = \frac{1}{2}\cdot\frac{1}{4} = \Pr\{S=0\}\cdot\Pr\{M=1\} \\
\Pr\{S=1 \text{ and } M=0\} &= \Pr\{TTH, THT, HTT\} \\
&= \frac{3}{8} = \frac{1}{2}\cdot\frac{3}{4} = \Pr\{S=1\}\cdot\Pr\{M=0\} \\
\Pr\{S=1 \text{ and } M=1\} &= \Pr\{HHH\} \\
&= \frac{1}{8} = \frac{1}{2}\cdot\frac{1}{4} = \Pr\{S=1\}\cdot\Pr\{M=1\}.
\end{aligned}$$

■

(c) Verify that no set of three variables including both M and H_i for any $i \in \{1,2,3\}$ is 3-wise independent.

Solution. We've seen that H_1, S, and M are not mutually independent. The same reasoning can be applied when substituting H_2 or H_3 for H_1. We've also seen that none of the variables is independent of C, so C cannot be one of the three variables in the set and still have the set be 3-wise independent.

That leaves sets of the form $\{M, H_i, H_j\}$, i and j not equal and between 1 and 3. If $H_i = H_j = M = 1$, then all coins are heads.

$$\begin{aligned}
\Pr\{H_i=1 \text{ and } H_j=1 \text{ and } M=1\} &= \Pr\{H_1=1 \text{ and } H_2=1 \text{ and } H_3=1\} \\
&= \Pr\{H_1=1\}\cdot\Pr\{H_2=1\}\cdot\Pr\{H_3=1\} \quad \text{(indep. of } H_i\text{'s)} \\
&= \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} \\
&\neq \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{4} \\
&= \Pr\{H_i=1\}\cdot\Pr\{H_j=1\}\cdot\Pr\{M=1\}.
\end{aligned}$$

■

Problem 14.6.3. Suppose you have three cards: $A\heartsuit$, $A\spadesuit$, and a Jack. From these, you choose a random hand (that is, each card is equally likely to be chosen) of two cards, and let K be the number of Aces in your hand. You then randomly pick one of the cards in the hand and reveal it.

(a) Describe a simple probability space (that is, outcomes and their probabilities) for this scenario, and list the outcomes in each of the following events:

1. $[K \geq 1]$, (that is, your hand has an Ace in it),
2. $A\heartsuit$ is in your hand,
3. the revealed card is an $A\heartsuit$,
4. the revealed card is an Ace.

Solution. Consider each outcome as a pair of cards, the first of which is the revealed card. Each outcome is equally likely (probability $1/6$).

The sets of outcomes are then as follows:

1. $[K \geq 1]$: all pairs: $\{(A\heartsuit, A\spadesuit), (A\heartsuit, \text{Jack}), (A\spadesuit, A\heartsuit), (A\spadesuit, \text{Jack}), (\text{Jack}, A\heartsuit), (\text{Jack}, A\spadesuit)\}$,
2. $A\heartsuit$ is in your hand: $\{(A\heartsuit, A\spadesuit), (A\heartsuit, \text{Jack}), (A\spadesuit, A\heartsuit), (\text{Jack}, A\heartsuit)\}$,
3. the revealed card is an $A\heartsuit$: $\{(A\heartsuit, A\spadesuit), (A\heartsuit, \text{Jack})\}$,
4. the revealed card is an Ace: $\{(A\heartsuit, A\spadesuit), (A\heartsuit, \text{Jack}), (A\spadesuit, A\heartsuit), (A\spadesuit, \text{Jack})\}$.

■

(b) Then calculate $\Pr\{K = 2 \mid E\}$ for E equal to each of the four events in part (a). Notice that most, but *not all*, of these probabilities are equal.

Solution. First, note that $\Pr\{K = 2\} = 1/3$.

1. $\Pr\{K = 2 \mid K \geq 1\} = \Pr\{K = 2\}/1 = 1/3$,
2. $\Pr\{K = 2 \mid A\heartsuit \text{ is in your hand}\} = \Pr\{K = 2\}/(2/3) = 1/2$,
3. $\Pr\{K = 2 \mid \text{the revealed card is an } A\heartsuit\} = \Pr\{(A\heartsuit, A\spadesuit)\}/(1/3) = 1/2$,
4. $\Pr\{K = 2 \mid \text{the revealed card is an Ace}\} = \Pr\{K = 2\}/(2/3) = 1/2$.

■

Now suppose you have a deck with d distinct cards, a different kinds of Aces (including an $A\heartsuit$), you draw a random hand with h cards, and then reveal a random card from your hand.

(c) Prove that $\Pr\{A\heartsuit \text{ is in your hand}\} = h/d$.

Solution. The number, N, of hands is

$$N = \binom{d}{h}.$$

$$\begin{aligned}
\Pr\{A\heartsuit \text{ is in your hand}\} &= \frac{\#\text{ hands with } A\heartsuit}{N} \\
&= \frac{\#\, h-1 \text{ card hands from a deck with no } A\heartsuit}{N} \\
&= \frac{\binom{d-1}{h-1}}{N} \\
&= \frac{(d-1)!h!(d-h)!}{(h-1)!(d-h)!d!} && \left(\text{def. of } \binom{m}{n}\right) \\
&= h/d. && \text{(simplification)}
\end{aligned}$$

■

(d) Prove that

$$\Pr\{K = 2 \mid A\heartsuit \text{ is in your hand}\} = \Pr\{K = 2\} \cdot \frac{2d}{ah}. \tag{14.11}$$

Solution.

$$\begin{aligned}
&\Pr\{K = 2 \mid A\heartsuit \text{ is in your hand}\} \\
&= \frac{\Pr\{K = 2 \text{ and } A\heartsuit \text{ is in your hand}\}}{\Pr\{A\heartsuit \text{ is in your hand}\}} \\
&= \frac{\Pr\{K = 2 \text{ and } A\heartsuit \text{ is in your hand}\}}{h/d} && \text{(part (c))} \\
&= \frac{\Pr\{K = 2\} \cdot \Pr\{A\heartsuit \text{ is in your hand} \mid K = 2\}}{h/d} \\
&= \frac{\Pr\{K = 2\} \cdot 2/a}{h/d} \\
&= \Pr\{K = 2\} \cdot \frac{2d}{ah}
\end{aligned}$$

■

(e) Conclude that $\Pr\{K = 2 \mid \text{the revealed card is an Ace}\} = \Pr\{K = 2 \mid A\heartsuit \text{ is in your hand}\}$.

Solution. Note that

$$\Pr\{\text{the revealed card is an Ace}\} = \frac{a}{d}, \tag{14.12}$$

since the probability of revealing an Ace from the random hand is simply the probability that a random card is an Ace. Now,

$$\begin{aligned}
&\Pr\{K = 2 \mid \text{the revealed card is an Ace}\} \\
&= \frac{\Pr\{K = 2 \text{ and the revealed card is an Ace}\}}{\Pr\{\text{the revealed card is an Ace}\}} \\
&= \frac{\Pr\{K = 2 \text{ and the revealed card is an Ace}\}}{a/d} && \text{(by (14.12))} \\
&= \frac{\Pr\{K = 2\} \Pr\{\text{the revealed card is an Ace} \mid K = 2\}}{a/d} \\
&= \frac{\Pr\{K = 2\}\,(2/h)}{a/d} \\
&= \frac{2d}{ah} \cdot \Pr\{K = 2\} \\
&= \Pr\{K = 2 \mid A\heartsuit \text{ is in your hand}\}. && \text{(by (14.11))}
\end{aligned}$$

■

Problem 14.6.4. Let's play a game! We repeatedly flip a fair coin. You have the sequence HHT, and I have the sequence HTT. If your sequence comes up first, then you win. If my sequence comes up first, then I win. For example, if the sequence of tosses is:

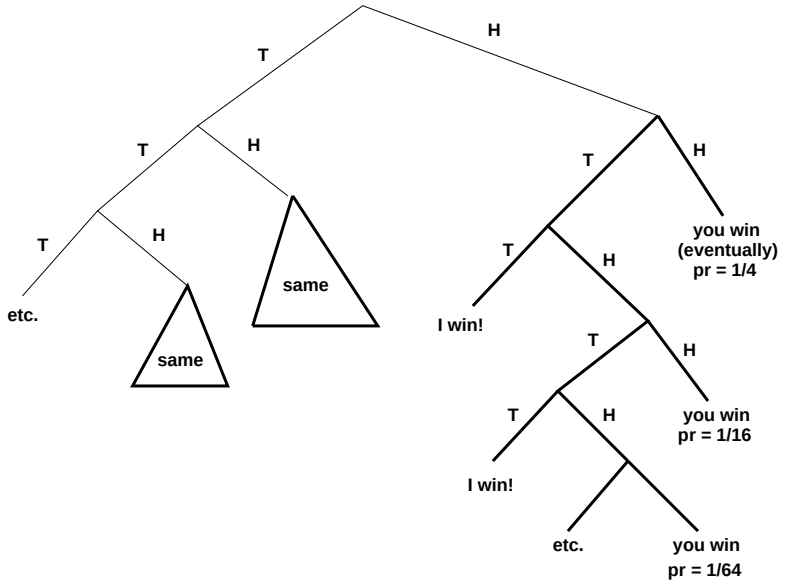
$$TTHTHT\underline{HHT}$$

then you win. What is the probability that you win? It may come as a surprise that the answer is very different from $1/2$.

This problem is tricky, because the game could go on for an arbitrarily long time. Draw enough of the tree diagram to see a pattern, and then sum up the probabilities of the (infinitely many) outcomes in which you win.

It turns out that for any sequence of three flips, there is another sequence that is likely to come up before it. So there is no sequence which turns up earliest! ...and given any sequence, knowing how to pick another sequence that comes up sooner more than half the time gives you a nice chance to fool people gambling at a bar :-)

Solution. A partial tree diagram is shown below. All edge probabilities are $1/2$.



Let's first focus on the subtree shown in bold. Note that if two heads are flipped in a row, then you are guaranteed to win eventually. The sum of the probabilities of all your winning outcomes in this subtree is:

$$\begin{aligned}\frac{1}{4}+\frac{1}{16}+\frac{1}{64}+\cdots &= \frac{1}{4}\cdot\frac{1}{1-1/4}\\ &= \frac{1}{3}\end{aligned}$$

The uppermost subtree marked **same** is identical to the one shown in bold, except that each outcome probability is reduced by $1/2$, because it is one edge farther from the root. Thus, the sum of your winning outcomes in this subtree is $1/6$. Similarly, the sum of your winning outcomes in the next subtree marked **same** is $1/12$, and so forth. Overall, your probability of winning is:

$$\begin{aligned}\frac{1}{3}+\frac{1}{6}+\frac{1}{12}+\cdots &= \frac{1}{3}\cdot\frac{1}{1-1/2}\\ &= \frac{2}{3}\end{aligned}$$

■

Problem 14.6.5. The Boston Globe, April 25, 2007 reported that the election between two candidates for Natick Town Moderator resulted in a tie after 4000 residents voted. The article went on to say

> Just what are the odds of a tie in a townwide election? Several election specialists said they could not think of one occurring previously in the state.
>
> "I'd suspect it's not astronomical in the cosmic sense, but it certainly is in the earthly sense," said Thomas Patterson, an elections specialist at Harvard's Kennedy School of Government.

(a) Suppose that each of the 4000 residents was equally likely to vote for either of the two candidates. Estimate the probability of a tie.

Solution.

$$\Pr\{tie\} \sim \sqrt{\frac{2}{\pi 4000}} \approx \frac{1}{80},$$

which is hardly unearthly. In fact, with all the elections in Natick size towns in recent years, we'd expect numerous ties to occur, **if** we accept this probability model of random voters. ■

(b) Comment on the plausibility of the "random vote" model of part (a). Can you suggest other models that seem more convincing?

Solution. This was meant as a discussion question to get you thinking about how probability models should be chosen. There's no compelling logical or empirical reason that we know of to commit to the random vote model. We'll report on any more convincing models that students suggest after this problem set is graded. ■

14.7 Miniquiz May 11

Problem 14.7.1. [A Baseball Series] The New York Yankees and the Boston Red Sox are playing a two-out-of-three series. (In other words, they play until one team has won two games. Then that team is declared the overall winner and the series ends.) Assume that the Red Sox win each game with probability $3/5$, regardless of the outcomes of previous games.

Answer the questions below. (Partial credit may be awarded based on work you show.)

(a) What is the probability that the Red Sox win the series?

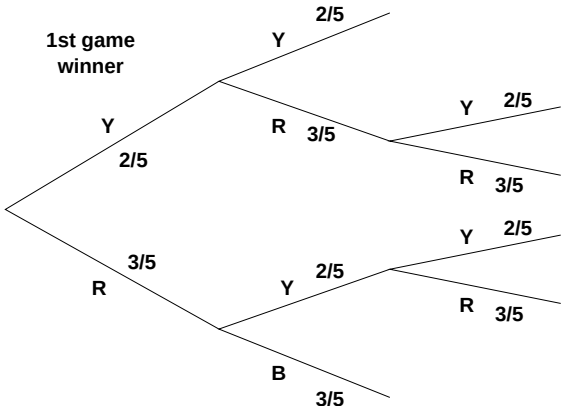
Solution. We can read the probability from the tree diagram below:

$$\Pr\{\text{Red Sox win}\} = \frac{18}{125} + \frac{18}{125} + \frac{9}{25} = \frac{81}{125}$$

■

(b) Given that the Red Sox won the series, what is the probability they lost the first game?

Solution. A tree diagram is worked out below.



outcome	3 games played?	winner lost first game?	correct team wins?	Probability
YY				4/25
YRY	X			12/125
YRR	X	X	X	18/125
RYY	X	X		12/125
RYR	X		X	18/125
RR			X	9/25

From the tree diagram, we get:

$$\frac{18/125}{81/125} = \frac{18}{81} = \frac{2}{9}.$$

We might also have calculated this without the tree:

$$\begin{aligned}
&\Pr\{\text{Red Sox lost the first game} \mid \text{Red Sox won the series}\} \\
&= \frac{\Pr\{\text{Red Sox lost the first game and won the series}\}}{\Pr\{\text{Red Sox won the series}\}} \\
&= \frac{\Pr\{\text{Red Sox lost the first game and won the next two games}\}}{81/125} \\
&= \frac{(2/5)(3/5)^2}{81/125} \\
&= 2/9.
\end{aligned}$$

■

Problem 14.7.2. There is a rare and serious disease called *Beaver Fever* which afflicts about 1 person in 1000. Victims eventually start telling math jokes at cocktail parties, thinking others will find the jokes amusing.

Doctor X says he can test for the disease.

- If a person has Beaver Fever, he says "yes" with probability 0.99.
- If a person doesn't have it, he says "no" with probability 0.97.

(a) What is the probability that Doctor X says "yes"?

Solution. Let D be the event that you have the disease, and Y be the event that the Doctor says you have the disease. By the Total Probability Law:

$$\Pr\{Y\} = \Pr\{Y \mid D\}\Pr\{D\} + \Pr\left\{Y \mid \overline{D}\right\}\Pr\left\{\overline{D}\right\} = 0.99(1/1000) + 0.03(1 - 1/1000) = 0.03096.$$

■

(b) If Doctor X says someone has the disease, what is the probability that person really does have the disease?

Solution.

$$\begin{aligned}
\Pr\{D \mid Y\} &= \frac{\Pr\{D \text{ and } Y\}}{\Pr\{Y\}} \\
&= \frac{\Pr\{Y \mid D\}\Pr\{D\}}{0.03096} \\
&= \frac{0.99(1/1000)}{0.03096} \\
&= \frac{99}{3096} \\
&= \frac{11}{344} \approx \frac{1}{32}.
\end{aligned}$$

■

Problem 14.7.3. You want to estimate the fraction of voters in the entire nation that will prefer Donald Duck in the upcoming elections. To do this, you pick a suitable number, n, perform n successive independent selections of a random voter, and ask each selected voter whether they will vote for the Donald. You find that a fraction of 0.53 of your selections will vote for the Donald. You also find that Mickey Mouse happened to be one of the voters you selected.

You also calculate that if you independently flip a fair coin n times, the fraction of Heads flipped will be between 0.49 and 0.51 with probability at least 0.97.

Circle all the following statements that are true:

- Assuming no voter was selected more than once, the probability is 0.53 that a randomly chosen voter *among those you selected* will vote for the Donald.
- The probability is 0.53 that a randomly chosen voter *among all the voters in the nation* will vote for the Donald.
- The probability is 0.53 that Mickey Mouse will vote for the Donald.
- You are 97% confident that the probability is 0.53 that Mickey Mouse will vote for the Donald.
- You are 97% confident that the Donald will get between 0.52 and 0.54 of the vote.
- Assuming no voter was selected more than once, the probability is at least 0.97 that the fraction of voters in the nation who will vote for the Donald is between 0.52 and 0.54.
- Even if several voters were sampled more than once, you can still say that the probability is at least 0.97 that the fraction of voters in the nation who will vote for the Donald is between 0.52 and 0.54.

Solution. The only true statements are the first (worth 1 point) and the fifth (worth 2 points). Each wrongly circled statement costs 1 point. ■

Problem 14.7.4. Independently flip two biased coins, each with probability p of coming up Heads, where $0 < p < 1$. Let H_1 be the indicator variable for a Head occurring on the first coin, and likewise H_2 for a Head on the second coin. Define $S ::= H_1 \oplus H_2$ where $\oplus$ denotes addition modulo 2.

Prove that if H_1 and S are independent, then $p = 1/2$.

Hint: $[H_1 = 1$ and $H_2 = 1]$ and $[H_1 = 1$ and $S = 0]$ are the same event.

Solution. Assume S and H_1 are independent. Then

$$\begin{aligned}
p^2 &= \Pr\{H_1 = 1\} \cdot \Pr\{H_2 = 1\} \\
&= \Pr\{H_1 = 1 \text{ and } H_2 = 1\} && \text{(indep of } H_1, H_2\text{)} \\
&= \Pr\{H_1 = 1 \text{ and } S = 0\} && \text{(hint)} \\
&= \Pr\{H_1 = 1\} \cdot \Pr\{S = 0\} && \text{(indep of } H_1, S\text{)} \\
&= p \cdot (p^2 + (1-p)^2).
\end{aligned}$$

Since $0 < p$, we can divide by p to obtain

$$\begin{aligned} p &= p^2 + (1-p)^2 \\ p - p^2 &= (1-p)^2 \\ p(1-p) &= (1-p)^2 \\ p &= 1-p && \text{(since } p < 1\text{)} \\ p &= \frac{1}{2}. \end{aligned}$$

■

Appendix

The probability of an event A given an event B, is

$$\Pr\{A \mid B\} ::= \frac{\Pr\{A \cap B\}}{\Pr\{B\}} \text{ where } \Pr\{B\} \neq 0$$

The Total Probability Law is

$$\Pr\{A\} = \Pr\{A \mid E\} \cdot \Pr\{E\} + \Pr\left\{A \mid \overline{E}\right\} \cdot \Pr\left\{\overline{E}\right\}.$$

The complement law says $\Pr\{E\} + \Pr\left\{\overline{E}\right\} = 1$.

The inclusion-exclusion formula says:

$$\Pr\{E_1 \cup \cdots \cup E_n\} = \sum_i \Pr\{E_i\} - \sum_{i,j} \Pr\{E_i \cap E_j\} + \sum_{i,j,k} \Pr\{E_i \cap E_j \cap E_k\} - \cdots.$$

Two events A and B are independent iff $\Pr\{A \cap B\} = \Pr\{A\} \Pr\{B\}$.

Random variables R_1 and R_2 are ***independent*** iff

$$\Pr\{R_1 = r_1 \text{ and } R_2 = r_2\} = \Pr\{R_1 = r_1\} \cdot \Pr\{R_2 = r_2\}$$

for all $r_1, r_2, \cdots \in \mathbb{R}$.

14.8 In-Class Problems Week 13, Mon.

Problem 14.8.1. Suppose $H_{n,p}$ and $H_{m,p}$ are independent binomially distributed random variables. What is $\Pr\{H_{n,p} + H_{m,p} = k\}$?

Solution. The pdf of $H_{n,p}$ is the probability of tossing k Heads out of n independent flips of a coin with bias p. Likewise for $H_{m,p}$ and m flips. Since $H_{n,p}$ and $H_{m,p}$ are independent, the pdf of $H_{n,p} + H_{m,p}$ corresponds to $n + m$ independent flips, that is, $H_{n,p} + H_{m,p}$ is a $(n + m, p)$-binomial variable. Hence,

$$\text{PDF}_{H_{n,p}+H_{m,p}}(k) = \binom{m+n}{k} p^k (1-p)^{(m+n)-k}.$$

■

Problem 14.8.2. **(a)** Prove that

$$\Pr\left\{H_{2n,1/2} = n\right\} \sim \frac{1}{\sqrt{\pi n}}. \tag{14.13}$$

Hint: Use Stirling's approximation (in the appendix). Note that the entropy function $H(\alpha)$ is 1 for $\alpha = 1/2$.

Solution. The probability that $H_{2n,1/2} = m$ is

$$\frac{\binom{2n}{m}}{2^{2n}}.$$

Using Stirling for an approximation:

$$\begin{aligned}
\binom{2n}{m} &\sim \frac{\sqrt{2\pi 2n}\left(\frac{2n}{e}\right)^{2n}}{\sqrt{2\pi m}\left(\frac{m}{e}\right)^{m}\sqrt{2\pi(2n-m)}\left(\frac{2n-m}{e}\right)^{2n-m}} \\
&= \frac{\sqrt{2n}\left(\frac{2n}{e}\right)^{m}\left(\frac{2n}{e}\right)^{2n-m}}{\sqrt{2\pi m(2n-m)}\left(\frac{m}{e}\right)^{m}\left(\frac{2n-m}{e}\right)^{2n-m}} \\
&= \sqrt{\frac{n}{\pi m(2n-m)}}\left(\frac{2n}{m}\right)^{m}\left(\frac{2n}{2n-m}\right)^{2n-m}.
\end{aligned} \tag{14.14}$$

Setting $m = n$ in (14.14) yields

$$\binom{2n}{n} \sim \frac{2^{2n}}{\sqrt{n\pi}}, \tag{14.15}$$

so

$$\Pr\left\{H_{2n,1/2} = n\right\} = \frac{\binom{2n}{n}}{2^{2n}} \sim \frac{1}{\sqrt{n\pi}}, \tag{14.16}$$

proving (14.13). ■

(b) Estimate the probability that the number of heads in 400 flips of a fair coin will be between 195 and 205, and likewise that in 10,000 flips it will be between 4980 and 5020.

Also, discuss writing a program to calculate the exact answer.

Solution. The numbers are all too large to calculate exactly by hand. Computing 400!, 190!, and 210! exactly will overflow on any calculator, so an exact computation by calculator would also be impractical. In fact, it will cause numerical overflow errors in most programming languages. Real Scheme has infinite precision rational arithmetic and can handle 400!, but typically overflows when computing 10,000!, so exact computation in the 10,000 case is unlikely to work in any language—remember the answer will have around 40,000 digits. But no applications require anywhere near such accuracy.

So we settle for an approximate solution, and make use of Stirling's formula to derive one. But there are still pitfalls: $(400/e)^{400}$ will overflow floating point arithmetic on calculators and essentially all computers, so the simplified version (14.14) would be needed. Since floating point is usually good for 8 to 10 places on most machines, getting six place accuracy should be manageable.

A calculator will yield the value of the righthand side of equation (14.13) at $n = 200$; it is ≈ 0.0399. But all the probabilities for 195 to 205 heads are about the same, so an offhand estimate would be $11 \times 0.0399 \approx 0.439$. The actual answer is 0.418.

Similarly, for $n = 5,000$, the value of (14.13) is about 0.00798, and all the probabilities for 4980 to 5020 heads are about the same, so an offhand estimate would be $41 \times 0.00798 \approx 0.327$. The actual answer is 0.318. ■

Appendix

Stirling's Approximation

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Binomial Variables

A random variable, J, is (n, p)-*binomial* for $n \in \mathbb{N}$ and $0 < p < 1$, if

$$J = \sum_{k=1}^{n} H_k$$

where $H_1, H_2, \ldots, H_n$ are mutually independent indicator variables with $\Pr\{H_i = 1\} = p$ for all i.

Equivalently, J is (n, p)-*binomial* iff PDF_J has the (n, p)-*binomial distribution*:

$$\text{PDF}_J(k) ::= \binom{n}{k} p^k (1-p)^{n-k},$$

for $0 \leq k \leq n$.

Binomial bounds

If J is an (n, p)-binomial variable, the following formula gives a fairly tight upper bound on PDF_J.

$$\text{PDF}_J(\alpha n) \leq \frac{2^{nH(\alpha)}}{\sqrt{2\pi\alpha(1-\alpha)n}} \cdot p^{\alpha n}(1-p)^{(1-\alpha)n} \tag{14.17}$$

where H is the entropy function,

$$H(\alpha) ::= -\left(\alpha \log_2 \alpha + (1-\alpha)\log_2(1-\alpha)\right).$$

The bounding formula is also asymptotically equal to PDF_J.

Chapter 15

Expectation & Variance

15.1 Expectation

15.1.1 Average & Expected Value

The ***expectation*** of a random variable is its average value, where each value is weighted according to the probability that it comes up. The expectation is also called the ***expected value*** or the ***mean*** of the random variable.

For example, suppose we select a student uniformly at random from the class, and let R be the student's quiz score. Then $\mathrm{E}[R]$ is just the class average —the first thing everyone wants to know after getting their test back! Similarly, the first thing you usually want to know about a random variable is its expected value.

Definition 15.1.1.

$$\begin{aligned}\mathrm{E}[R] ::= & \sum_{x \in \text{range}(R)} x \cdot \Pr\{R = x\} \\ = & \sum_{x \in \text{range}(R)} x \cdot \mathrm{PDF}_R(x).\end{aligned} \tag{15.1}$$

Let's work through an example. Let R be the number that comes up on a fair, six-sided die. Then by (15.1), the expected value of R is:

$$\begin{aligned}\mathrm{E}[R] &= \sum_{k=1}^{6} k \cdot \frac{1}{6} \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= \frac{7}{2}\end{aligned}$$

This calculation shows that the name "expected value" is a little misleading; the random variable might *never* actually take on that value. You can't roll a $3\frac{1}{2}$ on an ordinary die!

There is an even simpler formula for expectation:

Theorem 15.1.2. *If R is a random variable defined on a sample space, $\mathcal{S}$, then*

$$\mathrm{E}[R] = \sum_{\omega \in \mathcal{S}} R(\omega) \Pr\{\omega\} \tag{15.2}$$

The proof of Theorem 15.1.2, like many of the elementary proofs about expectation in these notes, follows by judicious regrouping of terms in the defining sum (15.1):

Proof.

$$\begin{aligned}
\mathrm{E}[R] ::= & \sum_{x \in \text{range}(R)} x \cdot \Pr\{R = x\} && \text{(Def 15.1.1 of expectation)} \\
= & \sum_{x \in \text{range}(R)} x \left(\sum_{\omega \in [R=x]} \Pr\{\omega\} \right) && \text{(def of } \Pr\{R = x\}) \\
= & \sum_{x \in \text{range}(R)} \sum_{\omega \in [R=x]} x \Pr\{\omega\} && \text{(distributing } x \text{ over the inner sum)} \\
= & \sum_{x \in \text{range}(R)} \sum_{\omega \in [R=x]} R(\omega) \Pr\{\omega\} && \text{(def of the event } [R = x]) \\
= & \sum_{\omega \in \mathcal{S}} R(\omega) \Pr\{\omega\}
\end{aligned}$$

The last equality follows because the events $[R = x]$ for $x \in \text{range}(R)$ partition the sample space, $\mathcal{S}$, so summing over the outcomes in $[R = x]$ for $x \in \text{range}(R)$ is the same as summing over $\mathcal{S}$. □

In general, the defining sum (15.1) is better for calculating expected values and has the advantage that it does not depend on the sample space, but only on the density function of the random variable. On the other hand, the simpler sum over all outcomes given in Theorem 15.1.2 is sometimes easier to use in proofs about expectation.

15.1.2 Expected Value of an Indicator Variable

The expected value of an indicator random variable for an event is just the probability of that event. (Remember that a random variable I_A is the indicator random variable for event A, if $I_A = 1$ when A occurs and $I_A = 0$ otherwise.)

Lemma 15.1.3. *If I_A is the indicator random variable for event A, then*

$$\mathrm{E}[I_A] = \Pr\{A\}.$$

Proof.

$$\begin{aligned}
\mathrm{E}[I_A] &= 1 \cdot \Pr\{I_A = 1\} + 0 \cdot \Pr\{I_A = 0\} \\
&= \Pr\{I_A = 1\} \\
&= \Pr\{A\}. && \text{(def of } I_A)
\end{aligned}$$

□

For example, if A is the event that a coin with bias p comes up heads, $\mathrm{E}[I_A] = \Pr\{I_A = 1\} = p$.

15.1.3 Mean Time to Failure

A computer program crashes at the end of each hour of use with probability p, if it has not crashed already. What is the expected time until the program crashes?

If we let C be the number of hours until the crash, then the answer to our problem is $\mathrm{E}[C]$. Now the probability that, for $i > 0$, the first crash occurs in the ith hour is the probability that it does not crash in each of the first $i-1$ hours and it does crash in the ith hour, which is $(1-p)^{i-1}p$. So from formula (15.1) for expectation, we have

$$\begin{aligned}
\mathrm{E}[C] &= \sum_{i\in\mathbb{N}} i \cdot \Pr\{R = i\} \\
&= \sum_{i\in\mathbb{N}^+} i(1-p)^{i-1}p \\
&= p\sum_{i\in\mathbb{N}^+} i(1-p)^{i-1} \\
&= p\frac{1}{(1-(1-p))^2} && \text{(by (15.3))} \\
&= \frac{1}{p}
\end{aligned}$$

As an alternative to applying the formula

$$\sum_{i\in\mathbb{N}^+} ix^{i-1} = \frac{1}{(1-x)^2} \tag{15.3}$$

from Notes 11 (which you remembered, right?), there is a useful trick for calculating expectations of nonegative integer valued variables:

Lemma 15.1.4. *If R is a nonegative integer-valued random variable, then:*

$$\mathrm{E}[R] = \sum_{i\in\mathbb{N}} \Pr\{R > i\} \tag{15.4}$$

Proof. Consider the sum:

$$\begin{array}{ccccccc}
\Pr\{R=1\} & + & \Pr\{R=2\} & + & \Pr\{R=3\} & + & \cdots \\
 & + & \Pr\{R=2\} & + & \Pr\{R=3\} & + & \cdots \\
 & & & + & \Pr\{R=3\} & + & \cdots \\
 & & & & & + & \cdots
\end{array}$$

The successive columns sum to $1\cdot\Pr\{R=1\}$, $2\cdot\Pr\{R=2\}$, $3\cdot\Pr\{R=3\}$, ... Thus, the whole sum is equal to:

$$\sum_{i\in\mathbb{N}} i\cdot\Pr\{R=i\}$$

which equals $\mathrm{E}[R]$ by (15.1). On the other hand, the successive rows sum to $\Pr\{R>0\}$, $\Pr\{R>1\}$, $\Pr\{R>2\}$, ... Thus, the whole sum is also equal to:

$$\sum_{i\in\mathbb{N}} \Pr\{R>i\},$$

which therefore must equal $\mathrm{E}[R]$ as well. □

Now $\Pr\{C > i\}$ is easy to evaluate: a crash happens later than the ith hour iff the system did not crash during the first i hours, which happens with probability $(1-p)^i$. Plugging this into (15.4) gives:

$$\begin{aligned}
\mathrm{E}[C] &= \sum_{i\in\mathbb{N}} (1-p)^i \\
&= \frac{1}{1-(1-p)} && \text{(sum of geometric series)} \\
&= \frac{1}{p}
\end{aligned}$$

So, for example, if there is a 1% chance that the program crashes at the end of each hour, then the expected time until the program crashes is $1/0.01 = 100$ hours. The general principle here is well-worth remembering: if a system fails at each time step with probability p, then the expected number of steps up to the first failure is $1/p$.

As a further example, suppose a couple really wants to have a baby girl. For simplicity assume there is a 50% chance that each child they have is a girl, and the genders of their children are mutually independent. If the couple insists on having children until they get a girl, then how many baby boys should they expect first?

This is really a variant of the previous problem. The question, "How many hours until the program crashes?" is mathematically the same as the question, "How many children must the couple have until they get a girl?" In this case, a crash corresponds to having a girl, so we should set $p = 1/2$. By the preceding analysis, the couple should expect a baby girl after having $1/p = 2$ children. Since the last of these will be the girl, they should expect just one boy.

Something to think about: If every couple follows the strategy of having children until they get a girl, what will eventually happen to the fraction of girls born in this world?

15.1.4 Linearity of Expectation

Expected values obey a simple, very helpful rule called *Linearity of Expectation*. Its simplest form says that the expected value of a sum of random variables is the sum of the expected values of the variables.

Theorem 15.1.5. *For any random variables R_1 and R_2,*

$$\mathrm{E}[R_1 + R_2] = \mathrm{E}[R_1] + \mathrm{E}[R_2].$$

Proof. Let $T ::= R_1 + R_2$. The proof follows straightforwardly by rearranging terms in the sum (15.2)

$$\begin{aligned}
\mathrm{E}[T] &= \sum_{\omega\in S} T(\omega)\cdot \Pr\{\omega\} && \text{(Theorem 15.1.2)} \\
&= \sum_{\omega\in S} (R_1(\omega) + R_2(\omega))\cdot \Pr\{\omega\} && \text{(def of } T\text{)} \\
&= \sum_{\omega\in S} R_1(\omega)\Pr\{\omega\} + \sum_{\omega\in S} R_2(\omega)\Pr\{\omega\} && \text{(rearranging terms)} \\
&= \mathrm{E}[R_1] + \mathrm{E}[R_2]. && \text{(Theorem 15.1.2)}
\end{aligned}$$

□

A small extension of this proof, which we leave to the reader, implies

Theorem 15.1.6 (Linearity of Expectation). *For random variables R_1, R_2 and constants $a_1, a_2 \in \mathbb{R}$,*

$$\mathrm{E}\left[a_1 R_1 + a_2 R_2\right] = a_1 \mathrm{E}\left[R_1\right] + a_2 \mathrm{E}\left[R_2\right].$$

In other words, expectation is a linear function. A routine induction extends the result to more than two variables:

Corollary 15.1.7. *For any random variables $R_1, \ldots, R_k$ and constants $a_1, \ldots, a_k \in \mathbb{R}$,*

$$\mathrm{E}\left[\sum_{i=1}^{k} a_i R_i\right] = \sum_{i=1}^{k} a_i \mathrm{E}\left[R_i\right].$$

The great thing about linearity of expectation is that *no independence is required*. This is really useful, because dealing with independence is a pain, and we often need to work with random variables that are not independent.

Expected Value of Two Dice

What is the expected value of the sum of two fair dice?

Let the random variable R_1 be the number on the first die, and let R_2 be the number on the second die. We observed earlier that the expected value of one die is 3.5. We can find the expected value of the sum using linearity of expectation:

$$\mathrm{E}\left[R_1 + R_2\right] = \mathrm{E}\left[R_1\right] + \mathrm{E}\left[R_2\right] = 3.5 + 3.5 = 7.$$

Notice that we did *not* have to assume that the two dice were independent. The expected sum of two dice is 7, even if they are glued together (provided the dice remain fair after the gluing). Proving that this expected sum is 7 with a tree diagram would be a bother: there are 36 cases. And if we did not assume that the dice were independent, the job would be really tough!

The Hat-Check Problem

There is a dinner party where n men check their hats. The hats are mixed up during dinner, so that afterward each man receives a random hat. In particular, each man gets his own hat with probability $1/n$. What is the expected number of men who get their own hat?

Letting G be the number of men that get their own hat, we want to find the expectation of G. But all we know about G is that the probability that a man gets his own hat back is $1/n$. There are many different probability distributions of hat permutations with this property, so we don't know enough about the distribution of G to calculate its expectation directly. But linearity of expectation makes the problem really easy.

The trick is to express G as a sum of indicator variables. In particular, let G_i be an indicator for the event that the ith man gets his own hat. That is, $G_i = 1$ if he gets his own hat, and $G_i = 0$ otherwise. The number of men that get their own hat is the sum of these indicators:

$$G = G_1 + G_2 + \cdots + G_n. \tag{15.5}$$

These indicator variables are *not* mutually independent. For example, if $n-1$ men all get their own hats, then the last man is certain to receive his own hat. But, since we plan to use linearity of expectation, we don't have worry about independence!

Now since G_i is an indicator, we know $1/n = \Pr\{G_i = 1\} = \mathrm{E}[G_i]$ by Lemma 15.1.3. Now we can take the expected value of both sides of equation (15.5) and apply linearity of expectation:

$$\begin{aligned}\mathrm{E}[G] &= \mathrm{E}[G_1 + G_2 + \cdots + G_n] \\ &= \mathrm{E}[G_1] + \mathrm{E}[G_2] + \cdots + \mathrm{E}[G_n] \\ &= \frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n} = n\left(\frac{1}{n}\right) = 1.\end{aligned}$$

So even though we don't know much about how hats are scrambled, we've figured out that on average, just one man gets his own hat back!

Expectation of a Binomial Distribution

Suppose that we independently flip n biased coins, each with probability p of coming up heads. What is the expected number that come up heads?

Let J be the number of heads after the flips, so J has the (n,p)-binomial distribution. Now let I_k be the indicator for the kth coin coming up heads. By Lemma 15.1.3, we have

$$\mathrm{E}[I_k] = p.$$

But

$$J = \sum_{k=1}^{n} I_k,$$

so by linearity

$$\mathrm{E}[J] = \mathrm{E}\left[\sum_{k=1}^{n} I_k\right] = \sum_{k=1}^{n} \mathrm{E}[I_k] = \sum_{k=1}^{n} p = pn.$$

In short, the expectation of an (n,p)-binomially distributed variable is pn.

The Coupon Collector Problem

Every time I purchase a kid's meal at Taco Bell, I am graciously presented with a miniature "Racin' Rocket" car together with a launching device which enables me to project my new vehicle across any tabletop or smooth floor at high velocity. Truly, my delight knows no bounds.

There are n different types of Racin' Rocket car (blue, green, red, gray, etc.). The type of car awarded to me each day by the kind woman at the Taco Bell register appears to be selected uniformly and independently at random. What is the expected number of kid's meals that I must purchase in order to acquire at least one of each type of Racin' Rocket car?

The same mathematical question shows up in many guises: for example, what is the expected number of people you must poll in order to find at least one person with each possible birthday? Here, instead of collecting Racin' Rocket cars, you're collecting birthdays. The general question is commonly called the ***coupon collector problem*** after yet another interpretation.

A clever application of linearity of expectation leads to a simple solution to the coupon collector problem. Suppose there are five different types of Racin' Rocket, and I receive this sequence:

blue green green red blue orange blue orange gray

Let's partition the sequence into 5 segments:

$$\underbrace{\text{blue}}_{X_0}\quad\underbrace{\text{green}}_{X_1}\quad\underbrace{\text{green}\quad\text{red}}_{X_2}\quad\underbrace{\text{blue}\quad\text{orange}}_{X_3}\quad\underbrace{\text{blue}\quad\text{orange}\quad\text{gray}}_{X_4}$$

The rule is that a segment ends whenever I get a new kind of car. For example, the middle segment ends when I get a red car for the first time. In this way, we can break the problem of collecting every type of car into stages. Then we can analyze each stage individually and assemble the results using linearity of expectation.

Let's return to the general case where I'm collecting n Racin' Rockets. Let X_k be the length of the kth segment. The total number of kid's meals I must purchase to get all n Racin' Rockets is the sum of the lengths of all these segments:

$$T = X_0 + X_1 + \cdots + X_{n-1}$$

Now let's focus our attention on X_k, the length of the kth segment. At the beginning of segment k, I have k different types of car, and the segment ends when I acquire a new type. When I own k types, each kid's meal contains a type that I already have with probability k/n. Therefore, each meal contains a new type of car with probability $1 - k/n = (n-k)/n$. Thus, the expected number of meals until I get a new kind of car is $n/(n-k)$ by the "mean time to failure" formula. So we have:

$$\mathrm{E}[X_k] = \frac{n}{n-k}$$

Linearity of expectation, together with this observation, solves the coupon collector problem:

$$\begin{aligned}
\mathrm{E}[T] &= \mathrm{E}[X_0 + X_1 + \cdots + X_{n-1}] \\
&= \mathrm{E}[X_0] + \mathrm{E}[X_1] + \cdots + \mathrm{E}[X_{n-1}] \\
&= \frac{n}{n-0} + \frac{n}{n-1} + \cdots + \frac{n}{3} + \frac{n}{2} + \frac{n}{1} \\
&= n\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{3} + \frac{1}{2} + \frac{1}{1}\right) \\
&\quad n\left(\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1} + \frac{1}{n}\right) \\
&= nH_n \sim n\ln n.
\end{aligned}$$

Let's use this general solution to answer some concrete questions. For example, the expected number of die rolls required to see every number from 1 to 6 is:

$$6H_6 = 14.7\ldots$$

And the expected number of people you must poll to find at least one person with each possible birthday is:

$$365H_{365} = 2364.6\ldots$$

The Number-Picking Game

Here is a game that you and I could play that reveals a strange property of expectation.

First, you think of a probability density function on the natural numbers. Your distribution can be absolutely anything you like. For example, you might choose a uniform distribution on $1, 2, \ldots, 6$, like the outcome of a fair die roll. Or you might choose a binomial distribution on $0, 1, \ldots, n$. You can even give every natural number a non-zero probability, provided that the sum of all probabilities is 1.

Next, I pick a random number z according to your distribution. Then, you pick a random number y_1 according to the same distribution. If your number is bigger than mine ($y_1 > z$), then the game ends. Otherwise, if our numbers are equal or mine is bigger ($z \geq y_1$), then you pick a new number y_2 with the same distribution, and keep picking values y_3, y_4, etc. until you get a value that is strictly bigger than my number, z. What is the expected number of picks that you must make?

Certainly, you always need at least one pick, so the expected number is greater than one. An answer like 2 or 3 sounds reasonable, though one might suspect that the answer depends on the distribution. Let's find out whether or not this intuition is correct.

The number of picks you must make is a natural-valued random variable, so from formula (15.4) we have:

$$\mathrm{E}\left[\#\text{ picks by you}\right] = \sum_{k \in \mathbb{N}} \Pr\left\{(\#\text{ picks by you}) > k\right\} \tag{15.6}$$

Suppose that I've picked my number z, and you have picked k numbers $y_1, y_2, \ldots, y_k$. There are two possibilities:

- If there is a unique largest number among our picks, then my number is as likely to be it as any one of yours. So with probability $1/(k+1)$ my number is larger than all of yours, and you must pick again.
- Otherwise, there are several numbers tied for largest. My number is as likely to be one of these as any of your numbers, so with probability greater than $1/(k+1)$ you must pick again.

In both cases, with probability at least $1/(k+1)$, you need more than k picks to beat me. In other words:

$$\Pr\left\{(\#\text{ picks by you}) > k\right\} \geq \frac{1}{k+1} \tag{15.7}$$

This suggests that in order to minimize your rolls, you should choose a distribution such that ties are very rare. For example, you might choose the uniform distribution on $\left\{1, 2, \ldots, 10^{100}\right\}$. In this case, the probability that you need more than k picks to beat me is very close to $1/(k+1)$ for moderate values of k. For example, the probability that you need more than 99 picks is almost exactly 1%. This sounds very promising for you; intuitively, you might expect to win within a reasonable number of picks on average!

Unfortunately for intuition, there is a simple proof that the expected number of picks that you need in order to beat me is *infinite*, regardless of the distribution! Let's plug (15.7) into (15.6):

$$\begin{aligned}\mathrm{E}\left[\#\text{ picks by you}\right] &= \sum_{i \in \mathbb{N}} \frac{1}{k+1} \\ &= \infty\end{aligned}$$

This phenomenon can cause all sorts of confusion! For example, suppose you have a communication network where each packet of data has a $1/k$ chance of being delayed by k or more steps. This sounds good; there is only a 1% chance of being delayed by 100 or more steps. But the *expected* delay for the packet is actually infinite!

There is a larger point here as well: not every random variable has a well-defined expectation. This idea may be disturbing at first, but remember that an expected value is just a weighted average. And there are many sets of numbers that have no conventional average either, such as:

$$\{1, -2, 3, -4, 5, -6, \dots\}$$

Strictly speaking, we should qualify virtually all theorems involving expectation with phrases such as "...provided all expectations exist." But we're going to leave that assumption implicit.

Random variables with infinite or ill-defined expectations are more the exception than the rule, but they do creep in occasionally.

15.1.5 The Expected Value of a Product

While the expectation of a sum is the sum of the expectations, the same is usually not true for products. But it is true in an important special case, namely, when the random variables are *independent*.

For example, suppose we throw two *independent*, fair dice and multiply the numbers that come up. What is the expected value of this product?

Let random variables R_1 and R_2 be the numbers shown on the two dice. We can compute the expected value of the product as follows:

$$\mathrm{E}[R_1 \cdot R_2] = \mathrm{E}[R_1] \cdot \mathrm{E}[R_2] = 3.5 \cdot 3.5 = 12.25. \tag{15.8}$$

Here the first equality holds because the dice are independent.

At the other extreme, suppose the second die is always the same as the first. Now $R_1 = R_2$, and we can compute the expectation, $\mathrm{E}\left[R_1^2\right]$, of the product of the dice explicitly, confirming that it is not equal to the product of the expectations.

$$\begin{aligned}
\mathrm{E}[R_1 \cdot R_2] &= \mathrm{E}\left[R_1^2\right] \\
&= \sum_{i=1}^{6} i^2 \cdot \Pr\left\{R_1^2 = i^2\right\} \\
&= \sum_{i=1}^{6} i^2 \cdot \Pr\{R_1 = i\} \\
&= \frac{1^2}{6} + \frac{2^2}{6} + \frac{3^2}{6} + \frac{4^2}{6} + \frac{5^2}{6} + \frac{6^2}{6} \\
&= 15\ 1/6 \\
&\neq 12\ 1/4 \\
&= \mathrm{E}[R_1] \cdot \mathrm{E}[R_2].
\end{aligned}$$

Theorem 15.1.8. *For any two* independent *random variables R_1, R_2,*

$$\mathrm{E}[R_1 \cdot R_2] = \mathrm{E}[R_1] \cdot \mathrm{E}[R_2].$$

Proof. The event $[R_1 \cdot R_2 = r]$ can be split up into events of the form $[R_1 = r_1 \text{ and } R_2 = r_2]$ where $r_1 \cdot r_2 = r$. So

$$\begin{aligned}
&\mathrm{E}[R_1 \cdot R_2] \\
&::= \sum_{r \in \text{range}(R_1 \cdot R_2)} r \cdot \Pr\{R_1 \cdot R_2 = r\} \\
&= \sum_{r_i \in \text{range}(R_i)} r_1 r_2 \cdot \Pr\{R_1 = r_1 \text{ and } R_2 = r_2\} \\
&= \sum_{r_1 \in \text{range}(R_1)} \sum_{r_2 \in \text{range}(R_2)} r_1 r_2 \cdot \Pr\{R_1 = r_1 \text{ and } R_2 = r_2\} && \text{(ordering terms in the sum)} \\
&= \sum_{r_1 \in \text{range}(R_1)} \sum_{r_2 \in \text{range}(R_2)} r_1 r_2 \cdot \Pr\{R_1 = r_1\} \cdot \Pr\{R_2 = r_2\} && \text{(indep. of } R_1, R_2) \\
&= \sum_{r_1 \in \text{range}(R_1)} \left(r_1 \Pr\{R_1 = r_1\} \cdot \sum_{r_2 \in \text{range}(R_2)} r_2 \Pr\{R_2 = r_2\} \right) && \text{(factoring out } r_1 \Pr\{R_1 = r_1\}) \\
&= \sum_{r_1 \in \text{range}(R_1)} r_1 \Pr\{R_1 = r_1\} \cdot \mathrm{E}[R_2] && \text{(def of } \mathrm{E}[R_2]) \\
&= \mathrm{E}[R_2] \cdot \sum_{r_1 \in \text{range}(R_1)} r_1 \Pr\{R_1 = r_1\} && \text{(factoring out } \mathrm{E}[R_2]) \\
&= \mathrm{E}[R_2] \cdot \mathrm{E}[R_1]. && \text{(def of } \mathrm{E}[R_1])
\end{aligned}$$

□

Theorem 15.1.8 extends routinely to a collection of mutually independent variables.

Corollary 15.1.9. *If random variables $R_1, R_2, \ldots, R_k$ are mutually independent, then*

$$\mathrm{E}\left[\prod_{i=1}^{k} R_i\right] = \prod_{i=1}^{k} \mathrm{E}[R_i].$$

15.1.6 Conditional Expectation

Just like event probabilities, expectations can be conditioned on some event.

Definition 15.1.10. The *conditional expectation*, $\mathrm{E}[R \mid A]$, of a random variable, R, given event, A, is:

$$\mathrm{E}[R \mid A] ::= \sum_{r \in \text{range}(R)} r \cdot \Pr\{R = r \mid A\}. \tag{15.9}$$

In other words, it is the average value of the variable R when values are weighted by their conditional probabilities given A.

For example, we can compute the expected value of a roll of a fair die, *given*, for example, that the number rolled is at least 4. We do this by letting R be the outcome of a roll of the die. Then by equation (15.9),

$$\mathrm{E}[R \mid R \geq 4] = \sum_{i=1}^{6} i \cdot \Pr\{R = i \mid R \geq 4\} = 1 \cdot 0 + 2 \cdot 0 + 3 \cdot 0 + 4 \cdot \tfrac{1}{3} + 5 \cdot \tfrac{1}{3} + 6 \cdot \tfrac{1}{3} = 5.$$

The power of conditional expectation is that it lets us divide complicated expectation calculations into simpler cases. We can find the desired expectation by calculating the conditional expectation in each simple case and averaging them, weighing each case by its probability.

For example, suppose that 49.8% of the people in the world are male and the rest female —which is more or less true. Also suppose the expected height of a randomly chosen male is $5'\,11''$, while the expected height of a randomly chosen female is $5'\,5''$. What is the expected height of a randomly chosen individual? We can calculate this by averaging the heights of men and women. Namely, let H be the height (in feet) of a randomly chosen person, and let M be the event that the person is male and F the event that the person is female. We have

$$\begin{aligned}\mathrm{E}[H] &= \mathrm{E}[H \mid M]\Pr\{M\} + \mathrm{E}[H \mid F]\Pr\{F\} \\ &= (5 + 11/12) \cdot 0.498 + (5 + 5/12) \cdot 0.502 \\ &= 5.665\end{aligned}$$

which is a little less that $5'\,8''$.

The Law of Total Expectation justifies this method.

Theorem 15.1.11 (Law of Total Expectation). *Let $A_1, A_2, \ldots$ be a partition of the sample space. Then*

$$\mathrm{E}[R] = \sum_i \mathrm{E}[R \mid A_i]\Pr\{A_i\}.$$

Proof.

$$\begin{aligned}
\mathrm{E}[R] ::= & \sum_{r \in \text{range}(R)} r \cdot \Pr\{R = r\} && \text{(Def 15.1.1 of expectation)} \\
= & \sum_r r \cdot \sum_i \Pr\{R = r \mid A_i\}\Pr\{A_i\} && \text{(Law of Total Probability)} \\
= & \sum_r \sum_i r \cdot \Pr\{R = r \mid A_i\}\Pr\{A_i\} && \text{(distribute constant } r\text{)} \\
= & \sum_i \sum_r r \cdot \Pr\{R = r \mid A_i\}\Pr\{A_i\} && \text{(exchange order of summation)} \\
= & \sum_i \Pr\{A_i\} \sum_r r \cdot \Pr\{R = r \mid A_i\} && \text{(factor constant } \Pr\{A_i\}\text{)} \\
= & \sum_i \Pr\{A_i\}\,\mathrm{E}[R \mid A_i]. && \text{(Def 15.1.10 of cond. expectation)}
\end{aligned}$$

□

Properties of Conditional Expectation

Many rules for conditional expectation correspond directly to rules for ordinary expectation.

For example, linearity of conditional expectation carries over with the same proof:

Theorem 15.1.12. *For any two random variables R_1, R_2, constants $a_1, a_2 \in \mathbb{R}$, and event A,*

$$\mathrm{E}\left[a_1 R_1 + a_2 R_2 \mid A\right] = a_1 \mathrm{E}\left[R_1 \mid A\right] + a_2 \mathrm{E}\left[R_2 \mid A\right].$$

Likewise,

Theorem 15.1.13. *For any two* independent *random variables R_1, R_2, and event, A,*

$$\mathrm{E}\left[R_1 \cdot R_2 \mid A\right] = \mathrm{E}\left[R_1 \mid A\right] \cdot \mathrm{E}\left[R_2 \mid A\right].$$

15.2 Expect the Mean

A random variable may never take a value anywhere near its expected value, so why is its expected value important? The reason is suggested by a property of gambling games that most people recognize intuitively. Suppose your gamble hinges on the roll of two dice, where you win if the sum of the dice is seven. If the dice are fair, the probabilty you win is $1/6$, which is also your expected number of wins in one roll. Of course there's no such thing as $1/6$ of a win in one roll, since either you win or you don't. But if you play *many times*, you would expect that the *fraction* of times you win would be close to $1/6$. In fact, if you played a lot of times and found that your fraction of wins wasn't pretty close to $1/6$, you would become pretty sure that the dice weren't fair.

More generally, if we independently sample a random variable many times and compute the average of the sample values, then we really can expect this average to be close to the expectation most of the time. In this section we work out a fundamental theorem about how repeated samples of a random variable *deviate from the mean*. This theorem provides an explanation of exactly how sampling can be used to test hypotheses and estimate unknown quantities.

15.2.1 Markov's Theorem

Markov's theorem is an easy result that gives a generally rough estimate of the probability that a random variable takes a value *much larger* than its mean.

The idea behind Markov's Theorem can be explained with a simple example of *intelligence quotient*,IQ. This quantity was devised so that the average IQ measurement would be 100. Now from this fact alone we can conclude that at most $1/3$ the population can have an IQ of 300 or more, because if more than a third had an IQ of 300, then the average would have to be *more* than $(1/3)300 = 100$, contradicting the fact that the average is 100. So the probability that a randomly chosen person has an IQ of 300 or more is at most $1/3$. Of course this is not a very strong conclusion; in fact no IQ of over 300 has ever been recorded. But by the same logic, we can also conclude that at most $2/3$ of the population can have an IQ of 150 or more. IQ's of over 150 have certainly been recorded, though again, a much smaller fraction than $2/3$ of the population actually has an IQ that high.

But although these conclusions about IQ are weak, they are actually the *strongest possible* general conclusions that can be reached about a nonnegative random variable using *only* the fact that its mean is 100. For example, if we choose a random variable equal to 300 with probability $1/3$, and 0 with probability $2/3$, then its mean is 100, and the probability of a value of 300 or more really is $1/3$. So we can't hope to get a better upper bound than $1/3$ on the probability of a value ≥ 300.

Theorem 15.2.1 (Markov's Theorem). *If R is a nonnegative random variable, then for all $x > 0$*

$$\Pr\{R \geq x\} \leq \frac{\mathrm{E}[R]}{x}.$$

Proof. We will show that $\mathrm{E}[R] \geq x \Pr\{R \geq x\}$. Dividing both sides by x gives the desired result.

So let I_x be the indicator variable for the event $[R \geq x]$, and consider the random variable xI_x. Note that

$$R \geq xI_x,$$

because at any sample point, w,

- if $R(\omega) \geq x$ then $R(\omega) \geq x = x \cdot 1 = xI_x(\omega)$, and
- if $R(\omega) < x$ then $R(\omega) \geq 0 = x \cdot 0 = xI_x(\omega)$.

Therefore,

$$\begin{aligned}
\mathrm{E}[R] &\geq \mathrm{E}[xI_x] && (\text{since } R \geq xI_x)\\
&= x\,\mathrm{E}[I_x] && (\text{linearity of } \mathrm{E}[\cdot])\\
&= x \Pr\{I_x = 1\} && (I_x \text{ is an indicator})\\
&= x \Pr\{R \geq x\}. && (\text{def of } I_x)
\end{aligned}$$

□

Markov's Theorem is often expressed in an alternative form, stated below as an immediate corollary.

Corollary 15.2.2. *If R is a nonnegative random variable, then for all $c \geq 1$*

$$\Pr\{R \geq c \cdot \mathrm{E}[R]\} \leq \frac{1}{c}.$$

Proof. In Markov's Theorem, set $x = c \cdot \mathrm{E}[R]$. □

Applying Markov's Theorem

Let's consider the Hat-Check problem again. Now we ask what the probability is that x or more men get the right hat, this is, what the value of $\Pr\{G \geq x\}$ is.

We can compute an upper bound with Markov's Theorem. Since we know $\mathrm{E}[G] = 1$, Markov's Theorem implies

$$\Pr\{G \geq x\} \leq \frac{\mathrm{E}[G]}{x} = \frac{1}{x}.$$

For example, there is no better than a 20% chance that 5 men get the right hat, regardless of the number of people at the dinner party.

The Chinese Appetizer problem is similar to the Hat-Check problem. In this case, n people are eating appetizers arranged on a circular, rotating Chinese banquet tray. Someone then spins the tray so that each person receives a random appetizer. What is the probability that everyone gets the same appetizer as before?

There are n equally likely orientations for the tray after it stops spinning. Everyone gets the right appetizer in just one of these n orientations. Therefore, the correct answer is $1/n$.

But what probability do we get from Markov's Theorem? Let the random variable, R, be the number of people that get the right appetizer. Then of course $\mathrm{E}[R] = 1$ (right?), so applying Markov's Theorem, we find:

$$\Pr\{R \geq n\} \leq \frac{\mathrm{E}[R]}{n} = \frac{1}{n}.$$

So for the Chinese appetizer problem, Markov's Theorem is tight!

On the other hand, Markov's Theorem gives the same $1/n$ bound for the probability everyone gets their hat in the Hat-Check problem in the case that all permutations are equally likely. But the probability of this event is $1/(n!)$. So for this case, Markov's Theorem gives a probability bound that is way off.

Markov's Theorem for Bounded Variables

Suppose we learn that the average IQ among MIT students is 150 (which is not true, by the way). What can we say about the probability that an MIT student has an IQ of more than 200? Markov's theorem immediately tells us that no more than $150/200$ or $3/4$ of the students can have such a high IQ. Here we simply applied Markov's Theorem to the random variable, R, equal to the IQ of a random MIT student to conclude:

$$\Pr\{R > 200\} \leq \frac{\mathrm{E}[R]}{200} = \frac{150}{200} = \frac{3}{4}.$$

But let's observe an additional fact (which may be true): no MIT student has an IQ less than 100. This means that if we let $T ::= R - 100$, then T is nonnegative and $\mathrm{E}[T] = 50$, so we can apply Markov's Theorem to T and conclude:

$$\Pr\{R > 200\} = \Pr\{T > 100\} \leq \frac{\mathrm{E}[T]}{100} = \frac{50}{100} = \frac{1}{2}.$$

So only half, not 3/4, of the students can be as amazing as they think they are. A bit of a relief!

More generally, we can get better bounds applying Markov's Theorem to $R - l$ instead of R for any lower bound $l > 0$ on R.

Similarly, if we have any upper bound, u, on a random variable, S, then $u - S$ will be a nonnegative random variable, and applying Markov's Theorem to $u - S$ will allow us to bound the probability that S is much *less* than its expectation.

15.2.2 Chebyshev's Theorem

We have separate versions of Markov's Theorem for the probability of deviation *above* the mean and *below* the mean, but often we want bounds that apply to *distance* from the mean in either direction, that is, bounds on the probability that $|R - \mathrm{E}[R]|$ is large.

It is a bit messy to apply Markov's Theorem directly to this problem, because it's generally not easy to compute $\mathrm{E}[\,|R - \mathrm{E}[R]|\,]$. However, since $|R|$ and hence $|R|^k$ are nonnegative variables for any R, Markov's inequality also applies to the event $[|R|^k \geq x^k]$. But this event is equivalent to the event $[|R| \geq x]$, so we have:

Lemma 15.2.3. *For any random variable R, any positive integer k, and any $x > 0$,*

$$\Pr\{|R| \geq x\} \leq \frac{\mathrm{E}\left[|R|^k\right]}{x^k}.$$

The special case of this Lemma for $k = 2$ can be applied to bound the random variable, $|R - \mathrm{E}[R]|$, that measures R's deviation from its mean. Namely

$$\Pr\{|R - \mathrm{E}[R]| \geq x\} = \Pr\left\{(R - \mathrm{E}[R])^2 \geq x^2\right\} \leq \frac{\mathrm{E}\left[(R - \mathrm{E}[R])^2\right]}{x^2}, \tag{15.10}$$

where the inequality (15.10) follows by applying Lemma 15.2.3 to the nonnegative random variable, $(R - \mathrm{E}[R])^2$. Assuming that the quantity $\mathrm{E}\left[(R - \mathrm{E}[R])^2\right]$ above is finite, we can conclude that the probability that R deviates from its mean by more than x is $O(1/x^2)$.

Definition 15.2.4. The *variance*, $\mathrm{Var}[R]$, of a random variable, R, is:

$$\mathrm{Var}[R] ::= \mathrm{E}\left[(R - \mathrm{E}[R])^2\right].$$

So we can restate (15.10) as

Theorem 15.2.5 (Chebyshev). *Let R be a random variable, and let x be a positive real number. Then*

$$\Pr\{|R - \mathrm{E}[R]| \geq x\} \leq \frac{\mathrm{Var}[R]}{x^2}.$$

The expression $\mathrm{E}\left[(R - \mathrm{E}[R])^2\right]$ for variance is a bit cryptic; the best approach is to work through it from the inside out. The innermost expression, $R - \mathrm{E}[R]$, is precisely the deviation of R above its mean. Squaring this, we obtain, $(R - \mathrm{E}[R])^2$. This is a random variable that is near 0 when R is close to the mean and is a large positive number when R deviates far above or below the mean. So if R is always close to the mean, then the variance will be small. If R is often far from the mean, then the variance will be large.

Variance in Two Gambling Games

The relevance of variance is apparent when we compare the following two gambling games.

Game A: We win \$2 with probability $2/3$ and lose \$1 with probability $1/3$.

Game B: We win \$1002 with probability $2/3$ and lose \$2001 with probability $1/3$.

Which game is better financially? We have the same probability, 2/3, of winning each game, but that does not tell the whole story. What about the expected return for each game? Let random variables A and B be the payoffs for the two games. For example, A is 2 with probability 2/3 and -1 with probability $1/3$. We can compute the expected payoff for each game as follows:

$$\begin{aligned} \mathrm{E}[A] &= 2 \cdot \frac{2}{3} + (-1) \cdot \frac{1}{3} = 1, \\ \mathrm{E}[B] &= 1002 \cdot \frac{2}{3} + (-2001) \cdot \frac{1}{3} = 1. \end{aligned}$$

The expected payoff is the same for both games, but they are obviously very different! This difference is not apparent in their expected value, but is captured by variance. We can compute the $\mathrm{Var}[A]$ by working "from the inside out" as follows:

$$\begin{aligned} A - \mathrm{E}[A] &= \begin{cases} 1 & \text{with probability } \frac{2}{3} \\ -2 & \text{with probability } \frac{1}{3} \end{cases} \\ (A - \mathrm{E}[A])^2 &= \begin{cases} 1 & \text{with probability } \frac{2}{3} \\ 4 & \text{with probability } \frac{1}{3} \end{cases} \\ \mathrm{E}\left[(A - \mathrm{E}[A])^2\right] &= 1 \cdot \frac{2}{3} + 4 \cdot \frac{1}{3} \\ \mathrm{Var}[A] &= 2. \end{aligned}$$

Similarly, we have for $\mathrm{Var}[B]$:

$$\begin{aligned} B - \mathrm{E}[B] &= \begin{cases} 1001 & \text{with probability } \frac{2}{3} \\ -2002 & \text{with probability } \frac{1}{3} \end{cases} \\ (B - \mathrm{E}[B])^2 &= \begin{cases} 1,002,001 & \text{with probability } \frac{2}{3} \\ 4,008,004 & \text{with probability } \frac{1}{3} \end{cases} \\ \mathrm{E}\left[(B - \mathrm{E}[B])^2\right] &= 1,002,001 \cdot \frac{2}{3} + 4,008,004 \cdot \frac{1}{3} \\ \mathrm{Var}[B] &= 2,004,002. \end{aligned}$$

The variance of Game A is 2 and the variance of Game B is more than two million! Intuitively, this means that the payoff in Game A is usually close to the expected value of \$1, but the payoff in Game B can deviate very far from this expected value.

High variance is often associated with high risk. For example, in ten rounds of Game A, we expect to make \$10, but could conceivably lose \$10 instead. On the other hand, in ten rounds of game B, we also expect to make \$10, but could actually lose more than \$20,000!

15.2.3 Standard Deviation

Because of its definition in terms of the square of a random variable, the variance of a random variable may be very far from a typical deviation from the mean. For example, in Game B above, the deviation from the mean is 1001 in one outcome and -2002 in the other. But the variance is a

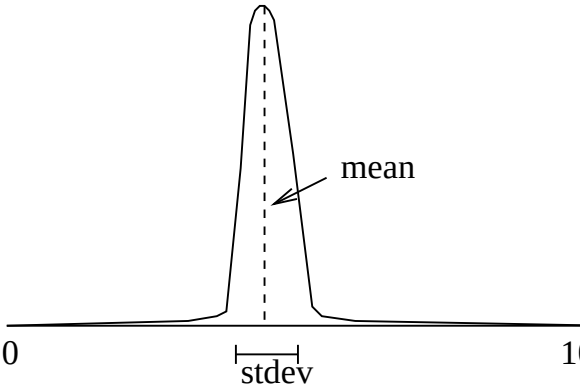


Figure 15.1: The standard deviation of a distribution indicates how wide the "main part" of it is.

whopping 2,004,002. From a dimensional analysis viewpoint, the "units" of variance are wrong: if the random variable is in dollars, then the expectation is also in dollars, but the variance is in square dollars. For this reason, people often describe random variables using standard deviation instead of variance.

Definition 15.2.6. The *standard deviation*, σ_R, of a random variable, R, is the square root of the variance:

$$\sigma_R ::= \sqrt{\text{Var}\,[R]} = \sqrt{\text{E}\,[(R - \text{E}\,[R])^2]}.$$

So the standard deviation is the square root of the mean of the square of the deviation, or the "root mean square" for short. It has the same units—dollars in our example—as the original random variable and as the mean. Intuitively, it measures the "expected (average) deviation from the mean," since we can think of the square root on the outside as canceling the square on the inside.

Example 15.2.7. The standard deviation of the payoff in Game B is:

$$\sigma_B = \sqrt{\text{Var}\,[B]} = \sqrt{2,004,002} \approx 1416.$$

The random variable B actually deviates from the mean by either positive 1001 or negative 2002; therefore, the standard deviation of 1416 describes this situation reasonably well.

Intuitively, the standard deviation measures the "width" of the "main part" of the distribution graph, as illustrated in Figure 15.1.

There is a useful, simple reformulation of Chebyshev's Theorem in terms of standard deviation.

Corollary 15.2.8. *Let R be a random variable, and let c be a positive real number.*

$$\Pr\{|R - \text{E}\,[R]| \geq c\sigma_R\} \leq \frac{1}{c^2}.$$

Here we see explicitly how the "likely" values of R are clustered in an $O(\sigma_R)$-sized region around $\text{E}\,[R]$, confirming that the standard deviation measures how spread out the distribution of R is around its mean.

Proof. Substituting $x = c\sigma_R$ in Chebyshev's Theorem gives:

$$\Pr\{|R - \mathrm{E}[R]| \geq c\sigma_R\} \leq \frac{\mathrm{Var}[R]}{(c\sigma_R)^2} = \frac{\sigma_R^2}{(c\sigma_R)^2} = \frac{1}{c^2}.$$

□

The IQ Example

Suppose that, in addition to the national average IQ being 100, we also know the standard deviation of IQ's is 10. How rare is an IQ of 300 or more?

Let the random variable, R, be the IQ of a random person. So we are supposing that $\mathrm{E}[R] = 100$, $\sigma_R = 10$, and R is nonnegative. We want to compute $\Pr\{R \geq 300\}$.

We have already seen that Markov's Theorem 15.2.1 gives a coarse bound, namely,

$$\Pr\{R \geq 300\} \leq \frac{1}{3}.$$

Now we apply Chebyshev's Theorem to the same problem:

$$\Pr\{R \geq 300\} = \Pr\{|R - 100| \geq 200\} \leq \frac{\mathrm{Var}[R]}{200^2} = \frac{10^2}{200^2} = \frac{1}{400}.$$

The purpose of the first step is to express the desired probability in the form required by Chebyshev's Theorem; the equality holds because R is nonnegative. Chebyshev's Theorem then yields the inequality.

So Chebyshev's Theorem implies that at most one person in four hundred has an IQ of 300 or more. We have gotten a much tighter bound using the additional information, namely the variance of R, than we could get knowing only the expectation.

15.2.4 Properties of Variance

The definition of variance of R as $\mathrm{E}\left[(R - \mathrm{E}[R])^2\right]$ may seem rather arbitrary. A direct measure of average deviation would be $\mathrm{E}[\,|R - \mathrm{E}[R]|\,]$. But variance has some valuable mathematical properties which the direct measure does not, as we explain below.

A Formula for Variance

Applying linearity of expectation to the formula for variance yields a convenient alternative formula.

Theorem 15.2.9.

$$\mathrm{Var}[R] = \mathrm{E}\left[R^2\right] - \mathrm{E}^2[R],$$

for any random variable, R.

Here we use the notation $\mathrm{E}^2[R]$ as shorthand for $(\mathrm{E}[R])^2$.

Proof. Let $\mu = \mathrm{E}[R]$. Then

$$\begin{aligned}
\mathrm{Var}[R] &= \mathrm{E}\left[(R - \mathrm{E}[R])^2\right] && \text{(Def 15.2.4 of variance)}\\
&= \mathrm{E}\left[(R-\mu)^2\right] && \text{(def of } \mu\text{)}\\
&= \mathrm{E}\left[R^2 - 2\mu R + \mu^2\right]\\
&= \mathrm{E}\left[R^2\right] - 2\mu\,\mathrm{E}[R] + \mu^2 && \text{(linearity of expectation)}\\
&= \mathrm{E}\left[R^2\right] - 2\mu^2 + \mu^2 && \text{(def of } \mu\text{)}\\
&= \mathrm{E}\left[R^2\right] - \mu^2\\
&= \mathrm{E}\left[R^2\right] - \mathrm{E}^2[R]\,.
\end{aligned}$$

□

For example, if B is a Bernoulli variable where $p ::= \Pr\{B = 1\}$, then

$$\mathrm{Var}[B] = p - p^2 = p(1-p). \tag{15.11}$$

Proof. Since B only takes values 0 and 1, we have $\mathrm{E}[B] = p \cdot 1 + (1-p)\cdot 0 = p$. Since $B = B^2$, we also have $\mathrm{E}\left[B^2\right] = p$, so (15.11) follows immediately from Theorem 15.2.9. □

Dealing with Constants

It helps to know how to calculate the variance of $aR + b$:

Theorem 15.2.10. *Let R be a random variable, and a a constant. Then*

$$\mathrm{Var}[aR] = a^2\,\mathrm{Var}[R]\,. \tag{15.12}$$

Proof. Beginning with the definition of variance and repeatedly applying linearity of expectation, we have:

$$\begin{aligned}
\mathrm{Var}[aR] &::= \mathrm{E}\left[(aR - \mathrm{E}[aR])^2\right]\\
&= \mathrm{E}\left[(aR)^2 - 2aR\,\mathrm{E}[aR] + \mathrm{E}^2[aR]\right]\\
&= \mathrm{E}\left[(aR)^2\right] - \mathrm{E}\left[2aR\,\mathrm{E}[aR]\right] + \mathrm{E}^2[aR]\\
&= a^2\,\mathrm{E}\left[R^2\right] - 2\,\mathrm{E}[aR]\,\mathrm{E}[aR] + \mathrm{E}^2[aR]\\
&= a^2\,\mathrm{E}\left[R^2\right] - a^2\mathrm{E}^2[R]\\
&= a^2\left(\mathrm{E}\left[R^2\right] - \mathrm{E}^2[R]\right)\\
&= a^2\,\mathrm{Var}[R] && \text{(by Theorem 15.2.9)}
\end{aligned}$$

□

It's even simpler to prove that adding a constant does not change the variance, as the reader can verify:

Theorem 15.2.11. *Let R be a random variable, and b a constant. Then*

$$\mathrm{Var}[R + b] = \mathrm{Var}[R]\,. \tag{15.13}$$

Proof.

$$\begin{aligned}\operatorname{Var}[R+b] &::= \mathrm{E}\left[((R+b)-\mathrm{E}[R+b])^2\right] \\ &= \mathrm{E}\left[((R+b)-(\mathrm{E}[R]+b))^2\right] \\ &= \mathrm{E}\left[(R-\mathrm{E}[R])^2\right] \\ &= \operatorname{Var}[R].\end{aligned}$$

□

Recalling that the standard deviation is the square root of variance, we immediately get:

Corollary 15.2.12. *The standard deviation of $aR+b$ equals a times the standard deviation of R:*

$$\sigma_{aR+b} = a\sigma_R.$$

Variance of a Sum

In general, the variance of a sum is not equal to the sum of the variances, but variances do add for *independent* variables. In fact, *mutual* independence is not necessary: *pairwise* independence will do. This is useful to know because there are some important situations involving variables that are pairwise independent but not mutually independent. Matching birthdays is an example of this kind, as we shall see below.

Theorem 15.2.13. *[Pairwise Independent Additivity of Variance] If $R_1, R_2, \ldots, R_n$ are* pairwise *independent random variables, then*

$$\operatorname{Var}[R_1+R_2+\cdots+R_n] = \operatorname{Var}[R_1]+\operatorname{Var}[R_2]+\cdots+\operatorname{Var}[R_n]. \tag{15.14}$$

Proof. We may assume that $\mathrm{E}[R_i] = 0$ for $i = 1,\ldots,n$, since we could always replace R_i by $(R_i - \mathrm{E}[R_i])$ in equation (15.14). This substitution preserves the independence of the variables, and by Theorem 15.2.11, does not change the variances.

Now by Theorem 15.2.9, $\operatorname{Var}[R_i] = \mathrm{E}\left[R_i^2\right]$ and $\operatorname{Var}[R_1+R_2+\cdots+R_n] = \mathrm{E}\left[(R_1+R_2+\cdots+R_n)^2\right]$, so we need only prove

$$\mathrm{E}\left[(R_1+R_2+\cdots+R_n)^2\right] = \mathrm{E}\left[R_1^2\right]+\mathrm{E}\left[R_2^2\right]+\cdots+\mathrm{E}\left[R_n^2\right] \tag{15.15}$$

But (15.15) follows from linearity of expectation and the fact that

$$\mathrm{E}[R_iR_j] = \mathrm{E}[R_i]\,\mathrm{E}[R_j] = 0\cdot 0 = 0 \tag{15.16}$$

for $i \neq j$, since R_i and R_j are independent.

$$\begin{aligned}
\mathrm{E}\left[(R_1 + R_2 + \cdots + R_n)^2\right] &= \mathrm{E}\left[\sum_{1\le i,j\le n} R_i R_j\right] \\
&= \sum_{1\le i,j\le n} \mathrm{E}\left[R_i R_j\right] \\
&= \sum_{1\le i\le n} \mathrm{E}\left[R_i^2\right] + \sum_{1\le i\neq j\le n} \mathrm{E}\left[R_i R_j\right] \\
&= \sum_{1\le i\le n} \mathrm{E}\left[R_i^2\right] + \sum_{1\le i\neq j\le n} 0 \qquad \text{(by (15.16))} \\
&= \mathrm{E}\left[R_1^2\right] + \mathrm{E}\left[R_2^2\right] + \cdots + \mathrm{E}\left[R_n^2\right].
\end{aligned}$$

□

Now we have a simple way of computing the expectation of a variable J which has an (n,p)-binomial distribution. We know that $J = \sum_{k=1}^{n} I_k$ where the I_k are mutually independent 0-1-valued variables with $\Pr\{I_k = 1\} = p$. The variance of each I_k is $p(1-p)$ by (15.11), so by linearity of variance, we have

Lemma (Variance of the Binomial Distribution). *If J has the (n,p)-binomial distribution, then*

$$\mathrm{Var}[J] = n\,\mathrm{Var}[I_k] = np(1-p). \tag{15.17}$$

15.2.5 Estimation by Random Sampling

Polling again

In Notes 12, we used bounds on the binomial distribution to determine confidence levels for a poll of voter preferences of Clinton vs. Giuliani. Now that we know the variance of the binomial distribution, we can use Chebyshev's Theorem as an alternative approach to calculate poll size.

The setup is the same as in Notes 12: we will poll n randomly chosen voters and let S_n be the total number in our sample who preferred Clinton. We use S_n/n as our estimate of the actual fraction, p, of all voters who prefer Clinton. We want to choose n so that our estimate will be within 0.04 of p at least 95% of the time.

Now S_n is binomially distributed, so from (15.17) we have

$$\mathrm{Var}[S_n] = n(p(1-p)) \le n \cdot \frac{1}{4} = \frac{n}{4}$$

The bound of $1/4$ follows from the easily verified fact that $p(1-p)$ is maximized when $p = 1-p$, that is, when $p = 1/2$.

Next, we bound the variance of S_n/n:

$$\begin{aligned}
\text{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \text{Var}\left[S_n\right] && \text{(by (15.12))} \\
&\leq \left(\frac{1}{n}\right)^2 \frac{n}{4} && \text{(by (15.2.5))} \\
&= \frac{1}{4n} && (15.18)
\end{aligned}$$

Now from Chebyshev and (15.18) we have:

$$\Pr\left\{\left|\frac{S_n}{n} - p\right| \geq 0.04\right\} \leq \frac{\text{Var}\left[S_n/n\right]}{(0.04)^2} = \frac{1}{4n(0.04)^2} = \frac{156.25}{n} \tag{15.19}$$

To make our our estimate with 95% confidence, we want the righthand side of (15.19) to be at most $1/20$. So we choose n so that

$$\frac{156.25}{n} \leq \frac{1}{20},$$

that is,

$$n \geq 3,125.$$

You may remember that in Notes 12 we calculated that it was actually sufficient to poll only 664 voters —many fewer than the 3,125 voters we derived using Chebyshev's Theorem. So the bound from Chebyshev's Theorem is not nearly as good as the bound we got earlier. This should not be surprising. In applying the Chebyshev Theorem, we used only a bound on the variance of S_n. In Notes 12, on the other hand, we used the fact that the random variable S_n was binomial (with known parameter, n, and unknown parameter, p). It makes sense that more detailed information about a distribution leads to better bounds. But even though the bound was not as good, this example nicely illustrates an approach to estimation using Chebyshev's Theorem that is more widely applicable than binomial estimations.

Birthdays again

There are important cases where the relevant distributions are not binomial because the mutual independence properties of the voter preference example do not hold. In these cases, estimation methods based on the Chebyshev bound may be the best approach. Birthday Matching is an example.

We've already seen that in a class of one hundred or more, there is a very high probability that some pair of students have birthdays on the same day of the month. We can also easily calculate the expected number of pairs of students with matching birthdays. But is it likely the number of matching pairs in a typical class will actually be close to the expected number? We can take the same approach to answering this question as we did in estimating voter preferences.

But notice that having matching birthdays for different pairs of students are not mutually independent events. For example, knowing that Alice and Bob have matching birthdays, and also that Ted and Alice have matching birthdays obviously implies that Bob and Ted have matching birthdays. On the other hand, knowing that Alice and Bob have matching birthdays tells us nothing

about whether Alice and Carol have matching birthdays, namely, these two events really are independent. So even though the events that various pairs of students have matching birthdays are not mutually independent, indeed not even three-way independent, they are *pairwise* independent.

This allows us to apply the same reasoning to Birthday Matching as we did for voter preference. Namely, let $B_1, B_2, \ldots, B_n$ be the birthdays of n independently chosen people, and let $E_{i,j}$ be the indicator variable for the event that the ith and jth people chosen have the same birthdays, that is, the event $[B_i = B_j]$. For simplicity, we'll assume that for $i \neq j$, the probability that $B_i = B_j$ is $1/365$. So the B_i's are mutually independent variables, and hence the $E_{i,j}$'s are *pairwise* independent variables, which is all we will need.

Let D be the number of matching pairs of birthdays among the n choices, that is,

$$D ::= \sum_{1 \leq i < j \leq n} E_{i,j}. \tag{15.20}$$

So by linearity of expectation

$$\mathrm{E}[D] = \mathrm{E}\left[\sum_{1 \leq i < j \leq n} E_{i,j}\right] = \sum_{1 \leq i < j \leq n} \mathrm{E}[E_{i,j}] = \binom{n}{2} \cdot \frac{1}{365}.$$

Also, by Theorem 15.2.13, the variances of pairwise independent variables are additive, so

$$\mathrm{Var}[D] = \mathrm{Var}\left[\sum_{1 \leq i < j \leq n} E_{i,j}\right] = \sum_{1 \leq i < j \leq n} \mathrm{Var}[E_{i,j}] = \binom{n}{2} \cdot \frac{1}{365}\left(1 - \frac{1}{365}\right).$$

Now for a class of $n = 100$ students, we have $\mathrm{E}[D] \approx 14$ and $\mathrm{Var}[D] < 14(1 - 1/365) < 14$. So by Chebyshev's Theorem

$$\Pr\{|D - 14| \geq x\} < \frac{14}{x^2}.$$

Letting $x = 6$, we conclude that there is a better than 50% chance that in a class of 100 students, the number of pairs of students with the same birthday will be between 8 and 20.

15.2.6 Pairwise Independent Sampling

The reasoning we used above to analyze voter polling and matching birthdays is very similar. We summarize it in slightly more general form with a basic result we call the Pairwise Independent Sampling Theorem. In particular, we do not need to restrict ourselves to sums of zero-one valued variables, or to variables with the same distribution. For simplicity, we state the Theorem for pairwise independent variables with possibly different distributions but with the same mean and variance.

Theorem (Pairwise Independent Sampling). *Let $G_1, \ldots, G_n$ be pairwise independent variables with the same mean, μ, and deviation, σ . Define*

$$S_n ::= \sum_{i=1}^{n} G_i. \tag{15.21}$$

Then

$$\Pr\left\{\left|\frac{S_n}{n}-\mu\right|\geq x\right\}\leq\frac{1}{n}\left(\frac{\sigma}{x}\right)^2.$$

Proof. We observe first that the expectation of S_n/n is μ:

$$\begin{aligned}
\mathrm{E}\left[\frac{S_n}{n}\right] &= \mathrm{E}\left[\frac{\sum_{i=1}^n G_i}{n}\right] && \text{(def of } S_n\text{)}\\
&= \frac{\sum_{i=1}^n \mathrm{E}\left[G_i\right]}{n} && \text{(linearity of expectation)}\\
&= \frac{\sum_{i=1}^n \mu}{n}\\
&= \frac{n\mu}{n} = \mu.
\end{aligned}$$

The second important property of S_n/n is that its variance is the variance of G_i divided by n:

$$\begin{aligned}
\mathrm{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \mathrm{Var}\left[S_n\right] && \text{(by (15.12))}\\
&= \frac{1}{n^2}\,\mathrm{Var}\left[\sum_{i=1}^n G_i\right] && \text{(def of } S_n\text{)}\\
&= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}\left[G_i\right] && \text{(pairwise independent additivity)}\\
&= \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}. && (15.22)
\end{aligned}$$

This is enough to apply Chebyshev's Bound and conclude:

$$\begin{aligned}
\Pr\left\{\left|\frac{S_n}{n}-\mu\right|\geq x\right\} &\leq \frac{\mathrm{Var}\left[S_n/n\right]}{x^2}. && \text{(Chebyshev's bound)}\\
&= \frac{\sigma^2/n}{x^2} && \text{(by (15.22))}\\
&= \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.
\end{aligned}$$

□

The Pairwise Independent Sampling Theorem provides a precise general statement about how the average of independent samples of a random variable approaches the mean. In particular, it proves what is known as the Law[1] of Large Numbers: by choosing a large enough sample size, n, we can get arbitrarily accurate estimates of the mean with confidence arbitrarily close to 100%.

[1]This is the *Weak* Law of Large Numbers. As you might suppose, there is also a Strong Law, but it's outside the scope of 6.042.

15.3 In-Class Problems Week 13, Wed.

Problem 15.3.1. Explaining Sampling to a Jury

In last lecture you learned why sampling 589 fish (or voters) will yield a fraction that, 95% of the time, will be within 0.04 of the actual fraction of contaminated fish (or voters who prefer Clinton over Giuliani). Notice that the size of the fish or voting population was never considered because *it did not matter*.

It seems remarkable that, whether there are a thousand, a million, or a billion voters in a country, polling only a few hundred is sufficient to be confident of an accurate estimation of average voter preference. Suppose you were going to serve as an expert witness in a trial. How would you explain why the number of people necessary to poll *does not depend on the population size*? Remember that juries do not understand formulas, so you have to provide an intuitive explanation, which is not quantitative.

Solution. This was intended to be a thought-provoking, conceptual question. In past terms, although most of the class could follow the derivations and crank through the formulas to calculate poll size and confidence levels, many students couldn't articulate, and indeed didn't really believe that the derived sample sizes were actually adequate to produce reliable estimates.

Here's a way to explain why we model polling people about Clinton as independent coin tosses that a jury might be able to follow:

> Of the approximately 100,000,000 voters in the US, there are some *unknown* number, say 51,000,000, who favor Clinton. So in this case, the *fraction* of voters favoring Clinton would be 51,000,000/100,000,000 = 0.51.
>
> To estimate this unknown fraction, we randomly select one person from the 100,000,000 in such a way that *everyone has an equal chance of being picked*. For example, we might get computer files from the Census Bureau listing all 100,000,000 voters in the US. Then we would generate a number between 1 and 100,000,000 by some physical or computational process that generated each number with equal probability, and then we would interview the person whose number came up. Picking a person this way amounts to flipping a coin that had a chance of coming up "Clinton" that was equal to the unknown fraction. In our example, there would be a 51% chance of the "coin" coming up "Clinton" and a 49% chance of coming up "Giuliani."
>
> After we have picked a person and learned their voting preference, we perform the procedure again, making sure that everyone is equally likely to be picked the second time, and so on, for picking a third, fourth, *etc*. person. Each pick is like flipping a coin whose probability of coming up "Clinton" is the same unknown fraction.
>
> Now we all understand that if we keep flipping a coin with a 51% chance of coming up Heads, then the more we flip, the closer the fraction of Heads flipped will be to 51%. Mathematical theory lets us calculate us how many times to flip coins to make the fraction of Heads very likely close to 51%, but we needn't go into the details of the calculation.

Now suppose we had two coins, say a penny and a nickel, which had the same 51% probability of coming up Heads. Then it's not going to make any difference which coin we use in our coin flips: the number of flips we need to get the fraction of Heads flipped being very likely close to 51% will be the same whether we flip pennies or nickels.

Different size populations correspond to different coins: the nickel might correspond to selecting a voter from a population 100,000,000 people, and the dime might correspond to selecting one from a population of 100. The same number of flips of pennies or nickels will allow us to estimate the probability of "Heads," and hence to estimate the fraction of voters favoring Clinton. All that mattered is that the "penny" population had the same probability of Heads, namely, 51,000,000 out of a population of 100,000,000, as the "nickel" population, namely, 51 out of 100.

So the number of "flips" needed does not depend on whether we're flipping a "51% penny" or a "51% nickel." That is, if two populations have the same fraction of voters favoring Clinton, then *the number of people we need to poll is the same*, even if the populations are of very different sizes.

■

Problem 15.3.2. An *International Journal of Epidemiology* has a policy that they will only publish the results of a drug trial when there were enough patients in the drug trial to be sure that the conclusions about the drug's effectiveness hold at the 95% confidence level. The editors of the Journal reason that under this policy, their readership can be confident that at most 5% of the published studies will be mistaken.

Later, the editors are astonished and embarrassed to learn that *every one* of the 20 drug trial results they published during the year was wrong. This happened even though the editors and reviewers had carefully checked the submitted data, and have no doubt that every one of the trials was *properly performed and reported* in the published paper.

The editors thought the probability of this was negligible (namely, $(1/20)^{20} < 10^{-25}$). Explain what's wrong with their reasoning and how it could be that all 20 published studies were wrong.

Solution. The editors have confused the statistical *confidence level* with *probability*. It's a mistake to think that because the conclusion of *particular* drug trial submitted to the journal holds at the 95% confidence level, this means its conclusion is wrong with probability only 1/20.

The conclusion of the particular submitted drug trial is right or wrong –period. An assertion of 95% confidence means that if very many trails were carried out, we expect that close to 95% of the trials would yield a correct conclusion. So if the results of all the many trials were all submitted for publication, and the editors selected 20 of these at random to publish, then they could reasonably expect that only one of them would be wrong.

But that's not what happens: not all the trials are written up and submitted, so the confidence level of the trial is not specially relevant. For example, there may be more than 400 worthless "alternative" drugs being tried by proponents who are genuinely honest, even if misguided. When

they conduct careful trials with a 95% confidence level, we can expect that in 1/20 of the 400 trials, worthless —even damaging —drugs will look helpful. The remaining 19/20 of the 400 trials would not be submitted for publication by honest proponents because the trials did not show positive results at the 95% level. But the 20 that mistakenly showed positive results might well all be submitted with no intention to mislead.

This is why, unless there is an explanation of *why* a therapy works, scientists and doctors usually doubt results claiming to confirm the efficacy of some mysterious therapy at a high confidence level. ■

Problem 15.3.3. Here is a fun game. You pick a number between 1 and 6. Then you roll three fair, independent dice.

- If your number never comes up, then you lose one dollar.
- If you number comes up once, then you win one dollar.
- If your number comes up twice, then you win two dollars.
- If your number comes up three times, then you win k dollars.

(a) Compute your expected payoff as a function of k.

Solution. Let the random variable P be your payoff. Then we can compute $\mathrm{E}[P]$ as follows:

$$\begin{aligned}\mathrm{E}[P] &= -1 \cdot \Pr\{\text{0 matches}\} + 1 \cdot \Pr\{\text{1 match}\} + 2 \cdot \Pr\{\text{2 matches}\} + k \cdot \Pr\{\text{3 matches}\} \\ &= -1 \cdot \left(\frac{5}{6}\right)^3 + 1 \cdot 3\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^2 + 2 \cdot 3\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right) + k \cdot \left(\frac{1}{6}\right)^3 \\ &= \frac{-125 + 75 + 30 + k}{216}\end{aligned}$$

■

(b) For what value of k is this game fair?

Solution. The game is fair when $\mathrm{E}[P] = 0$. This happens when $k = 20$. ■

Problem 15.3.4. Find the expectation of a variable, J, with an (n, p)-binomial distribution:

$$\mathrm{PDF}_J(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Hint: Consider

$$\frac{d\,(x+y)^n}{dx}$$

Solution. Let's begin from the hint.

By the binomial theorem

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

so

$$\begin{aligned} n(x+y)^{n-1} &= \frac{d\,(x+y)^n}{dx} \\ &= \frac{d\,\left(\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}\right)}{dx} \\ &= \sum_{k=0}^{n}\binom{n}{k}\frac{d\,x^k}{dx}y^{n-k} \\ &= \sum_{k=1}^{n}\binom{n}{k}kx^{k-1}y^{n-k} \\ &= (1/x)\sum_{k=1}^{n}k\binom{n}{k}x^k y^{n-k} \end{aligned} \tag{15.23}$$

Now

$$\mathrm{E}[J] = \sum_{k=0}^{n} k \Pr\{J=k\} = \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k}$$

so if we let $x = p$ and $y = 1-p$ in (15.23), we obtain

$$n = n(p+(1-p))^{n-1} = (1/p)\,\mathrm{E}[J]$$

and so

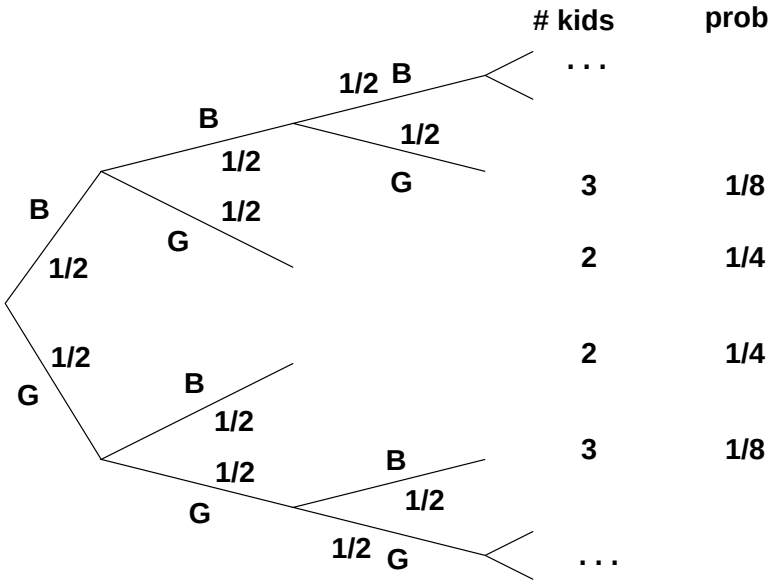
$$pn = \mathrm{E}[J].$$

■

Problem 15.3.5. A couple decides to have children until they have both a boy and a girl. What is the expected number of children that they'll end up with? Assume that each child is equally likely to be a boy or a girl and genders are mutually independent.

Solution. There are many ways to solve this problem. We'll do it from first principles.

Suppose that a couple has children until they have both a boy and a girl. A tree diagram for this experiment is shown below.



Let the random variable R be the number of children the couple has. From the definition of expectation, we have:

$$\begin{aligned}
\mathrm{E}[R] &= \sum_{w \in S} R(w) \cdot \Pr\{w\} \\
&= \left(2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 4 \cdot \frac{1}{16} + \ldots\right) + \left(2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 4 \cdot \frac{1}{16} + \ldots\right) \\
&= 2\left(2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 4 \cdot \frac{1}{16} + \ldots\right). \qquad (15.24)
\end{aligned}$$

The only difficulty is evaluating the sum. We can use the general formula

$$1 + 2r + 3r^2 + 4r^3 + \ldots = \frac{1}{(1-r)^2}$$

which is obtained by differentiating the formula for the sum of an infinite geometric series. Setting $r = 1/2$ gives:

$$1 + 2 \cdot \frac{1}{2} + 3 \cdot \frac{1}{4} + 4 \cdot \frac{1}{8} + \ldots = 4$$

We have to tweak this a little to get the sum we're interested in. Subtracting 1 from each side and then dividing both sides by 2 does the trick:

$$2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 4 \cdot \frac{1}{16} + \ldots = \frac{4-1}{2} = \frac{3}{2}$$

So from (15.24) we have

$$\mathrm{E}[R] = 2\left(\frac{3}{2}\right) = 3.$$

A much simpler approach uses the fact that the "mean time to failure" is $1/p$ where p is the probability of failure in one step. If we consider having a child of opposite sex to the first a "failure" of that child, then the mean time to failure is the expected number of children after the first until the couple has both a boy and a girl. But the probability of a failure at the kth child after the first is $1/2$ for all $k \geq 1$. So the expected number of children after the first is $1/(1/2) = 2$, and the expected number of children including the first is 1+2 =3. ■

Appendix

The ***expected value*** of a random variable R defined on a sample space, $\mathcal{S}$, is:

$$\mathrm{E}[R] = \sum_{w \in \mathcal{S}} R(w) \Pr\{w\}$$

Another helpful formula for expected values is:

$$\mathrm{E}[R] = \sum_{x \in \text{range}(R)} x \cdot \Pr\{R = x\}$$

Mean Time to Failure: If a biased coin with probability, p, of Heads is repeatedly flipped until a Head comes up, where the flips are mutually independent, then the expected number of flips is $1/p$.

15.4 In-Class Problems Week 13, Fri.

Problem 15.4.1. Here are seven propositions:

$$\begin{array}{rcrcr} x_1 & \vee & x_3 & \vee & \neg x_7 \\ \neg x_5 & \vee & x_6 & \vee & x_7 \\ x_2 & \vee & \neg x_4 & \vee & x_6 \\ \neg x_4 & \vee & x_5 & \vee & \neg x_7 \\ x_3 & \vee & \neg x_5 & \vee & \neg x_8 \\ x_9 & \vee & \neg x_8 & \vee & x_2 \\ \neg x_3 & \vee & x_9 & \vee & x_4 \end{array}$$

Note that:

1. Each proposition is the disjunction (OR) of three terms of the form x_i or the form $\neg x_i$.
2. The variables in the three terms in each proposition are all different.

Suppose that we assign true/false values to the variables $x_1, \ldots, x_9$ independently and with equal probability.

(a) What is the expected number of true propositions?

Hint: Let T_i be an indicator for the event that the i-th proposition is true.

Solution. Each proposition is true unless all three of its terms are false. Thus, each proposition is true with probability:

$$1 - \left(\frac{1}{2}\right)^3 = \frac{7}{8}$$

Let T_i be an indicator for the event that the i-th proposition is true. Then the number of true propositions is $T_1 + \cdots + T_7$ and the expected number is:

$$\begin{aligned} \mathrm{E}\left[T_1 + \cdots + T_7\right] &= \mathrm{E}\left[T_1\right] + \cdots + \mathrm{E}\left[T_7\right] \\ &= \frac{7}{8} + \cdots + \frac{7}{8} \\ &= \frac{49}{8} = 6\frac{1}{8} \end{aligned}$$

■

(b) Use your answer to prove that for *any* set of 7 propositions satisfying the conditions 1. and 2., there is an assignment to the variables that makes all 7 of the propositions true.

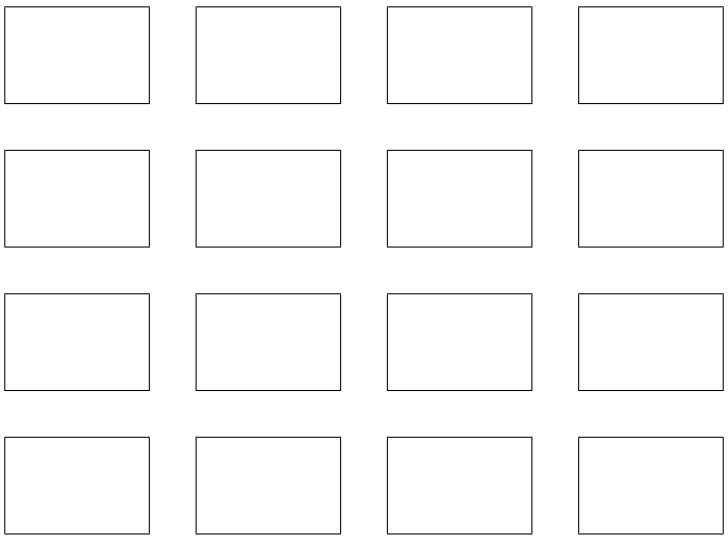
Solution. The calculation that the expected number of true propositions is $6\frac{1}{8}$ only used the fact that there were 7 propositions satisfying the conditions 1. and 2. So for *any* such set of 7 propositions, the expected number of true propositions is the same.

But a random variable can not always be less than its expectation, so there must be some assignment such that

$$T_1 + \ldots T_7 \geq 6\frac{1}{8},$$

which implies that $T_1 + \cdots + T_7 = 7$ for at least one outcome. This outcome is an assignment to the variables such that all of the propositions are true. ■

Problem 15.4.2. A classroom has sixteen desks arranged as shown below.



If there is a girl in front, behind, to the left, or to the right of a boy, then the two of them *flirt*. One student may be in multiple flirting couples; for example, a student in a corner of the classroom can flirt with up to two others, while a student in the center can flirt with as many as four others. Suppose that desks are occupied by boys and girls with equal probability and mutually independently. What is the expected number of flirting couples?

Solution. First, let's count the number of pairs of adjacent desks. There are three in each row and three in each column. Since there are four rows and four columns, there are $3 \cdot 4 + 3 \cdot 4 = 24$ pairs of adjacent desks.

Number these pairs of adjacent desks from 1 to 24. Let F_i be an indicator for the event that occupants of the desks in the i-th pair are flirting. The probability we want is then:

$$\begin{aligned} \mathrm{E}\left[\sum_{i=1}^{24} F_i\right] &= \sum_{i=1}^{24} \mathrm{E}\left[F_i\right] && \text{(linearity of } \mathrm{E}\left[\cdot\right]\text{)} \\ &= \sum_{i=1}^{24} \Pr\left\{F_i = 1\right\} && (F_i \text{ is an indicator}) \end{aligned}$$

The occupants of adjacent desks are flirting iff they are of opposite sexes, which happens with probability $1/2$, that is, $\Pr\{F_i = 1\} = 1/2$. Plugging this into the previous expression gives:

$$\mathrm{E}\left[\sum_{i=1}^{24} F_i\right] = \sum_{i=1}^{24} \Pr\left\{F_i = 1\right\} = 24 \cdot \frac{1}{2} = 12$$

■

Problem 15.4.3. Let R_1 and R_2 be random variables on a sample space, $\mathcal{S}$.

(a) Prove that

$$\mathrm{E}[R_1 + R_2] = \mathrm{E}[R_1] + \mathrm{E}[R_2].$$

Hint:

$$\mathrm{E}[R] = \sum_{\omega \in \mathcal{S}} R(\omega) \Pr\{\omega\} \tag{15.25}$$

Solution. *Proof.* Let $T ::= R_1 + R_2$. The proof follows straightforwardly by rearranging terms in the sum (15.25):

$$\begin{aligned}
\mathrm{E}[T] &= \sum_{\omega \in \mathcal{S}} T(\omega) \Pr\{\omega\} && \text{(by (15.25))} \\
&= \sum_{\omega \in \mathcal{S}} (R_1(\omega) + R_2(\omega)) \Pr\{\omega\} && \text{(def of } T\text{)} \\
&= \sum_{\omega \in \mathcal{S}} R_1(\omega) \Pr\{\omega\} + \sum_{\omega \in \mathcal{S}} R_2(\omega) \Pr\{\omega\} && \text{(rearranging terms)} \\
&= \mathrm{E}[R_1] + \mathrm{E}[R_2]. && \text{(by (15.25))}
\end{aligned}$$

□

■

(b) Justify each line of the following proof that if R_1 and R_2 are *independent*, then

$$\mathrm{E}[R_1 \cdot R_2] = \mathrm{E}[R_1] \cdot \mathrm{E}[R_2].$$

Proof.

$$
\begin{aligned}
&\mathrm{E}\left[R_1 \cdot R_2\right] \\
&= \sum_{r \in \text{range}(R_1 \cdot R_2)} r \cdot \Pr\left\{R_1 \cdot R_2 = r\right\} \\
&= \sum_{r_i \in \text{range}(R_i)} r_1 r_2 \cdot \Pr\left\{R_1 = r_1 \text{ and } R_2 = r_2\right\} \\
&= \sum_{r_1 \in \text{range}(R_1)} \sum_{r_2 \in \text{range}(R_2)} r_1 r_2 \cdot \Pr\left\{R_1 = r_1 \text{ and } R_2 = r_2\right\} \\
&= \sum_{r_1 \in \text{range}(R_1)} \sum_{r_2 \in \text{range}(R_2)} r_1 r_2 \cdot \Pr\left\{R_1 = r_1\right\} \cdot \Pr\left\{R_2 = r_2\right\} \\
&= \sum_{r_1 \in \text{range}(R_1)} \left(r_1 \Pr\left\{R_1 = r_1\right\} \cdot \sum_{r_2 \in \text{range}(R_2)} r_2 \Pr\left\{R_2 = r_2\right\} \right) \\
&= \sum_{r_1 \in \text{range}(R_1)} r_1 \Pr\left\{R_1 = r_1\right\} \cdot \mathrm{E}\left[R_2\right] \\
&= \mathrm{E}\left[R_2\right] \cdot \sum_{r_1 \in \text{range}(R_1)} r_1 \Pr\left\{R_1 = r_1\right\} \\
&= \mathrm{E}\left[R_2\right] \cdot \mathrm{E}\left[R_1\right].
\end{aligned}
$$

□

Solution. *Proof.*

$$
\begin{aligned}
&\mathrm{E}\left[R_1 \cdot R_2\right] \\
&::= \sum_{r \in \text{range}(R_1 \cdot R_2)} r \cdot \Pr\left\{R_1 \cdot R_2 = r\right\} && \text{(by definition)} \\
&= \sum_{r_i \in \text{range}(R_i)} r_1 r_2 \cdot \Pr\left\{R_1 = r_1 \text{ and } R_2 = r_2\right\} \\
&\qquad \text{(event } [R_1 \cdot R_2 = r] \text{ splits into events} \\
&\qquad\ [R_1 = r_1 \text{ and } R_2 = r_2] \text{ such that } r_1 r_2 = r) \\
&= \sum_{r_1 \in \text{range}(R_1)} \sum_{r_2 \in \text{range}(R_2)} r_1 r_2 \cdot \Pr\left\{R_1 = r_1 \text{ and } R_2 = r_2\right\} && \text{(ordering terms in the sum)} \\
&= \sum_{r_1 \in \text{range}(R_1)} \sum_{r_2 \in \text{range}(R_2)} r_1 r_2 \cdot \Pr\left\{R_1 = r_1\right\} \cdot \Pr\left\{R_2 = r_2\right\} && \text{(independence of } R_1, R_2) \\
&= \sum_{r_1 \in \text{range}(R_1)} \left(r_1 \Pr\left\{R_1 = r_1\right\} \cdot \sum_{r_2 \in \text{range}(R_2)} r_2 \Pr\left\{R_2 = r_2\right\} \right) && \text{(factor out } r_1 \Pr\{R_1 = r_1\}) \\
&= \sum_{r_1 \in \text{range}(R_1)} r_1 \Pr\left\{R_1 = r_1\right\} \cdot \mathrm{E}\left[R_2\right] && \text{(def of } \mathrm{E}[R_2]) \\
&= \mathrm{E}\left[R_2\right] \cdot \sum_{r_1 \in \text{range}(R_1)} r_1 \Pr\left\{R_1 = r_1\right\} && \text{(factor out } \mathrm{E}[R_2]) \\
&= \mathrm{E}\left[R_2\right] \cdot \mathrm{E}\left[R_1\right]. && \text{(def of } \mathrm{E}[R_1])
\end{aligned}
$$

□

■

Problem 15.4.4. **(a)** Compute the expected value of the number rolled on a fair, six-sided die, given that the outcome is even.

Solution. Let the random variable R be the number rolled, and let E be the event that the number rolled is even.

$$
\begin{aligned}
\mathrm{E}\left[R \mid E\right] &= \sum_{x=1}^{6} x \cdot \Pr\left\{R = x \mid E\right\} \\
&= 1 \cdot 0 + 2 \cdot \frac{1}{3} + 3 \cdot 0 + 4 \cdot \frac{1}{3} + 5 \cdot 0 + 6 \cdot \frac{1}{3} \\
&= 4
\end{aligned}
$$

■

(b) Define the random variable R using the following procedure. Roll two fair, independent dice and let the numbers rolled be D_1 and D_2. Then pick a random card from a standard deck. If the suit of the card is spades, let the value of R be D_1. Otherwise, let R be $D_1 \cdot D_2$. What is the expected value of R?

Solution. From the total expectation law we have:

$$\mathrm{E}[R] = \mathrm{E}[R \mid \spadesuit] \cdot \Pr\{\spadesuit\} + \mathrm{E}[R \mid \neg\spadesuit] \cdot \Pr\{\neg\spadesuit\}$$

Now

$$\mathrm{E}[R \mid \spadesuit] = \mathrm{E}[D_1] = \frac{7}{2}$$

Similarly,

$$\begin{aligned}\mathrm{E}[R \mid \neg\spadesuit] &= \mathrm{E}[D_1 \cdot D_2] \\ &= \mathrm{E}[D_1] \cdot \mathrm{E}[D_2] && (D_1, D_2 \text{ independent}) \\ &= \left(\frac{7}{2}\right)^2 = \frac{49}{4}\end{aligned}$$

So

$$\mathrm{E}[R] = \frac{7}{2}\frac{1}{4} + \frac{49}{4}\frac{3}{4} = \frac{161}{16} = 10\ \tfrac{1}{16}$$

■

Appendix

For random variable, R,

$$\mathrm{E}[R] ::= \sum_{x \in \text{range}(R)} x \cdot \Pr\{R = x\}.$$

If R is defined on a sample space, $\mathcal{S}$, then

$$\mathrm{E}[R] = \sum_{\omega \in \mathcal{S}} R(\omega) \Pr\{\omega\}.$$

The expected value of a random variable, R, *given event* A, is

$$\mathrm{E}[R \mid A] ::= \sum_{x \in \text{range}(R)} x \cdot \Pr\{R = x \mid A\}$$

So $\mathrm{E}[R] = \mathrm{E}[R \mid \mathcal{S}]$ where $\mathcal{S}$ is the sample space for R.

[The Law of Total Expectation] Let $A_1, A_2, \ldots$ be a partition of the sample space. Then

$$\mathrm{E}[R] = \sum_i \mathrm{E}[R \mid A_i] \Pr\{A_i\}.$$

15.5 In-Class Problems Week 14, Mon.

Problem 15.5.1. A herd of cows is stricken by an outbreak of *cold cow disease*. The disease lowers the normal body temperature of a cow, and a cow will die if its temperature goes below 90 degrees F. The disease epidemic is so intense that it lowered the average temperature of the herd to 85 degrees. Body temperatures as low as 70 degrees, **but no lower**, were actually found in the herd.

(a) Based solely on the information above, use Markov's bound to state an upper bound on the probability that a randomly chosen cow from the herd will have a high enough temperature to survive. Try to make the bound as small as possible.

Solution. Let T be the temperature of a random cow. Apply Markov's Bound to $T - 70$:

$$\Pr\{T \geq 90\} = \Pr\{T - 70 \geq 20\} \leq \mathrm{E}[T - 70]/20 = (85 - 70)/20 = 3/4.$$

■

(b) Suppose there are 400 cows in the herd. Give an example set of temperatures for the cows so that the probability that a randomly chosen cow will have a high enough temperature to survive is as large as possible, and explain why it cannot be larger.

Solution. Let 100 cows have 70 degrees and 300 have 90 degrees. So the probability that a random cow has a high enough temperature to survive is exactly 3/4. Also, the mean temperature is

$$(1/4)70 + (3/4)90 = 85.$$

So this distribution of temperatures satisfies the conditions under which the Markov bound implies that the probability of having a high enough temperature to survive cannot be larger than 3/4. ■

Problem 15.5.2. A gambler plays 120 hands of draw poker, 60 hands of black jack, and 20 hands of stud poker per day. He wins a hand of draw poker with probability 1/6, a hand of black jack with probability 1/2, and a hand of stud poker with probability 1/5. Assume the outcomes of the card games are mutually independent.

(a) What is the expected number of hands the gambler wins in a day?

Solution. $120(1/6) + 60(1/2) + 20(1/5) = 54$. ■

(b) What would the Markov bound be on the probability that the gambler will win at least 108 hands on a given day?

Solution. The expected number of games won is 54, so by Markov, $\Pr\{R \geq 108\} \leq 54/108 = 1/2$.
■

(c) What is the variance in the number of hands won per day?

Solution. The variance can also be calculated using linearity of variance. For an individual hand the variance is $p(1-p)$ where p is the probability of winning. Therefore the variance is

$$120(1/6)(5/6) + 60(1/2)(1/2) + 20(1/5)(4/5) = 523/15 = 34\,\tfrac{13}{15}.$$

■

(d) What would the Chebyshev bound be on the probability that the gambler will win at least 108 hands on a given day? You may answer with a numerical expression that is not completely evaluated.

Solution.

$$\Pr\{R - 54 \geq 54\} \leq \Pr\{|R - 54| \geq 54\} \leq \frac{V}{54^2} = \frac{523}{15(54)^2} \approx 0.01196.$$

■

Problem 15.5.3. The hat-check staff has had a long day serving at a party, and at the end of the party they simply return people's hats at random. Assume that n people checked hats at the party.

(a) What is the expected number of people who get their own hat back?

Solution. Let $X_i = 1$ be the indicator variable for the ith person getting their own hat back. Let $S_n = \sum_{i=1}^{n} X_i$ be the number of people who get their own hat back. By linearity of expectation,

$$\mathrm{E}[S_n] = \sum_{1}^{n} \mathrm{E}[X_i].$$

Since hats are returned at random, the each person has an equal chance of getting any of the hats, so the probability they get their own hat is $1/n$, that is, $1/n = \Pr\{X_i = 1\}$, and since X_i is an indicator, we have $\mathrm{E}[X_i] = 1/n$. By linearity of expectation,

$$\mathrm{E}[S_n] = \sum_{1}^{n} \mathrm{E}[X_i] = n \cdot \frac{1}{n} = 1.$$

■

Let $X_i = 1$ be the indicator variable for the ith person getting their own hat back. Let $S_n = \sum_{i=1}^{n} X_i$, so S_n is the total number of people who get their own hat back.

(b) Write a simple formula for $\mathrm{E}[X_i X_j]$ for $i \neq j$. *Hint:* What is $\Pr\{X_j = 1 \mid X_i = 1\}$?

Solution. We observed above that $\Pr\{X_i = 1\} = 1/n$. Also, given that the ith person got their own hat, each other person has an equal chance of getting their own hat among the remaining $n-1$ hats. So

$$\Pr\{X_j = 1 \mid X_i = 1\} = \frac{1}{n-1},$$

for $j \neq i$. Therefore,

$$\Pr\{X_i = 1 \text{ and } X_j = 1\} = \Pr\{X_j = 1 \mid X_i = 1\} \cdot \Pr\{X_i = 1\} = \frac{1}{n(n-1)}.$$

But $X_i = 1$ and $X_j = 1$ iff $X_i X_j = 1$, so

$$\mathrm{E}[X_i X_j] = \Pr\{X_i X_j = 1\} = \Pr\{X_i = 1 \text{ and } X_j = 1\},$$

and hence

$$\mathrm{E}[X_i X_j] = \frac{1}{n(n-1)}.$$

■

(c) Explain why you cannot use the variance of sums formula to calculate $\mathrm{Var}[S_n]$.

Solution. The principle of additivity of variances requires the variables be pairwise independent, but the indicator variables for people getting their hats back are not pairwise independent, since $\Pr\{X_j = 1 \mid X_i = 1\} = 1/(n-1) \neq 1/n = \Pr\{X_j = 1\}$ for $i \neq j$. ■

(d) Show that $\mathrm{E}\left[S_n^2\right] = 2$. *Hint:* $X_i^2 = X_i$.

Solution.

$$\begin{aligned}
\mathrm{E}\left[S_n^2\right] &= \mathrm{E}\left[\sum_i X_i^2 + \sum_i \sum_{j \neq i} X_i X_j\right] && \text{(expanding the sum for } S_n\text{)} \\
&= \sum_i \mathrm{E}\left[X_i^2\right] + \sum_i \sum_{j \neq i} \mathrm{E}[X_i X_j] && \text{(linearity of } \mathrm{E}[\cdot]\text{)} \\
&= \sum_i \mathrm{E}[X_i] + \sum_i \sum_{j \neq i} \frac{1}{n(n-1)} && \text{(since } X_i^2 = X_i\text{)} \\
&= n \cdot \frac{1}{n} + n(n-1) \cdot \frac{1}{n(n-1)} \\
&= 2.
\end{aligned}$$

■

(e) What is the variance of S_n?

Solution.

$$\begin{aligned}
\mathrm{Var}[S_n] &= \mathrm{E}\left[S_n^2\right] - \mathrm{E}^2[S_n] \\
&= 2 - 1^2 \\
&= 1.
\end{aligned}$$

■

(f) Use Chebyshev's bound to show that the probability that 11 or more people get their own hat back is at most 0.01.

Solution.

$$\begin{aligned}\Pr\{S_n \geq 11\} &= \Pr\{S_n - \mathrm{E}[S_n] \geq 11 - \mathrm{E}[S_n]\} \\ &= \Pr\{S_n - \mathrm{E}[S_n] \geq 10\} \\ &\leq \Pr\{|S_n - \mathrm{E}[S_n]| \geq 10\} \\ &\leq \frac{\mathrm{Var}[S_n]}{10^2} = .01\end{aligned}$$

■

Problem 15.5.4. This problem is a review of the derivation of Chebyshev's Theorem from Markov's Theorem.

(a) Explain why the following corollary of Markov's Theorem holds:

Corollary. *For any random variable R, any positive integer k, and any $x > 0$,*

$$\Pr\{|R| \geq x\} \leq \frac{\mathrm{E}\left[|R|^k\right]}{x^k}.$$

Solution. This can be seen by letting the random variable in Markov's Theorem be $|R|^k$, which is nonnegative for any random variable R and positive integer k. Notice as well that since $|R| \geq x$ iff $|R|^k \geq x^k$, the probabilities of the two events are the same. Combining these facts we get,

$$\Pr\{|R| \geq x\} = \Pr\left\{|R|^k \geq x^k\right\} \leq \frac{\mathrm{E}\left[|R|^k\right]}{x^k}.$$

■

(b) Use the above corollary to prove the following:

Theorem (Chebyshev). *Let R be a random variable, and let x be a positive real number. Then*

$$\Pr\{|R - \mathrm{E}[R]| \geq x\} \leq \frac{\mathrm{Var}[R]}{x^2}.$$

Solution. Let $T ::= R - \mathrm{E}[R]$

$$\begin{aligned}\Pr\{|R - \mathrm{E}[R]| \geq x\} &= \Pr\{|T| \geq x\} \\ &\leq \frac{\mathrm{E}\left[|T|^2\right]}{x^2} && \text{(Corollary above with } k = 2) \\ &= \frac{\mathrm{E}\left[T^2\right]}{x^2} && \text{(since } |T|^2 = T^2) \\ &= \frac{\mathrm{E}\left[(R - \mathrm{E}[R])^2\right]}{x^2} && \text{(def of } T) \\ &= \frac{\mathrm{Var}[R]}{x^2}. && \text{(def of Var } [R])\end{aligned}$$

■

Problem 15.5.5. **(a)** Prove from the definition of variance (in the Appendix) that

$$\text{Var}[R] = \text{E}\left[R^2\right] - \text{E}^2[R].$$

Solution. *Proof.* Write $\mu = \text{E}[R]$. Then

$$\begin{aligned}
\text{Var}[R] &= \text{E}\left[(R-\mu)^2\right] && \text{(def of } \mu\text{)} \\
&= \text{E}\left[R^2 - 2R\cdot\mu + \mu^2\right] \\
&= \text{E}\left[R^2\right] - 2\,\text{E}[R]\cdot\mu + \text{E}\left[\mu^2\right] && \text{(linearity of E}[\cdot]\text{)} \\
&= \text{E}\left[R^2\right] - 2\,\text{E}[R]\cdot\mu + \mu^2 && \text{(expectation of a constant)} \\
\\
&= \text{E}\left[R^2\right] - 2\text{E}^2[R] + \text{E}^2[R] && \text{(def of } \mu\text{)} \\
&= \text{E}\left[R^2\right] - \text{E}^2[R].
\end{aligned}$$

□

■

(b) Prove that $\text{Var}[X+Y+Z] = \text{Var}[X] + \text{Var}[Y] + \text{Var}[Z]$, if X, Y, Z are *pairwise* independent.

Solution. Let $R = X+Y+Z$. We can compute $\text{E}^2[R]$ as follows:

$$\begin{aligned}
\text{E}^2[R] &= (\text{E}[X+Y+Z])^2 \\
&= (\text{E}[X] + \text{E}[Y] + \text{E}[Z])^2 \\
&= \text{E}^2[X] + \text{E}^2[Y] + \text{E}^2[Z] + 2\,\text{E}[X]\,\text{E}[Y] + 2\,\text{E}[X]\,\text{E}[Z] + 2\,\text{E}[Y]\,\text{E}[Z]
\end{aligned}$$

Computing $\text{E}\left[R^2\right]$ we get:

$$\begin{aligned}
\text{E}\left[R^2\right] &= \text{E}\left[(X+Y+Z)^2\right] \\
&= \text{E}\left[X^2 + Y^2 + Z^2 + 2XY + 2XZ + 2YZ\right] \\
&= \text{E}\left[X^2\right] + \text{E}\left[Y^2\right] + \text{E}\left[Z^2\right] + 2\,\text{E}[XY] + 2\,\text{E}[XZ] + 2\,\text{E}[YZ] \\
&= \text{E}\left[X^2\right] + \text{E}\left[Y^2\right] + \text{E}\left[Z^2\right] + 2\,\text{E}[X]\,\text{E}[Y] + 2\,\text{E}[X]\,\text{E}[Z] + 2\,\text{E}[Y]\,\text{E}[Z]
\end{aligned}$$

Notice that the last step is only valid because the random variables are pairwise independent. Finally, we can compute $\text{Var}[R]$. We can begin immediately by cancelling out the cross terms, leaving us with:

$$\begin{aligned}
\text{Var}[R] &= \text{E}\left[R^2\right] - \text{E}^2[R] \\
&= \text{E}\left[X^2\right] + \text{E}\left[Y^2\right] + \text{E}\left[Z^2\right] - (\text{E}^2[X] + \text{E}^2[Y] + \text{E}^2[Z]) \\
&= \text{E}\left[X^2\right] - \text{E}^2[X] + \text{E}\left[Y^2\right] - \text{E}^2[Y] + \text{E}\left[Z^2\right] - \text{E}^2[Z] \\
&= \text{Var}[X] + \text{Var}[Y] + \text{Var}[Z],
\end{aligned}$$

thus concluding the proof. ■

Appendix

The *expectation* of a random variable, R, is:

$$\mathrm{E}[R] ::= \sum_{r \in \mathrm{range}(R)} r \cdot \Pr\{R = r\}$$

Theorem (Markov's Theorem). *If R is a nonnegative random variable, then for all $x > 0$*

$$\Pr\{R \geq x\} \leq \frac{\mathrm{E}[R]}{x}.$$

The *variance*, $\mathrm{Var}[R]$, of a random variable, R, is:

$$\mathrm{Var}[R] ::= \mathrm{E}\left[(R - \mathrm{E}[R])^2\right].$$

It is easy to show that

$$\mathrm{Var}[R] = \mathrm{E}\left[R^2\right] - \mathrm{E}^2[R].$$

[Variance of an index variable], I, with $\Pr\{I = 1\} = p$:

$$\mathrm{Var}[I] = pq$$

where $q ::= 1 - p$.

[Variance and constants] For constants, a, b,

$$\mathrm{Var}[aR + b] = a^2 \mathrm{Var}[R].$$

[Variance Additivity] If $R_1, R_2, \ldots, R_n$ are *pairwise* independent variables, then

$$\mathrm{Var}[R_1 + R_2 + \cdots + R_n] = \mathrm{Var}[R_1] + \mathrm{Var}[R_2] + \cdots + \mathrm{Var}[R_n]$$

Theorem (Chebyshev). *Let R be a random variable, and let x be a positive real number. Then*

$$\Pr\{|R - \mathrm{E}[R]| \geq x\} \leq \frac{\mathrm{Var}[R]}{x^2}.$$

15.6 In-Class Problems Week 14, Wed.

Problem 15.6.1. Among the 73 students who reported their birthdays to us this term, we found there were 5 pairs with the same birthday (there were no triples).

Use the Chebyshev Bound to prove that in a class of 73, the probability that there are between 2 and 12 pairs of students with the same birthday is at least $3/4$. (Assume all birthdays are equally likely and a year has 365 days.)

Hint: Let $S_{i,j}$ be the indicator variable for the ith and jth students having the same birthday. Let $M = \sum_{1\le i<j\le 73} S_{i,j}$ be the number of pairs with the same birthday. Calculate $\mathrm{E}[M]$ and $\mathrm{Var}[M]$.

Solution. We have observed in previous lectures that $\mathrm{E}[S_{ij}] = 1/365$, so $\mathrm{Var}[S_{ij}] = 364/(365)^2 \approx 1/365$. Now

$$\mathrm{E}[M] = \#\text{pairs} \cdot \mathrm{E}[S_{ij}] = \binom{73}{2}/365 = 7.2\ .$$

We have also observed that S_{ij} and S_{ik} are independent for $j \neq k$. That is, the S_{ij}'s are pairwise independent, and so their variances add:

$$\mathrm{Var}[M] = \#\text{pairs} \cdot \mathrm{Var}[S_{ij}] \approx \binom{73}{2}/365 = 7.2,$$

so

$$\sigma_M = \sqrt{\mathrm{Var}[M]} \approx 2.7\ .$$

By Chebyshev

$$1 - \Pr\{|M - 7.2| \ge 2\sigma_M\} \ge 1 - \frac{1}{4}.$$

That is,

$$\Pr\{1.84 < M < 12.6\} \ge \frac{3}{4}.$$

Since M is integer-valued, this last probability is the same as $\Pr\{2 \le M \le 12\}$. ■

Problem 15.6.2. For any random variable, R, with mean, μ, and standard deviation, σ, the Chebyshev Bound says that for any real number $x > 0$,

$$\Pr\{|R - \mu| \ge x\} \le \left(\frac{\sigma}{x}\right)^2.$$

Show that for any real number, μ, and real numbers $x \ge \sigma > 0$, there is an R for which the Chebyshev Bound is tight, that is,

$$\Pr\{|R| \ge x\} = \left(\frac{\sigma}{x}\right)^2. \tag{15.26}$$

Hint: First assume $\mu = 0$ and let R only take values $0, -x$, and x.

Solution. From the hint, we aim to find an R with $\mathrm{E}[R] = 0$ and $\mathrm{Var}[R] = \sigma^2$ that satisfies equation (15.26).

Using the further hint that R takes only values $0, -x, x$, we have

$$0 = \mathrm{E}[R] = x\Pr\{R = x\} - x\Pr\{R = -x\} = x(\Pr\{R = x\} - \Pr\{R = -x\})$$

so

$$\Pr\{R = x\} = \Pr\{R = -x\}, \tag{15.27}$$

since $x > 0$. Also,

$$\sigma^2 = \mathrm{E}\left[R^2\right] = x^2\Pr\{R = -x\} + x^2\Pr\{R = x\} = 2x^2\Pr\{R = x\},$$

so

$$\Pr\{R = x\} = \frac{\sigma^2}{2x^2}.$$

This implies

$$\Pr\{R = 0\} = 1 - 2\Pr\{R = x\} = 1 - \left(\frac{\sigma}{x}\right)^2,$$

which completely determines the distribution of R. Moreover,

$$\Pr\{|R| \geq x\} = \Pr\{R = -x\} + \Pr\{R = x\} = 2\Pr\{R = x\} = \left(\frac{\sigma}{x}\right)^2$$

which confirms (15.26).

Finally, given μ, x, and σ, if we let $R' ::= R + \mu$, then R' will be the desired random variable for which the Chebyshev Bound is tight. ■

Problem 15.6.3. The proof of the Pairwise Independent Sampling Theorem you just saw in lecture (it's also in Notes 14) was given for a sequence $R_1, R_2, \ldots$ of pairwise independent random variables with the same mean and variance. The proof is repeated in the Appendix.

We can generalize the Theorem to sequences of pairwise independent random variables, possibly with *different* distributions, as long as all their variances are bounded by some constant.

Theorem (Generalized Pairwise Independent Sampling). *Let $X_1, X_2, \ldots$ be a sequence of pairwise independent random variables such that $\mathrm{Var}[X_i] \leq b$ for some $b \geq 0$ and all $i \geq 1$. Let*

$$A_n ::= \frac{X_1 + X_2 + \cdots + X_n}{n},$$
$$\mu_n ::= \mathrm{E}[A_n].$$

Then for every $\epsilon > 0$,

$$\Pr\{|A_n - \mu_n| > \epsilon\} \leq \frac{b}{\epsilon^2} \cdot \frac{1}{n}. \tag{15.28}$$

(a) Prove the Generalized Pairwise Independent Sampling Theorem.

Solution. Essentially identical to the proof in the Appendix, except that R gets replaced by X and Var $[X_i]$ by b, with the equality where the b is first used becoming $\leq$. ■

(b) Conclude that the following holds:

Corollary (Generalized Weak Law of Large Numbers). *For every $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pr\{|A_n - \mu_n| \leq \epsilon\} = 1.$$

Solution.

$$\begin{aligned}\Pr\{|A_n - \mu_n| \leq \epsilon\} &= 1 - \Pr\{|A_n - \mu_n| > \epsilon\} \\ &\geq 1 - b/(n\epsilon^2) \qquad \text{(by (15.28))},\end{aligned}$$

and for any fixed ϵ, this last term approaches 1 as n approaches infinity. ■

Problem 15.6.4. **(a)** A computer program crashes at the end of each hour of use with probability p, if it has not crashed already. We know that the expected number of hours, $\mathrm{E}[H]$, until the program crashes is $1/p$. What is the variance of the number of hours until the program crashes? *Hint:* From Notes 11:

$$\sum_{1}^{\infty} k^2 x^k = \frac{x(1+x)}{(1-x)^3}$$

Solution. We have

$$\Pr\{H = k\} = q^{k-1}p$$

where $q ::= 1 - p$. Now

$$\mathrm{Var}[H] = \mathrm{E}\left[H^2\right] - \mathrm{E}^2[H],$$

But

$$\begin{aligned}\mathrm{E}\left[H^2\right] ::= &\sum_{k\in\mathbb{N}^+} k^2 q^{k-1} p \\ &= \frac{p}{q}\sum_{k\in\mathbb{N}^+} k^2 q^k \\ &= \frac{p}{q}\left(\frac{q(1+q)}{(1-q)^3}\right) \qquad \text{(hint)} \\ &= \frac{1+q}{p^2},\end{aligned}$$

so

$$\mathrm{Var}[H] = \frac{(1+q)}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{q}{p^2}.$$

■

(b) What is the Chebyshev bound on

$$\Pr\{|H - (1/p)| > x/p\}$$

where $x > 0$?

Solution.

$$\frac{\text{Var}[H]}{(x/p)^2} = \frac{q}{p^2(x/p)^2} = \frac{q}{x^2}.$$

■

(c) Conclude from part (b) that for $a \geq 2$,

$$\Pr\{H > a/p\} \leq \frac{q}{(a-1)^2}$$

Hint: Check that $|H - (1/p)| > (a-1)/p$ iff $H > a/p$.

Solution. Note that if $H \leq 1/p$, then $|H - (1/p)| = (1/p) - H$, and since $H > 0$, we have $(1/p) - H < 1/p \leq (a-1)/p$. It follows that $|H - (1/p)| > (a-1)/p$ iff $H - (1/p) > (a-1)/p$ iff $H > a/p$. So

$$\begin{aligned}\Pr\{H > a/p\} &= \Pr\{|H - (1/p)| > (a-1)/p\} \\ &\leq q/(a-1)^2 \qquad \text{(by part (b)).}\end{aligned}$$

■

(d) What actually is

$$\Pr\{H > a/p\}?$$

Conclude that for any fixed $p > 0$, the probability that $H > a/p$ is an asymptotically smaller function of a than the Chebyshev bound of part (c).

Solution.

$$\Pr\{H > a/p\} = \sum_{k > a/p} q^{k-1}p = pq^{\lfloor a/p \rfloor} \sum_{0}^{\infty} q^j = q^{\lfloor a/p \rfloor}.$$

So

$$\Pr\{H > a/p\} \leq q^{(a-p)/p} = \left((1-p)^{1/p}\right)^{a-p} < \left(\left(e^{-p}\right)^{1/p}\right)^{a-p} = e^{p-a}.$$

But Chebyshev gives $\Pr\{H > a/p\} \leq q/(a-1)^2 = \Theta(1/a^2)$. Since $e^{p-a} = o(1/a^2)$, we conclude that the actual probability is much smaller than we got from the Chebyshev Bound. ■

Appendix

The *variance*, $\text{Var}[R]$, of a random variable, R, is:

$$\text{Var}[R] ::= \text{E}\left[(R - \text{E}[R])^2\right].$$

Variance can also be equivalently defined as:

$$\text{Var}[R] ::= \text{E}\left[R^2\right] - \text{E}^2[R].$$

Lemma. *For $a, b \in \mathbb{R}$,*

$$\text{Var}[aR + b] = a^2 \text{Var}[R] \tag{15.29}$$

Theorem. *If $R_1, R_2, \ldots, R_n$ are* pairwise *independent random variables, then*

$$\text{Var}\left[R_1 + R_2 + \cdots + R_n\right] = \text{Var}\left[R_1\right] + \text{Var}\left[R_2\right] + \cdots + \text{Var}\left[R_n\right].$$

Theorem (Chebyshev). *Let R be a random variable, and let x be a positive real number. Then*

$$\Pr\left\{|R - \text{E}\left[R\right]| \geq x\right\} \leq \frac{\text{Var}\left[R\right]}{x^2}. \tag{15.30}$$

Theorem (Pairwise Independent Sampling). *Let*

$$A_n ::= \frac{\sum_{i=1}^{n} R_i}{n}$$

where $R_1, \ldots, R_n$ are pairwise independent random variables with the same mean, μ, and deviation, σ. Then

$$\Pr\left\{|A_n - \mu| > x\right\} \leq \left(\frac{\sigma}{x}\right)^2 \cdot \frac{1}{n}. \tag{15.31}$$

Proof. By linearity of expectation,

$$\text{E}\left[A_n\right] = \frac{\text{E}\left[\sum_{i=1}^{n} R_i\right]}{n} = \frac{\sum_{i=1}^{n} \text{E}\left[R_i\right]}{n} = \frac{n\mu}{n} = \mu.$$

Since the R_i's are pairwise independent, their variances will also add, so

$$\begin{aligned}
\text{Var}\left[A_n\right] &= \left(\frac{1}{n}\right)^2 \text{Var}\left[\sum_{i=1}^{n} R_i\right] && \text{(by (15.29))} \\
&= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \text{Var}\left[R_i\right] && \text{(additivity)} \\
&= \left(\frac{1}{n}\right)^2 n\sigma^2 \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

Now letting R be A_n in Chebyshev's Bound (15.30) yields (15.31), as required.

□