Mathematics for Computer Science

Eric Lehman and Tom Leighton

2004

Contents

Chapter 1

What is a Proof?

A ***proof*** is a method of establishing truth. This is done in many different ways in everyday life:

Jury trial. Truth is ascertained by twelve people selected at random.

Word of God. Truth is ascertained by communication with God, perhaps via a third party.

Experimental science. The truth is guessed and the hypothesis is confirmed or refuted by experiments.

Sampling. The truth is obtained by statistical analysis of many bits of evidence. For example, public opinion is obtained by polling only a representative sample.

Inner conviction. "*My* program is perfect. I know this to be true."

"I don't see why not..." Claim something is true and then shift the burden of proof to anyone who disagrees with you.

Intimidation. Truth is asserted by someone with whom disagrement seems unwise.

Mathematics its own notion of "proof". In mathematics, a ***proof*** is a verification of a ***proposition*** by a chain of ***logical deductions*** from a base set of ***axioms***. Each of the three highlighted terms in this definition is discussed in a section below. The last section contains some complete examples of proofs.

1.1 Propositions

A *proposition* is a statement that is either true or false. This definition sounds very general and is a little vague, but it does exclude sentences such as, "What's a surjection, again?" and "Learn logarithms!" Here are some examples of propositions.

Proposition 1. *2 + 3 = 5*

This proposition happens to be true.

Proposition 2. $\forall\, n \in \mathbb{N} \quad n^2 + n + 41$ *is a prime number.*

This proposition is more complicated. The symbol $\forall$ is read "for all", and the symbol $\mathbb{N}$ stands for the set of natural numbers, $\{0, 1, 2, 3, \ldots\}$. (There is some disagreement about whether 0 is a natural number; in this course, it is.) So this proposition asserts that the final phrase is true for all natural numbers n. That phrase is actually a proposition in its own right:

$$\text{"}n^2 + n + 41 \text{ is a prime number"}$$

In fact, this is a special kind of proposition called a ***predicate***, which is a proposition whose truth depends on the value of one or more variables. This predicate is certainly true for *many* natural numbers n:

$\mathbf{n}$	$\mathbf{n^2 + n + 41}$	**prime or composite?**
0	41	prime
1	43	prime
2	47	prime
3	53	prime
...	...	(all prime)
20	461	prime
39	1601	prime

Experimental data like this can be useful in mathematics, but can also be misleading. In this case, when $n = 40$, we get $n^2 + n + 41 = 40^2 + 40 + 41 = 41 \cdot 41$, which is not prime. So Proposition 2 is actually false!

Proposition 3. $a^4 + b^4 + c^4 = d^4$ *has no solution when* $a, b, c, d \in \mathbb{N}^+$.

Here $\mathbb{N}^+$ denotes the *positive* natural numbers, $\{1, 2, 3, \ldots\}$. In 1769, Euler conjectured that this proposition was true. But the it was proven false 218 years later by Noam Elkies at the liberal arts school up Mass Ave. He found the solution $a = 95800, b = 217519, c = 414560, d = 422481$. We could write his assertion symbolically as follows:

$$\exists\, a, b, c, d \in \mathbb{N}^+ \quad a^4 + b^4 + c^4 = d^4$$

The $\exists$ symbol is read "there exists". So, in words, the expression above says that there exist positive natural numbers a, b, c, and d such that $a^4 + b^4 + c^4 = d^4$.

Proposition 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. This counterexample could never have been found by a brute-force computer search!

The symbols $\forall$ ("for all") and $\exists$ ("there exists") are called ***quantifiers***. A quantifier is always followed by a variable (and perhaps an indication of what values that variable can take on) and then a predicate that typically involves that variable. The predicate may itself involve more quantifiers. Here are a couple examples of statements involving quantifiers:

$$\exists\, x \in \mathbb{R} \quad x^2 - x + 1 = 0$$
$$\forall\, y \in \mathbb{R}^+ \quad \exists\, z \in \mathbb{R} \quad e^z = y$$

The first statement asserts that the equation $x^2 - x + 1 = 0$ has a real solution, which is false. The second statement says that as z ranges over the real numbers, e^z takes on every positive, real value at least once.

Proposition 5. *In every map, the regions can be colored with 4 colors so that adjacent regions have different colors.*

This proposition was conjectured by Guthrie in 1853. The proposition was "proved" in 1879 by Kempe. His argument relied on pictures and— as is often the case with picture-proofs— contained a subtle error, which Heawood found 11 years later. In 1977 Appel and Haken announced a proof that relied on a computer to check an enormous number of cases. However, many mathematicians remained unsatisfied because no human could hand-check the computer's work and also because of doubts about other parts of the argument. In 1996, Robertson, Sanders, Seymour, and Thomas produced a rigorous proof that still relied on computers. Purported proofs of the Four Color Theorem continue to stream in. For example, I. Cahit unveiled his 12-page solution in August 2004, but here is his proof of Lemma 4: "Details of this lemma is left to the reader (see Fig. 7)." Don't try that on your homework! Even if this one doesn't hold up, some day a simple argument may be found.

Proposition 6. *Every even integer greater than 2 is the sum of two primes.*

For example, $24 = 11 + 13$ and $26 = 13 + 13$. This is called the Goldbach Conjecture, after Christian Goldbach who first stated the proposition in 1742. Even today, no one knows whether the conjecture is true or false. Every integer ever checked is a sum of two primes, but just one exception would disprove the proposition.

Proposition 7. $\forall\, n \in \mathbb{Z} \quad (n \geq 2) \Rightarrow (n^2 \geq 4)$

The symbol $\mathbb{Z}$ denotes the set of integers, $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$. There is predicate nested inside this proposition:

$$(n \geq 2) \Rightarrow (n^2 \geq 4)$$

This is an example of an ***implication***, a proposition of the form $P \Rightarrow Q$. This expression is read "P implies Q" or "if P, then Q". The proposition correctly asserts that this particular implication is true for every integer n. In general, *the implication $P \Rightarrow Q$ is true when P is false or Q is true.* Another way of saying how implication works is with a ***truth table***:

P	Q	$P \Rightarrow Q$
T	T	T
T	F	F
F	T	T
F	F	T

In general, a truth table indicates whether a compound proposition is true or false for every possible truth setting of the constituent propositions. The second line of this table, for example, says that the implication $P \Rightarrow Q$ is false when P is true and Q is false.

Just now we used variables (P and Q) to denote arbitrary propositions. We'll often use such ***Boolean variables*** in place of specific propositions. These are variables that can take on only two possible values, true or false, just as the propositions they represent could be either true or false.

Here another example of an implication:

"If pigs fly, then you will understand the Chernoff Bound."

This is no insult! It's a true proposition, even if you're planning to sleep like a baby through the entire Chernoff Bound lecture. The reason is that the first part of the implication ("pigs fly") is false. And the last two lines of the truth table say that *$P \Rightarrow Q$ is always true when P is false.* This might not be the way you interpret if-then statements in everyday speech, but it's the accepted convention in mathematical discussions.

Proposition 8. $\forall\, n \in \mathbb{Z} \quad (n \geq 2) \Leftrightarrow (n^2 \geq 4)$

A proposition of the form $P \Leftrightarrow Q$ is read "P if and only if Q". (Sometimes "if and only if" is abbreviated "iff".) This proposition is true provided *P and Q are both true or both false*. Put another way, $P \Leftrightarrow Q$ is true provided *$P \Rightarrow Q$ and $Q \Rightarrow P$ are both true.* Here is a truth table that compares all these kinds of implication:

P	Q	$P \Rightarrow Q$	$Q \Rightarrow P$	$P \Leftrightarrow Q$
T	T	T	T	T
T	F	F	T	F
F	T	T	F	F
F	F	T	T	T

The predicate $(n \geq 2) \Leftrightarrow (n^2 \geq 4)$ is true when $n = 1$ (because both sides are false) and true when $n = 3$ (because both sides are true) but false when $n = -3$ (because the left side is false, but the right side is true). Therefore, Proposition 8 as a whole is false.

1.2 Axioms

An ***axiom*** is a proposition that is assumed to be true, because you believe it is somehow reasonable. Here are some examples:

Axiom 1. *If $a = b$ and $b = c$, then $a = c$.*

This seems very reasonable! But, of course, there is room for disagreement about what consitutes a reasonable axiom. For example, one of Euclid's axioms for geometry is equivalent to the following:

Axiom 2 (Parallel Postulate). *Given a line l and a point p not on l, there is exactly one line through p parallel to l.*

In the 1800's several mathematicians realized that the Parallel Postulate could be replaced with a couple alternatives. This axiom leads to "spherical geometry":

Axiom 3. *Given a line l and a point p not on l, there is* no *line through p parallel to l.*

And this axiom generates "hyperbolic geometry".

Axiom 4. *Given a line l and a point p not on l, there are* infinitely many *lines through p parallel to l.*

Arguably, no one of these axioms is really better than the other two. Of course, a different choice of axioms makes different propositions true. And axioms should not be chosen carelessly. In particular, there are two basic properties that one wants in a set of axioms: they should be consistent and complete.

A set of axioms is ***consistent*** if no proposition can be proved both true and false. This is an absolute must. One would not want to spend years proving a proposition true only to have it proved false the next day! Proofs would become meaningless if axioms were inconsistent.

A set of axioms is ***complete*** if every proposition can be proved or disproved. Completeness is very desirable; we would like to believe that any proposition could be proved or disproved with sufficient work and insight.

Surprisingly, making a complete, consistent set of axioms is not easy. Bertrand Russell and Alfred Whitehead tried during their entire careers to find such axioms for basic arithmetic and failed. Then Kurt Gödel proved that *no* finite set of axioms for arithmetic can be both consistent and complete! This means that any set of consistent axioms is necessarily incomplete; there will be true statements that can not be proved. For example, it might be that Goldbach's conjecture is true, but there is no proof!

In this class, we will not dwell too much on the precise set of axioms underpinning our proofs. Generally, we'll regard familiar facts from high school as axioms. You may

find this imprecision regarding the axioms troublesome at times. For example, in the midst of a proof, you may find yourself wondering, "Must I prove this little fact or can I assume it?" Unfortunately, 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 Logical Deductions

Logical deductions or *inference rules* are used to combine axioms and true propositions in order to form more true propositions.

One fundamental inference rule is ***modus ponens***. This rule says that if P is true and $P \Rightarrow Q$ is true, then Q is also true. Inference rules are sometimes written in a funny notation. For example, modus ponens is written:

$$\frac{P \qquad P \Rightarrow Q}{Q}$$

This says that if you know that the statements above the line are true, then you can infer that the statement below the line is also true.

Modus ponens is closely related to the proposition $(P \wedge (P \Rightarrow Q)) \Rightarrow Q$. Both in some sense say, "if P and $P \Rightarrow Q$ are true, then Q is true". This proposition is an example of a ***tautology***, because it is true for every setting of P and Q. The difference is that this tautology is a *single proposition*, whereas modus ponens is an inference rule that allows us to *deduce new propositions from old ones*. However, if we accept modus ponens, then a general theorem of logic says that for each tautological implication there is an associated inference rule. For example, $((P \Rightarrow Q) \wedge (Q \Rightarrow R)) \Rightarrow (P \Rightarrow R)$ and $((P \Rightarrow Q) \wedge \neg Q) \Rightarrow \neg P$ are both tautologies, as one can verify with truth tables, and here are the analogous inference rules:

$$\frac{P \Rightarrow Q \qquad Q \Rightarrow R}{P \Rightarrow R} \qquad\qquad \frac{P \Rightarrow Q \qquad \neg Q}{\neg P}$$

As with axioms, we won't say exactly what inference rules are legal in this class. Each step in a proof should be clear and "logical"; in particular, you should make clear what previously proved facts are used to derive each new conclusion.

1.4 Examples of Proofs

Let's put these ideas together and make some complete proofs.

1.4.1 A Tautology

Theorem 9. *The following proposition is a tautology:*

$$(X \Rightarrow Y) \quad \Leftrightarrow \quad (\neg Y \Rightarrow \neg X)$$

The expression on the right is called the ***contrapositive*** of $X \Rightarrow Y$. This theorem is asserting that an implication is true if and only if its contrapositive is true. As an everyday example, the implication:

"If you are wise, then you attend recitation."

is logically equivalent to its contrapositive:

"If you do not attend recitation, then you are not wise."

The simplest way to prove a statement involving a small number of Boolean variables, like Theorem 9, is to check all possible cases. In particular, we need to verify that the proposition is true for every setting of the Boolean variables X and Y. A truth table can help you organize such a proof and work systematically through all the cases.

Proof. We show that the left side is logically equivalent to the right side for every setting of the variables X and Y.

X	Y	$X \Rightarrow Y$	$\neg Y \Rightarrow \neg X$
T	T	T	T
T	F	F	F
F	T	T	T
F	F	T	T

Thus, the proposition $(X \Rightarrow Y) \Leftrightarrow (\neg Y \Rightarrow \neg X)$ is true in every case, which implies that it is a tautology. □

The since the tautological implication in Theorem 9 runs both ways, there are two corresponding inference rules (although they amount to about the same thing):

$$\frac{P \Rightarrow Q}{\neg Q \Rightarrow \neg P} \qquad\qquad \frac{\neg Q \Rightarrow \neg P}{P \Rightarrow Q}$$

These rules are quite useful. Sometimes when you set out to prove an implication $P \Rightarrow Q$, proving the contrapositive $\neg Q \Rightarrow \neg P$ turns out to be a bit easier or clearer. If you prove the contrapositive, then the original implication immediately follows by the second inference rule shown above.

1.4.2 A Proof by Contradiction

The three preceding theorems were established by ***direct proofs***; that is, we combined axioms and previously-proved theorems in a straightforward way until we reached the desired conclusion. Sometimes an ***indirect proof*** (also known as a ***proof by contradiction***) is easier. The idea is to assume that the desired conclusion is *false* and then show that that assumption leads to an absurdity or contradiction. This means that the assumption must be wrong, and so the desired conclusion is actually true.

In logical terms, indirect proof relies on the following inference rule:

$$\frac{\neg P \Rightarrow \text{false}}{P}$$

In words, if $\neg P$ implies some falsehood, then P must actually be true. We can verify this inference rule by checking that the corresponding implication is a tautology:

P	$(\neg P \Rightarrow \text{false}) \Rightarrow P$
T	T
F	T

Sure enough. Now let's see how indirect proof works in practice.

Theorem 10. *$\sqrt{2}$ is an irrational number.*

This theorem was first proved by the Pythagoreans, a secretive society dating back to about 500 BC that intertwined mysticism and mathematics. The irrationality of $\sqrt{2}$ and the existence of a twelve-sided regular polyhedron (the dodecahedron) were among their prized secrets.

Proof. In order to obtain a contradiction, assume that $\sqrt{2}$ is rational. Then we can write $\sqrt{2} = a/b$ where a and b are integers, b is nonzero, and the fraction is 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. As a result, a^2 is a multiple of 4. Because of the equality $2b^2 = a^2$, $2b^2$ 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, and so the assumption that $\sqrt{2}$ is rational must be false. □

When you use indirect proof, state this clearly and specify what assumption you are making in order to obtain a contradiction. Also, remember that the intermediate statements in an indirect proof may very well be false, because they derive from a false assumption. A common mistake is to forget this and later regard these statements as true!

Chapter 2

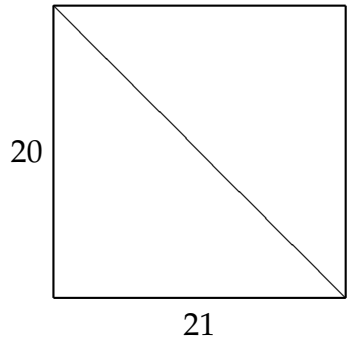
Induction I

2.1 A Warmup Puzzle

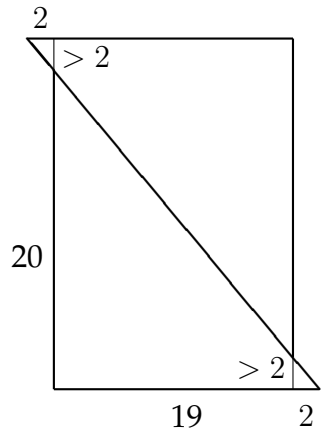
In principle, a proof should establish the truth of a proposition with absolute certainty. In practice, however, many purported proofs contain errors: overlooked cases, logical slips, and even algebra mistakes. But in a well-written proof, even if there is a bug, one should at least be able to pinpoint a specific statement that does not logically follow. See if you can find the first error in the following argument.

False Theorem 11. $420 > 422$

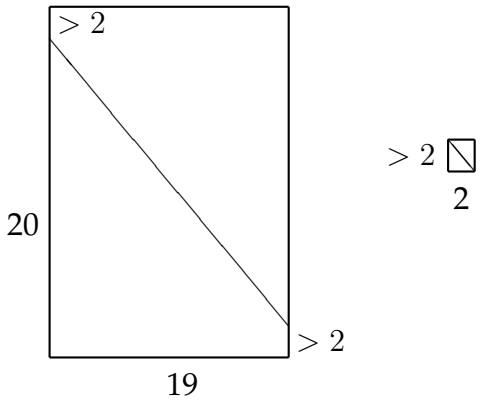
Proof. We will demonstrate this fact geometrically. We begin with a 20×21 rectangle, which has area 420:



Now we cut along the diagonal as indicated above and slide the upper piece parallel to the cut until it has moved exactly 2 units leftward. This leaves a couple stray corners, which are 2 units wide and just over 2 units high.



Finally, we snip off the two corners and place them together to form an additional small rectangle:



Now we have two rectangles, a large one with area just over $(20 + 2) \times 19 = 418$ and a small one with area just over $2 \times 2 = 4$. Thus, the total area of the resulting figure is a bit over $418 + 4 = 422$. By conservation of area, 420 is equal to just a little bit more than 422. □

Where is the error?

2.2 Induction

A professor 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. For all $n \in \mathbb{N}$, if student n gets candy, then student $n + 1$ also gets candy.

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, and so forth.

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 reasoning generalizes to a principle called ***induction***:

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. Mathematicians find this intuition so compelling that induction is always either taken as an axiom or else proved from more primitive axioms, which are themselves specifically designed so that induction *is* provable. In any case, the induction principle is a core truth of mathematics.

2.3 Using Induction

Induction is by far the most important proof technique in computer science. Generally, induction is used to prove that some statement holds for all natural values of a variable. For example, here is a classic formula:

Theorem 12. *For all $n \in \mathbb{N}$:*

$$1 + 2 + 3 + \ldots + n = \frac{n(n+1)}{2}$$

The left side of the equation represents the sum of all the numbers from 1 to n. You're supposed to guess the pattern and mentally replace the ... with the other terms. We could eliminate the need for guessing by rewriting the left side with ***summation notation***:

$$\sum_{i=1}^{n} i \qquad \text{or} \qquad \sum_{1 \le i \le n} i \qquad \text{or} \qquad \sum_{i \in \{1,\ldots,n\}} i$$

Each of these expressions denotes 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. The meaning of the sum in Theorem 12 is not so obvious in a couple special cases:

- If $n = 1$, then there is only one term in the summation, and so $1+2+3+\ldots+n = 1$. Don't be misled by the appearance of 2 and 3 and the suggestion that 1 and n are distinct terms!
- If $n \le 0$, then there are no terms at all in the summation, and so $1+2+3+\ldots+n = 0$.

The ... notation is convenient, but watch out for these special cases where the notation is misleading!

Now let's use the induction principle to prove Theorem 12. Suppose that we define predicate $P(n)$ to be "$1 + 2 + 3 + \ldots + n = n(n+1)/2$". 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 $P(n)$:

$$1 + 2 + 3 + \ldots + n = \frac{n(n+1)}{2}$$

in order to prove $P(n+1)$:

$$1 + 2 + 3 + \ldots + n + (n+1) = \frac{(n+1)(n+2)}{2}$$

These two equations are quite similar; in fact, adding $(n+1)$ to both sides of the first equation and simplifying the right side gives the second equation:

$$\begin{aligned} 1+2+3+\ldots+n+(n+1) &= \frac{n(n+1)}{2}+(n+1) \\ &= \frac{(n+2)(n+1)}{2} \end{aligned}$$

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. And so the theorem is proved!

A Template for Induction Proofs

The proof of Theorem 12 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".

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".

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 more clear.

A Clean Writeup

The proof of Theorem 12 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. Let $P(n)$ be the predicate:

$$1 + 2 + 3 + \ldots + n = \frac{n(n+1)}{2}$$

Base case: $P(0)$ is true, because both sides of the equation are zero.

Inductive step: Assume that $P(n)$ is true, where n is any natural number. Then $P(n+1)$ is also true, because:

$$\begin{aligned} 1 + 2 + 3 + \ldots + n + (n+1) &= \frac{n(n+1)}{2} + (n+1) \\ &= \frac{(n+1)(n+2)}{2} \end{aligned}$$

The first step uses the assumption $P(n)$, and the second follows by simplification.

Therefore, $P(n)$ is true for all natural n by induction, and the theorem is proved. □

Induction was helpful for *proving the correctness* of this summation formula, but not helpful for *discovering* the formula in the first place. There are some tricks for finding such formulas, which we'll show you in a few weeks.

2.4 A Divisibility Theorem

An integer a ***divides*** an integer b if b is a multiple of a. This is denoted $a \mid b$. For example, $3 \mid (5^3 - 5)$, since $5^3 - 5 = 120$ is a multiple of 3. More generally, we have the following theorem:

Theorem 13. $\forall\, n \in \mathbb{N} \quad 3 \mid (n^3 - n)$

Let's try to prove this with induction. The first challenge is always selecting the right induction hypothesis, $P(n)$. Your first instinct should be to lift the induction hypothesis directly from the claim. Thus, in this case, we should first try letting $P(n)$ be the predicate "$3 \mid (n^3 - n)$". (This doesn't always work out – as we'll see in a later example – but it does work in this case.)

Now we must address the base case by proving that $P(0)$ is true. As is often the case, this is easy: $3 \mid (0^3 - 0)$, since 0 is a multiple of 3. (Specifically, $3 \cdot 0 = 0$.)

Our next task, the inductive step, is typically the most difficult part of an induction proof. We must show that $P(n)$ implies $P(n+1)$. Thus, as usual when proving an implication, we *assume* $P(n)$ in order to *prove* $P(n+1)$. Writing out what these two expressions actually mean is often helpful. In this case, we assume $P(n)$:

$$3 \mid (n^3 - n)$$

in order to prove $P(n+1)$:

$$3 \mid ((n+1)^3 - (n+1))$$

These two statements look somewhat similar, but how can we use the first to prove the second? For lack of any other ideas, let's multiply out the expression in the second statement:

$$\begin{aligned} 3 \mid ((n+1)^3 - (n+1)) &\Leftrightarrow 3 \mid (n^3 + 3n^2 + 3n + 1 - n - 1) \\ &\Leftrightarrow 3 \mid (n^3 + 3n^2 + 2n) \end{aligned}$$

Aha! Notice that the last expression is equal to $n^3 - n$ plus $3n^2 + 3n$. Since $3n^2 + 3n$ is a multiple of 3 and $n^3 - n$ is a multiple of 3 by assumption, their sum must also be a multiple of 3. Therefore, $((n+1)^3 - (n+1)$ must also be a multiple of 3.

Playing around with $P(n)$ and $P(n+1)$ in this way, trying to understand how the two are related, is pretty typical when one is searching for an induction argument. However, what we've done so far is only scratchwork. What remains is to organize our reasoning into a clear proof.

Proof. We use induction. Let $P(n)$ be the proposition that $3 \mid (n^3 - n)$.

Base case: $P(0)$ is true because $3 \mid 0^3 - 0$.

Inductive step: Assume that $P(n)$ is true, where n is any natural number. Then:

$$\begin{aligned} 3 \mid (n^3 - n) &\Rightarrow 3 \mid (n^3 - n) + 3(n^2 + n) \\ &\Rightarrow 3 \mid n^3 + 3n^2 + 3n + 1 - n - 1 \\ &\Rightarrow 3 \mid (n+1)^3 - (n+1) \end{aligned}$$

The first implication relies on the fact that $3(n^2 + n)$ is divisible by 3. The remaining implications involve only rewriting the expression on the right. The last statement is $P(n+1)$, so we've proved that $P(n)$ implies $P(n+1)$ for all $n \in \mathbb{N}$.

By the principle of induction, $P(n)$ is true for all $n \in \mathbb{N}$, which proves the claim. □

This proof would look quite mysterious to anyone not privy to the scratchwork we did beforehand. In particular, one might ask how we had the foresight to introduce the magic term $3(n^2 + n)$. Of course, this was not foresight at all; we just worked backward initially!

2.5 A Faulty Induction Proof

Sometimes we want to prove that a statement is true for, say, all integers $n \geq 1$ rather than all integers $n \geq 0$. In this circumstance, we can 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.

False Theorem 14. *All horses are the same color.*

Proof. The proof is by induction. Let $P(n)$ be the proposition that in every set of n horses, all are the same color.

Base case: $P(1)$ is true, because all horses in a set of 1 must be the same color.

Inductive step: Assume that $P(n)$ is true, where n is a positive integer; 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:

$$h_1,\ \underbrace{h_2,\ \ldots,\ h_n,\ h_{n+1}}_{\text{same color}}$$

Therefore, 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 $n \geq 1$. The theorem is a special case where n is equal to the number of horses in the world. □

We've proved something false! Is math broken? Should we all become poets?

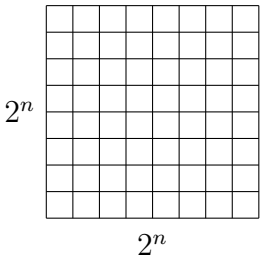
The error in this argument is in the sentence that begins, "Therefore, horses h_1, h_2, …, h_n, h_{n+1} must all be the same color." The "…" notation creates the impression that the sets $h_1, h_2, \ldots, h_n$ and $h_2, \ldots, h_n, h_{n+1}$ overlap. However, this is not true when $n = 1$. In that case, the first set is just h_1 and the second is h_2, and these do not overlap at all!

This mistake knocks a critical link out of our induction argument. We proved $P(1)$ and we proved $P(2) \Rightarrow P(3)$, $P(3) \Rightarrow P(4)$, etc. But we failed to prove $P(1) \Rightarrow P(2)$, and so everything falls apart: we can not conclude that $P(3)$, $P(4)$, etc. are true. And, of course, these propositions are all false; there are horses of a different color.

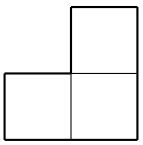
2.6 Courtyard Tiling

Induction served purely as a proof technique in the preceding examples. But induction sometimes can serve as a more general reasoning tool.

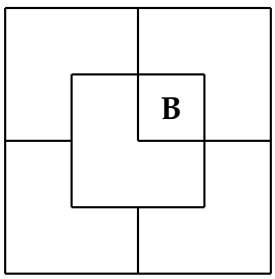
MIT recently constructed a new computer science building. As the project went further and further over budget, there were some radical fundraising ideas. One plan was to install a big courtyard with dimensions $2^n \times 2^n$:



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, insisted that only special L-shaped tiles be used:



A courtyard meeting these constraints exsists, 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 15. *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 does not help tile a larger courtyard with Bill in the center. We can not bridge the gap between $P(n)$ and $P(n+1)$. The usual recipe for finding an inductive proof will not work!

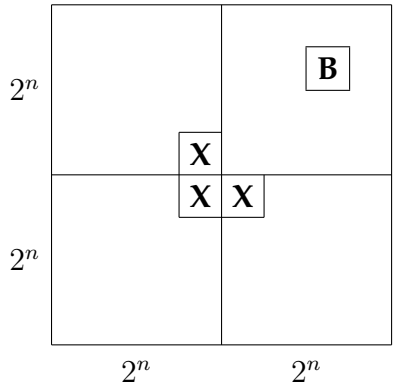
When this happens, your first fallback should be to look for a *stronger* induction hypothesis; that is, one which implies your previous hypothesis. For example, we could make $P(n)$ the proposition that for *every* location of Bill in a $2^n \times 2^n$ courtyard, there exists a tiling of the remainder.

This advice may sound bizzare: "If you can't prove something, try to prove something more grand!" But for induction arguments, this makes sense. In the inductive step, where you have to prove $P(n) \Rightarrow P(n+1)$, you're in better shape because you can *assume* $P(n)$, which is now a more general, more useful 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: 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 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"[1]. 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!

2.7 Another Faulty Proof

False Theorem 16. *I can lift all the sand on the beach.*

Proof. The proof is by induction. Let $P(n)$ be the predicate, "I can lift n grains of sand." The base case $P(1)$ is true because I can certainly lift one grain of sand. In the inductive step, assume that I can lift n grains of sand to prove that I can lift $n+1$ grains of sand. If I can lift n grains of sand, then surely I can lift $n+1$; one grain of sand will not make any difference. Therefore $P(n) \Rightarrow P(n+1)$. By induction, $P(n)$ is true for all $n \geq 1$. The theorem is the special case where n is equal to the number of grains of sand on the beach. □

The flaw here is in the bogus assertion that I can lift $n+1$ grains of sand because I can lift n grains. It is hard to say for exactly which n this is false, but certainly there is some value!

There is a field of mathematics called "fuzzy logic" in which truth is not a 0/1 thing, but is rather represented by a real value between 0 and 1. There is an analogue of induction in which the truth value decreases a bit with each implication. That might better model the situation here, since my lifts would probably gradually grow more and more sorry-looking as n approached my maximum. We will not be using fuzzy logic in this class, however. At least not intentionally.

[1] 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.

Chapter 3

Induction II

3.1 Good Proofs and Bad Proofs

In a purely technical sense, a mathematical proof is verification of a proposition by a chain of logical deductions from a base set of axioms. But the *purpose* of a proof is to provide readers with compelling evidence for the truth of an assertion. To serve this purpose effectively, more is required of a proof than just logical correctness: a good proof must also be clear. These goals are complimentary; a well-written proof is more likely to be a correct proof, since mistakes are harder to hide. Here are some tips on writing good proofs:

State your game plan. A good proof begins by explaining the general line of reasoning, e.g. "We use induction" 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 judiciously across the page. This is not good. The steps of your argument should follow one another in a clear, sequential order.

Explain your reasoning. Many students initially write proofs the way they compute integrals. The result is a long sequence of expressions without explantion. 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.

Simplify. Long, complicated proofs take the reader more time and effort to understand and can more easily conceal errors. So a proof with fewer logical steps is a better proof.

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 this 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 program are usually broken into a heirarchy 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 a preliminary lemmas. Also, if you are repeating essentially the same argument over and over, try to capture that argument in a general lemma and then repeatedly cite that instead.

Don't bully. Words such as "clearly" and "obviously" serve no logical function. Rather, they almost always signal an attempt to bully the reader into accepting something which the author is having trouble justifying rigorously. Don't use these words in your own proofs and go on the alert whenever you read one.

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 right conclusions. 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 system. When algorithms and protocols only "mostly work" due to reliance on handwaving 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, a buggy electronic voting system credited presidential candidate Al Gore with *negative* 16,022 votes in one county. 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.

3.2 A Puzzle

Here is a puzzle. There are 15 numbered tiles and a blank square arranged in a 4×4 grid. Any numbered tile adjacent to the blank square can be slid into the blank. For example, a sequence of two moves is illustrated below:

1	2	3	4
5	6	7	8
9	10	11	12
13	15	14	

→

1	2	3	4
5	6	7	8
9	10	11	12
13	15		**14**

→

1	2	3	4
5	6	7	8
9	10		12
13	15	**11**	14

In the leftmost configuration shown above, the 14 and 15 tiles are out of order. Can you find a sequence of moves that puts these two numbers in the correct order, but returns every other tile to its original position? Some experimentation suggests that the answer is probably "no", so let's try to prove that.

We're going to take an approach that is frequently used in the analysis of software and systems. We'll look for an ***invariant***, a property of the puzzle that is always maintained, no matter how you move the tiles around. If we can then show that putting the 14 and 15 tiles in the correct order would violate the invariant, then we can conclude that this is impossible.

Let's see how this game plan plays out. Here is the theorem we're trying to prove:

Theorem 17. *No sequence of moves transforms the board below on the left into the board below on the right.*

1	2	3	4
5	6	7	8
9	10	11	12
13	**15**	**14**	

1	2	3	4
5	6	7	8
9	10	11	12
13	**14**	**15**	

We'll build up a sequence of observations, stated as lemmas. Once we achieve a critical mass, we'll assemble these observations into a complete proof.

Define a ***row move*** as a move in which a tile slides horizontally and a ***column move*** as one in which a tile slides vertically. Assume that tiles are read top-to-bottom and left-to-right like English text; so when we say two tiles are "out of order", we mean that the larger number precedes the smaller number in this order.

Our difficulty is that one pair of tiles (the 14 and 15) is out of order initially. An immediate observation is that row moves alone are of little value in addressing this problem:

Lemma 18. *A row move does not change the order of the tiles.*

Usually a lemma requires a proof. However, in this case, there are probably no more-compelling observations that we could use as the basis for such an argument. So we'll let this statement stand on its own and turn to column moves.

Lemma 19. *A column move changes the relative order of exactly 3 pairs of tiles.*

For example, the column move shown below changes the relative order of the pairs (j, g), (j, h), and (j, i).

a	b	c	d
e	f		g
h	i	**j**	k
l	m	n	o

$\rightarrow$

a	b	c	d
e	f	**j**	g
h	i		k
l	m	n	o

Proof. Sliding a tile down moves it after the next three tiles in the order. Sliding a tile up moves it before the previous three tiles in the order. Either way, the relative order changes between the moved tile and each of the three it crosses. □

These observations suggest that there are limitations on how tiles can be swapped. Some such limitation may lead to the invariant we need. In order to reason about swaps more precisely, let's define a term: a pair of tiles i and j is ***inverted*** if $i < j$, but i appears after j in the puzzle. For example, in the puzzle below, there are four inversions: (12, 11), (13, 11), (15, 11), and (15, 14).

1	2	3	4
5	6	7	8
9	10		12
13	15	11	14

Let's work out the effects of row and column moves in terms of inversions.

Lemma 20. *A row move never changes the parity of the number of inversions. A column move always changes the parity of the number of inversions.*

The "parity" of a number refers to whether the number is even or odd. For example, 7 and -5 have odd parity, and 18 and 0 have even parity.

Proof. By Lemma 18, a row move does not change the order of the tiles; thus, in particular, a row move does not change the number of inversions.

By Lemma 19, a column move changes the relative order of exactly 3 pairs of tiles. Thus, an inverted pair becomes uninverted and vice versa. Thus, one exchange flips the total number of inversions to the opposite parity, a second exhange flips it back to the original parity, and a third exchange flips it to the opposite parity again. □

This lemma implies that we must make an *odd* number of column moves in order to exchange just one pair of tiles (14 and 15, say). But this is problematic, because each column move also knocks the blank square up or down one row. So after an *odd* number of column moves, the blank can not possibly be back in the last row, where it belongs! Now we can bundle up all these observations and state an invariant, a property of the puzzle that never changes, no matter how you slide the tiles around.

Lemma 21. *In every configuration reachable from the position shown below, the parity of the number of inversions is different from the parity of the row containing the blank square.*

row 1	1	2	3	4
row 2	5	6	7	8
row 3	9	10	11	12
row 4	13	15	14	

Proof. We use induction. Let $P(n)$ be the proposition that after n moves, the parity of the number of inversions is different from the parity of the row containing the blank square.

Base case: After zero moves, exactly one pair of tiles is inverted (14 and 15), which is an odd number. And the blank square is in row 4, which is an even number. Therefore, $P(0)$ is true.

Inductive step: Now we must prove that $P(n)$ implies $P(n+1)$ for all $n \geq 0$. So assume that $P(n)$ is true; that is, after n moves the parity of the number of inversions is different from the parity of the row containing the blank square. There are two cases:

1. Suppose move $n+1$ is a row move. Then the parity of the total number of inversions does not change by Lemma 20. The parity of the row containing the blank square does not change either, since the blank remains in the same row. Therefore, these two parities are different after $n+1$ moves as well, so $P(n+1)$ is true.

2. Suppose move $n+1$ is a column move. Then the parity of the total number of inversions changes by Lemma 20. However, the parity of the row containing the blank square also changes, since the blank moves up or down one row. Thus, the parities remain different after $n+1$ moves, and so $P(n+1)$ is again true.

Thus, $P(n)$ implies $P(n+1)$ for all $n \geq 0$.

By the principle of induction, $P(n)$ is true for all $n \geq 0$. □

The theorem we originally set out to prove is restated below. With this invariant in hand, the proof is simple.

Theorem. *No sequence of moves transforms the board below on the left into the board below on the right.*

1	2	3	4
5	6	7	8
9	10	11	12
13	**15**	**14**	

1	2	3	4
5	6	7	8
9	10	11	12
13	**14**	**15**	

Proof. In the target configuration on the right, the total number of inversions is zero, which is even, and the blank square is in row 4, which is also even. Therefore, by Lemma 21, the target configuartion is unreachable. □

If you ever played with Rubik's cube, you know that there is no way to rotate a single corner, swap two corners, or flip a single edge. All these facts are provable with invariant arguments like the one above. In the wider world, invariant arguments are used in the analysis of complex protocols and systems. For example, in analyzing the software and physical dynamics a nuclear power plant, one might want to prove an invariant to the effect that the core temperature never rises high enough to cause a meltdown.

3.3 Unstacking

Here is another wildly fun 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
10										
5	5									25 points
5	3	2								6
4	3	2	1							4
2	3	2	1	2						4
2	2	2	1	2	1					2
1	2	2	1	2	1	1				1
1	1	2	1	2	1	1	1			1
1	1	1	1	2	1	1	1	1		1
1	1	1	1	1	1	1	1	1	1	1
								total score	=	45 points

Can you find a better strategy?

3.3.1 Strong Induction

We'll analyze the unstacking game using a variant of induction 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 $n \in \mathbb{N}$, $P(0)$, $P(1)$, ..., $P(n)$ 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 $P(0)$, $P(1)$, ..., $P(n-1)$, and $P(n)$ are *all* true when you go to prove $P(n+1)$. These extra assumptions can only make your job easier.

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 can also be proved with ordinary induction. However, an apeal to the strong induction principle can make some proofs a bit simpler. On the other hand, if $P(n)$ is easily sufficient to prove $P(n+1)$, then use ordinary induction for simplicity.

3.3.2 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 22. *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 $P(1), \ldots, P(n-1)$ imply $P(n)$ for all $n \geq 2$ 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 $P(1)$, ..., $P(n-1)$ imply $P(n)$ for all $n \geq 2$. So assume that $P(1)$, ..., $P(n-1)$ are all true and that we have a stack of n blocks. The

first move must split this stack into substacks with sizes k and $n-k$ for some k strictly between 0 and 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:

$$\begin{aligned}
\text{total score} &= (\text{score for 1st move}) \\
&\quad + (\text{score for unstacking } k \text{ blocks}) \\
&\quad + (\text{score for unstacking } n-k \text{ blocks}) \\
&= k(n-k) + \frac{k(k-1)}{2} + \frac{(n-k)(n-k-1)}{2} \\
&= \frac{2nk - 2k^2 + k^2 - k + n^2 - nk - n - nk + k^2 + k}{2} \\
&= \frac{n(n-1)}{2}
\end{aligned}$$

The second equation uses the assumptions $P(k)$ and $P(n-k)$ and the rest is simplification. This shows that $P(1)$, $P(2)$, …, $P(n)$ imply $P(n+1)$.

Therefore, the claim is true by strong induction. □

Top 10 Proof Techniques NOT Allowed in 6.042

10. **Proof by throwing in the kitchen sink:** The author writes down every theorem or result known to mankind and then adds a few more just for good measure. When questioned later, the author correctly observes that the proof contains all the key facts needed to actually prove the result. Very popular strategy on 6.042 exams. Known to result in extra credit with sufficient whining.

9. **Proof by example:** The author gives only the case n = 2 and suggests that it contains most of the ideas of the general proof.

8. **Proof by vigorous handwaving:** A faculty favorite. Works well in any classroom or seminar setting.

7. **Proof by cumbersome notation:** Best done with access to at least four alphabets and special symbols. Helps to speak several foreign languages.

6. **Proof by exhaustion:** An issue or two of a journal devoted to your proof is useful. Works well in combination with proof by throwing in the kitchen sink and proof by cumbersome notation.

5. **Proof by omission:**
   "The reader may easily supply the details."
   "The other 253 cases are analogous."
   "..."

4. **Proof by picture:** A more convincing form of proof by example. Combines well with proof by omission.

3. **Proof by vehement assertion:** It is useful to have some kind of authority in relation to the audience.

2. **Proof by appeal to intuition:** Cloud-shaped drawings frequently help here. Can be seen on 6.042 exams when there was not time to include a complete proof by throwing in the kitchen sink.

1. **Proof by reference to eminent authority:**
   "I saw Fermat in the elevator and he said he had a proof . . ."

Here are some other common proof techniques that can be very useful, but which are not recommended for this class.

- **Proof by intimidation:** Can involve phrases such as: "Any moron knows that..." or "You know the Zorac Theorem of Hyperbolic Manifold Theory, right?" Sometimes seen in 6.042 tutorials.
- **Proof by intimidation (alternate form):** Consists of a single word: "Trivial." Often used by faculty who don't know the proof.
- **Proof by reference to inaccessible literature:** The author cites a simple corollary of a theorem to be found in a privately circulated memoir of the Slovenian Philological Society, 1883. It helps if the issue has not been translated.
- **Proof by semantic shift:** Some standard but inconvenient definitions are changed for the statement of the result.
- **Proof by cosmology:** The negation of the proposition is unimaginable or meaningless. Popular for proofs of the existence of God.
- **Proof by obfuscation:** A long plotless sequence of true and/or meaningless syntactically related statements.
- **Proof by wishful citation:** The author cites the negation, converse, or generalization of a theorem from the literature to support his claims.
- **Proof by funding:** How could three different government agencies be wrong?
- **Proof by personal communication:**
"$x^n + y^n \neq z^n$ for $n > 2$" [Fermat, personal communication].
- **Proof by importance:** A large body of useful consequences all follow from the proposition in question.
- **Proof by accumulated evidence:** Long and diligent search has not revealed a counterexample.
- **Proof by mutual reference:** In reference A, Theorem 5 is said to follow from Theorem 3 in reference B, which is shown from Corollary 6.2 in reference C, which is an easy consequence of Theorem 5 in reference A.
- **Proof by ghost reference:** Nothing even remotely resembling the cited theorem appears in the reference given.
- **Proof by forward reference:** Reference is usually to a forthcoming paper of the author, which is often not as forthcoming as the first.
- **Proof by metaproof:** A method is given to construct the desired proof. The correctness of the method is proved by any of the above techniques.

Chapter 4

Number Theory I

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 it dates to 1936, a full decade before any 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. It was sort of stupid.

4.1 A Theory of the Integers

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, and there's the negatives. Which one don't you understand? And who cares, anyway? 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?

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 forsaw that preserving military secrets would be vital in the coming conflict and proposed a way *to encrypt communications using number theory*. This was 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.

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 possibilities. But, first, we need some number theory.

All quantities discussed in this lecture are integers with a few exception that are noted explicitly.

4.2 Divisibility

We say that a ***divides*** b if there is an integer k such that $ak = b$. This is denoted $a \mid b$. For example, $7 \mid 63$, because $7 \cdot 9 = 63$. A consequence of this definition is that every number divides zero. If a divides b, then b is a ***multiple*** of a. For example, 63 is a multiple of 7.

Divisibility is a simple idea, but sufficient to dispel the notion that number theory lacks depth. The ancient Greeks considered a number ***perfect*** if it equalled the sum of its positive divisors, excluding itself. For example, $6 = 1+2+3$ and $28 = 1+2+4+7+14$ are perfect numbers. 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 a odd perfect number waiting just over the horizon. Number theory is full of questions like this: easy to pose, but incredibly difficult to answer.

The lemma below states some basic facts about divisibility that are *not* difficult to prove:

Lemma 23. *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 only prove part (2); the other proofs are similar. 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***. 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 some important contexts, such as the Fundamental Theorem of Arithmetic.

Here is one more essential fact about divisibility and primes.

Theorem 24. *Let p be a prime. If $p \mid a_1a_2 \ldots a_n$, then p divides some a_i.*

For example, if you know that $19 \mid 403 \cdot 629$, then you know that either $19 \mid 403$ or $19 \mid 629$, though you might not know which! The proof of this theorem takes some work, which we'll defer to avoid bogging down in minutiae here at the outset.

4.2.1 Turing's Code (Version 1.0)

Now let's look at Turing's scheme for using number theory to encrypt messages. First, the message must be translated to a prime number. 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.), string all the digits together to form one huge number, and then append digits as needed to produce a prime. For example, the message "victory" could be translated this way:

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 solved the problem in 1994, after seven years of working in secrecy and isolation in his attic.

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!

Prime Number Theorem How are the primes distributed? Let $\pi(x)$ denote the number of primes less than or equal to x. Primes are very irregularly distributed, so this is a complicated function. However, the Prime Number Theorem states that $\pi(x)$ is very nearly $x/\ln x$. The theorem was conjectured by Legendre in 1798 and proved a century later by de la Vallée Poussin and Hadamard in 1896. However, after his death, a notebook of Gauss was found to contain the conjecture, which he apparently made in 1791 at age 14.

Primality Testing Is there an efficient way to determine whether n is prime? An amazing simple, yet efficient method was finally discovered in 2002 by Agrawal, Kayal, and Saxena. Their paper began with a quote from Gauss emphasizing the importance and antiquity of the problem even in his time— two centuries ago.

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 seive", which runs in time proportional to:

$$e^{1.9(\ln n)^{1/3}(\ln \ln n)^{2/3}}$$

This is infeasible when n has a couple hundred digits or more.

	"v	i	c	t	o	r	y"	
→	22	09	03	20	15	18	25	13

Appending the digits 13 gives the number 2209032015182513, which is a prime.

Now here is how the encryption process works. In the description below, the variable m denotes the unencoded message (which we want to keep secret), and m' denotes the encrypted message (which the Nazis may intercept).

Beforehand The sender and receiver agree on a secret key, which is a large prime p.

Encryption The sender encrypts the message m by computing:

$$m' = m \cdot p$$

Decryption The receiver decrypts m' by computing:

$$\frac{m'}{p} = \frac{m \cdot p}{p} = m$$

For example, suppose that the secret key is the prime number 22801763489 and the message m is "victory". Then the encrypted message is:

$$\begin{aligned} m' &= m \cdot p \\ &= 2209032015182513 \cdot 22801763489 \\ &= 50369825549820718594667857 \end{aligned}$$

Turing's code raises a couple immediate questions.

1. How can the sender and receiver ensure that m and p 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. This definitively placed primality testing in the class of "easy" computational problems at last. Amazingly, the description of their algorithm is only thirteen lines!

2. Is Turing's code secure?

   The Nazis see only the encrypted message $m' = m \cdot p$, 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 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 p are sufficiently large, the Nazis seem to be out of luck!

Nevertheless, there is a major flaw in Turing's code. Can you find it? We'll reveal the weakness after saying a bit more about divisibility.

4.2.2 The Division Algorithm

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. This basic fact is the subject of a useful theorem:

Theorem 25 (Division Algorithm). *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 $0 \leq r < d$.*

Proof. We must prove that the integers q and r exist and that they are unique.

For existence, we use the well-ordering principle. First, we show that the equation $n = qd + r$ holds for *some* $r \geq 0$. If n is positive, then the equation holds when $q = 0$ and $r = n$. If n is not positive, then the equation holds when $q = n$ and $r = n(1-d) \geq 0$. Thus, by the well-ordering principle, there must exist a *smallest* $r \geq 0$ such that the equation holds. Furthermore, r must be less than d; otherwise, $b = (q+1)d + (r-d)$ would be another solution with a smaller nonnegative remainder, contradicting the choice of r.

Now we show uniqueness. Suppose that that there exist two different pairs of integers q_1, r_1 and q_2, r_2 such that:

$$\begin{aligned} n &= q_1 d + r_1 && \text{(where } 0 \leq r_1 < d\text{)} \\ n &= q_2 d + r_2 && \text{(where } 0 \leq r_2 < d\text{)} \end{aligned}$$

Subtracting the second equation from the first gives:

$$0 = (q_1 - q_2)d + (r_1 - r_2)$$

The absolute difference between the remainders r_1 and r_2 must be less than d, since $0 \leq r_1, r_2 < d$. This implies that the absolute value of $(q_1 - q_2)d$ must also be less than d, which means that $q_1 - q_2 = 0$. But then the equation above implies that $r_1 - r_2 = 0$ as well. Therefore, the pairs q_1, r_1 and q_2, r_2 are actually the same, which is a contradiction. So the quotient and remainder are unique. □

This theorem is traditionally called the "Division Algorithm", even though it is really a statement *about* the long-division procedure. 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 Algorithm is often denoted b ***rem*** a. In other words, $b \text{ rem } a$ is the remainder when b is divided by a. For example, $32 \text{ rem } 5$ is the remainder when 32 is divided by 5, which is 2. Similarly, $-11 \text{ rem } 7 = 3$, since $-11 = (-2) \cdot 7 + 3$.

Some people use the notation "mod" (which is short for "modulo") instead of "rem". This is unfortunate, because "mod" has been used by mathematicians for centuries in a confusingly similar context, and we shall do so here as well. Until *those* people are, shall we say, liquidated, you'll have to cope with the confusion.

Many programming languages have a remainder or modulus operator. For example, the expression "32 % 5" evaluates to 2 in Java, C, and C++. However, all these languages treat negative numbers strangely.

4.2.3 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:

$$m_1' = m_1 \cdot p \qquad \text{and} \qquad m_2' = m_2 \cdot p$$

The ***greatest common divisor (gcd)*** of a and b is the largest integer c such that $c \mid a$ and $c \mid b$. For example, the greatest common divisor of 9 and 15 is $\gcd(9, 15) = 3$. In this case, the greatest common divisor of the two encrypted messages, m_1' and m_2', is the secret key p. And, as we'll see, the $\gcd$ of two numbers can be computed very efficiently. So after the second message is sent, the Nazis can read 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.

4.3 Modular Arithmetic

On page 1 of his masterpiece on number theory, *Disquisitiones Arithmeticae*, C. F. 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*** c if $c \mid (a - b)$. This is denoted $a \equiv b \pmod{c}$. For example:

$$29 \equiv 15 \pmod{7} \quad \text{because } 7 \mid (29 - 15).$$

Intuitively, the $\equiv$ symbol is sort of like an $=$ sign, and the (mod 7) describes the specific sense in which 29 is equal-ish to 15. Thus, even though (mod 7) appears on the right side, it is in no sense more strongly associated with the 15 than the 29; in fact, it is actually associated with the $\equiv$ sign.

4.3.1 Congruence and Remainders

There is a strong relationship between congruence and remainders. In particular, *two numbers are congruent modulo c if and only if they have the same remainder when divided by c.* For example, 19 and 32 are congruent modulo 13, because both leave a remainder of 6 when divided by 13. We state this as a Lemma:

Lemma 26.

$$a \equiv b \pmod{c} \qquad \textit{if and only if} \qquad (a \text{ rem } c) = (b \text{ rem } c)$$

We'll prove this in a moment, but an example is probably more convincing. Some integers are listed on the first line of the table below. The second line lists the remainders when those integers are divided by 3.

$\#:$	$\ldots$	-4	-3	-2	-1	0	1	2	3	4	5	$\ldots$
$\#$ rem $3:$	$\ldots$	2	0	1	2	0	1	2	0	1	2	$\ldots$

Notice that numbers on the first line which differ by a multiple of 3 have the same remainder. For example, -2 and 4 differ by 6 and both leave remainder 1. This is precisely what the lemma asserts for $c = 3$; $a \equiv b \pmod 3$ means that a and b differ by a multiple of 3, and $(a \text{ rem } 3) = (b \text{ rem } 3)$ means that they leave the same remainder when divided by 3.

Proof. By the division algorithm, there exist unique pairs of integers q_1, r_1 and q_2, r_2 such that:

$$\begin{aligned} a &= q_1 c + r_1 && \text{(where } 0 \le r_1 < c\text{)} \\ b &= q_2 c + r_2 && \text{(where } 0 \le r_2 < c\text{)} \end{aligned}$$

In these terms, $(a \text{ rem } c) = r_1$ and $(b \text{ rem } c) = r_2$. Subtracting the second equation from the first gives:

$$a - b = (q_1 - q_2)c + (r_1 - r_2) \qquad \text{(where } -c < r_1 - r_2 < c\text{)}$$

Now $a \equiv b \pmod c$ if and only if c divides the left side. This is true if and only if c divides the right side, which holds if and only if $r_1 - r_2$ is a multiple of c. Given the bounds on $r_1 = r_2$, this happens precisely when $r_1 = r_2$, which is equivalent to $(a \text{ rem } c) = (b \text{ rem } c)$. □

Mathematics done with congruences instead of traditional equations is usually called "modular arithmetic". If anything, the importance of modular arithmetic has grown since the time of Gauss. For him, congruences were a reasoning tool. These days, computer hardware works with fixed-sized chunks of data, so the arbitrarily large integers that can come up in ordinary arithmetic are problematic. A standard solution is to design computers to do modular arithmetic instead. For example, a computer with 64-bit internal registers typically does integer arithmetic modulo 2^{64}. Thus, an instruction to add the contents of registers A and B actually computes $(A + B) \text{ rem } 2^{64}$.

4.3.2 Facts about rem and mod

Many familiar rules remain valid when one works modulo an integer $n \ge 1$. For example, we can add a constant to both sides of a congruence:

$$a \equiv b \pmod n \qquad \text{implies} \qquad a + c \equiv b + c \pmod n$$

Whenever you are unsure about a relationship involving congruences or remainders, go back to the definitions. For example, $a \equiv b \pmod{n}$ means that $n \mid (a-b)$. We can rewrite this as $n \mid ((a+c)-(b+c))$, which means that $a+c \equiv b+c \pmod{n}$ as we claimed above.

There is one glaring difference between traditional arithmetic and modular arithmetic. You can cancel multiplicative terms on opposite sides of an ordinary equation:

$$a \not{c} = b \not{c} \quad \text{implies} \quad a = b \quad (\text{provided } c \neq 0)$$

However, you can not always cancel such multiplicative terms in a congruence. Here is an example where a true statement becomes a false statement after cancelling:

$$2 \cdot \not{3} \equiv 4 \cdot \not{3} \pmod{6} \quad \leftarrow \text{This is an error!}$$

We'll revisit this issue of cancellation. Meanwhile, let's get more practice with rem and mod by proving some basic facts:

Lemma 27. *The following assertions hold for all $n \geq 1$:*

1. *If $a_1 \equiv b_1 \pmod{n}$ and $a_2 \equiv b_2 \pmod{n}$, then $a_1 a_2 \equiv b_1 b_2 \pmod{n}$.*
2. *$(a \text{ rem } n) \equiv a \pmod{n}$*
3. *$(a_1 \text{ rem } n) \cdot (a_2 \text{ rem } n) \cdots (a_k \text{ rem } n) \equiv a_1 \cdot a_2 \cdots a_k \pmod{n}$*

Proof. We prove each part separately.

1. The conditions $a_1 \equiv b_1 \pmod{n}$ and $a_2 \equiv b_2 \pmod{n}$ are equivalent to the assertions $n \mid (a_1 - b_1)$ and $n \mid (a_2 - b_2)$. By part 3 of Theorem 23, we know that:

   $$n \mid a_2(a_1 - b_1) + b_1(a_2 - b_2)$$

   Simplifying the right side gives $n \mid a_1 a_2 - b_1 b_2$. Thus, $a_1 a_2 \equiv b_1 b_2 \pmod{n}$ as claimed.

2. Recall that $a \text{ rem } n$ is equal to $a - qn$ for some quotient q. We can reason as follows:

   $$\begin{aligned} & n \mid qn \\ \Rightarrow \quad & n \mid a - (a - qn) \\ \Rightarrow \quad & n \mid a - (a \text{ rem } n) \end{aligned}$$

   The last statement is equivalent to $(a \text{ rem } n) \equiv a \pmod{n}$.

3. (sketch) We can extend the congruence in part 1 to k variables using induction. This general assertion that products are congruent if their terms are congruent, together with part 2, proves the claim.

□

4.3.3 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-boat" 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 have been something more, perhaps related to Turing's life after the war, what the government did to him, and his tragic end.

We'll come back to Turing's story shortly. Meanwhile, let's consider an alternative interpretation of his code. Perhaps we had the basic idea right (multiply the message by the key), but erred in using *conventional* arithmetic instead of *modular* arithemtic. Maybe this is what Turing meant:

Beforehand The sender and receiver agree on a large prime p, which may be made public. 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 $\{1, 2, \ldots, p-1\}$. The sender encrypts the message m to produce m' by computing:

$$m' = mk \text{ rem } p \qquad (*)$$

Decryption The receiver decrypts m' by finding a message m such that equation $(*)$ holds.

The decryption step is troubling. How can the receiver find a message m satisfying equation $(*)$ except by trying every possibility? That could be a lot of effort! Addressing this defect and understanding why Turing's code works at all requires a bit more number theory.

4.3.4 Cancellation Modulo a Prime

An immediate question about Turing's code is whether there could be two different messages with the same encoding. For example, perhaps the messages "Fight on!" and "Surrender!" actually encrypt to the same thing. This would be a disaster, because the receiver could not possibly determine which of the two was actually sent! The following lemma rules out this unpleasant possibility.

Lemma 28. *Suppose p is a prime and k is not a multiple of p. If*

$$ak \equiv bk \pmod{p}$$

then:

$$a \equiv b \pmod{p}$$

Proof. If $ak \equiv bk \pmod{p}$, then $p \mid (ak - bk)$ by the defintion of congruence, and so $p \mid k(a-b)$. Therefore, p divides either k or $a-b$ by Theorem 24. The former case is ruled out by assumption, so $p \mid (a-b)$, which means $a \equiv b \pmod{p}$. □

To understand the relevance of this lemma to Turing's code, regard a and b as two messages. Their encryptions are the same only if:

$$(ak \text{ rem } p) = (bk \text{ rem } p)$$

or, equivalently:

$$ak \equiv bk \pmod{p}$$

But then the lemma implies that $a \equiv b \pmod{p}$. Since the messages a and b are drawn from the set $\{1, 2, \ldots, p-1\}$, this means that $a = b$. In short, two messages encrypt to the same thing only if they are themselves identical.

In the bigger picture, Lemma 28 says that the encryption operation in Turing's code *permutes the space of messages*. This is stated more precisely in the following corollary.

Corollary 29. *Suppose p is a prime and k is not a multiple of p. Then the sequence:*

$$(0 \cdot k) \text{ rem } p, \quad (1 \cdot k) \text{ rem } p, \quad (2 \cdot k) \text{ rem } p, \quad \ldots, \quad \big((p-1) \cdot k\big) \text{ rem } p$$

is a permutation 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. By Lemma 28, no two of these are congruent modulo p and thus no two are equal. Therefore, the first sequence must be *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:

$$\underbrace{(0 \cdot 3) \text{ rem } 5}_{=0}, \quad \underbrace{(1 \cdot 3) \text{ rem } 5}_{=3}, \quad \underbrace{(2 \cdot 3) \text{ rem } 5}_{=1}, \quad \underbrace{(3 \cdot 3) \text{ rem } 5}_{=4}, \quad \underbrace{(4 \cdot 3) \text{ rem } 5}_{=2}$$

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 message space is permuted by the process of encryption and thus can't read encoded messages.

Lemma 28 also has a broader significance: it identifies one situation in which we *can* safely cancel a multiplicative term in a congruence. For example, if we know that:

$$8x \equiv 8y \pmod{37}$$

then we can safely conclude that:

$$x \equiv y \pmod{37}$$

because 37 is prime, and 8 is not a multiple of 37. We'll come back to this issue later and prove a more general theorem about cancellation.

4.3.5 Multiplicative Inverses

The real numbers have a nice quality that the integers lack. Namely, every nonzero real number r has a ***multiplicative inverse*** r^{-1} such that $r \cdot r^{-1} = 1$. For example, the multiplicative inverse of -3 is $-1/3$. Multiplicative inverses provide a basis for division: a/b can be defined as $a \cdot b^{-1}$. In contrast, most integers do not have multiplicative inverses within the set of integers. For example, no integer can be multiplied by 5 to give 1. As a result, if we want to divide integers, we are forced to muck about with remainders.

Remarkably, this defect of the integers vanishes *when we work modulo a prime number* p. In this setting, most integers do have multiplicative inverses! For example, if we are working modulo 11, then the multiplicative inverse of 5 is 9, because:

$$5 \cdot 9 \equiv 1 \pmod{11}$$

The only exceptions are multiples of the modulus p, which lack inverses in much the same way as 0 lacks an inverse in the real numbers. The following corollary makes this observation precise.

Corollary 30. *Let p be a prime. If k is not a multiple of p, then there exists an integer $k^{-1} \in \{1, 2, \ldots, p-1\}$ such that:*

$$k \cdot k^{-1} \equiv 1 \pmod{p}$$

Proof. Corollary 29 says that the expression $(m \cdot k \text{ rem } p)$ takes on all values in the set $\{1, 2, \ldots, p-1\}$ as m ranges over all values in the same set. Thus, in particular, $(m \cdot k \text{ rem } p) = 1$ for some m, which means $m \cdot k \equiv 1 \pmod{p}$. Let $k^{-1} = m$. □

The existence of multiplicative inverses has far-ranging consequences. Many theorems that hold for the real numbers (from linear algebra, say) have direct analogues that hold for the integers modulo a prime.

Multiplicative inverses also have practical significance in the context of Turing's code. Since we encode by multiplying the message m by the secret key k, we can decode by multiplying by the encoded message m' by the inverse k^{-1}. Let's justify this formally:

$$\begin{aligned} m' \cdot k^{-1} \text{ rem } p &\equiv m' \cdot k^{-1} \pmod{p} && \text{by part 2 of Lemma 27} \\ &\equiv (mk \text{ rem } p) \cdot k^{-1} \pmod{p} && \text{by definition of } m' \\ &\equiv mkk^{-1} \pmod{p} && \text{by parts 1 and 2 of Lemma 27} \\ &\equiv m \pmod{p} && \text{by definition of } k^{-1} \end{aligned}$$

Therefore, if the receiver can compute the multiplicative inverse of the secret key k modulo p, then he can decrypt with a single multiplication rather than an exhaustive search! The only remaining problem is finding the multiplicative inverse k^{-1} in the first place. Fermat's Theorem provides a way.

4.3.6 Fermat's Theorem

We can now prove a classic result known as Fermat's Theorem, which is much easier than his famous Last Theorem— and also more useful.

Theorem 31 (Fermat's Theorem). *Suppose p is a prime and k is not a multiple of p. Then:*

$$k^{p-1} \equiv 1 \pmod{p}$$

Proof.

$$\begin{aligned} 1 \cdot 2 \cdot 3 \cdots (p-1) &\equiv (k \text{ rem } p) \cdot (2k \text{ rem } p) \cdot (3k \text{ rem } p) \cdots ((p-1)k \text{ rem } p) \pmod{p} \\ &\equiv k \cdot 2k \cdot 3k \cdots (p-1)k \pmod{p} \\ &\equiv (p-1)! \cdot k^{p-1} \pmod{p} \end{aligned}$$

The expressions on the first line are actually equal, by Corollary 29, so they are certainly congruent modulo p. The second step uses part (3) of Lemma 27. In the third step, we rearrange terms in the product.

Now $(p-1)!$ can not be a multiple of p by Theorem 24, since p is a prime and does not divide any of 1, 2, …, $(p-1)$. Therefore, we can cancel $(p-1)!$ from the first expression and the last by Lemma 28, which proves the claim. □

4.3.7 Finding Inverses with Fermat's Theorem

Fermat's Theorem suggests an efficient procedure for finding the multiplicative inverse of a number modulo a large prime, which is just what we need for fast decryption in Turing's code. Suppose that 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 $6^{15} \text{ rem } 17$, 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, $6^{15} \text{ rem } 17 = 3$. 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. (We'll see another way to find inverses later.)

4.3.8 Breaking Turing's Code— Again

German weather reports were *not* encrypted 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:

$$\begin{aligned} m^{p-2} \cdot m' &\equiv m^{p-2} \cdot mk \pmod{p} \\ &\equiv m^{p-1} \cdot k \pmod{p} \\ &\equiv k \pmod{p} \end{aligned}$$

The last step relies on Fermat's Theorem. 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.

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.

Chapter 5

Number Theory II

5.1 Die Hard

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: Shit. Every cop within 50 miles is running his ass off and I'm out here playing kids games in the park.

Samuel: Hey, you want to focus on the problem at hand?

This is from the movie *Die Hard 3: With a Vengance*. Bruce Willis and Samuel L. Jackson are cops trying to disarm Simon's diabolical bomb. In the nick of time, they find a solution. On the surface, *Die Hard 3* is just a B-grade action movie; however, I think the inner message of the film is that everyone should learn at least a little number theory.

5.1.1 Death by Induction

Apparently, *Die Hard 4: Die Hardest* is planned for release in 2005. In this new film, Bruce just happens to stumble into another terrorist plot while on vacation. But the villian is even more devious this time: Bruce is given a 3-gallon jug and a 6-gallon jug and must measure out exactly 4 gallons. Some scratchwork suggests that the task is impossible. But surely we'd all rest easier if we could mathematically prove that there can be no *Die Hard 5*.

Let's try an inductive proof that Bruce can not produce exactly 4 gallons of water using 3 and 6-gallon jugs. The first difficulty is that induction always proves that some statement holds for all values of a natural variable n, but no such natural variable appears in the problem statement.

In general, however, in problems where we want to show that some condition always holds, a good choice is to let n be a number of steps or operations. In this case, we can let n be the number of times water is poured. Then we can try to prove that for all $n \geq 0$, neither jug contains exactly 4 gallons of water after n steps. At least we've now recast the problem in a form that makes induction applicable. Let's see how the argument works out.

Theorem 32. *Bruce dies.*

Proof. The proof is by induction. Let $P(n)$ be the proposition, "Neither jug contains 4 gallons after n steps." In the base case, $P(0)$ is true because both jugs are initially empty. Now we assume that $P(n)$ is true in order to prove $P(n+1)$; that is, we assume that neither jug contains 4 gallons after n steps and prove that neither jug contains 4 gallons after $n+1$ steps. Umm... □

We're stuck! The proof can not be completed. The fact that neither jug contains 4 gallons of water after n steps is not sufficient to prove that neither jug can contain 4 gallons after $n+1$ steps. For example, after n steps each jug might hold 2 gallons of water. But then combining their contents in the 6-gallon jug would produce 4 gallons. In logical terms, we can't hope to *prove* that $P(n)$ implies $P(n+1)$, because that's not even true: $P(n)$ does *not* imply $P(n+1)$!

Our first recourse when the inductive step falls apart is to strengthen the induction hypothesis. For example, further scratchwork suggests that the number of gallons in each jug is always a multiple of 3. So we might try again with the induction hypothesis, "After n steps, both jugs contains a multiple of 3 gallons." In fact, this approach goes through.

But we're going to use some number theory to knock off all these water jug problems. For example, what if we want to get 3 gallons using a 17-gallon jug and a 19-gallon jug? Or 1 gallon with 777 and 1001-gallon jugs?

5.1.2 A General Theorem

Suppose that we have jugs with capacities a and b. Let's carry out a few arbitrary operations and see what happens. The state of the system at each step 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.

$(0,0) \to (a,0)$	fill first jug
$\to (a-b,b)$	fill second jug from first
$\to (a-b,0)$	empty second jug
$\to (a-2b,b)$	fill second jug from first
$\to (a-2b,0)$	empty second jug
$\to (0,a-2b)$	empty first jug into second
$\to (a,a-2b)$	fill first jug
$\to (2a-3b,b)$	fill second jug from first

Of course, we're making some assumptions about the relative capacities of the two jugs here. But another point leaps out: at every step, the amount of water in each jug is of the form

$$s \cdot a + t \cdot b$$

for some integers s and t. An expression of this form is called a ***linear combination*** of a and b. Furthermore, at least one of the jugs is empty or full at all times. This sounds like an assertion that we might be able to prove by induction!

Lemma 33. *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 and at least one jug is either empty or full.*

Proof. We use induction. Let $P(n)$ be the proposition that after n steps, the amount of water in each jug is a linear combination of a and b and at least one jug is either empty or full. For the base case, $P(0)$ is true, because both jugs are empty, and $0 \cdot a + 0 \cdot b = 0$. Now we assume that $P(n)$ is true and prove the claim for $P(n+1)$. There are two cases:

- If we fill a jug from the fountain or empty a jug into the fountain, then that jug is empty or full. The amount in the other jug remains a linear combination of a and b. So $P(n+1)$ holds.

- Otherwise, we pour water from one jug to another until one is empty or the other is full. By the assumption $P(n)$, the amount in each jug is a linear combination of a and b before we begin pouring:

$$j_1 = s_1 \cdot a + t_1 \cdot b$$
$$j_2 = s_2 \cdot a + t_2 \cdot b$$

  After pouring, one jug is either empty (contains 0 gallons) or full (contains a or b gallons). Thus, the other jug contains either $j_1 + j_2$ gallons, $j_1 + j_2 - a$, or $j_1 + j_2 - b$ gallons, all of which are linear combinations of a and b.

Thus, $P(n+1)$ holds in both cases, and the claim follows by the principle of induction. □

Corollary 34. *Bruce dies.*

Proof. Since Bruce has water jugs with capacities 3 and 6, the amount in each jug is always of the form $3s + 6t$. This is always a multiple of 3, so he can not measure out 4 gallons. □

So Bruce is toast. But Lemma 33 isn't a very satisfying answer to the general water jug problem. Can we get 3 gallons with 17 and 19 gallon jugs or not? Part of the difficulty is that linear combinations are not so easy to reason about. (Is 3 a linear combination of 17 and 19?) Fortunately, linear combinations are closely related to a more familiar concept.

5.1.3 The Greatest Common Divisor

The ***greatest common divisor*** of a and b is exactly what you'd guess: the largest number that is a divisor of both a and b. It is denoted $\gcd(a, b)$. For example, $\gcd(18, 24) = 6$.

Probably some junior high math teacher made you compute greatest common divisors for no apparent reason until you were blue in the face. But, amazingly, the greatest common divisor actually turns out to be quite useful for reasoning about the integers. Specifically, the quantity $\gcd(a, b)$ is a crucial piece of information about the relationship between the numbers a and b.

The theorem below describes an alternative characterization of the greatest common divisor in terms of linear combinations, which is often easier to work with in proofs.

Theorem 35. *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. Sure enough, 4 is a linear combination of 52 and 44:

$$6 \cdot 52 + (-7) \cdot 44 = 4$$

And no linear combination of 52 and 44 is equal to a smaller positive integer.

Proof. Let m be the smallest positive linear combination of a and b. 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 $\gcd(a, b) \leq m$. By the definition of common divisor, $\gcd(a, b) \mid a$ and $\gcd(a, b) \mid b$. Therefore, for every pair of integers s and t:

$$\gcd(a, b) \mid sa + tb$$

Thus, in particular, $\gcd(a, b)$ divides m, and so $\gcd(a, b) \leq m$.

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:

$$a = q \cdot m + r \qquad \text{(where } 0 \leq r < m\text{)}$$

Now $m = sa + tb$ for some integers s and t. Subtituting in for m and rearranging terms gives:

$$\begin{aligned} a &= q \cdot (sa + tb) + r \\ r &= (1 - qs)a + (-qt)b \end{aligned}$$

We've just expressed r is 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 $\gcd(a, b)$. In the reverse direction, 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 36. *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 37. *Suppose that we have water jugs with capacities a and b. Then the amount of water in each jug is always a multiple of $\gcd(a, b)$ and at least one jug is either empty or full.*

Of course, now it would help to know a thing or two about the greatest common divisor.

5.1.4 Properties of the Greatest Common Divisor

prove some properties of the greatest common divisor.

Lemma 38. *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. $\gcd(a, b) = \gcd(a \text{ rem } b, b)$.

Proof. We prove only parts (3) and (4). For part (3), the assumptions together with Theorem 35 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 $a(asu + btu + csv) + bc(tv)$. This is a linear combination of a and bc that is equal to 1, so $\gcd(a, bc) = 1$ by Theorem 35.

For part (4), observe that $a \mid ac$ trivially and $a \mid bc$ by assumption. Therefore, a divides every linear combination of ac and bc. Theorem 35 says that $\gcd(ac, bc)$ is equal to one such linear combination. Therefore, a divides $\gcd(ac, bc) = c \cdot \gcd(a, b) = c$. The first equality uses part (2) of this lemma, and the second uses the assumption that $\gcd(a, b) = 1$. □

Part (5) of this lemma is useful for quickly computing the greatest common divisor of two numbers. For example, we could compute the greatest common divisor of 1001 and 777 like this:

$$\begin{aligned}
\gcd(1001, 777) &= \gcd(\underbrace{1001 \text{ rem } 777}_{=224}, 777) \\
&= \gcd(\underbrace{777 \text{ rem } 224}_{=105}, 224) \\
&= \gcd(\underbrace{224 \text{ rem } 105}_{=14}, 105) \\
&= \gcd(\underbrace{105 \text{ rem } 14}_{=7}, 14) \\
&= \gcd(14, 7) \\
&= 7
\end{aligned}$$

(This is called ***Euclid's algorithm***.) This calculation, together with Corollary 37, implies that there is no way to measure out 1 gallon of water using jugs with capacities 777 and 1001; we can only obtains multiples of 7 gallons. On the other hand, we might be able to get 3 gallons using jugs with capacities 17 and 19, since $\gcd(17, 19) = 1$ and 3 is a multiple of 1. We leave the question of whether or not this is really possible to you!

5.2 The Fundamental Theorem of Arithemtic

We're now almost ready to prove something that you probably already know.

Theorem 39 (Fundamental Theorem of Arithmetic). *Every positive integer n can be written in a unique way as a product of primes:*

$$n = p_1 \cdot p_2 \cdots p_j \qquad (p_1 \leq p_2 \leq \ldots \leq p_j)$$

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 the 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 fact that we just assumed earlier.

Lemma 40. *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 38. □

A routine induction argument extends this statement to the fact we assumed last time:

Lemma 41. *Let p be a prime. If $p \mid a_1 a_2 \ldots a_n$, then p divides some a_i.*

Now we're ready to prove the Fundamental Theorem of Arithemtic.

Proof. We must prove two things: (1) every positive integer can be expressed as a product of primes, and (2) this expression is unique.

First, we use strong induction to prove that every positive integer n is a product of primes. As a base case, $n = 1$ is the product of the empty set of primes. For the inductive step, suppose that every $k < n$ is a product of primes. We must show that n is also a product of primes. If n is itself prime, then this is true trivially. Otherwise, $n = ab$ for some $a, b < n$. By the induction assumption, a and b are both products of primes. Therefore, $a \cdot b = n$ is also a product of primes. Thus, the claim is proved by induction.

Second, we use the well-ordering principle to prove that every positive integer can be written as a product of primes in a unique way. 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 41 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. □

5.3 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 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 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.

5.3.1 Relative Primality and Phi

First, we need a new definition. Integers a and b are ***relatively prime*** if $\gcd(a, b) = 1$. For example, 8 and 15 are relatively prime, since $\gcd(8, 15) = 1$. Note that every integer is

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 the most famous unsolved problem in mathematics. The formula for the sum of an infinite geometric series says:

$$1 + x + x^2 + x^3 + \ldots = \frac{1}{1-x}$$

Substituting $x = 2s$, $x = 3s$, $x = 5s$, and so on for each prime number gives a sequence of equations:

$$1 + \frac{1}{2^s} + \frac{1}{2^{2s}} + \frac{1}{2^{3s}} + \ldots = \frac{1}{1-1/2^s}$$
$$1 + \frac{1}{3^s} + \frac{1}{3^{2s}} + \frac{1}{3^{3s}} + \ldots = \frac{1}{1-1/3^s}$$
$$1 + \frac{1}{5^s} + \frac{1}{5^{2s}} + \frac{1}{5^{3s}} + \ldots = \frac{1}{1-1/5^s}$$
$$\text{etc.}$$

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:

Theorem. *Every nontrivial zero of the zeta function $\zeta(s)$ lies on the line $s = 1/2 + ci$ in the complex plane.*

A proof of the Riemann Hypothesis would immediately imply, among other things, a strong form of the Prime Number Theorem— and earn the prover a $1 million prize!

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 42. *The function ϕ obeys the following relationships:*

1. *If a and b are relatively prime, then $\phi(ab) = \phi(a)\phi(b)$.*
2. *If p is a prime, then $\phi(p^k) = p^k - p^{k-1}$ for $k \geq 1$.*

This is not a terribly difficult theorem, but we'll hold off on the proof for a few weeks. In the meanwhile, here's an example of how we might use Theorem 42 to compute $\phi(300)$:

$$\begin{aligned}\phi(300) &= \phi(2^2 \cdot 3 \cdot 5^2) \\ &= \phi(2^2) \cdot \phi(3) \cdot \phi(5^2) \\ &= (2^2 - 2^1)(3^1 - 3^0)(5^2 - 5^1) \\ &= 80\end{aligned}$$

We factor 300 in the first step, use part (1) of Theorem 42 twice in the second step, use part (2) in the third step, and then simplify.

5.3.2 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, if k is relatively prime to n, then k has a multiplicative inverse modulo n:

Lemma 43. *Let n be a positive integer. If k is relatively prime to n, then there exists an integer k^{-1} such that:*

$$k \cdot k^{-1} \equiv 1 \pmod{n}$$

Proof. There exist integers s and t such that $sk + tn = \gcd(k, n) = 1$ by Theorem 35. Rearranging terms gives $tn = 1 - sk$, which implies that $n \mid 1 - sk$ and $sk \equiv 1 \pmod{n}$. Define k^{-1} to be s. □

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 44. *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.

5.3.3 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 45. *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:*

$$(k_1 \cdot k) \text{ rem } n, \quad (k_2 \cdot k) \text{ rem } n, \quad (k_3 \cdot k) \text{ rem } n, \quad \ldots, \quad (k_r \cdot k) \text{ rem } n$$

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 $k_i k \text{ rem } n = k_j k \text{ rem } n$. This is equivalent to $k_i k \equiv k_j k \pmod{n}$, which implies $k_i \equiv k_j \pmod{n}$ by Corollary 44. 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 only equal to itself.

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

$$\gcd(k_i k, n) = \gcd(k_i k \text{ rem } n, n) = 1$$

by part (3) of Lemma 38. Therefore, $k_i k \text{ rem } n$ is relatively prime to n and is in the range from 0 to $n-1$ by the definition of rem. The second sequence is defined to consist of all such integers. □

We can now prove Euler's Theorem:

Theorem 46 (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:

$$\begin{aligned} & k_1 \cdot k_2 \cdot k_3 \cdots k_r \\ & \quad \equiv \left(k_1 \cdot k \text{ rem } n\right) \cdot \left(k_2 \cdot k \text{ rem } n\right) \cdot \left(k_3 \cdot k \text{ rem } n\right) \cdots \left(k_r \cdot k \text{ rem } n\right) \pmod{n} \\ & \quad \equiv (k_1 \cdot k) \cdot (k_2 \cdot k) \cdot (k_3 \cdot k) \cdots (k_r \cdot k) \pmod{n} \\ & \quad \equiv (k_1 \cdot k_2 \cdot k_3 \cdots k_r) \cdot k^r \pmod{p} \end{aligned}$$

The first two expressions are actually equal by Lemma 45; therefore, they are certainly congruent modulo n. The second step uses a property of mod and rem that we proved earlier. In the third step, we rearrange terms in the product.

Part (3) of Lemma 38 implies that $k_1 \cdot k_2 \cdot k_3 \cdots k_r$ is prime relative to n. Therefore, we can cancel this product from the first expression and the last by Corollary 44. This proves the claim. □

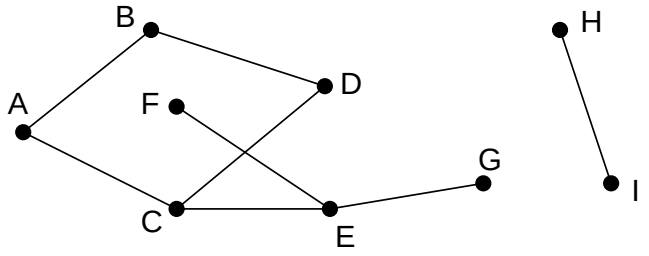
We can find multiplicative inverses using Euler's theorem as we did with Fermat's theorem: if k is relatively prime to n, then $k^{\phi(n)-1}$ is a multiplicative inverse of k modulo n. However, this approach requires computing $\phi(n)$. Our best method for doing so requires factoring n, which can be quite difficult in general. Fortunately, there is a tool that finds multiplicative inverses and solves many other number-theoretic problems as well. You'll see this— and RSA— in recitation.

Chapter 6

Graph Theory

6.1 Introduction

Informally, a ***graph*** is a bunch of dots connected by lines. Here is an example of a graph:



Sadly, this definition is not precise enough for mathematical discussion. Formally, a graph is a pair of sets (V, E), where:

- V is a nonempty set whose elements are called ***vertices***.
- E is a collection of two-element subsets of V called ***edges***.

The vertices correspond to the dots in the picture, and the edges correspond to the lines. Thus, the dots-and-lines diagram above is a pictorial representation of the graph (V, E) where:

$$V = \{A, B, C, D, E, F, G, H, I\}$$
$$E = \{\{A, B\}, \{A, C\}, \{B, D\}, \{C, D\}, \{C, E\}, \{E, F\}, \{E, G\}, \{H, I\}\}.$$

6.1.1 Definitions

A nuisance in first learning graph theory is that there are so many definitions. They all correspond to intuitive ideas, but can take a while to absorb. Some ideas have multiple names. For example, graphs are sometimes called ***networks***, vertices are sometimes called ***nodes***, and edges are sometimes called ***arcs***. Even worse, no one can agree on the exact meanings of terms. For example, in our definition, every graph must have at least one vertex. However, other authors permit graphs with no vertices. (The graph with no vertices is the single, stupid counterexample to many would-be theorems— so we're banning it!) This is typical; everyone agrees more-or-less what each term means, but disagrees about weird special cases. So do not be alarmed if definitions here differ subtly from definitions you see elsewhere. Usually, these differences do not matter.

Hereafter, we use A—B to denote an edge between vertices A and B rather than the set notation $\{A, B\}$. Note that A—B and B—A are the same edge, just as $\{A, B\}$ and $\{B, A\}$ are the same set.

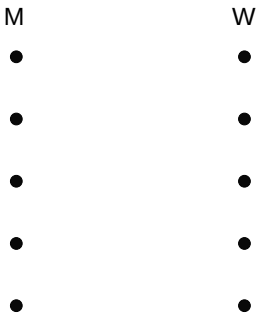
Two vertices in a graph are said to be ***adjacent*** if they are joined by an edge, and an edge is said to be ***incident*** to the vertices it joins. The number of edges incident to a vertex is called the ***degree*** of the vertex. For example, in the graph above, A is adjacent to B and B is adjacent 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.

Deleting some vertices or edges from a graph leaves a ***subgraph***. Formally, a subgraph of $G = (V, E)$ is a graph $G' = (V', E')$ where V' is a nonempty subset of V and E' is a subset of E. Since a subgraph is itself a graph, the endpoints of every edge in E' must be vertices in V'.

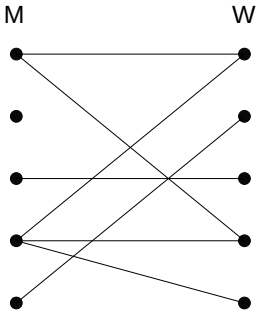
6.1.2 Sex in America

A 1994 University of Chicago study entitled *The Social Organization of Sexuality* found that on average men have 74% more opposite-gender partners than women.

Let's recast this observation in graph theoretic terms. Let $G = (V, E)$ be a graph where the set of vertices V consists of everyone in America. Now each vertex either represents either a man or a woman, so we can partition V into two subsets: M, which contains all the male vertices, and W, which contains all the female vertices. Let's draw all the M vertices on the left and the W vertices on the right:



Now, without getting into a lot of specifics, *sometimes an edge appears* between an M vertex and a W vertex:



Since we're only considering opposite-gender relationships, every edge connects an M vertex on the left to a W vertex on the right. So the sum of the degrees of the M vertices must equal the sum of the degrees of the W vertices:

$$\sum_{x \in M} \deg(x) = \sum_{y \in W} \deg(y)$$

Now suppose we divide both sides of this equation by the product of the sizes of the two sets, $|M| \cdot |W|$:

$$\left(\frac{\sum_{x \in M} \deg(x)}{|M|} \right) \cdot \frac{1}{|W|} = \left(\frac{\sum_{y \in W} \deg(y)}{|W|} \right) \cdot \frac{1}{|M|}$$

The terms above in parentheses are the *average degree of an M vertex* and the *average degree of a W* vertex. So we know:

$$\frac{\text{Avg. deg in } M}{|W|} = \frac{\text{Avg. deg in } W}{|M|}$$

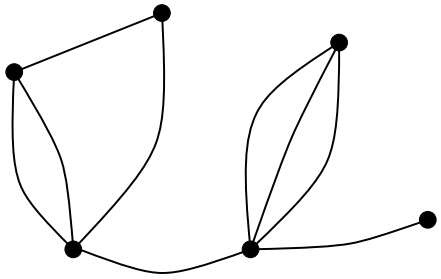
$$\text{Avg. deg in } M = \frac{|W|}{|M|} \cdot \text{Avg. deg in } W$$

Now the Census Bureau reports that there are slightly more women than men in America; in particular $|W| \,/\, |M|$ is about 1.035. So— assuming the Census Bureau is correct—

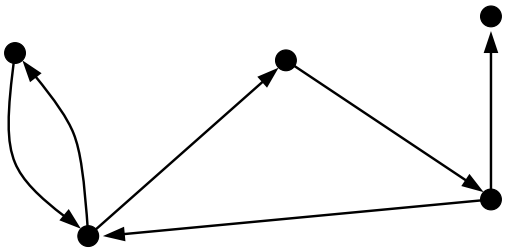
we've just proved that the University of Chicago study got bad data! On average, men have 3.5% more opposite-gender partners. Furthermore, this is totally unaffected by differences in sexual practices between men and women; rather, it is completely determined by the relative number of men and women!

6.1.3 Graph Variations

There are many variations on the basic notion of a graph. Three particularly common variations are described below. In a ***multigraph***, there may be more than one edge between a pair of vertices. Here is an example:

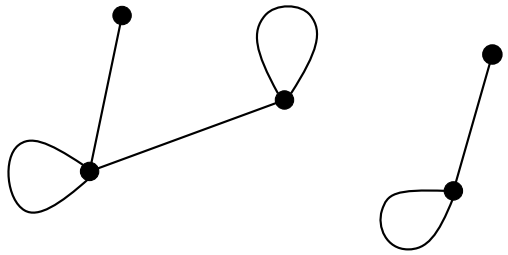


The edges in a ***directed graph*** are arrows pointing to one endpoint or the other. Here is an example:



Directed graphs are often called ***digraphs***. We denote an edge from vertex A to vertex B in a digraph by $A \longrightarrow B$. Formally, the edges in a directed graph are ordered pairs of vertices rather than sets of two vertices. The number of edges directed into a vertex is called the ***indegree*** of the vertex, and the number of edged directed out is called the ***outdegree***.

One can also allow ***self-loops***, edges with both endpoints at one vertex. Here is an example of a graph with self-loops:



Combinations of these variations are also possible; for example, one could work with directed multigraphs with self-loops.

Except where stated otherwise, the word "graph" in this course refers to a graph without multiple edges, directed edges, or self-loops.

6.1.4 Applications of Graphs

Graphs are the most useful mathematical objects in computer science. You can model an enormous number of real-world systems and phenomena using graphs. Once you've created such a model, you can tap the vast store of theorems about graphs to gain insight into the system you're modeling. Here are some practical situations where graphs arise:

Data Structures Each vertex represents a data object. There is a directed edge from one object to another if the first contains a pointer or reference to the second.

Attraction Each vertex represents a person, and each edge represents a romantic attraction. The graph could be directed to model the unfortunate asymmetries.

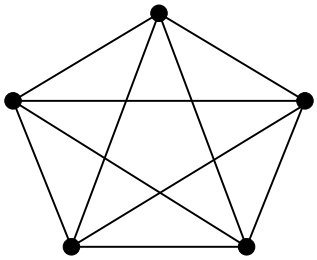
Airline Connections Each vertex represents an airport. If there is a direct flight between two airports, then there is an edge between the corresponding vertices. These graphs often appear in airline magazines.

The Web Each vertex represents a web page. Directed edges between vertices represent hyperlinks.

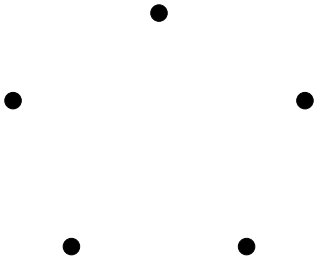
People often put numbers on the edges of a graph, put colors on the vertices, or add other ornaments that capture additional aspects of the phenomenon being modeled. For example, a graph of airline connections might have numbers on the edges to indicate the duration of the corresponding flight. The vertices in the attraction graph might be colored to indicate the person's gender.

6.1.5 Some Common Graphs

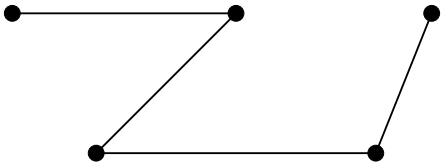
Some graphs come up so frequently that they have names. The ***complete graph*** on n vertices, also called K_n, has an edge between every pair of vertices. Here is K_5:



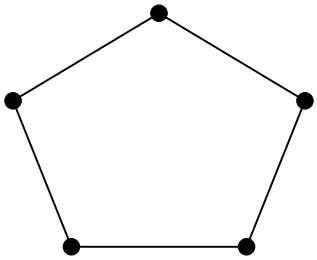
The ***empty graph*** has no edges at all. Here is the empty graph on 5 vertices:



Here is a ***path*** with 5 vertices:



And here is a ***cycle*** with 5 vertices, which is typically denoted C_5:



Paths and cycles are going to be particularly important, so let's define them precisely. A ***path*** is a graph $P = (V, E)$ of the form

$$V = \{v_1, v_2, \ldots, v_n\} \qquad E = \{v_1\text{—}v_2, v_2\text{—}v_3, \ldots, v_{n-1}\text{—}v_n\}$$

where $n \geq 1$ and vertices $v_1, \ldots, v_n$ are all distinct. Vertices v_1 and v_n are the ***endpoints*** of the path. Note that a path may consist of a single vertex, in which case both endpoints are

the same. We'll often say that there is a path from u to v in a graph G; this is a shorthand for saying that a path with endpoints u and v is a subgraph of G.

Similarly, a cycle is a graph $C = (V, E)$ of the form

$$V = \{v_1, v_2, \ldots, v_n\} \qquad E = \{v_1\text{—}v_2, v_2\text{—}v_3, \ldots, v_{n-1}\text{—}v_n, v_n\text{—}v_1\}$$

where $n \geq 3$ and $v_1, \ldots, v_n$ are all distinct. The ***length*** of a path or cycle is the number of edges it contains. For example, a path with 5 vertices has length 4, but a cycle with 5 vertices has length 5.

6.1.6 Isomorphism

Two graphs that look the same might actually be different in a formal sense. For example, the two graphs below are both 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" and use that as our main notion of equivalence between graphs. Graphs G_1 and G_2 are ***isomorphic*** if there exists a one-to-one correspondence between vertices in G_1 and 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 correspondence 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 correspondence itself is called an ***isomorphism***.

Two isomorphic graphs may be drawn to look quite different. For example, here are two different ways of drawing C_5:

Isomorphic graphs share a great many properties, such as the number of vertices, number of edges, and the pattern of vertex degrees. Thus, two graphs can be proved *nonisomorphic* by identifying some property that one possesses that the other does not. For example, if one graph has two vertices of degree 5 and another has three vertices of degree 5, then the graphs can not be isomorphic.

6.2 Connectivity

In the diagram below, the graph on the left has two pieces, while the graph on the right has just one.

Let's put this observation in rigorous terms. A graph is ***connected*** if for every pair of vertices u and v, the graph contains a path with endpoints u and v as a subgraph. The graph on the left is not connected because there is no path from any of the top three vertices to either of the bottom two vertices. However, the graph on the right is connected, because there is a path between every pair of vertices.

A maximal, connected subgraph is called a ***connected component***. (By "maximal", we mean that including any additional vertices would make the subgraph disconnected.) The graph on the left has two connected components, the triangle and the single edge. The graph on the right is entirely connected and thus has a single connected component.

6.2.1 A Simple Connectivity Theorem

The following theorem says that a graph with few edges must have many connected components.

Theorem 47. *Every graph $G = (V, E)$ has at least $|V| - |E|$ connected components.*

Proof. We use induction on the number of edges. Let $P(n)$ be the proposition that every graph $G = (V, E)$ with $|E| = n$ has at least $|V| - n$ connected components.

Base case: In a graph with 0 edges, each vertex is itself a connected component, and so there are exactly $|V| - 0 = |V|$ connected components.

Inductive step: Now we assume that the induction hypothesis holds for every n-edge graph in order to prove that it holds for every $(n + 1)$-edge graph, where $n \geq 0$. Consider a graph $G = (V, E)$ with $n + 1$ edges. Remove an arbitrary edge u—v and call the resulting graph G'. By the induction assumption, G' has at least $|V| - n$ connected components. Now add back the edge u—v to obtain the original graph G. If u and v were in the same connected component of G', then G has the same number of connected components as G', which is at least $|V| - n$. Otherwise, if u and v were in different connected components of G', then these two components are merged into one in G, but all other components remain. Therefore, G has at least $|V| - n - 1 = |V| - (n + 1)$ connected components.

The theorem follows by induction. □

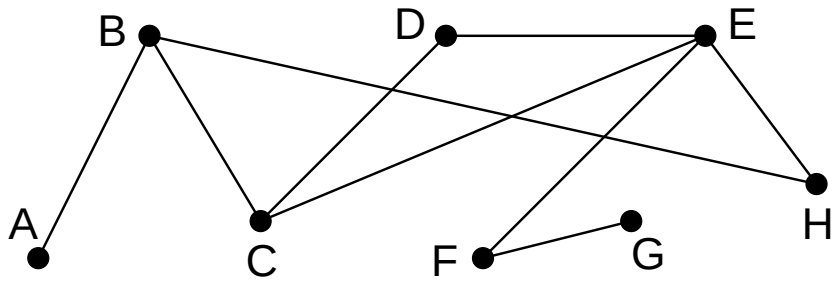
Corollary 48. *Every connected graph with n vertices has at least $n - 1$ edges.*

A couple points about the proof of Theorem 47 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 approachs should be among the first you consider. Don't try induction on other variables that crop up in the problem unless these two strategies seem hopeless.

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 n-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 very nasty logical error in similar arguments. (You'll see an example in recitation.) Always use shrink-down, grow-back arguments, and you'll never fall into this trap.

6.2.2 Distance and Diameter

The ***distance*** between two vertices in a graph is the length of the shortest path between them. For example, the distance between two vertices in a graph of airline connections is the minimum number of flights required to travel between two cities.



In this graph, the distance between C and H is 2, the distance between G and C is 3, and the distance between A and itself is 0. If there is *no* path between two vertices, then the distance between them is said to be "infinity".

The ***diameter*** of a graph is the distance between the two vertices that are farthest apart. The diameter of the graph above is 5. The most-distant vertices are A and G, which are at distance 5 from one another.

Six Degrees of Separation

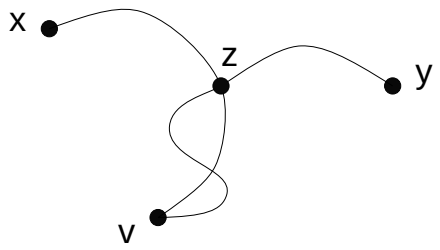
There is an old claim that the world has only "six degrees of separation". In other words, if you pick any two people on the planet— say a hermit in Montana and a random person off the street in Beijing— then the hermit knows someone who knows someone who knows someone who knows the Chinese pedestrian, where the word "knows" appears at most six times.

We can recast this in graph-theoretic terms. Consider a graph where the vertices are all the people on the planet, and there is an edge between two people if and only if they know each other. Then the "six degrees of separation" claim amounts to the assertion that the diameter of this graph is at most 6.

There is little hope of proving or disproving the claim, since people are constantly being born, meeting one another, and dying and no one can keep track of who-knows-who. However, precise data does exist for something similar. The University of Virginia maintains the *Oracle of Bacon* website. This is based on an "acting graph" where the vertices are actors and actresses, and there is an edge between two performers if they appeared in a movie together. The website reports that everyone is within distance 8 of Kevin Bacon. (This excludes a few actors who are completely disconnected.) This allows us to at least obtain an upper bound on the diameter of the acting graph.

Theorem 49. *Let v be an arbitrary vertex in a graph G. If every vertex is within distance d of v, then the diameter of the graph is at most $2d$.*

Proof. Let x and y be arbitrary vertices in the graph. Then there exists a path of length at most d from x to v, and there exists a path of length at most d from v to y.



Let z be the vertex that lies on both the x-to-v and v-to-y paths and is closest to x. (We know that such a vertex exists, since z could be v, at least.) Joining the x-to-z segment to the z-to-y segment gives a path from x to y of length at most $2d$. Therefore, every vertex is within distance $2d$ of every other. □

Data elsewhere on the Oracle of Bacon site shows that the diameter of the acting graph is at least 15, so the upper bound isn't far off.

6.2.3 Walks

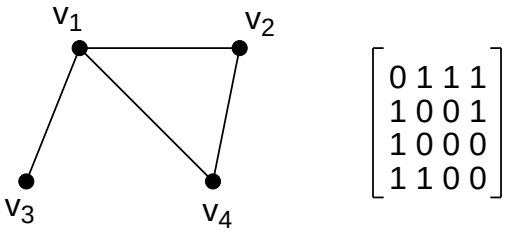
A ***walk*** in a graph G is an alternating sequence of vertices and edges of the form:

$$v_0\ v_0\text{—}v_1\ v_1\ v_1\text{—}v_2\ v_2\ \ldots\ v_{n-1}\ v_{n-1}\text{—}v_n\ v_n$$

If $v_0 = v_n$, then the walk is ***closed***. Walks are similar to paths. However, a walk can cross itself, traverse the same edge multiple times, etc. There is a walk between two vertices if and only if there is a path between the vertices.

6.3 Adjacency Matrices

A graph can be represented by an ***adjacency matrix***. In particular, if a graph has vertices $v_1, \ldots, v_n$, then the adjacency matrix is $n \times n$. The entry in row i, column j is 1 if there is an edge v_i—v_j and is 0 if there is no such edge. For example, here is a graph and its adjacency matrix:



The adjacency matrix of an undirected graph is always symmetric about the diagonal line running from the upper left entry to the lower right. The adjacency matrix of a directed graph need not be symmetric, however. Entries on the diagonal of an adjacency matrix are nonzero only if the graph contains self-loops.

Adjacency matrices are useful for two reasons. First, they provide one way to represent a graph in computer memory. Second, by mapping graphs to the world of matrices, one can bring all the machinery of linear algebra to bear on the study of graphs. For example, one can analyze a highly-prized quality of graphs called "expansion" by looking at eigenvalues of the adjacency matrix. (In a graph with good expansion, the number of edges departing each subset of vertices is at least proportional to the size of the subset. This is not so easy to achieve when the graph as a whole as few edges, say $|E| = 3|V|$.) Here we prove a simpler theorem in this vein. If M is a matrix, then M_{ij} denotes the entry in row i, column j. Let M^k denote the k-th power of M. As a special case, M^0 is the identity matrix.

Theorem 50. *Let G be a digraph (possibly with self-loops) with vertices $v_1, \ldots, v_n$. Let M be the adjacency matrix of G. Then M^k_{ij} is equal to the number of length-k walks from v_i to v_j.*

Proof. We use induction on k. The induction hypothesis is that M^k_{ij} is equal to the number of length-k walks from v_i to v_j, for all i, j.

Each vertex has a length-0 walk only to itself. Since $M^0_{ij} = 1$ if and only if $i = j$, the hypothesis holds for $k = 0$.

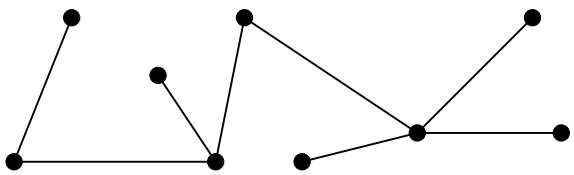
Now suppose that the hypothesis holds for some $k \geq 0$. We prove that it also holds for $k + 1$. Every length-$(k + 1)$ walk from v_i to v_j consists of a length k walk from v_i to some intermediate vertex v_m followed by an edge v_m—v_j. Thus, the number of length-$(k + 1)$ walks from v_i to v_j is equal to:

$$M^k_{iv_1} M_{v_1 j} + M^k_{iv_2} M_{v_2 j} + \ldots + M^k_{iv_n} M_{v_n j}$$

This is precisely the value of M^{k+1}_{ij}, so the hypothesis holds for $k + 1$ as well. The theorem follows by induction. □

6.4 Trees

A connected, acyclic graph is called a ***tree***. (A graph is ***acyclic*** if no subgraph is a cycle.) Here is an example of a tree:



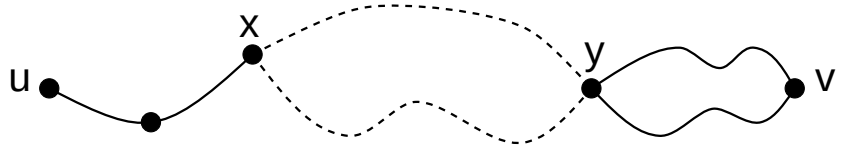
A vertex of degree one is called a ***leaf***. In this example, there are 5 leaves.

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 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 51. *Every tree $T = (V, E)$ has the following properties:*

1. *There is a unique path between every pair of vertices.*
2. *Adding any edge creates a cycle.*
3. *Removing any edge disconnects the graph.*
4. *Every tree with at least two vertices has at least two leaves.*
5. $|V| = |E| + 1$.

Proof. 1. There is at least one path between every pair of vertices, because the graph is connected. Suppose that there are two different 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 paths from x to y with no common edges, which defines a cycle. This is a contradiction, since trees are acyclic. Therefore, there is exactly one path between every pair of vertices.



2. An additional edge u—v together with the unique path between u and v forms a cycle.
3. 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.
4. Let $v_1, \ldots, v_m$ be the sequence of vertices on a longest path in T. Then $m \geq 2$, since a tree with two vertices must contain at least one edge. There can not be an edge v_1—v_i for $2 < i \leq m$; otherwise, vertices $v_1, \ldots, v_i$ would from a cycle. Furthermore, there can not 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.

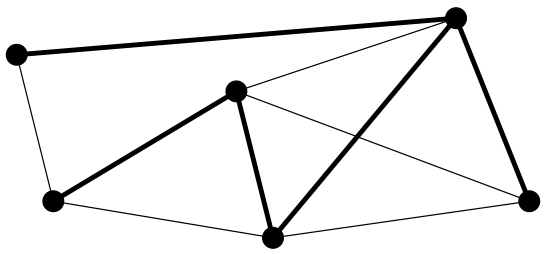
5. We use induction on $|V|$. For a tree with a single vertex, the claim holds since $|E| + 1 = 0 + 1 = 1$. Now suppose that the claim holds for all n-vertex trees and consider an $(n+1)$-vertex tree T. Let v be a leaf of the tree. Deleting v and its incident edge gives a smaller tree for which the equation $|V| = |E| + 1$ holds by induction. If we add back the vertex v and its incident edge, then the equation still holds because the number of vertices and number of edges both increased by 1. Thus, the claim holds for T and, by induction, for all trees.

□

Many subsets of the properties above, together with connectedness and lack of cycles, are sufficient to characterize all trees. For example, a connected graph that satisfies $|V| = |E| + 1$ is necessarily a tree, though we won't prove this fact.

6.4.1 Spanning Trees

Trees are everywhere. In fact, every connected graph $G = (V, E)$ contains a ***spanning tree*** $T = (V, E')$ as a subgraph. (Note that original graph G and the spanning tree T have the same set of vertices.) For example, here is a connected graph with a spanning tree highlighted.



Theorem 52. *Every connected graph $G = (V, E)$ contains a spanning tree.*

Proof. Let $T = (V, E')$ be a connected subgraph of G with the smallest number of edges. We show that T is acyclic by contradiction. So suppose that T has a cycle with the following edges:

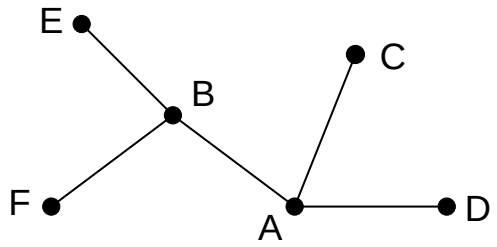
$$v_0—v_1, v_1—v_2, \ldots, v_n—v_0$$

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. This is a contradiction, since T was defined to be a connected subgraph of G with the smallest number of edges. Therefore, T is acyclic. □

6.4.2 Tree Variations

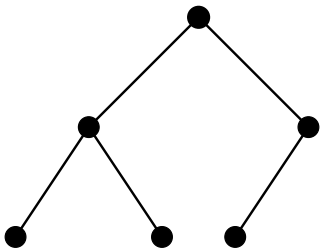
Trees come up often in computer science. For example, information is often stored in tree-like data structures and the execution of many recursive programs can be regarded as a traversal of a tree.

There are many varieties of trees. For example, a ***rooted tree*** is a tree with one vertex identified as the ***root***. Let u—v be an edge in a rooted tree such that u is closer to the root than v. Then u is the ***parent*** of v, and v is a ***child*** of u.

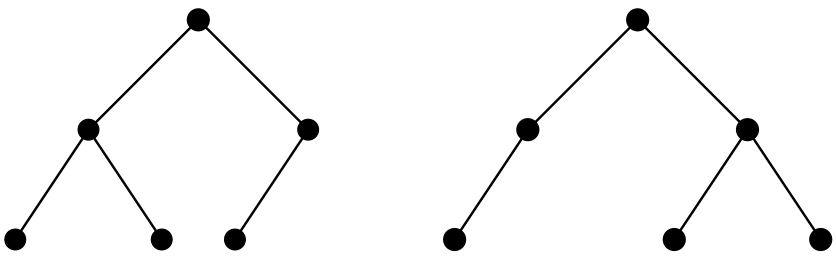


In the tree above, suppose that we regard vertex A as the root. Then E and F are the children of B, and A is the parent of B, C, and D.

A ***binary*** tree is a rooted tree in which every vertex has at most two children. Here is an example, where the topmost vertex is the root.



In an ***ordered, binary*** tree, the children of a vertex v are distinguished. One is called the ***left child*** of v, and the other is called the ***right child***. For example, if we regard the two binary trees below as unordered, then they are equivalent. However, if we regard these trees as ordered, then they are different.



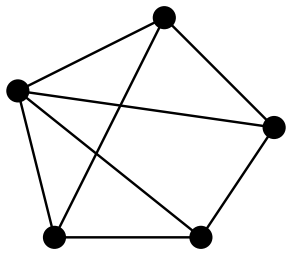
Chapter 7

Graph Theory II

7.1 Coloring Graphs

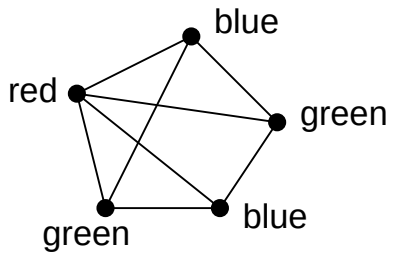
Each term, the MIT Schedules Office must assign a time slot for each final exam. This is not easy, because some students are taking several classes with finals, and a student can take only one test during a particular time slot. The Schedules Office wants to avoid all conflicts, but to make the exam period as short as possible.

We can recast this scheduling problem as a question about coloring the vertices of a graph. Create a vertex for each course with a final exam. Put an edge between two vertices if some student is taking both courses. For example, the scheduling graph might look like this:



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 (red), two Monday afternoon (blue), and two Tuesday morning (green).

7.1.1 k-Coloring

Many other resource allocation problems boil down to coloring some graph. In general, a graph G is ***k-colorable*** if each vertex can be assigned one of k colors so that adjacent vertices get different colors. The smallest sufficient number of colors is called the ***chromatic number*** of G. The chromatic number of a graph is generally difficult to compute, but the following theorem provides an upper bound:

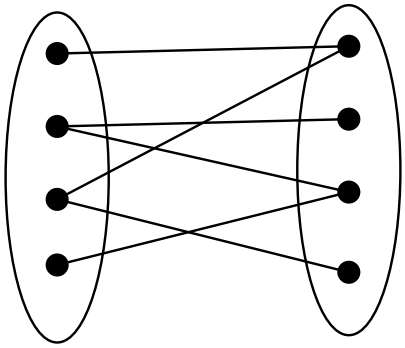
Theorem 53. *A graph with maximum degree at most k is $(k+1)$-colorable.*

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. A 1-vertex graph has maximum degree 0 and is 1-colorable, so $P(1)$ is true.

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, leaving an n-vertex graph G'. The maximum degree of G' is at most k, and so G' is $(k+1)$-colorable by our assumption $P(n)$. Now add back vertex v. We can assign v a color different from all adjacent vertices, since v has degree at most k and $k+1$ colors are available. Therefore, G is $(k+1)$-colorable. The theorem follows by induction. □

7.1.2 Bipartite Graphs

The 2-colorable graphs are important enough to merit a special name; they are called ***bipartite graphs***. Suppose that G is bipartite. This means we can color every vertex in G either black or white so that adjacent vertices get different colors. Then we can put all the black vertices in a clump on the left and all the white vertices in a clump on the right. Since every edge joins differently-colored vertices, every edge must run between the two clumps. Therefore, every bipartite graph looks something like this:



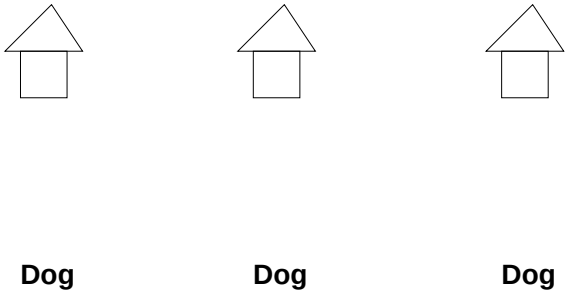
Bipartite graphs are both useful and common. For example, every path, every tree, and every even-length cycle is bipartite. In turns out, in fact, that every graph not containing an odd cycle is bipartite and vice verse.

Theorem 54. *A graph is bipartite if and only if it contains no odd cycle.*

We'll help you prove this on the homework.

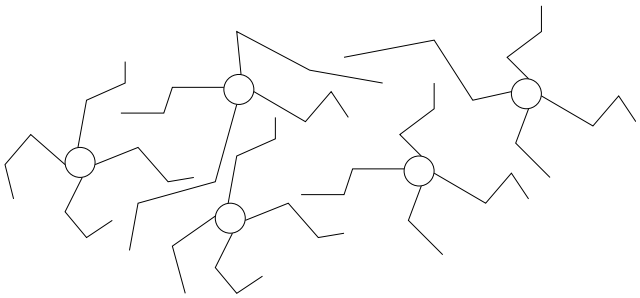
7.2 Planar Graphs

Here are three dogs and three houses.



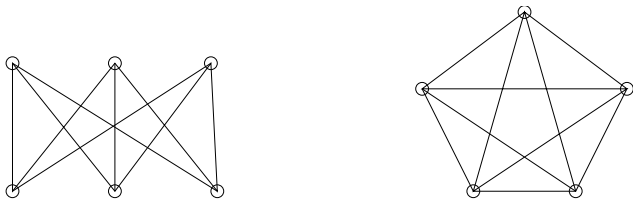
Can you find a path from each dog to each house such that no two paths intersect?

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. Thus, these two puzzles are asking whether the graphs below are planar; that is, whether they can be redrawn so that no edges cross.

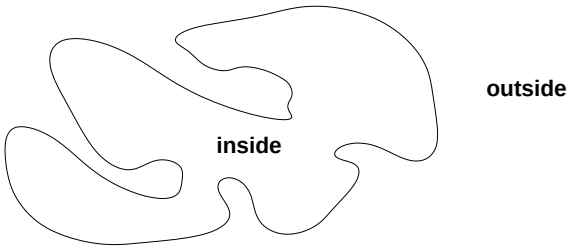


In each case, the answer is, "No— but almost!" In fact, each drawing would be possible if any single edge were removed.

More precisely, graph is planar if it has a ***planar embedding*** (or ***drawing***). This is a way of associating each vertex with a distinct point in the plane and each edge with a continuous, non-self-intersecting curve such that:

- The endpoints of the curve associated with an edge (u, v) are the points associated with vertices u and v.
- The curve associated with an edge (u, v) contains no other vertex point and intersects no other edge curve, except at its endpoints.

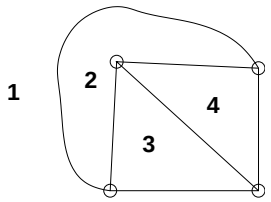
This scary definition hints at a theoretical problem associated with planar graphs: while the idea seems intuitively simple, rigorous arguments about planar graphs require some heavy-duty math. The classic example of the difficulties that can arise is the Jordan Curve Theorem. This states that every simple, closed curve separates the plane into two regions, an inside and an outside, like this:



Up until the late 1800's, mathematicians considered this obvious and implicitly treated it as an axiom. However, in 1887 Jordan pointed out that, in principle, this could be a theorem proved from simpler axioms. Actually nailing down such a proof required more than 20 years of effort. (It turns out that there are some *nasty* curves that defy simple arguments.) Planar graphs come up all the time and are worth learning about, but a several-month diversion into topology isn't in the cards. So when we need an "obvious" geometric fact, we'll handle it the old fashioned way: we'll assume it!

7.2.1 Euler's Formula

A drawing of a planar graph divides the plane into ***faces***, regions bounded by edges of the graph. For example, the drawing below has four faces:



Face 1, which extends off to infinity in all directions, is called the ***outside face***. It turns out that the number of vertices and edges in a connected planar graph determine the number of faces in every drawing:

Theorem 55 (Euler's Formula). *For every drawing of a connected planar graph*

$$v - e + f = 2$$

where v is the number of vertices, e is the number of edges, and f is the number of faces.

For example, in the drawing above, $|V| = 4$, $|E| = 6$, and $f = 4$. Sure enough, $4-6+4 = 2$, as Euler's Formula claims.

Proof. We use induction on the number of edges in the graph. Let $P(e)$ be the proposition that $v - e + f = 2$ for every drawing of a graph G with e edges.

Base case: A connected graph with $e = 0$ edges has $v = 1$ vertices, and every drawing of the graph has $f = 1$ faces (the outside face). Thus, $v - e + f = 1 - 0 + 1 = 2$, and so $P(0)$ is true.

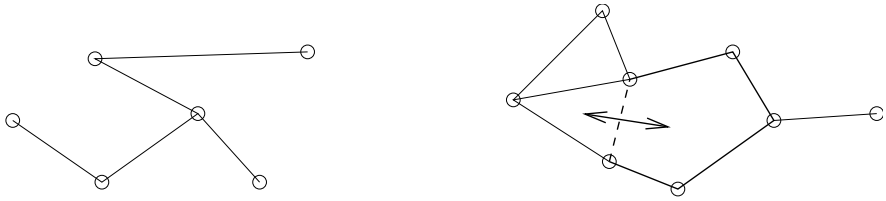
Inductive step: Now we assume that $P(e)$ is true in order to prove $P(e + 1)$ where $e \geq 0$. Consider a connected graph G with $e + 1$ edges. There are two cases:

1. If G is acylic, then the graph is a tree. Thus, there are $e+2$ vertices and every drawing has only the outside face. Since $(e + 2) - (e + 1) + 1 = 2 - 1 + 1 = 2$, $P(n + 1)$ is true.

2. Otherwise, G has at least one cycle. Select a spanning tree and an edge (u, v) in the cycle, but not in the tree. (The spanning tree can not contain all edges in the cycle, since trees are acyclic.) Removing (u, v) merges the two faces on either side of the edge and leaves a graph G' with only e edges and some number of vertices v and faces f. Graph G' is connected, because there is a path between every pair of vertices within the spanning tree. So $v - e + f = 2$ by the induction assumption $P(e)$. Thus, the original graph G had v vertices, $e + 1$ edges, and $f + 1$ faces. Since $v - (e + 1) + (f + 1) = v - e + f = 2$, $P(n + 1)$ is again true.

The theorem follows by the principle of induction. □

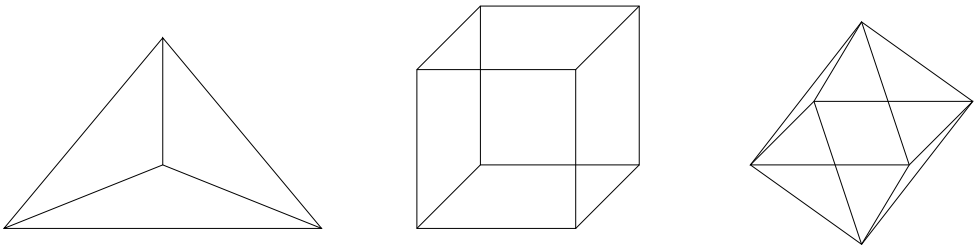
In this argument, we implicitly assumed two geometric facts: a drawing of a tree can not have multiple faces and removing an edge on a cycle merges two faces into one.



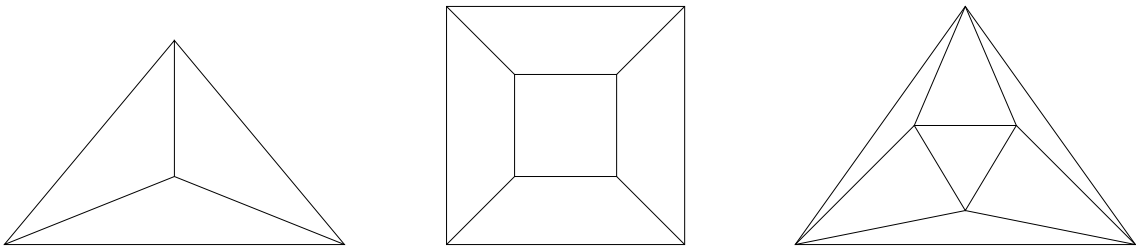
7.2.2 Classifying Polyhedra

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 tetraheron, 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 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 non-degenerate polygon has at least 3 sides, so $n \geq 3$. And at least 3 polygons must meet to from 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 $\frac{1}{3} + \frac{1}{6} = \frac{1}{2}$, which is less than the right side. Checking the finitely-many cases that remain turns up 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:

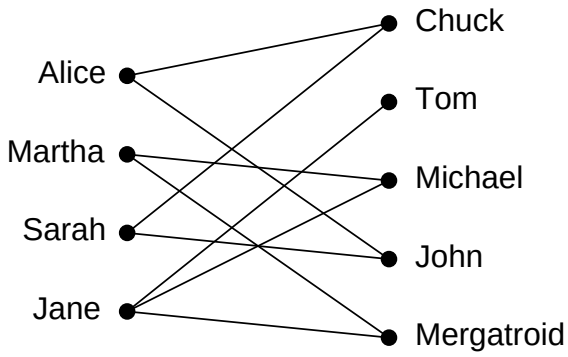
n	m	v	e	f	polyhedron
3	3	4	6	4	tetrahedron
4	3	8	12	6	cube
3	4	6	12	8	octahedron
3	5	12	30	20	icosahedron
5	3	20	30	12	dodecahedron

The last polyhedron in this list, the dodecahedron, was the other great mathematical secret of the Pythagorean sect!

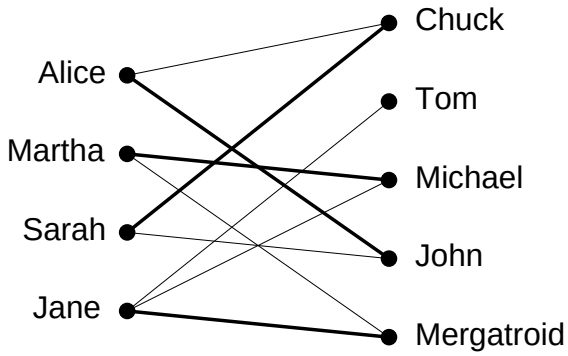
7.3 Hall's Marriage Theorem

A class contains some girls and some boys. Each girl likes some boys and does not like others. Under what conditions can each girl be paired up with a boy that she likes?

We can model the situation with a bipartite graph. Create a vertex on the left for each girl and a vertex on the right for each boy. If a girl likes a boy, put an edge between them. For example, we might obtain the following graph:



In graph terms, our goal is to find a ***matching*** for the girls; that is, a subset of edges such that exactly one edge is incident to each girl and at most one edge is incident to each boy. For example, here is one possible matching for the girls:



Hall's Marriage Theorem states necessary and sufficient conditions for the existence of a matching in a bipartite graph. Hall's Theorem is a remarkably useful mathematical tool, a hammer that bashes many problems. Moreover, it is the tip of a conceptual iceberg, a special case of the "max-flow, min-cut theorem", which is in turn a byproduct of "linear programming duality", one of the central ideas of algorithmic theory.

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 ***marriage 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 marriage condition holds, then a matching exists.

Theorem 56. *A matching for a set of girls G with a set of boys B can be found if and only if the marriage condition holds.*

Proof. First, let's suppose that a matching exists and show that the marriage 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 marriage condition holds.

Next, let's suppose that the marriage condition holds and show that a matching exists. We use strong induction on $|G|$, the number of girls. If $|G| = 1$, then the marriage 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 marriage 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 marriage 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 marriage condition, we know:

   $$|X \cup X'| \leq |Y \cup Y'|$$

   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.

7.3.1 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..." 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, but not actually in S.

Theorem 57 (Hall's Theorem). *Let $G = (L \cup R, E)$ be a bipartite graph such that every edge has one endpoint in L and one endpoint in R. There is a matching for the L vertices if and only if $|N(S)| \geq |S|$ for every set $S \subseteq L$.*

Chapter 8

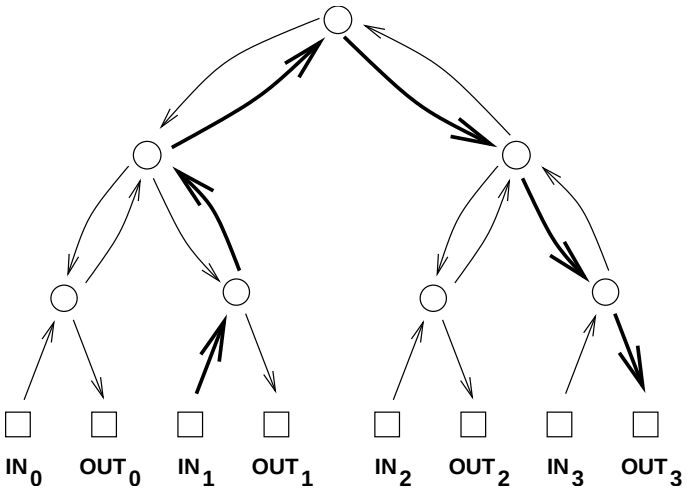
Communication Networks

This is a new topic in 6.042, so these lecture notes are more likely to contain errors. If you suspect you've found an error or just find something particularly confusing, send email to e_lehman@mit.edu and I'll try to correct the problem.

Today we'll explore an important application of graphs in computer science: modeling communication networks. Generally, vertices will represent computers, processors, and switches and 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 this lecture, we'll look at some of the nicest and most commonly used communication networks.

8.1 Complete Binary Tree

Let's start with a ***complete binary tree***. Here is an example with 4 inputs and 4 outputs.



The basic function of the communication networks we consider today is 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 outoing 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 path between every pair of vertices in an undirected 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 analogous directed path. For example, the route of a packet traveling from input 1 to output 3 is shown in bold.

8.1.1 Latency and Diameter

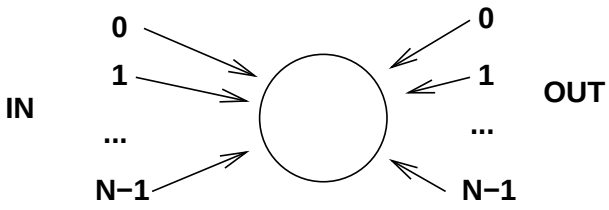
Latency is a critical issue in communication networks. This is the time required for a packet to travel from an input to an output. One measure of latency is the number of switches that a packet must pass through when traveling between the most distant input and output. For example, in the complete binary tree example, the packet traveling from input 1 to output 3 crosses 5 switches. We could measure latency some other way, such as summing wire lengths, but switches usually have the biggest impact on network speed.

The ***diameter*** of a network is the number of switches on the shortest path between the input and output that are farthest apart. Thus, diameter is an approximate measure of worst-case latency. (Notice that the diameter of a *communications network* is defined somewhat differently from the diameter of an *undirected graph*. Specifically, distance in a communication network is measured by counting switches, but distance in an undirected graph is measured by counting edges. Furthermore, in the context of a communication network we're only interested in the distance between inputs and outputs, not between arbitrary pairs of switches.) Since input 1 and output 3 are as far apart as possible in the complete binary tree, the diameter is 5.

We're going to consider several different communication networks today. 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 $2 \log N + 1$. (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 $2 \log(2^{10}) + 1 = 21$.

8.1.2 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.

8.1.3 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, the total number of switches is $2N - 1$, which is nearly the best possible with 3×3 switches.

8.1.4 Congestion

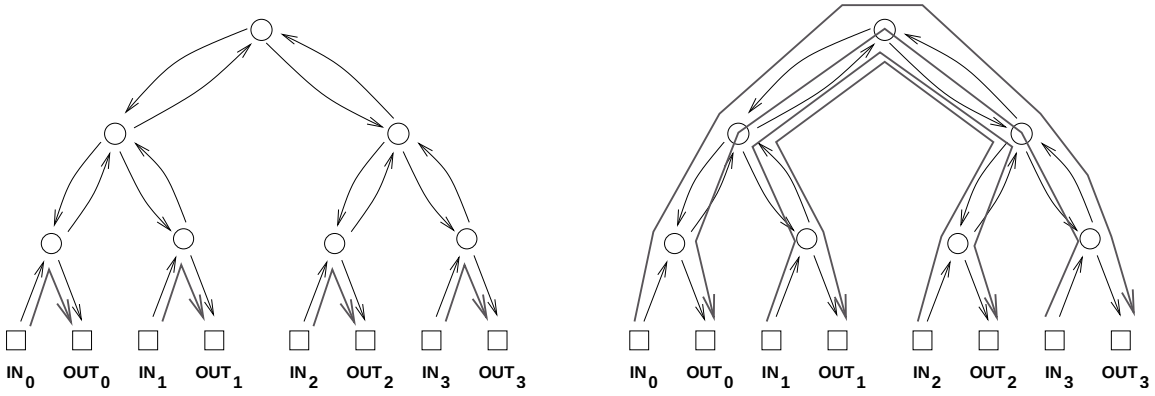
The complete binary tree has a fatal drawback: the root switch is a bottleneck. At best, this switch must handle an enormous amount of traffic: every packet traveling from the left side of the network to the right or vice-versa. Passing all these packets through a single switch could take a long time. At worst, if this switch fails, the network is broken into two equal-sized pieces. We're going to develop a single statistic called "max congestion" that quantifies bottleneck problems in communication networks. But we'll need 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, $\pi(i) = \pi(j)$ if and only if $i = j$. For example, $\pi(i) = i$ is one permutation (called the ***identity permutation***) and $\pi(i) = (N-1) - i$ is another.

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 that *maximum* over all permutations π of the *minimium* over all paths $P_{i,\pi(i)}$ of the congestion of the paths.

You may find it easier to think about max congestion in terms of a competition. Imagine that you've designed a spiffy, new communication network. Your worst enemy devises a permutation routing problem; that is, she decides which input terminal sends a packet to which output terminal. However, you are free to choose the precise path that each packet takes through your network so as to avoid overloading any one switch. Assuming that you both do your absolute best, the largest number of packets that end up passing through any switch is the max congestion of the network.

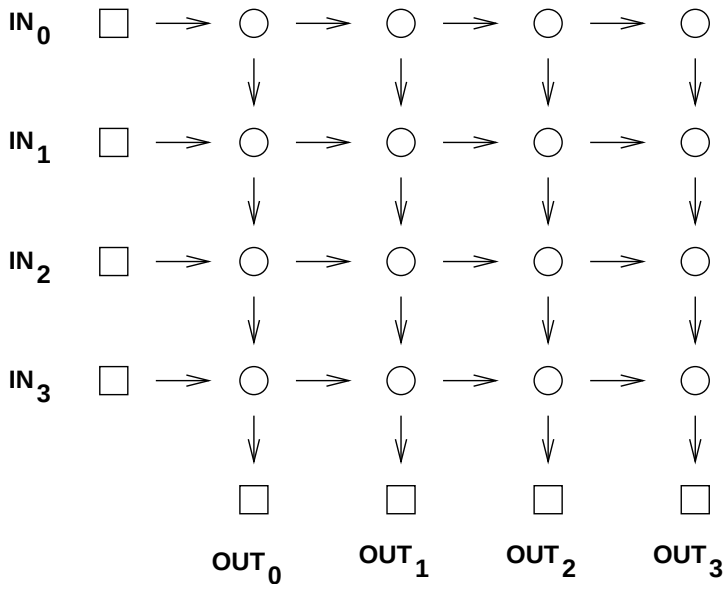
For example, if your enemy were trying to defeat the complete binary tree, she would choose a permutation like $\pi(i) = (N-1) - i$. Then for *every* packet i, you would be forced to select a path $P_{i,\pi(i)}$ passing through the root switch. Thus, the max congestion of the complete binary tree is N— which is horrible!

Let's tally the results of our analysis so far:

network	diameter	switch size	# switches	congestion
complete binary tree	$2\log N + 1$	3×3	$2N - 1$	N

8.2 2-D Array

Let's look an another communication network. This one is called a ***2-dimensional array*** or ***grid*** or ***crossbar***.



Here there are four inputs and four outputs, so $N = 4$.

The diameter in this example is 7, which is the number of switches between input 0 and output 3. More generally, the diameter of an array with N inputs and outputs is $2N - 1$, which is much worse than the diameter of $2\log N + 1$ in the complete binary tree. On the other hand, replacing a complete binary tree with an array almost eliminates congestion.

Theorem 58. *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 j and then downward to output $\pi(i)$. Thus, the switch in row i and column j transmits at most two packets: the packet originating at input i and the packet destined for column j.

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. □

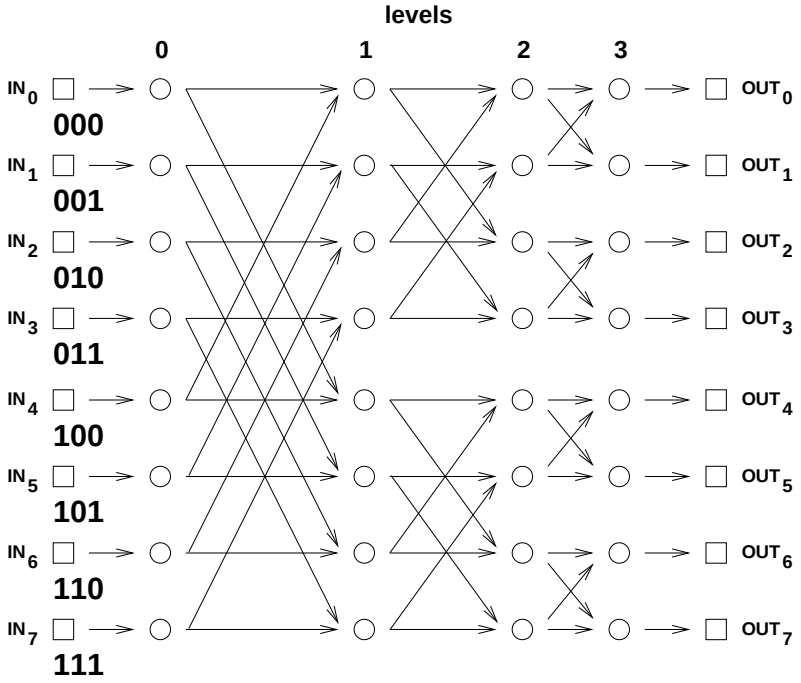
Now we can record the characteristics of the 2-D array.

network	diameter	switch size	# switches	congestion
complete binary tree	$2\log N + 1$	3×3	$2N - 1$	N
2-D array	$2N - 1$	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.

8.3 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.

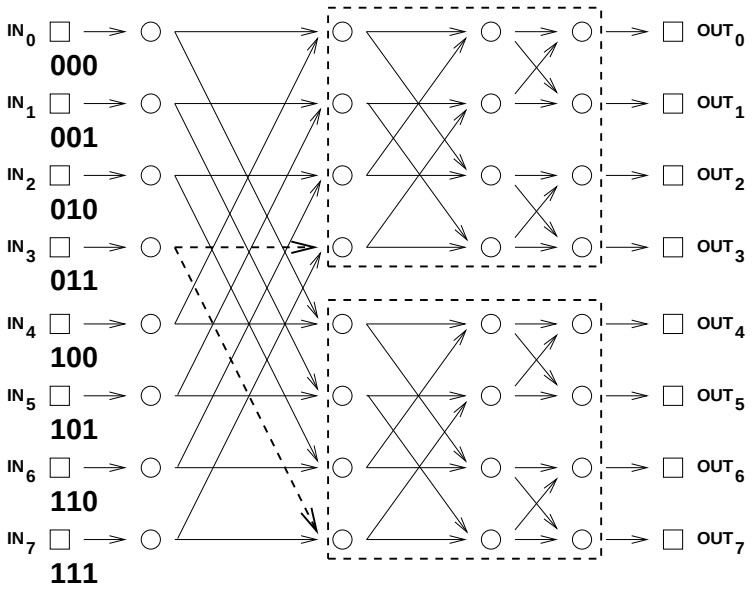
Between the inputs and the outputs, there are $\log(N) + 1$ levels of switches, numbered from 0 to $\log N$. Each level consists of a column of N switches, one per row. Thus, each switch in the network is uniquely identified by a sequence $(b_1, b_2, \ldots, b_{\log N}, l)$, where $b_1 b_2 \ldots b_{\log N}$ 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:

$$(b_1, b_2, \ldots, b_{l+1}, \ldots, b_{\log N}, l) \begin{array}{l} \nearrow \\ \searrow \end{array} \begin{array}{l} (b_1, b_2, \ldots, b_{l+1}, \ldots, b_{\log N}, l+1) \\ (b_1, b_2, \ldots, \overline{b_{l+1}}, \ldots, b_{\log N}, l+1) \end{array}$$

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 in 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.)

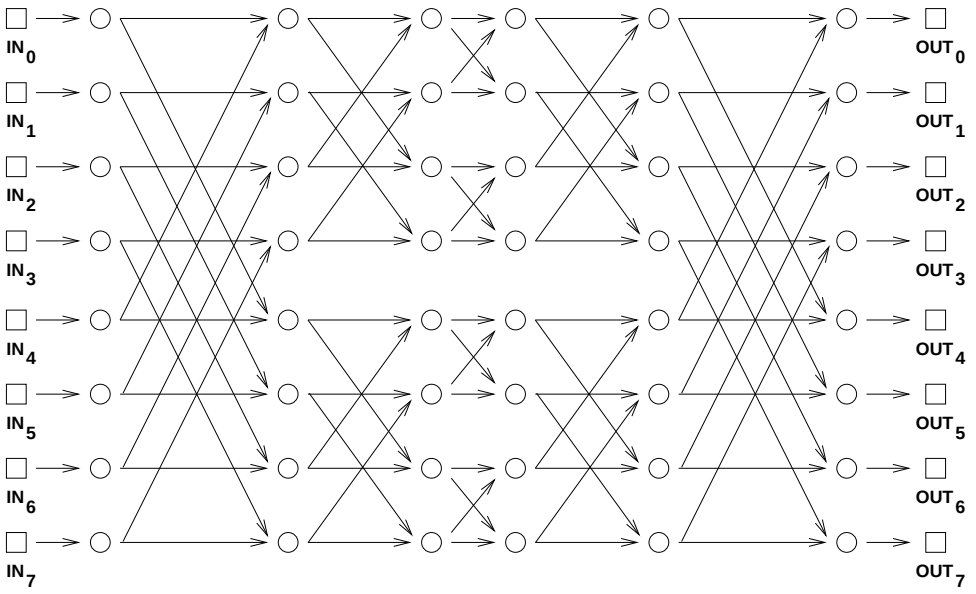
Let's add the butterfly data to our comparison table:

network	**diameter**	**switch size**	**# switches**	**congestion**
complete binary tree	$2 \log N + 1$	3×3	$2N - 1$	N
2-D array	$2N - 1$	2×2	N^2	2
butterfly	$\log N + 1$	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 compromise somewhere between the two. So our quest for the Holy Grail of routing networks goes on.

8.4 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:

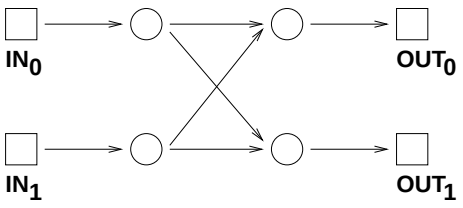
network	diameter	switch size	# switches	congestion
complete binary tree	$2\log N + 1$	3×3	$2N - 1$	N
2-D array	$2N - 1$	2×2	N^2	2
butterfly	$\log N + 1$	2×2	$N(\log(N) + 1)$	$\sqrt{N}$ or $\sqrt{N/2}$
Beneš	$2\log N$	2×2	$2N \log N$	1

The Beneš network is small, compact, and completely eliminates congestion. The Holy Grail of routing networks is in hand!

Theorem 59. *The congestion of the N-input Beneš network is 1, where $N = 2^a$ for some $a \geq 1$.*

Proof. We use induction. Let $P(a)$ be the propositon 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:

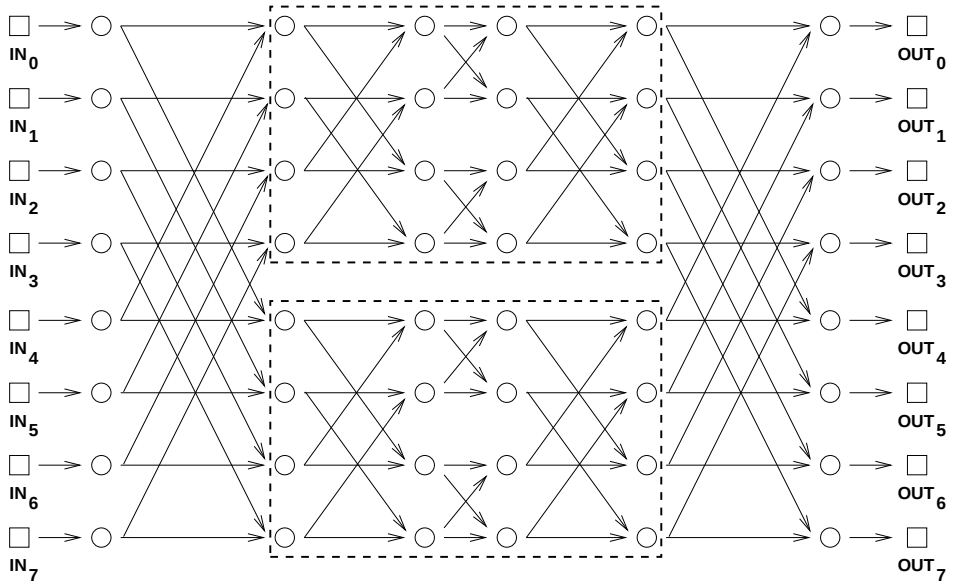


There are only two possible permutation routing problems for a 2-input network. If $\pi(0) = 0$ and $\pi(1) = 1$, then we can route both packets along the straight edges. On the

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.

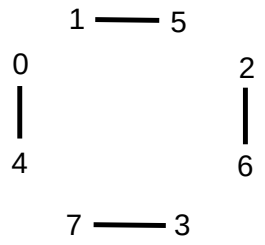


By the inductive assumption, the subnetworks can each route an arbitrary permution with congestion 1. So if we can guide packets safely through just the first and last levels, then we can rely on induction for the rest! Let's see how this works in an example. Consider the following permutation routing problem:

$$\begin{aligned} \pi(0) &= 1 & \pi(4) &= 3 \\ \pi(1) &= 5 & \pi(5) &= 6 \\ \pi(2) &= 4 & \pi(6) &= 0 \\ \pi(3) &= 7 & \pi(7) &= 2 \end{aligned}$$

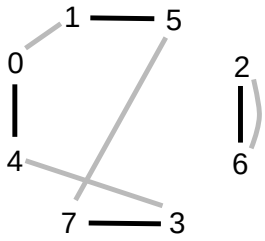
We can route each packet to its destination through either the upper subnetwork or the lower subnetwork. However, the choice for one packet may constrain the choice for another. For example, we can not route both packet 0 *and* packet 4 through the same network since that would cause two packets to collide 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:



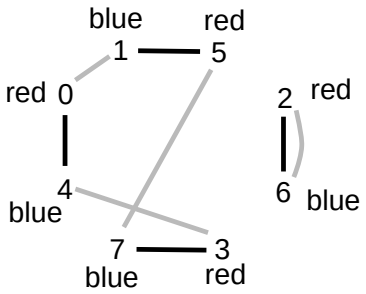
Notice that at most one new edge is incident to each vertex. The two lines drawn between vertices 2 and 6 reflect the two different reasons why these packets must be routed through different networks. However, we intend this to be a simple graph; the two lines still signify a single edge.

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. This follows from a fact proved in homework:

Theorem 60. *If the 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 $|\pi(u) - \pi(v)| = N$. By Theorem 60, 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. We can complete the routing within each subnetwork by the induction hypothesis $P(a)$. □

Chapter 9

Relations

A "relation" is a mathematical tool used to describe relationships between set elements. Relations are widely used in computer science, especially in databases and and scheduling applications.

Formally, a ***relation*** from a set A to a set B is a subset $R \subseteq A \times B$. For example, suppose that A is a set of students, and B is a set of classes. Then we might consider the relation R consisting of all pairs (a, b) such that student a is taking class b:

$$R = \{(a, b) \mid \text{student } a \text{ is taking class } b\}$$

Thus, student a is taking class b if and only if $(a, b) \in R$. There are a couple common, alternative ways of writing $(a, b) \in R$ when we're working with relations: aRb and $a \sim_R b$. The motivation for these alternative notations will become clear shortly.

9.0.1 Relations on One Set

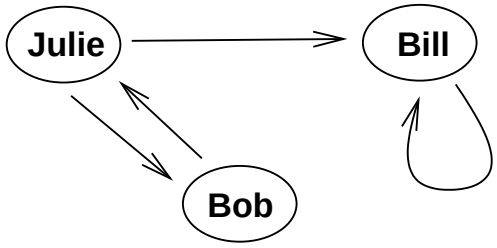
We're mainly going to focus on relationships between elements of a single set; that is, relations from a set A to a set B where $A = B$. Thus, a ***relation*** on a set A is a subset $R \subseteq A \times A$. Here are some examples:

- Let A be a set of people and the relation R describe who likes whom; that is, $(x, y) \in R$ if and only if x likes y.
- Let A be cities. Then we can define a relation R such that xRy if and only if there is a nonstop flight from city x to city y.
- Let $A = \mathbb{Z}$, and let xRy hold if and only if $x \equiv y \pmod 5$.
- Let $A = \mathbb{N}$, and let xRy hold if and only if $x \mid y$.
- Let $A = \mathbb{N}$, and let xRy hold if and only if $x \leq y$.

The last examples clarify the reason for using xRy or $x \sim_R y$ to indicate that the relation R holds between x and y: many common relations ($<$, $\leq$, $=$, $|$, $\equiv$) are expressed with the relational symbol in the middle.

9.0.2 Relations and Directed Graphs

A relation can be modeled very nicely with a directed graph. For example, the directed graph below describes the "likes" relation on a set of three people:



From this directed graph, we can conclude that:

- Julie likes Bill and Bob, but not herself.
- Bill likes only himself.
- Bob likes Julie, but not himself.

In fact, everything about the "likes" relation is conveyed by this graph, and nothing more. This is no coincidence; a set A together with a relation R is precisely the same thing as a directed graph $G = (V, E)$ with vertex set $V = A$ and edge set $E = R$.

9.1 Properties of Relations

Many relations that arise in practice possess some standard properties. A relation R on set A is:

1. ***reflexive*** if xRx for all x in A.
   (Everyone likes themself.)
2. ***symmetric*** if for all $x, y \in A$, xRy implies yRx.
   (If x likes y, then y likes x.)
3. ***antisymmetric*** if for all $x, y \in A$, xRy and yRx imply that $x = y$.
   (If x likes y and y likes x, then x and y are the same person.)

4. ***transitive*** if for all $x, y, z \in A$, xRy and yRz imply xRz.

   (If x likes y and y likes z, then x also likes z.)

Let's see which of these properties hold for some of the relations we've considered so far:

	reflexive?	symmetric?	antisymmetric?	transitive?
$x \equiv y \pmod 5$	yes	yes	no	yes
$x \mid y$	yes	no	yes	yes
$x \leq y$	yes	no	yes	yes

The two different yes/not patterns in this table are both extremely common. A relation with the first pattern of properties (like $\equiv$) is called an "equivalence relation", and a relation with the second pattern (like $\mid$ and $\leq$) is called a "partially-ordered set". The rest of this lecture focuses on just these two types of relation.

9.2 Equivalence Relations

A relation is an ***equivalence relation*** if it is reflexive, symmetric, and transitive. Congruence modulo n is a excellent example of an equivalence relation:

- It is reflexive because $x \equiv x \pmod n$.
- It is symmetric because $x \equiv y \pmod n$ implies $y \equiv x \pmod n$.
- It is transitive because $x \equiv y \pmod n$ and $y \equiv z \pmod n$ imply that $x \equiv z \pmod n$.

There is an even more well-known example of an equivalence relation: equality itself. Thus, an equivalence relation is a relation that shares some key properties with $=$.

9.2.1 Partitions

There is another way to think about equivalence relations, but we'll need a couple definitions to understand this alternative perspective.

Suppose that R is an equivalence relation on a set A. Then the ***equivalence class*** of an element $x \in A$ is the set of all elements in A related to x by R. The equivalence class of x is denoted $[x]$. Thus, in symbols:

$$[x] = \{y \mid xRy\}$$

For example, suppose that $A = \mathbb{Z}$ and xRy means that $x \equiv y \pmod 5$. Then:

$$[7] = \{\ldots, -3, 2, 7, 12, 17, 22, \ldots\}$$

Notice that 7, 12, 17, etc. all have the same equivalence class; that is, $[7] = [12] = [17] = \ldots$.

A ***partition*** of a set A is a collection of disjoint, nonempty subsets $A_1, A_2, \ldots, A_n$ whose union is all of A. For example, one possible partition of $A = \{a, b, c, d, e\}$ is:

$$A_1 = \{a, c\} \qquad A_2 = \{b, e\} \qquad A_3 = \{d\}$$

These subsets are usually called the ***blocks*** of the partition.[1]

Here's the connection between all this stuff: there is an exact correspondence between *equivalence relations on* A and *partitions of* A. We can state this as a theorem:

Theorem 61. *The equivalence classes of an equivalence relation on a set A form a partition of A.*

We won't prove this theorem (too dull even for us!), but let's look at an example. The congruent-mod-5 relation partitions the integers into five equivalance classes:

$$\{\ldots, -5, 0, 5, 10, 15, 20, \ldots\}$$
$$\{\ldots, -4, 1, 6, 11, 16, 21, \ldots\}$$
$$\{\ldots, -3, 2, 7, 12, 17, 22, \ldots\}$$
$$\{\ldots, -2, 3, 8, 13, 18, 23, \ldots\}$$
$$\{\ldots, -1, 4, 9, 14, 19, 24, \ldots\}$$

In these terms, $x \equiv y \pmod 5$ is equivalent to the assertion that x and y are both in the same block of this partition. For example, $6 \equiv 16 \pmod 5$, because they're both in the second block, but $2 \not\equiv 9 \pmod 5$ because 2 is in the third block while 9 is in the last block.

In social terms, if "likes" were an equivalence relation, then everyone would be partitioned into cliques of friends who all like each other and no one else.

9.3 Partial Orders

A relation is a ***partial order*** if it is reflexive, antisymmetric, and transitive. In terms of properties, the only difference between an equivalence relation and a partial order is that the former is symmetric and the latter is antisymmetric. But this small change makes a big difference; equivalence relations and partial orders are very different creatures.

An an example, the "divides" relation on the natural numbers is a partial order:

- It is reflexive because $x \mid x$.
- It is antisymmetric because $x \mid y$ and $y \mid x$ implies $x = y$.
- It is transitive because $x \mid y$ and $y \mid z$ implies $x \mid z$.

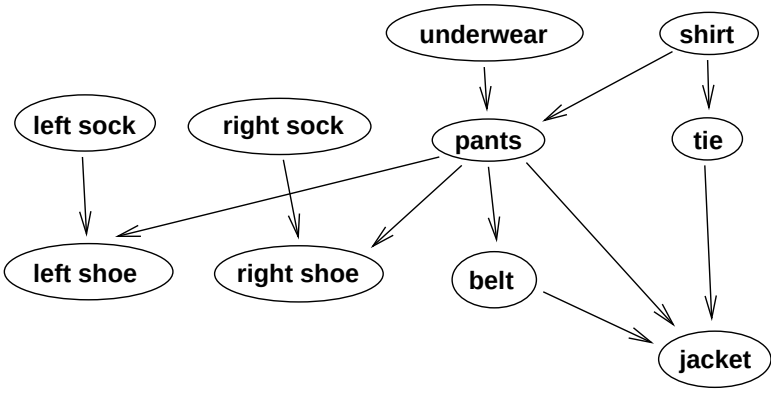
[1] I think they should be called the ***parts*** of the partition. Don't you think that makes a lot more sense?

The $\leq$ relation on the natural numbers is also a partial order. However, the $<$ relation is not a partial order, because it is not reflexive; no number is less than itelf. [2]

Often a partial order relation is denoted with the symbol $\preceq$ instead of a letter, like R. This makes sense from one perspective since the symbol calls to mind $\leq$, which is one of the most common partial orders. On the other hand, this means that $\preceq$ actually denotes the *set* of all related pairs. And a set is usually named with a letter like R instead of a cryptic squiggly symbol. (So $\preceq$ is kind of like Prince.)

Anyway, if $\preceq$ is a partial order on the set A, then the pair $(A, \preceq)$ is called a ***partially-ordered set*** or ***poset***. Mathematically, a poset *is* a directed graph with vertex set A and edge set $\preceq$. So posets can be drawn just like directed graphs.

An an example, here is a poset that describes how a guy might get dressed for a formal occasion:



In this poset, the *set* is all the garments and the *partial order* specifies which items must precede others when getting dressed. Not every edge appears in this diagram; for example, the shirt must be put on before the jacket, but there is no edge to indicate this. This edge would just clutter up the diagram without adding any new information; we already know that the shirt must precede the jacket, because the tie precedes the jacket and the shirt precedes the tie. We've also not not bothered to draw all the self-loops, even though $x \preceq x$ for all x by the definition of a partial order. Again, we know they're there, so the self-loops would just add clutter.

In general, a ***Hasse diagram*** for a poset $(A, \preceq)$ is a directed graph with vertex set A and edge set $\preceq$ minus all self-loops and edges implied by transitivity. The diagram above is *almost* a Hasse diagram, except we've left in one extra edge. Can you find it?

[2] Some sources omit the requirement that a partial order be reflexive and thus would say that $<$ *is* a partial order. The convention in this course, however, is that a relation *must* be reflexive to be a partial order.

9.3.1 Directed Acyclic Graphs

Notice that there are no ***directed cycles*** in the getting-dressed poset. In other words, there is no sequence of $n \geq 2$ distinct elements $a_1, a_2, \ldots, a_n$ such that:

$$a_1 \preceq a_2 \preceq a_3 \preceq \ldots \preceq a_{n-1} \preceq a_n \preceq a_1$$

This is a good thing; if there were such a cycle, you could never get dressed and would have to spend all day in bed reading books and eating fudgesicles. This lack of directed cycles is a property shared by all posets.

Theorem 62. *A poset has no directed cycles other than self-loops.*

Proof. We use proof by contradiction. Let $(A, \preceq)$ be a poset. Suppose that there exist $n \geq 2$ distinct elements $a_1, a_2, \ldots, a_n$ such that:

$$a_1 \preceq a_2 \preceq a_3 \preceq \ldots \preceq a_{n-1} \preceq a_n \preceq a_1$$

Since $a_1 \preceq a_2$ and $a_2 \preceq a_3$, transitivity implies $a_1 \preceq a_3$. Another application of transitivity shows that $a_1 \preceq a_4$ and a routine induction argument establishes that $a_1 \preceq a_n$. Since we also know that $a_n \preceq a_1$, antisymmetry implies $a_1 = a_n$ contradicting the supposition that $a_1, \ldots, a_n$ are distinct and $n \geq 2$. Thus, there is no such directed cycle. □

Thus, deleting the self-loops from a poset leaves a directed graph without cycles, which makes it a ***directed acyclic graph*** or ***DAG***.

9.3.2 Partial Orders and Total Orders

A partially-ordered set is "partial" because there can be two elements with no relation between them. For example, in the getting-dressed poset, there is no relation between the left sock and the right sock; you could put them on in either order. In general, elements a and b of a poset are ***incomparable*** if neither $a \preceq b$ nor $b \preceq a$. Otherwise, if $a \preceq b$ or $b \preceq a$, then a and b are ***comparable***.

A ***total order*** is a partial order in which every pair of elements is comparable. For example, the natural numbers are totally ordered by the relation $\leq$; for every pair of natural numbers a and b, either $a \leq b$ or $b \leq a$. On the other hand, the natural numbers are *not* totally ordered by the "divides" relation. For example, 3 and 5 are incomparable under this relation; 3 does not divide 5 and 5 does not divide 3. The Hasse diagram of a total order is distinctive:

A total order defines a complete ranking of elements, unlike other posets. Still, for every poset there exists some ranking of the elements that is consistent with the partial order, though that ranking might not be unique. For example, you can put your clothes on in several different orders that are consistent with the getting-dressed poset. Here are a couple:

underwear	left sock
pants	shirt
belt	tie
shirt	underwear
tie	right sock
jacket	pants
left sock	right shoe
right sock	belt
left shoe	jacket
right shoe	left shoe

A total order consistent with a partial order is called a "topological sort". More precisely, a ***topological sort*** of a poset $(A, \preceq)$ is a total order $(A, \preceq_T)$ such that:

$$x \preceq y \qquad \text{implies} \qquad x \preceq_T y$$

So the two lists above are topological sorts of the getting-dressed poset. We're going to prove that *every* finite poset has a topological sort. You can think of this as a mathematical proof that you *can* get dressed in the morning (and then show up for 6.042 lecture).

Theorem 63. *Every finite poset has a topological sort.*

We'll prove the theorem constructively. The basic idea is to pull the "smallest" element a out of the poset, find a topological sort of the remainder recursively, and then add a back into the topological sort as an element smaller than all the others.

The first hurdle is that "smallest" is not such a simple concept in a set that is only partially ordered. In a poset $(A, \preceq)$, an element $x \in A$ is ***minimal*** if there is no other element $y \in A$ such that $y \preceq x$. For example, there are *four* minimal elements in the getting-dressed poset: left sock, right sock, underwear, and shirt. (It may seem odd that the minimal elements are at the top of the Hasse diagram rather than the bottom. Some people adopt the opposite convention. If you're not sure whether minimal elements are on the top or bottom in a particular context, ask.) Similarly, an element $x \in A$ is ***maximal*** if there is no other element $y \in A$ such that $x \preceq y$.

Proving that every poset *has* a minimal element is extremely difficult, because this is actually false. For example the poset $(\mathbb{Z}, \leq)$ has no minimal element. However, there is at least one minimal element in every *finite* poset.

Lemma 64. *Every finite poset has a minimal element.*

Proof. Let $(A, \preceq)$ be an arbitrary poset. Let $a_1, a_2, \ldots, a_n$ be a maximum-length sequence of distinct elements in A such that:

$$a_1 \preceq a_2 \preceq \ldots \preceq a_n$$

The existence of such a maximum-length sequence follows from the well-ordering principle and the fact that A is finite. Now $a_0 \preceq a_1$ can not hold for any element $a_0 \in A$ not in the chain, since the chain already has maximum length. And $a_i \preceq a_1$ can not hold for any $i \geq 2$, since that would imply a cycle

$$a_i \preceq a_1 \preceq a_2 \preceq \ldots \preceq a_i$$

and no cycles exist in a poset by Theorem 62. Therefore, a_1 is a minimal element. □

Now we're ready to prove Theorem 63, which says that every finite poset has a topological sort. The proof is rather intricate; understanding the argument requires a clear grasp of all the mathematical machinery related to posets and relations!

Proof. (of Theorem 63) We use induction. Let $P(n)$ be the proposition that every n-element poset has a topological sort.

Base case. Every 1-element poset is already a total order and thus is its own topological sort. So $P(1)$ is true.

Inductive step. Now we assume $P(n)$ in order to prove $P(n+1)$ where $n \geq 1$. Let $(A, \preceq)$ be an $(n+1)$-element poset. By Lemma 64, there exists a minimal element $a \in A$. Remove a and all pairs in $\preceq$ involving a to obtain an n-element poset $(A', \preceq')$. This has a topological sort $(A', \preceq'_T)$ by the assumption $P(n)$. Now we construct a total order $(A, \preceq_T)$ by adding back a as an element smaller than all the others. Formally, let:

$$\preceq_T \quad = \quad \preceq'_T \cup \{(a, z) \mid z \in A\}$$

All that remains is the check that this total order is consistent with the original partial order $(A, \preceq)$; that is, we must show that:

$$x \preceq y \qquad \text{implies} \qquad x \preceq_T y$$

We assume that the left side is true and show that the right side follows. There are two cases:

Case 1: If $x = a$, then $a \preceq_T y$ holds, because $a \preceq_T z$ for all $z \in A$.

Case 2: If $x \neq a$, then y can not equal a either, since a is a minimal element in the partial order $\preceq$. Thus, both x and y are in A' and so $x \preceq' y$. This means $x \preceq'_T y$, since $\preceq'_T$ is a topological sort of the partial order $\preceq'$. And this implies $x \preceq_T y$, since $\preceq_T$ contains $\preceq'_T$.

Thus, $(A, \preceq_T)$ is a topological sort of $(A, \preceq)$. This shows that $P(n)$ implies $P(n+1)$ for all $n \geq 1$. Therefore, $P(n)$ is true for all $n \geq 1$ by the principle of induction, which prove the theorem. □

Chapter 10

Sums, Approximations, and Asymptotics

10.1 The Value of an Annuity

If you won the lottery, would you prefer a million dollars today or $50,000 a year for the rest of your life? 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 actually worth. For many reasons, $50,000 a year for 20 years is worth much less than a million dollars right now. For example, consider the last $50,000 installment. If you had that $50,000 right now, then you could start earning interest, invest the money in the stock market, or just buy something fun. However, if you don't get the $50,000 for another 20 years, then someone else is earning all the interest or investment profit. Furthermore, prices are likely to gradually rise over the next 20 years, so you when you finally get the money, you won't be able to buy as much. Furthermore, people only live so long; if you were 90 years old, a payout 20 years in the future would be worth next to nothing!

But what if your choice were between $50,000 a year for 20 years and a *half* million dollars today?

10.1.1 The Future Value of Money

In order to address such questions, we have to make an assumption about the future value of money. Let's put most of the complications aside and think about this from a simple perspective. Suppose you invested $1 today at an annual interest rate of $p\%$. Then $1 today would become $1 + p$ dollars in a year, $(1 + p)^2$ dollars in two years and so forth. A reasonable estimate for p these days is about 6%.

Looked at another way, a dollar paid out a year from now is worth $1/(1+p)$ dollars today, a dollar paid in two years is worth only $1/(1+p)^2$ today, etc. Now we can work out the value of an annuity that pays m dollars at the start of each year for the next n years:

payments	current value
$\$m$ today	m
$\$m$ in 1 year	$\frac{m}{1+p}$
$\$m$ in 2 years	$\frac{m}{(1+p)^2}$
$\ldots$	$\ldots$
$\$m$ in $n-1$ years	$\frac{m}{(1+p)^{n-1}}$
Total current value:	$V = \sum_{k=1}^{n} \frac{m}{(1+p)^{k-1}}$

So to compute the value of the annuity, we need only evaluate this sum. We *could* plug in values for m, n, and p, compute each term explicitly, and then add them all up. However, this particular sum has an equivalent "closed form" that makes the job easier. In general, a ***closed form*** is a mathematical expression that can be evaluated with a fixed number of basic operations (addition, multiplication, exponentiation, etc.) In contrast, evaluating the sum above requires a number of operations proportional to n.

10.1.2 A Geometric Sum

Our goal is to find a closed form equivalent to:

$$V = \sum_{k=1}^{n} \frac{m}{(1+p)^{k-1}}$$

Let's begin by rewriting the sum:

$$V = \sum_{j=0}^{n-1} \frac{m}{(1+p)^j} \qquad \text{(substitute } j = k-1\text{)}$$

$$= m \sum_{j=0}^{n-1} x^j \qquad \text{(where } x = \frac{1}{1+p}\text{)}$$

The goal of these substitutions is to put the summation into a special form so that we can bash it with a general theorem. In particular, the terms of the sum

$$\sum_{j=0}^{n-1} x^j = 1 + x + x^2 + x^3 + \ldots + x^{n-1}$$

form a ***geometric series***, which means that each term is a constant times the preceding term. (In this case, the constant is x.) And we've already encountered a theorem about geometric sums:

Theorem 65. *For all $n \geq 1$ and all $z \neq 1$:*

$$\sum_{i=0}^{n} z^j = \frac{1 - z^{n+1}}{1 - z}$$

This theorem can be proved by induction, but that proof gives no hint how the formula might be found in the first place. Here is a more insightful derivation based on the ***perturbation method***. First, we let S equal the value of the sum and then "perturb" it by multiplying by z.

$$\begin{array}{rcccccccccc} S & = & 1 & + & z & + & z^2 & + & \ldots & + & z^n \\ zS & = & & & z & + & z^2 & + & \ldots & + & z^n & + & z^{n+1} \end{array}$$

The difference between the original sum and the perturbed sum is not so great, because there is massive cancellation on the right side:

$$S - zS = 1 - z^{n+1}$$

Now solving for S gives the expression in Theorem 65:

$$S = \frac{1 - z^{n+1}}{1 - z}$$

You can derive a passable number of summation formulas by mimicking the approach used above. We'll look at some other methods for evaluting sums shortly.

10.1.3 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:

$$\begin{aligned} V &= m \sum_{j=0}^{n-1} x^j \qquad \left(\text{where } x = \frac{1}{1+p}\right) \\ &= m \cdot \frac{1 - x^n}{1 - x} \\ &= m \cdot \frac{1 - \left(\frac{1}{1+p}\right)^n}{1 - \left(\frac{1}{1+p}\right)} \end{aligned}$$

We apply Theorem 65 on the second line, and undo the the earlier substitution $x = 1/(1+p)$ on the last line.

The last expression is a closed form; it can be evaluated with a fixed number of basic operations. 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 $p = 0.6$ gives $V \approx \$607,906$. The deferred payments are worth more than a half million dollars today, but not by much!

10.1.4 Infinite Sums

Would you prefer a million dollars today or $50,000 a year *forever*? This might seem like an easy choice— when infinite money is on offer, why worry about inflation?

This is a question about an ***infinite sum***. In general, the value of an infinite sum is defined as the limit of a finite sum as the number of terms goes to infinity:

$$\sum_{k=0}^{\infty} z_k \qquad \text{means} \qquad \lim_{n\to\infty} \sum_{k=0}^{n} z_k$$

So the value of an annuity with an infinite number of payments is given by our previous answer in the limit as n goes to infinity:

$$\begin{aligned} V &= \lim_{n\to\infty} m \cdot \frac{1 - \left(\frac{1}{1+p}\right)^n}{1 - \left(\frac{1}{1+p}\right)} \\ &= m \cdot \frac{1}{1 - \left(\frac{1}{1+p}\right)} \\ &= m \cdot \frac{1+p}{p} \end{aligned}$$

In the second step, notice that the $1/(1+p)^n$ term in the numerator goes to zero in the limit. The third equation follows by simplifying.

Plugging in $m = \$50,000$ and $p = 0.6$ into this formula gives $V \approx \$883,333$. This means that getting $\$50,000$ per year *forever* is still not as good as a million dollars today! Then again, if you had a million dollars today in the bank earning $p = 6\%$ interest, you could take out and spend $\$60,000$ a year forever. So this answer makes some sense.

More generally, we can get a closed form for infinite geometric sums from Theorem 65 by taking a limit.

Corollary 66. *If $|z| < 1$, then:*

$$\sum_{i=0}^{\infty} z^i = \frac{1}{1-z}$$

Proof.

$$\begin{aligned}\sum_{i=0}^{\infty} z^i &= \lim_{n\to\infty} \sum_{i=0}^{n} z^i \\ &= \lim_{n\to\infty} \frac{1-z^{n+1}}{1-z} \\ &= \frac{1}{1-z}\end{aligned}$$

The first equation uses the definition of an infinite limit, and the second uses Theorem 65. In the limit, the term z^{n+1} in the numerator vanishes since $|z| < 1$. □

We now have closed forms for both finite and infinite geometric series. Some examples are given below. In each case, the solution follows immediately from either Theorem 65 (for finite series) or Corollary 66 (for infinite series).

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots = \sum_{i=0}^{\infty} (1/2)^i = \frac{1}{1-(1/2)} = 2$$

$$1 - \frac{1}{2} + \frac{1}{4} - \frac{1}{8} + \ldots = \sum_{i=0}^{\infty} (-1/2)^i = \frac{1}{1-(-1/2)} = \frac{2}{3}$$

$$1 + 2 + 4 + 8 + \ldots + 2^{n-1} = \sum_{i=0}^{n-1} 2^i = \frac{1-2^n}{1-2} = 2^n - 1$$

$$1 + 3 + 9 + 27 + \ldots + 3^{n-1} = \sum_{i=0}^{n-1} 3^i = \frac{1-3^n}{1-3} = \frac{3^n-1}{2}$$

Here is a good rule of thumb: *the sum of a geometric series is approximately equal to the term with greatest absolute value*. In the first two examples, the largest term is equal to 1 and the sums are 2 and 2/3, which are both relatively close to 1. In the third example, the sum is about twice the largest term. In the final example, the largest term is 3^{n-1} and the sum is $(3^n - 1)/2$, which is 1.5 times greater.

10.2 Variants of Geometric Sums

You now know everything about geometric series. But one often encounters sums that cannot be transformed by simple variable substitutions to the form $\sum z^i$. A useful way to obtain new summation formulas from old is by *differentiating* or *integrating* with respect to z.

For example, consider the following series:

$$\sum_{i=1}^{n} iz^i = z + 2z^2 + 3z^3 + \ldots + nz^n$$

This is not a geometric series, since the ratio between successive terms is not constant. So our formula for the sum of a geometric series cannot be directly applied. But suppose that we differentiate both sides of that equation:

$$\begin{aligned}
\frac{d}{dz}\sum_{i=0}^{n} z^i &= \frac{d}{dz}\frac{1-z^{n+1}}{1-z} \\
\sum_{i=0}^{n} iz^{i-1} &= \frac{-(n+1)z^n(1-z)-(-1)(1-z^{n+1})}{(1-z)^2} \\
&= \frac{-(n+1)z^n+(n+1)z^{n+1}+1-z^{n+1}}{(1-z)^2} \\
&= \frac{1-(n+1)z^n+nz^{n+1}}{(1-z)^2}
\end{aligned}$$

Often differentiating or integrating messes up the exponent of z in every term. In this case, we now have a formula for a sum of the form $\sum iz^{i-1}$, but we want a formula for $\sum iz^i$. The solution is to multiply both sidez by z. Let's bundle up the result as a theorem:

Theorem 67. *For all $n \geq 0$ and all $z \neq 1$:*

$$\sum_{i=0}^{n} iz^i = \frac{z-(n+1)z^{n+1}+nz^{n+2}}{(1-z)^2}$$

If $|z| < 1$, then the sum converges to a finite value even if there are infinitely many terms. Taking the limit as n tends infinity gives:

Corollary 68. *If $|z| < 1$, then:*

$$\sum_{i=0}^{\infty} iz^i = \frac{z}{(1-z)^2}$$

As a consequence, suppose you're offered \$50,000 at the end of this year, \$100,000 at the end of next year, \$150,000 after the following year, and so on. How much is this worth? Surely *this* is infinite money! Let's work out the answer:

$$\begin{aligned}
V &= \sum_{i=1}^{n} \frac{m}{(1+p)^i} \\
&= m \cdot \frac{\frac{1}{1+p}}{\left(1-\frac{1}{1+p}\right)^2} \\
&= m \cdot \frac{1+p}{p^2}
\end{aligned}$$

The second line follows from Corollary 68. The last step is simplification.

Setting $m = \$50,000$ and $p = 0.06$ gives value of the annuity: \$14,722,222. Intuitively, even though payments increase every year, they increase only *linearly* with time. In contrast, dollars paid out in the future decrease in value *exponentially* with time. As a result, payments in the distant future are almost worthless, so the value of the annuity is still finite!

Integrating a geometric sum gives yet another summation formula. Let's start with the formula for an infinite geometric sum:

$$\sum_{i=0}^{\infty} z^i = \frac{1}{1-z} \qquad (|z| < 1)$$

Now we integrate both sides from 0 to x:

$$\int_0^x \sum_{i=0}^{\infty} z^i \, dz = \int_0^x \frac{1}{1-z} \, dz$$

$$\sum_{i=0}^{\infty} \frac{z^{i+1}}{i+1} \Bigg|_0^x = -\ln(1-z) \Big|_0^x$$

$$\sum_{i=0}^{\infty} \frac{x^{i+1}}{i+1} = -\ln(1-x)$$

Reindexing on the left side with the substitution $j = i + 1$ gives the summation formula:

$$\sum_{j=1}^{\infty} \frac{x^j}{j} = -\ln(1-x)$$

You might have seen this before: this is the the Taylor expansion for $-\ln(1-x)$.

10.3 Sums of Powers

Long ago, in the before-time, we verified the formula:

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

The source of this formula is still a mystery! Sure, we can prove this statement by induction, but where did the expression on the right come from in the first place? Even more inexplicable is the summation formula for consecutive squares:

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

Here is one way we might discover such formulas. Remember that sums are the discrete cousins of integrals. So we might guess that the sum of a degree-k polynomial is a degree-$(k+1)$ polynomial, just as if we were integrating. If this guess is correct, then

$$\sum_{i=1}^{n} i^2 = an^3 + bn^2 + cn + d$$

for some constants a, b, c, and d. All that remains is to determine these constants. We can do this by plugging a few values for n into the equation above. Each value gives a linear equation in a, b, c, and d. For example:

$$\begin{aligned} n = 0 \quad &\Rightarrow \quad 0 = d \\ n = 1 \quad &\Rightarrow \quad 1 = a + b + c + d \\ n = 2 \quad &\Rightarrow \quad 5 = 8a + 4b + 2c + d \\ n = 3 \quad &\Rightarrow \quad 14 = 27a + 9b + 3c + d \end{aligned}$$

We now have four equations in four unknowns. Solving this system gives $a = 1/3$, $b = 1/2$, $c = 1/6$, and $d = 0$ and so it is tempting to conclude that:

$$\sum_{i=1}^{n} i^2 = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6} \tag{2}$$

Be careful! This equation is valid only if we were correct in our initial guess at the form of the solution. If we were wrong, then the equation above might not hold all n! The only way to be sure is to verify this formula with an induction proof. In fact, we did guess correctly; the expressions on the right sides of equations (1) and (2) are equal.

10.4 Approximating Sums

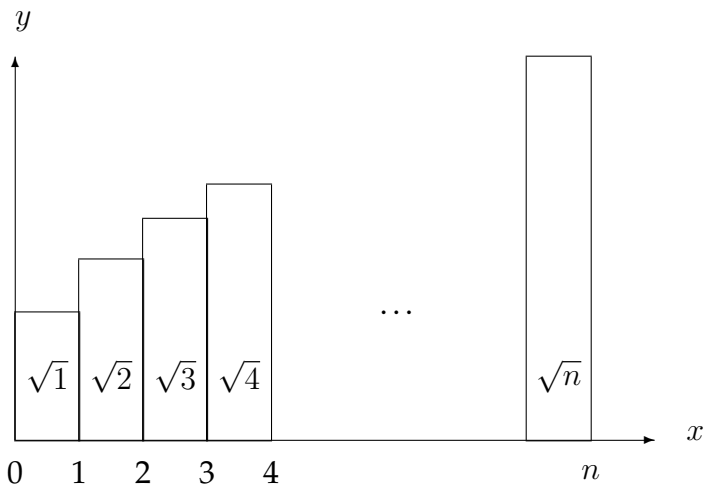
Unfortunately, there is no equivalent closed-form for many of the sums that arise in practice. Here is an example that we'll gnaw on for quite a while:

$$\sum_{i=1}^{n} \sqrt{i} \tag{*}$$

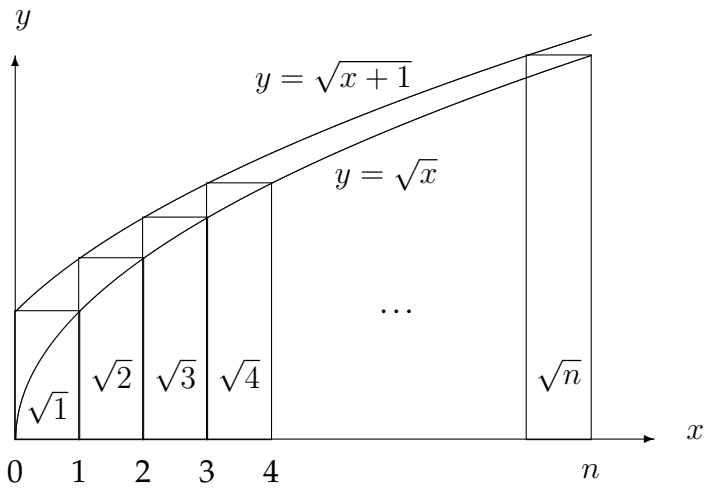
There is no closed form exactly equal to this sum, which we'll refer to as (*) hereafter. However, there do exist good closed-form upper and lower bounds. And for many practical problems, bounds and approximations are good enough.

10.4.1 Integration Bounds

One way to cope with sums is pass over to the continuous world and apply integration techniques. Suppose that we represent the sum (*) with a "bar graph" on the coordinate plane:



The i-th bar from the left has height $\sqrt{i}$ and width 1. So the area of the i-th bar is equal to the value of the i-th term in the sum. Thus, *the area of the bar graph is equal the value of the sum* (*). Now the graph of the function $y = \sqrt{x+1}$ runs just above the bar graph, and the graph of $y = \sqrt{x}$ runs just below:



So the areas beneath these curves are upper and lower bounds on the area of the bar graph and thus are upper and lower bounds on sum (*). In symbols:

$$\int_0^n \sqrt{x}\,dx \quad \leq \quad \sum_{i=1}^{n} \sqrt{i} \quad \leq \quad \int_0^n \sqrt{x+1}\,dx$$

Now we can obtain closed-form bounds by integration:

$$\frac{x^{3/2}}{3/2}\Bigg|_0^n \quad \leq \quad \sum_{i=1}^{n} \sqrt{i} \quad \leq \quad \frac{(x+1)^{3/2}}{3/2}\Bigg|_0^n$$

$$\frac{2}{3}n^{3/2} \quad \leq \quad \sum_{i=1}^{n} \sqrt{i} \quad \leq \quad \frac{2}{3}\big((n+1)^{3/2}-1\big)$$

These are pretty good bounds. For example, if $n = 100$, then we get:

$$666.7 \quad \leq \quad \sqrt{1}+\sqrt{2}+\ldots+\sqrt{100} \quad \leq \quad 676.1$$

In order to determine exactly *how good* these bounds are in general, we'll need a better understanding of the rather complicated upper bound, $\frac{2}{3}((n+1)^{3/2}-1)$.

10.4.2 Taylor's Theorem

In practice, a simple, approximate answer is often better than an exact, complicated answer. A great way to get such approximations is to dredge Taylor's Theorem up from the dark, murky regions of single-variable calculus.

Theorem 69 (Taylor's Theorem). *If f is a real-valued function, then*

$$f(x) = \frac{f(0)}{0!} + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \ldots + \frac{f^{(n-1)}(0)}{(n-1)!}x^{n-1} + \underbrace{\frac{f^{(n)}(z)}{n!}x^n}_{\text{"error term"}}$$

for some $z \in [0,x]$, provided f is continuous on $[0,x]$ and n derivatives exist on $(0,x)$.

The idea is to approximate the function f with a short Taylor series. We can then analyze the approximation by studying the "error term". Let's see how this plays out for the function that came up in the previous section:

$$\frac{2}{3}\big((n+1)^{3/2}-1\big)$$

The crux of this expression is $(n+1)^{3/2}$. However, Taylor's Theorem usually approximates a function $f(x)$ best for x close to zero. But n could be a big number, so we're in some trouble. But we can rewrite the expression into a more tractable form:

$$\frac{2}{3}\left(n^{3/2}\cdot\left(1+\frac{1}{n}\right)^{3/2}-1\right)$$

Now the nasty bit is $(1+1/n)^{3/2}$. If we let $x = 1/n$, then our job is to approximate $(1+x)^{3/2}$ where x *is close to zero*. This is a proper target for Taylor's Theorem. First, we compute a few derivatives and their values at 0.

$$\begin{aligned} f(x) &= (1+x)^{3/2} & f(0) &= 1 \\ f'(x) &= \frac{3}{2}\cdot(1+x)^{1/2} & f'(0) &= \frac{3}{2} \\ f''(x) &= \frac{3}{2}\cdot\frac{1}{2}\cdot(1+x)^{-1/2} & f''(0) &= \frac{3}{4} \end{aligned}$$

Now Taylor's Theorem says

$$\begin{aligned} f(x) &= \frac{f(0)}{0!} + \frac{f'(0)}{1!}x + \frac{f''(z)}{2!}x^2 \\ &= 1 + \frac{3}{2}x + \frac{3}{8}(1+z)^{-1/2}x^2 \end{aligned}$$

for some $z \in [0, x]$. This expression is maximized when $z = 0$, so we get the upper bound:

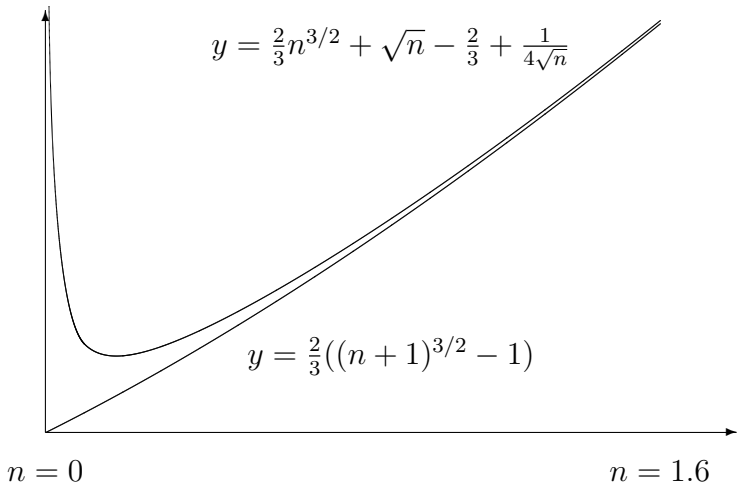
$$(1+x)^{3/2} \le 1 + \frac{3}{2}x + \frac{3}{8}x^2$$

Now we can upper bound our original expression.

$$\begin{aligned} \frac{2}{3}\big((n+1)^{3/2} - 1\big) &= \frac{2}{3}\left(n^{3/2}\cdot\left(1+\frac{1}{n}\right)^{3/2} - 1\right) \\ &\le \frac{2}{3}\left(n^{3/2}\cdot\left(1+\frac{3}{2}\cdot\frac{1}{n}+\frac{3}{8}\cdot\frac{1}{n^2}\right) - 1\right) \\ &= \frac{2}{3}n^{3/2} + \sqrt{n} - \frac{2}{3} + \frac{1}{4\sqrt{n}} \end{aligned}$$

There are a lot of terms here, so this might not seem like much of an improvement! However, each term has a simple form. And simpler terms are easier to cancel and combine in subsequent calculations.

Applying Taylor's Theorem involved a lot of grungy math, but you can see the result in a picture. This is a plot of the original function compared to the upper bound we just derived:



The picture shows that our upper bound is *extremely* close to the original function, except for very small values of n.

10.4.3 Back to the Sum

Before we got to yammering on about Taylor's Theorem, we proved bounds on the sum (*):

$$\frac{2}{3}n^{3/2} \quad \leq \quad \sum_{i=1}^{n} \sqrt{i} \quad \leq \quad \frac{2}{3}\big((n+1)^{3/2} - 1\big)$$

Let's rewrite the right side using the upper bound we just derived:

$$\frac{2}{3}n^{3/2} \quad \leq \quad \sum_{i=1}^{n} \sqrt{i} \quad \leq \quad \frac{2}{3}n^{3/2} + \sqrt{n} - \frac{2}{3} + \frac{1}{4\sqrt{n}}$$

Now the dominant term in both the lower and upper bounds is $\frac{2}{3}n^{2/3}$, so we can compare the two more readily. As you can see, the gap between the bounds is at most $\sqrt{n}$, which is small relative to the sum as a whole. In fact, the gap between upper and lower bounds that we actually observed for $n = 100$ was very nearly $\sqrt{100} = 10$.

$$666.7 \quad \leq \quad \sqrt{1} + \sqrt{2} + \ldots + \sqrt{100} \quad \leq \quad 676.1$$

There is an elaborate system of notation for describing the quality of approximations. One simple convention is:

$$f(n) \sim g(n) \qquad \text{means} \qquad \lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$$

In terms of this notation, we've just shown that:

$$\sum_{i=1}^{n} \sqrt{i} \quad \sim \quad \frac{2}{3}n^{3/2}$$

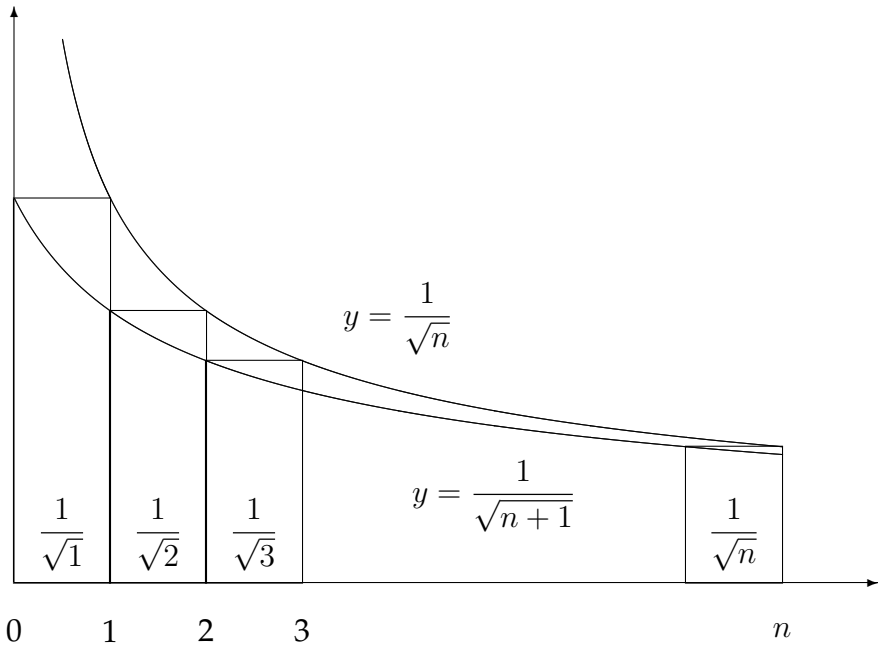
Thus, with a little loss of accuracy, we can replace this vicious sum with a simple closed form.

10.4.4 Another Integration Example

Let's work through one more example of using integrals to bound a nasty sum. This time our target is:

$$\sum_{i=1}^{n} \frac{1}{\sqrt{n}}$$

As before, we construct a bar graph whose area is equal to the value of the sum. Then we find curves that lie just above and below the bars:



In this way, we get the bounds:

$$\int_0^n \frac{1}{\sqrt{x+1}}\,dx \quad \leq \quad \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \quad \leq \quad \int_0^n \frac{1}{\sqrt{x}}\,dx$$

Now the picture shows that most of the error arises because of the large gap between the two curves for small values of n. In particular, the upper curve goes off to infinity as n approaches zero!

We can eliminate some of this error by summing a couple terms explicitly, and bounding the remaining terms with integrals. For example, summing two terms explicitly gives:

$$\frac{1}{\sqrt{1}} + \frac{1}{\sqrt{2}} + \int_2^n \frac{1}{\sqrt{x+1}}\,dx \quad \leq \quad \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \quad \leq \quad \frac{1}{\sqrt{1}} + \frac{1}{\sqrt{2}} + \int_2^n \frac{1}{\sqrt{x}}\,dx$$

Evaluating the integrals and simplifying, we find:

$$\frac{2+\sqrt{2}-4\sqrt{3}}{2}+2\sqrt{n+1} \quad \leq \quad \sum_{i=1}^{n}\frac{1}{\sqrt{n}} \quad \leq \quad \frac{2-3\sqrt{2}}{2}+2\sqrt{n}$$

Computing the values of the constants gives:

$$2\sqrt{n+1}-1.75 \quad \leq \quad \sum_{i=1}^{n}\frac{1}{\sqrt{n}} \quad \leq \quad 2\sqrt{n}-1.12$$

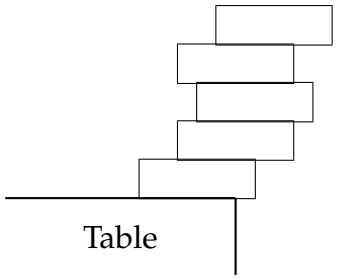
These bounds are quite tight; they always differ by less than 1. So once again, we've gotten a good handle on a sum that has no exact closed-form equivalent.

Chapter 11

Sums, Approximations, and Asymptotics II

11.1 Block Stacking

How far can a stack of identical blocks overhang the end of a table without toppling over? Can a block be suspended *beyond* the table's edge?

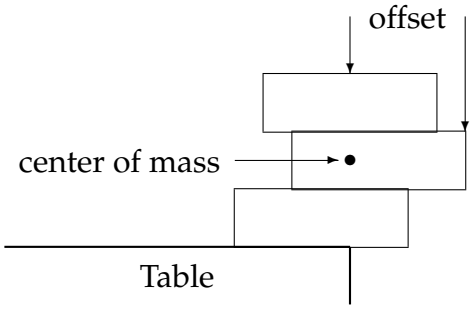


The stack falls off the desk if its center of mass lies beyond the desk's edge. Moreover, the center of mass of the top k blocks must lie above the $(k+1)$-st block; otherwise, the top k fall off. In order to find the best configuration of blocks, we'll need a fact from physics about centers of mass.

Fact 1. *If two objects have masses m_1 and m_2 and centers-of-mass at positions x_1 and x_2, then the center of mass of the two objects together is at position:*

$$\frac{z_1 m_1 + z_2 m_2}{m_1 + m_2}$$

For this problem, only the horizonal dimension is relevant, and we'll use the width of a block as our measure of distance. Define the *offset* of a particular configuration of blocks to be the horizonal distance from its center of mass to its rightmost edge. The offset measures how far the configuration can extend beyond the desk since at best the center of mass lies right at the desk's edge.



We can find the greatest possible offset of a stack of n blocks with an inductive argument. This is an instance where induction not only serves as a proof technique, but also turns out to be a great tool for reasoning about the problem.

Theorem 70. *The greatest possible offset of a stack of $n \geq 1$ blocks is:*

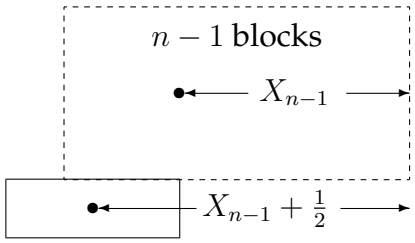
$$X_n = \frac{1}{2} + \frac{1}{4} + \frac{1}{6} + \ldots + \frac{1}{2n}$$

Proof. [1] We use induction on n, the number of blocks. Let $P(n)$ be the proposition that the greatest possible offset of a stack of $n \geq 1$ blocks is $1/2 + 1/4 + \ldots + 1/(2n)$.

Base case: For a single block, the center of mass is distance $X_1 = 1/2$ from its rightmost edge. So $P(1)$ is true.

Inductive step: For $n \geq 2$, assume that $P(n-1)$ is true in order to prove $P(n)$. A stack of n blocks consists of the bottom block together with a stack of $n-1$ blocks on top.

In order to acheive the greatest possible offset with n blocks, the top $n-1$ blocks should have the greatest possible offset, which is X_{n-1}; otherwise, we could do better by replacing the top $n-1$ blocks with a different configuration that has greater offset. Furthermore, the center of mass of the top $n-1$ blocks should lie directly above the right edge of the bottom block; otherwise, we could do better by sliding the top $n-1$ blocks farther to the right.



[1] A different analysis was presented in lecture.

Thus, by the physics fact, the maximum possible offset of a stack of n blocks is:

$$\begin{aligned} X_n &= \frac{X_{n-1} \cdot (n-1) + (X_{n-1} + \frac{1}{2}) \cdot 1}{n} \\ &= X_{n-1} + \frac{1}{2n} \\ &= \frac{1}{2} + \frac{1}{4} + \frac{1}{6} + \ldots + \frac{1}{2n} \end{aligned}$$

We use the assumption $P(n-1)$ in the last step. This proves $P(n)$.

The theorem follows by the principle of induction. □

11.1.1 Harmonic Numbers

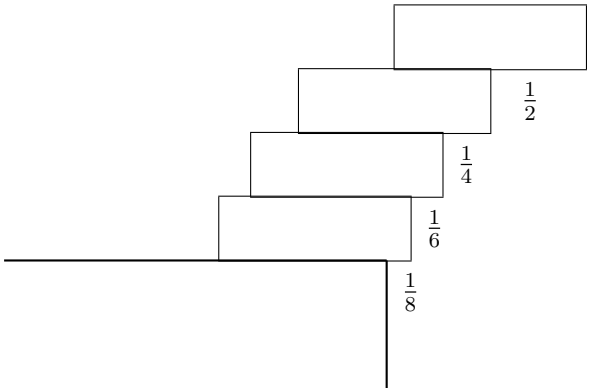
Sums similar to the one in Theorem 70 come up all the time in computer science. In particular,

$$\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}$$

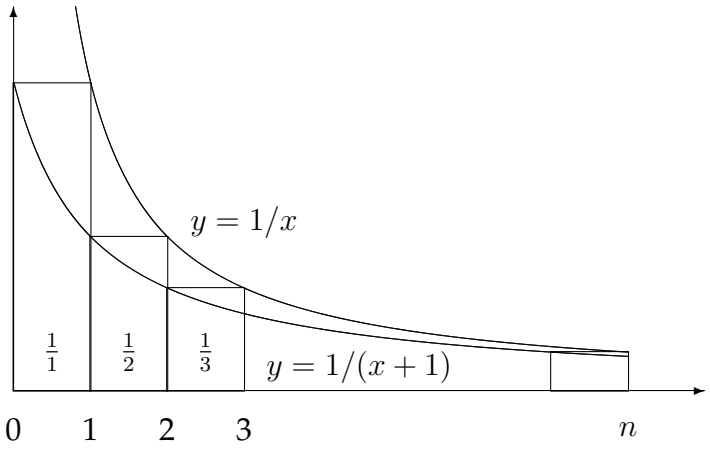
is called a ***harmonic sum***. Its value is called the **n-*th harmonic number*** and is denoted H_n. In these terms, the greatest possible offset of a stack of n blocks is $\frac{1}{2}H_n$. We can tabulate the greatest overhang achievable with $n = 1$, 2, 3 and 4 blocks by computing harmonic numbers:

# of blocks	maximum overhang
1	$\frac{1}{2}H_1 = \frac{1}{2}(\frac{1}{1}) = \frac{1}{2}$
2	$\frac{1}{2}H_2 = \frac{1}{2}(\frac{1}{1} + \frac{1}{2}) = \frac{3}{4}$
3	$\frac{1}{2}H_3 = \frac{1}{2}(\frac{1}{1} + \frac{1}{2} + \frac{1}{3}) = \frac{11}{12}$
4	$\frac{1}{2}H_4 = \frac{1}{2}(\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}) = \frac{25}{24} > 1$

The last line reveals that we can suspend the *fourth* block beyond the edge of the table! Here's the configuration that does the trick:



We'll have to study harmonic sums more closely to determine what can be done with large numbers of blocks. Let's use integration to bound the n-th harmonic number. A picture is extremely helpful for getting the right functions and limits of integration.



We can not integrate the function $1/x$ starting from zero. So, for the upper bound on H_n we'll take the first term explicitly $(1/1)$ and then upper bound the remaining terms normally. This gives:

$$\begin{aligned}
\int_0^n \frac{1}{x+1}\,dx &\leq H_n \leq 1 + \int_1^n \frac{1}{x}\,dx \\
\ln(x+1)\Big|_0^n &\leq H_n \leq 1 + \left(\ln x\Big|_1^n\right) \\
\ln(n+1) &\leq H_n \leq 1 + \ln n
\end{aligned}$$

There are good bounds; the difference between the upper and lower values is never more than 1.

Suppose we had a *million* blocks. Then the overhang we could achieve— assuming no breeze or deformation of the blocks— would be $\frac{1}{2}H_{1000000}$. According to our bounds, this is:

$$\text{at least} \qquad \frac{1}{2}\ln(1000001) = 6.907\ldots$$

$$\text{at most} \qquad \frac{1}{2}(1+\ln(1000000)) = 7.407\ldots$$

So the top block would extend about 7 lengths past the end of the table! In fact, since the lower bound or $\frac{1}{2}\ln(n+1)$ grows arbitrarily large, there is *no limit* on how far the stack can overhang!

Mathematicians have worked out some extremely precise approximations for the n-th harmonic number. For example:

$$H_n = \ln(n) + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + \frac{\epsilon(n)}{120n^4}$$

where $\gamma = 0.577215664\ldots$ is ***Euler's constant*** and $\epsilon(n)$ is between 0 and 1. Interestingly, no knows whether Euler's constant is rational or irrational.

11.2 Products

We've now looked at many techniques for coping with sums, but no methods for dealing with products. Fortunately, we don't need to develop an entirely new set of tools. Instead, we can first convert any product into a sum by taking a logarithm:

$$\ln\left(\prod f(n)\right) = \sum \ln f(n)$$

Then we can apply our summing tools and exponentiate at the end to undo the logarithm.

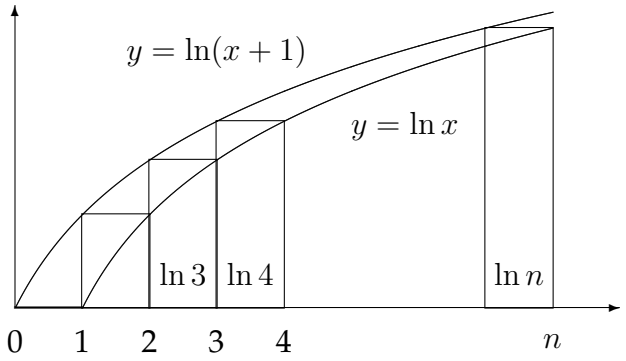
Let's apply this strategy to a product that you'll encounter almost daily hereafter:

$$n! = 1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n$$

First, we take a logarithm:

$$\ln n! = \ln 1 + \ln 2 + \ln 3 + \ldots + \ln n$$

This sum is rather nasty, but we can still get bounds by integrating. First, we work out appropriate functions and limits of integration with a picture.



Now we integrate to get bounds on $\ln n!$.

$$\int_1^n \ln x \, dx \;\leq\; \ln n! \;\leq\; \int_0^n \ln(x+1)\, dx$$

$$x\ln x - x\Big|_1^n \;\leq\; \ln n! \;\leq\; (x+1)\ln(x+1) - (x+1)\Big|_1^n$$

$$n\ln n - n \;\leq\; \ln n! \;\leq\; (n+1)\ln(n+1) - n$$

Finally, we exponentiate to get bounds on $n!$.

$$\frac{n^n}{e^n} \;\leq\; n! \;\leq\; \frac{(n+1)^{(n+1)}}{e^n}$$

This gives some indication how big $n!$ is: about $(n/e)^n$. This estimate is often good enough. However, as with the harmonic numbers, more precise bounds are known.

Fact 2.

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{1/(12n+1)} \;\leq\; n! \;\leq\; \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{1/(12n)}$$

These bounds are ridiculously close. For example, if $n = 100$, then we get:

$$100! \geq \left(\frac{100}{e}\right)^{100} \sqrt{200\pi}\; \underbrace{e^{1/1201}}_{=1.000832\ldots}$$

$$100! \leq \left(\frac{100}{e}\right)^{100} \sqrt{200\pi}\; \underbrace{e^{1/1200}}_{=1.000833\ldots}$$

The upper bound is less than 7 hundred-thousandths of 1% greater than the lower bound! Taken together, these bounds imply ***Stirling's Formula***:

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Stirling's formula is worth committing to memory; we'll often use it to rewrite expressions involving $n!$. Now, one might object that the expression on the *left* actually looks a lot better than the expression on the *right*. But that's just an artifact of notation. If you actually wanted to compute $n!$, you'd need $n-1$ multiplications. However, the expression on the right is a closed form; evaluating it requires only a handful of basic operations, regardless of the value of n. Furthermore, when $n!$ appears inside a larger expression, you usually can't do much with it. It doesn't readily cancel or combine with other terms. In contrast, the expression on the right looks scary, but melds nicely into larger formulas. So don't be put off; the expression on the right is your true friend.

Stepping back a bit, Stirling's formula is fairly amazing. Who would guess that the product of the first n positive *integers* could be so precisely described by a formula involving both e and π?

11.3 Asymptotic Notation

Approximation is a powerful tool. It lets you sweep aside irrelevant detail without losing track of the big picture.

Approximations are particularly useful in the analysis of computer systems and algorithms. For example, suppose you wanted to know how long multiplying two $n \times n$

matrices takes. You *could* tally up all the multiplications and additions and loop variable increments and comparisons and perhaps factor in hardware-specific considerations such as page faults and cache misses and branch mispredicts and floating-point unit availability and all this would give you one sort of answer.

On the other hand, each of the n^2 entries in the product matrix takes about n steps to compute. So the running time is proportional to n^3. This answer is certainly less precise. However, high-precision answers are rarely needed in practice. And this approximate answer is independent of tiny implementation and hardware details; it remains valid even after you upgrade your computer.

Computer scientists make heavy use of a specialized ***asymptotic notation*** to describe the growth of functions approximately. The notation involves six weird little symbols. We've already encountered one:

$$f(n) \sim g(n) \qquad \text{means} \qquad \lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$$

Less formally, $f \sim g$ means that the functions f and g grow at essentially the same rate. We've already derived two important examples:

$$H_n \sim \ln n \qquad\qquad n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Here are the other five symbols in the asymptotic notation system:

O	Ω	Θ	o	ω
oh	omega	theta	little-oh	little-omega

We'll focus on the most important one, O. Here's the definition: given functions $f, g : \mathbb{R} \to \mathbb{R}$, we say that

$$f(x) = O(g(x))$$

if there exist constants x_0 and $c > 0$ such that $|f(x)| \leq c \cdot g(x)$ for all $x \geq x_0$. Now this definition is pretty hairy. But what it's trying to say, with all its cryptic little constants, is that f grows no faster than g. A bit more precisely, it says that f is at most a constant times greater than g, except maybe for small values of x. For example:

$$5x + 100 = O(x)$$

This holds because the left side is only about 5 times larger than the right. Of course, for small values of x (like $x = 1$) the left side is many times larger than the right, but the definition of O is cleverly designed to sweep such inconvenient details under the rug.

Let's work through a sequence of examples carefully to better understand the definition.

Claim 71. $5x + 100 = O(x)$

Proof. We must show that there exist constants x_0 and $c > 0$ such that $|5x + 100| \leq c \cdot x$ for all $x \geq x_0$. Let $c = 10$ and $x_0 = 20$ and note that:

$$|5x + 100| \leq 5x + 5x = 10x \qquad \text{for all } x \geq 20$$

□

Claim 72. $x = O(x^2)$

Proof. We must show that there exist constants x_0 and $c > 0$ such that $|x| \leq c \cdot x^2$ for all $x \geq x_0$. Let $c = 1$ and $x_0 = 1$ and note that

$$|x| \leq 1 \cdot x^2 \qquad \text{for all } x \geq 1$$

□

What about the reverse? Is $x^2 = O(x)$? On an informal basis, this means x^2 grows no faster than x, which is false. Let's prove this formally.

Claim 73. $x^2 \neq O(x)$

Proof. We argue by contradiction; suppose that there exist constants x_0 and c such that:

$$|x^2| \leq c \cdot x \qquad \text{for all } x \geq x_0$$

Dividing both sides of the inequality by x gives:

$$x \leq c \qquad \text{for all } x \geq x_0$$

But this is false when $x = \max(x_0, c + 1)$. □

We can show that $x^2 \neq O(100x)$ by essentially the same argument; intuitively, x^2 grows quadratically, while $100x$ grows only linearly. Generally, changes in multiplicative constants do not affect the validity of an assertion involving O. However, constants in exponentials are critical:

Claim 74.

$$4^x \neq O(2^x)$$

Proof. We argue by contradiction; suppose that there exist constants x_0 and $c > 0$ such that:

$$|4^x| \leq c \cdot 2^x \qquad \text{for all } x \geq x_0$$

Dividing both sides by 2^x gives:

$$2^x \leq c \qquad \text{for all } x \geq x_0$$

But this is false when $x = \max(x_0, 1 + \log c)$. □

While asymptotic notation is useful for sweeping aside irrelevant detail, it can be abused. For example, there are ingenious algorithms for multiplying matrices that are asymptotically faster than the naive algorithm. The "best" requires only $O(n^{2.376})$ steps. However interesting theoretically, these algorithms are useless in practice because the constants hidden by the O notation are gigantic!

For more information about asymptotic notation see the "Asymptotic Cheat Sheet".

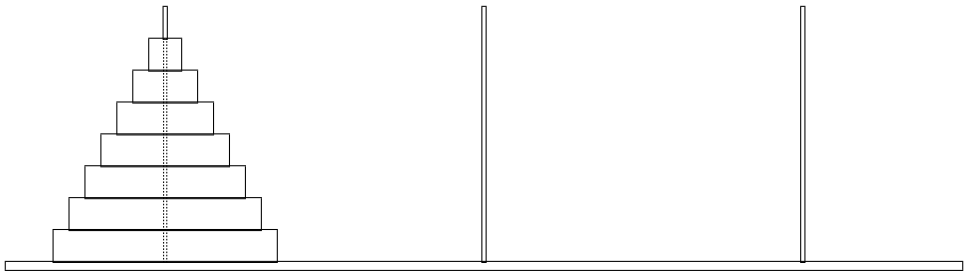
Chapter 12

Recurrences I

This is the first of two lectures about solving recurrences and recurrent problems. Needless to say, recurrent problems come up again and again. In particular, recurrences often arise in the analysis of recursive algorithms.

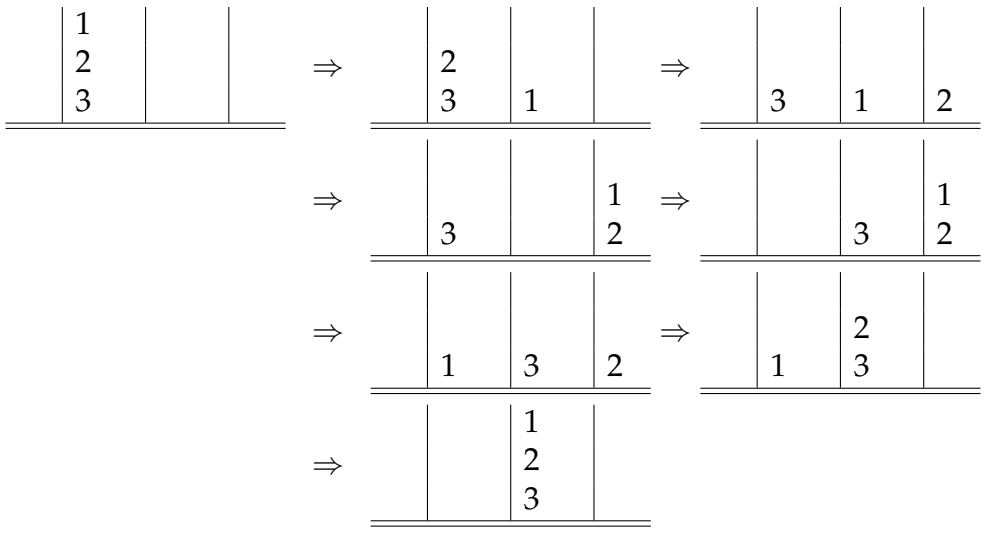
12.1 The Towers of Hanoi

In the Towers of Hanoi problem, there are three posts and seven disks of different sizes. Each disk has a hole through the center so that it fits on a post. At the start, all seven disks are on post #1 as shown below. The disks are arranged by size so that the smallest is on top and the largest is on the bottom. The goal is to end up with all seven disks in the same order, but on a different post. This is not trivial because of two restrictions. First, the only permitted action is removing the top disk from a post and dropping it onto another post. Second, a larger disk can never lie above a smaller disk on any post. (These rules imply, for example, that it is no fair to pick up the whole stack of disks at once and then to drop them all on another post!)



It is not immediately clear that a solution to this problem exists; maybe the rules are so stringent that the disks cannot all be moved to another post!

One approach to this problem is to consider a simpler variant with only three disks. We can quickly exhaust the possibilities of this simpler puzzle and find a 7-move solution such as the one shown below. (The disks on each post are indicated by the numbers immediately to the right. Larger numbers correspond to larger disks.)



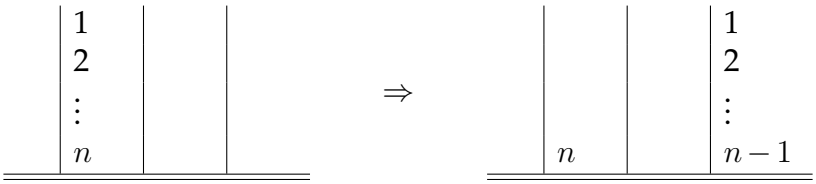
This problem was invented in 1883 by the French mathematician Edouard Lucas. In his original account, there were 64 disks of solid gold. At the beginning of time, all 64 were placed on a single post, and monks were assigned the task of moving them to another post according to the rules described above. According to legend, when the monks complete their task, the Tower will crumble and the world will end!

The questions we must answer are, "Given sufficient time, can the monks succeed?" and if so, "How long until the world ends?" and, most importantly, "Will this happen before the 6.042 final?"

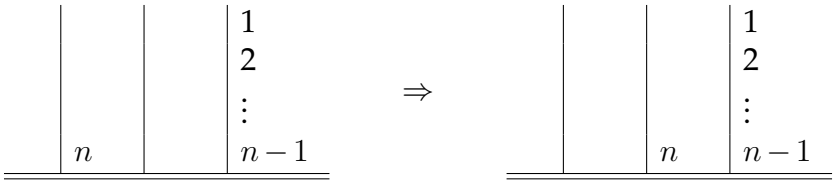
12.1.1 Finding a Recurrence

The Towers of Hanoi problem can be solved recursively as follows. Let T_n be the minimum number of steps needed to move an n-disk tower from one post to another. For example, a bit of experimentation shows that $T_1 = 1$ and $T_2 = 3$. For 3 disks, the solution given above proves that $T_3 \leq 7$. We can generalize the approach used for 3 disks to the following recursive algorithm for n disks.

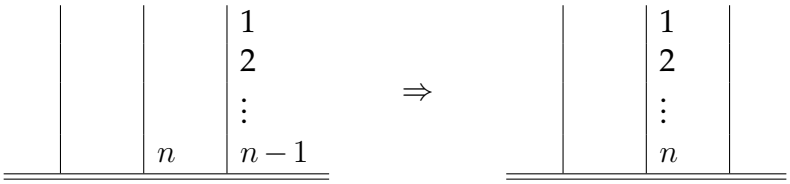
Step 1. Move the top $n-1$ disks from the first post to the third post. This can be done in T_{n-1} steps.



Step 2. Move the largest disk from the first post to the to the second post. This requires just 1 step.



Step 3. Move the $n-1$ disks from the third post onto the second post. Again, T_{n-1} steps are required.



This algorithm shows that T_n, the number of steps required to move n disks to a different post, is at most $2T_{n-1}+1$. We can use this fact to compute upper bounds on the number of steps required for various numbers of disks:

$$\begin{aligned} T_3 &\leq 2 \cdot T_2 + 1 \\ &= 7 \\ T_4 &\leq 2 \cdot T_3 + 1 \\ &\leq 15 \end{aligned}$$

The algorithm described above answers our first question: given sufficient time, the monks will finish their task and end the world. (Which is a shame. After all that effort they'd probably want to smack a few high-fives and go out for burgers and ice cream, but nope— world's over.)

12.1.2 A Lower Bound for Towers of Hanoi

We can not yet compute the exact number of steps that the monks need to move the 64 disks; we can only show an upper bound. Perhaps— having pondered the problem since the beginning of time— the monks have devised a better algorithm.

In fact, there is no better algorithm, and here is why. At some step, the monks must move the n-th disk from the first post to a different post. For this to happen, the $n-1$ smaller disks must all be stacked out of the way on the only remaining post. Arranging the $n-1$ smaller disks this way requires at least T_{n-1} moves. After the largest disk is moved, at least another T_{n-1} moves are required to pile the $n-1$ smaller disks on top.

This argument shows that the number of steps required is at least $2T_{n-1} + 1$. Since we gave an algorithm using exactly that number of steps, we now have a recurrence for T_n, the number of moves required to complete the Tower of Hanoi problem with n disks:

$$\begin{aligned} T_1 &= 1 \\ T_n &= 2T_{n-1} + 1 \qquad\qquad (\text{for } n \geq 2) \end{aligned}$$

We can use this recurrence to conclude that $T_2 = 3, T_3 = 7, T_4 = 15, \ldots$.

12.1.3 Guess-and-Verify

Computing T_{64} from the recurrence would require a lot of work. It would be nice to have a closed form expression for T_n, so that we could quickly compute the number of steps required to solve the Towers of Hanoi problem for any given number of disks. (For example, we might want to know how much sooner the world would end if the monks melted down one disk to purchase burgers and ice cream *before* the end of the world.)

There are several different methods for solving recurrences. The simplest method is to *guess* the solution and then to *verify* that the guess is correct, usually with an induction proof. This method is called **guess-and-verify** or "substitution". As a basis for a good guess, let's tabulate T_n for small values of n:

n	T_n
1	1
2	3
3	7
4	15
5	31
6	63

Based on this table, a natural guess is that $T_n = 2^n - 1$.

Whenever you guess a solution to a recurrence, you should always verify it with a proof by induction or by some other technique; after all, your guess might be wrong. (But why bother to verify in this case? After all, if we're wrong, its not the end of the…no, let's check.)

Claim. *If:*

$$\begin{aligned} T_1 &= 1 \\ T_n &= 2T_{n-1} + 1 \qquad\qquad (\textit{for } n \geq 2) \end{aligned}$$

then:

$$T_n = 2^n - 1$$

Proof. The proof is by induction on n. Let $P(n)$ be the proposition that $T_n = 2^n - 1$.

Base case: $P(1)$ is true because $T_1 = 1 = 2^1 - 1$.

Inductive step: Now we assume $T_n = 2^n - 1$ to prove that $T_{n+1} = 2^{n+1} - 1$, where $n \geq 1$.

$$\begin{aligned} T_{n+1} &= 2T_n + 1 \\ &= 2(2^n - 1) + 1 \\ &= 2^{n+1} - 1 \end{aligned}$$

The first equality is the recurrence relation, and the second equation follows by the assumption $P(n)$. The last step is simplification. □

Our guess is now verified. This shows, for example, that the 7-disk puzzle will require $2^7 - 1 = 127$ moves to complete. We can also now resolve our remaining questions about the 64-disk puzzle. Since $T_{64} = 2^{64} - 1$, the monks must complete more than 18 billion billion steps before the world ends. Better study for the final.

12.1.4 The Plug-and-Chug Method

In general, guess-and-verify is a great way to solve recurrences. The only problem with the method is guessing the right solution. This was easy in the Towers of Hanoi example, but sometimes the solution has a strange form that is quite hard to guess. Practice helps, of course, but so can some other methods.

Plug-and-chug is one such alternative method for solving recurrences. Plug-and-chug is also sometimes called "expansion", "iteration", or "brute force". The method consists of four calculation-intensive steps. These are described below and illustrated with the Tower of Hanoi examle.

Step 1: Plug and Chug

Expand the recurrence equation by alternately "plugging" (applying the recurrence equation) and "chugging" (simplifying the resulting expression).

$$\begin{aligned} T_n &= 1 + 2T_{n-1} && \\ &= 1 + 2(1 + 2T_{n-2}) && \text{plug} \\ &= 1 + 2 + 4T_{n-2} && \text{chug} \\ &= 1 + 2 + 4(1 + 2T_{n-3}) && \text{plug} \\ &= 1 + 2 + 4 + 8T_{n-3} && \text{chug} \\ &= 1 + 2 + 4 + 8(1 + 2T_{n-4}) && \text{plug} \\ &= 1 + 2 + 4 + 8 + 16T_{n-4} && \text{chug} \end{aligned}$$

Be careful in the "chug" stage; too much simplification can obscure an emerging pattern. For example, summing $1 + 2 + 4 + \ldots$ at every stage would have concealed the geometric series. The rule to remember— indeed, a rule applicable to the whole of college life— is *chug in moderation*.

Step 2: Identify and Verify a Pattern

Identify a pattern for the recurrence equation after i rounds of plugging and chugging. Verify that this pattern is correct by carrying out one additional round of plug and chug. In the Towers of Hanoi example, a strong pattern emerges: T_n is always a sum of consecutive powers of two together with an earlier T term:

$$T_n = 1 + 2 + 4 + \ldots + 2^{i-1} + 2^i T_{n-i}$$

We do one last round of plug-and-chug to confirm that the pattern is correct. This is amounts to the inductive step of a proof that we have the right general form.

$$\begin{aligned} T_n &= 1 + 2 + 4 + \ldots + 2^{i-1} + 2^i(1 + 2T_{n-(i+1)}) && \text{plug} \\ &= 1 + 2 + 4 + \ldots + 2^{i-1} + 2^i + 2^{i+1}T_{n-(i+1)} && \text{chug} \end{aligned}$$

Step 3: Express n-th Term Using Early Terms

Substitute a value of i into the pattern so that T_n is expressed as a function of just the base cases. Substitute values for these terms to obtain an ordinary, non-recurrent expression for T_n. For the Towers of Hanoi recurrence, substituting $i = n - 1$ into the general form determined in Step 2 gives:

$$\begin{aligned} T_n &= 1 + 2 + 4 + \ldots + 2^{n-2} + 2^{n-1}T_1 \\ &= 1 + 2 + 4 + \ldots + 2^{n-2} + 2^{n-1} \end{aligned}$$

The second step uses the base case $T_1 = 1$. Now we have an ordinary, non-recurrent expression for T_n.

Step 4: Find a Closed Form for the n-th Term

All that remains is to reduce the ordinary expression for T_n to a closed form. We are fortunate in this case, because T_n is the sum of a geometric series. We learned how to tackle these last week!

$$\begin{aligned} T_n &= 1 + 2 + 4 + 8 + \ldots + 2^{n-2} + 2^{n-1} \\ &= \sum_{i=0}^{n-1} 2^i \\ &= 2^n - 1 \end{aligned}$$

We're done! When using plug-and-chug method, you might want to verify your solution with induction. It is easy to make a mistake when observing the general pattern.

12.2 Merge Sort

There are many algorithms for sorting a list of n items; in fact, you will see about dozen of them in 6.046. One of the most popular sorting algorithms is Merge Sort.

12.2.1 The Algorithm

Here is how Merge Sort works. The input is a list of $n \geq 1$ items $x_1, x_2, \ldots, x_n$. If $n = 1$, then the algorithm returns the single item x_1. If $n > 1$, then the original list is broken into two lists, $x_1, \ldots, x_{n/2}$ and $x_{n/2+1}, \ldots, x_n$. Both of these lists are sorted recursively, and then they are merged to form a complete, sorted list of the original n items.

Let's work through an example. Suppose we want to sort this list:

$$10, 7, 23, 5, 2, 4, 3, 9$$

Since there is more than one item, we divide into two lists; one is 10, 7, 23, 5, and the other is 2, 4, 3, 9. Each list is sorted recursively. The results are:

$$5, 7, 10, 23$$
$$2, 3, 4, 9$$

Now we must merge these two small sorted lists into one big sorted list. We start with an empty big list and add one item at a time. At each step, we compare the first items in the small lists. We move the smaller of these two to the end of the big list. This process repeats until one of the small lists becomes empty. At that point, the remaining small list is appended to the big list and we are done. For the example, the contents of the three lists after each step are shown in the table below. The next items to move are underlined.

small list #1	small list #2	big list
$5, 7, 10, 23$	$\underline{2}, 3, 4, 9$	
$5, 7, 10, 23$	$\underline{3}, 4, 9$	2
$5, 7, 10, 23$	$\underline{4}, 9$	$2, 3$
$\underline{5}, 7, 10, 23$	9	$2, 3, 4$
$\underline{7}, 10, 23$	9	$2, 3, 4, 5$
$10, 23$	$\underline{9}$	$2, 3, 4, 5, 7$
$\underline{10, 23}$		$2, 3, 4, 5, 7, 9$
		$2, 3, 4, 5, 7, 9, 10, 23$

Because we keep dividing up the original list recursively until only 1 item remains, all the work is in the merging!

12.2.2 Finding a Recurrence

In the analysis of a sorting algorithm, a traditional question is, "What is the maximum number comparisons used in sorting n items?" The number of comparisons is taken as an estimate of the running time. In the case of Merge Sort, we can find a recurrence for this quantity. Solving this recurrence will allow us to study the asymptotic behavior of the algorithm.

To make the analysis easier, assume for now that the number of items we are sorting is a power of 2. This ensures that we can divide the original list of items exactly in half at every stage of the recursion.

Let $T(n)$ be the maximum number of comparisons used by Merge Sort in sorting a list of n items. If there is only one item, then no comparisons are required, so $T(1) = 0$. If $n > 1$, then $T(n)$ is the sum of:

- The number of comparisions used in sorting both halves of the list, which is at most $2T(n/2)$.
- The number of comparisons used in merging two lists of length n. This is at most $n-1$ because one item is appended to the big list after each comparison, and at least one additional item is appended to the big list in the final step when one small list becomes empty. Since the big list eventually contains n items, there can be at most $n-1$ comparisons. (There might be fewer comparisons if one small list empties out quickly, but we are analyzing the worst case.)

Therefore, the number of comparisons required to sort n items is at most:

$$T(n) = 2T(n/2) + n - 1$$

12.2.3 Solving the Recurrence

Now we need a closed form for the number of comparisons used by Merge Sort in sorting a list of n items. This requires solving the recurrence:

$$\begin{aligned} &T(1) = 0 \\ &T(n) = 2T(n/2) + n - 1 \qquad\qquad (\text{for } n > 1) \end{aligned}$$

Let's first compute a few terms and try to apply guess-and-verify:

n	$T(n)$
1	0
2	1
4	5
8	17
16	49

There is no obvious pattern. We could compute more values and look harder, but let's try our other method, plug-and-chug.

Step 1: Plug and Chug

First, we alternately plug and chug until a pattern emerges:

$$\begin{aligned}
T(n) &= n - 1 + 2T(n/2) \\
&= (n-1) + 2(n/2 - 1 + 2T(n/4)) && \text{plug} \\
&= (n-1) + (n-2) + 4T(n/4) && \text{chug} \\
&= (n-1) + (n-2) + 4(n/4 - 1 + 2T(n/8)) && \text{plug} \\
&= (n-1) + (n-2) + (n-4) + 8T(n/8) && \text{chug} \\
&= (n-1) + (n-2) + (n-4) + 8(n/8 - 1 + 2T(n/16)) && \text{plug} \\
&= (n-1) + (n-2) + (n-4) + (n-8) + 16T(n/16) && \text{chug}
\end{aligned}$$

Note that too much simplification would have obscured the pattern that has now emerged.

Step 2: Identify and Verify a Pattern

Now we identify the general pattern and do one more round of plug-and-chug to verify that it is maintained:

$$\begin{aligned}
T(n) &= (n-1) + (n-2) + \ldots + (n - 2^{i-1}) + 2^i T(n/2^i) \\
&= (n-1) + (n-2) + \ldots + (n - 2^{i-1}) + 2^i(n/2^i - 1 + 2T(n/2^{i+1})) && \text{plug} \\
&= (n-1) + (n-2) + \ldots + (n - 2^{i-1}) + (n - 2^i) + 2^{i+1} T(n/2^{i+1}) && \text{chug}
\end{aligned}$$

Step 3: Express n-th Term Using Early Terms

Now we substitute a value for i into the pattern so that $T(n)$ depends on only base cases. A natural choice is $i = \log n$, since then $T(n/2^i) = T(1)$. This substitution makes $T(n)$ dependent only on $T(1)$, which we know is 0.

$$\begin{aligned}
T(n) &= (n-1) + (n-2) + \ldots + (n - 2^{\log(n)-1}) + 2^{\log n} T(n/2^{\log n}) \\
&= (n-1) + (n-2) + \ldots + (n - n/2) + nT(1) \\
&= (n-1) + (n-2) + \ldots + (n - n/2)
\end{aligned}$$

Step 4: Find a Closed-Form for the n-th Term

Now we have an ordinary, non-recurrent expression for $T(n)$. We can reduce this to a closed form by summing a series.

$$\begin{aligned}
T(n) &= (n-1) + (n-2) + \ldots + (n - n/2) \\
&= n \log n - (1 + 2 + 4 + \ldots + n/2) \\
&= n \log n - n + 1 \\
&\sim n \log n
\end{aligned}$$

What a weird answer— no one would ever guess that![1] As a check, we can verify that the formula gives the same values for $T(n)$ that we computed earlier:

n	$n\log_2 n - n + 1$
1	$1\log_2 1 - 1 + 1 = 0$
2	$2\log_2 2 - 2 + 1 = 1$
4	$4\log_2 4 - 4 + 1 = 5$
8	$8\log_2 8 - 8 + 1 = 17$
16	$16\log_2 16 - 16 + 1 = 49$

The values match! If we wanted certainty, we could verify this solution with an induction proof.

12.3 More Recurrences

Let's compare the Tower of Hanoi and Merge Sort recurrences.

$$\begin{array}{rll} \text{Hanoi} & T(n) = 2T(n-1) + 1 & \Rightarrow T(n) \sim 2^n \\ \text{Merge Sort} & T(n) = 2T(n/2) + (n-1) & \Rightarrow T(n) \sim n\log n \end{array}$$

Though the recurrence equations are quite similar, the solutions are radically different!

At first glance each recurrence has one strength and one weakness. In particular, in the Towers of Hanoi, we broke a problem of size n into two subproblem of size $n-1$ (which is large), but needed only 1 additional step (which is small). In Merge Sort, we divided the problem of size n into two subproblems of size $n/2$ (which is small), but needed $(n-1)$ additional steps (which is large). Yet, Merge Sort is faster by a mile! The take-away point is that generating smaller subproblems is far more important to algorithmic speed than reducing the additional steps per recursive call.

12.3.1 A Speedy Algorithm

Let's try one more recurrence. Suppose we have a speedy algorithm with the best properties of both earlier algorithms; that is, at each stage the problem is divided in half *and* we do only one additional step. Then the run time is described by the following recurrence:

$$\begin{aligned} &S(1) = 0 \\ &S(n) = 2S(n/2) + 1 \qquad\qquad \text{(for } n \geq 2\text{)} \end{aligned}$$

[1] Except for the couple people in lecture who actually did. Oh well.

Let's first try guess-and-verify. As usual, we tabulate a few values of $S(n)$. As before, assume that n is a power of two.

n	$S(n)$
1	0
2	$2S(1)+1=1$
4	$2S(2)+1=3$
8	$2S(4)+1=7$
16	$2S(8)+1=15$

The obvious guess is that $S(n) = n - 1$. Let's try to verify this.

Claim. *Suppose:*

$$\begin{aligned} S(1) &= 0 \\ S(n) &= 2S(n/2) + 1 \qquad \textit{(for } n \geq 2\textit{)} \end{aligned}$$

If n is a power of 2, then:

$$S(n) = n - 1$$

Proof. The proof is by strong induction. Let $P(n)$ be the proposition that if n is a power of 2, then $S(n) = n - 1$.

Base case: $P(1)$ is true because $S(1) = 1 - 0 = 0$.

Inductive step: Now assume $P(1), \ldots, P(n-1)$ in order to prove that $P(n)$, where $n \geq 2$. If n is not a power of 2, then $P(n)$ is vacuously true. Otherwise, we can reason as follows:

$$\begin{aligned} S(n) &= 2S(n/2) + 1 \\ &= 2(n/2 - 1) + 1 \\ &= n - 1 \end{aligned}$$

The first equation is the recurrence. The second equality follows from assumption $P(n/2)$, and the last step is simplification only. □

Thus, the running time of this speedy algorithm is $S(n) \sim n$. This is better than the $T(n) \sim n \log n$ running time of Merge Sort, but only slightly so. This is consistent with the idea that decreasing the number of additional steps per recursive call is much less important than reducing the size of subproblems.

12.3.2 A Verification Problem

Sometimes verifying the solution to a recurrence using induction can be tricky. For example, suppose that we take the recurrence equation from the speedy algorithm, but we only try to prove that $S(n) \leq n$. This is true, but the proof goes awry!

Claim 75. *If n is a power of two, then $S(n) \leq n$.*

Proof. (failed attempt) The proof is by strong induction. Let $P(n)$ be the proposition that if n is a power of two, then $S(n) \leq n$.

Base case: $P(1)$ is true because $S(1) = 1 - 0 < 1$.

Inductive step: For $n \geq 2$, assume $P(1)$, $P(2)$, ..., $P(n-1)$ to prove $P(n)$. If n is not a power of two, then $P(n)$ is vacuously true. Otherwise, we have:

$$\begin{aligned} S(n) &= 2S(n/2) + 1 \\ &\leq 2(n/2) + 1 \\ &= n + 1 \\ &\not\leq n \end{aligned}$$

The first equation is the recurrence. The second equality follows by the assumption $P(n/2)$. The third step is a simplification, and in the fourth step we crash and burn spectacularly. □

We know that the result is true, but the proof did not work! The natural temptation is to ease up and try to prove something *weaker*, say $S(n) \leq 2n$. Bad plan! Here's what would happen in the inductive step:

$$\begin{aligned} S(n) &= 2S(n/2) + 1 \\ &\leq 2n + 1 \\ &\not\leq 2n \end{aligned}$$

We're still stuck! As with other induction proofs, the key is to use a *stronger* induction hypothesis such as $S(n) = n - 1$ (as above) or $S(n) \leq n - 1$.

12.3.3 A False Proof

What happens if we try an even *stronger* induction hypothesis? Shouldn't the proof work out even *more* easily? For example, suppose our hypothesis is that $S(n) \leq n - 2$. This hypothesis is false, since we proved that $S(n) = n - 1$. But let's see where the proof breaks. Here again is the crux of the argument:

$$\begin{aligned} S(n) &= 2S(n/2) + 1 \\ &\leq 2(n/2 - 2) + 1 \\ &= n - 3 \\ &\leq n - 2 \end{aligned}$$

Something is wrong; we proved a false statement! The problem is that we were lazy and did not write out the full proof; in particular, we ignored the base case. Since $S(1) = 0 \not\leq -1$, the induction hypothesis is actually false in the base case. This is why we cannot construct a valid proof with a "too strong" induction hypothesis.

12.3.4 Altering the Number of Subproblems

Some variations of the Merge Sort recurrence have truly peculiar solutions! The main difference in these variants is that we replace the constant 2 (arising because we create 2 subproblems) by a parameter a.

$$\begin{aligned} T(1) &= 1 \\ T(n) &= aT(n/2) + n \end{aligned}$$

Intuitively, a is the number of subproblems of size $n/2$ generated at each stage of the algorithm; however, a is actually not required to be an integer. This recurrence can be solved by plug-and-chug, but we'll omit the details. The solution depends strongly on the value of a:

$$T(n) \sim \begin{cases} \dfrac{2n}{2-a} & \text{for } 0 \le a < 2, \\ n \log n & \text{for } a = 2, \\ \dfrac{an^{\log a}}{a-2} & \text{for } a > 2. \end{cases}$$

The examples below show that the Merge Sort recurrence is extremely sensitive to the multiplicative term, especially when it is near 2.

$$\begin{aligned} a &= 1.99 && \Rightarrow T(n) = \Theta(n) \\ a &= 2 && \Rightarrow T(n) = \Theta(n \log n) \\ a &= 2.01 && \Rightarrow T(n) = \Theta(n^{1.007\ldots}) \end{aligned}$$

The constant $1.007\ldots$ is equal to $\log 2.01$.

12.4 The Akra-Bazzi Method

The Merge Sort recurrence and all the variations we considered are called divide-and-conquer recurrences because they arise all the time in the analysis of divide-and-conquer algorithms. In general, a ***divide-and-conquer*** recurrence has the form:

$$T(x) = \begin{cases} \text{is defined} & \text{for } 0 \le x \le x_0 \\ \displaystyle\sum_{i=1}^{k} a_i T(b_i x) + g(x) & \text{for } x > x_0 \end{cases}$$

Here x is any nonnegative real number; it need not be a power of two or even an integer. In addition, $a_1, \ldots a_k$ are positive constants, $b_1, \ldots, b_k$ are constants between 0 and 1, and x_0 is "large enough" to ensure that $T(x)$ is well-defined. (This is a technical issue that we'll not go into in any greater depth.)

This general form describes all recurrences in this lecture, except for the Towers of Hanoi recurrence. (We'll learn a method for solving that type of problem in the next lecture.) Some hideous recurrences are also in the divide-and-conquer class. Here is an example:

$$T(x) = 2T(x/2) + 8/9T(3x/4) + x^2$$

In this case, $k = 2$, $a_1 = 2$, $a_2 = 8/9$, $b_1 = 1/2$, $b_2 = 3/4$, and $g(x) = x^2$.

12.4.1 Solving Divide and Conquer Recurrences

A few years ago, two guys in Beirut named Akra and Bazzi discovered an elegant way to solve *all* divide-and-conquer recurrences.

Theorem 76 (Akra-Bazzi, weak form). *Suppose that:*

$$T(x) = \begin{cases} \textit{is defined} & \textit{for } 0 \le x \le x_0 \\ \displaystyle\sum_{i=1}^{k} a_i T(b_i x) + g(x) & \textit{for } x > x_0 \end{cases}$$

where:

- $a_1, \ldots, a_k$ *are positive constants*
- $b_1, \ldots, b_k$ *are constants between 0 and 1*
- x_0 *is "large enough" in a technical sense we leave unspecified*
- $|g'(x)| = O(x^c)$ *for some* $c \in \mathbb{N}$

Then:

$$T(x) = \Theta\left(x^p \left(1 + \int_1^x \frac{g(u)}{u^{p+1}}\, du\right)\right)$$

where p *satisfies the equation* $\sum_{i=1}^{k} a_i b_i^p = 1$.

We won't prove this here, but let's apply the theorem to solve the nasty recurrence from above:

$$T(x) = 2T(x/2) + 8/9T(3x/4) + x^2$$

The first step is to find p, which is defined by the equation:

$$2\left(\frac{1}{2}\right)^p + \frac{8}{9}\left(\frac{3}{4}\right)^p = 1$$

Equations of this form don't always have closed-form solutions. But, in this case, the solution is simple: $p = 2$. Next, we have to check that $g'(x)$ does not grow too fast:

$$|g'(x)| = |2x| = O(x)$$

Finally, we can compute the solution to the recurrence by integrating:

$$\begin{aligned} T(x) &= \Theta\left(x^2\left(1+\int_1^x \frac{u^2}{u^3}\,du\right)\right) \\ &= \Theta\left(x^2\left(1+\log x\right)\right) \\ &= \Theta(x^2 \log x) \end{aligned}$$

The Akra-Bazzi method an be frustrating, because you do a lot of inexplicable intermediate calculations and then the answer just pops out of an integral. However, it goes through divide-and-conquer recurrences like a Weed Wacker.

Let's try one more example. Suppose that the following recurrence holds for all sufficiently large x:

$$T(x) = T(x/3) + T(x/4) + x$$

Here $k = 2$, $a_1 = 1$, $a_2 = 1$, $b_1 = 1/3$, $b_2 = 1/4$, and $g(x) = x$. Note that $|g'(x)| = 1 = O(1)$, so the Akra-Bazzi theorem applies. The next job is to compute p, which is defined by the equation:

$$\left(\frac{1}{3}\right)^p + \left(\frac{1}{4}\right)^p = 1$$

We're in trouble: there is no closed-form expression for p. But at least we can say $p < 1$, and this turns out to be enough. Let's plow ahead and see why. The Akra-Bazzi theorem says:

$$\begin{aligned} T(x) &= \Theta\left(x^p\left(1+\int_1^x \frac{u}{u^{p+1}}\,du\right)\right) \\ &= \Theta\left(x^p\left(1+\int_1^x u^{-p}\,du\right)\right) \\ &= \Theta\left(x^p\left(1+\left(\frac{u^{1-p}}{1-p}\Big|_{u=1}^{x}\right)\right)\right) \\ &= \Theta\left(x^p\left(1+\frac{x^{1-p}-1}{1-p}\right)\right) \\ &= \Theta\left(x^p + \frac{x}{1-p} - \frac{x^p}{1-p}\right) \\ &= \Theta\left(x\right) \end{aligned}$$

In the last step, we use the fact that $x^p = o(x)$ since $p < 1$; in other words, the term involving x dominates the terms involving x^p, so we can ignore the latter. Overall, this calculation show that we don't need to know the exact value of p because it cancels out!

In recitation we'll go over a slightly more general version of the Akra-Bazzi theorem. This generalization says that *small* changes in the sizes of subproblems do not affect the solution. This means that some apparent complications are actually irrelevant, which is nice.

Chapter 13

Recurrences II

13.1 Asymptotic Notation and Induction

We've seen that asymptotic notation is quite useful, particularly in connection with recurrences. And induction is our favorite proof technique. But mixing the two is risky business; there is great potential for subtle errors and false conclusions!

False Claim 77. *If*

$$T(1) = 1$$
$$T(n) = 2T(n/2) + n$$

then $T(n) = O(n)$.

This claim is false; the Akra-Bazzi theorem implies that the correct solution is $T(n) = \Theta(n \log n)$. But where does the following "proof" go astray?

Proof. The proof is by strong induction. Let $P(n)$ be the proposition that $T(n) = O(n)$.

Base case: $P(1)$ is true because $T(1) = 1 = O(1)$.

Inductive step: For $n \geq 2$ assume $P(1), P(2), \ldots P(n-1)$ to prove $P(n)$. We have:

$$\begin{aligned} T(n) &= 2 \cdot T(n/2) + n \\ &= 2 \cdot O(n/2) + n \\ &= O(n) \end{aligned}$$

The first equation is the recurrence, the second uses the assumption $P(n/2)$, and the third is a simplification. □

Where's the bug? The proof is already far off the mark in the second sentence, which defines the induction hypothesis. The statement "$T(n) = O(n)$" is either true or false; its

validity does not depend on a particular value of n. Thus, the very idea of trying to prove that the statement holds for $n = 0, 1, 2, \ldots$ is wrong-headed.

The safe way to reason inductively about asymptotic phenomena is to *work directly with the definition of the notation*. Let's try to prove the claim above in this way. Remember that $f(n) = O(n)$ means that there exist constants n_0 and $c > 0$ such that $|f(n)| \leq cn$ for all $n \geq n_0$. (Let's not worry about the absolute value for now.) If all goes well, the proof attempt should fail in some blatantly obvious way, instead of in a subtle, hard-to-detect way like the earlier argument. Since our perverse goal is to demonstrate that the proof won't work for *any* constants n_0 and c, we'll leave these as variables and assume only that they're chosen so that the base case holds; that is, $T(n_0) \leq cn$.

Proof Attempt. We use strong induction. Let $P(n)$ be the proposition that $T(n) \leq cn$.

Base case: $P(n_0)$ is true, because $T(n_0) \leq cn$.

Inductive step: For $n > n_0$, assume that $P(n_0), \ldots, P(n-1)$ are true in order to prove $P(n)$. We reason as follows:

$$\begin{aligned} T(n) &= 2T(n/2) + n \\ &\leq 2c(n/2) + n \\ &= cn + n \\ &= (c+1)n \\ &\not\leq cn \end{aligned}$$

The first equation is the recurrence. Then we use induction and simplify until the argument collapses!

13.2 Linear Recurrences

Last lecture we saw how to solve Divide and Conquer recurrences. In this lecture, we'll consider another family of recurrences, called "linear recurrences", that also frequently arise in computer science and other disciplines. We'll first see how to solve a specific linear recurrence and then generalize our method to work for all linear recurrences.

13.2.1 Graduate Student Job Prospects

In a new academic field (say computer science), there are only so many faculty positions available in all the universities of the world. Initially, there were not enough qualified candidates, and many positions were unfilled. But, over time, new graduates are filling the positions, making job prospects for later computer science students ever more bleak. Worse, the increasing number of professors are able to train an increasing number of graduate students, causing positions to fill ever more rapidly. Eventually, the universities will be saturated; new computer science graduates will have no chance at an academic

career. Our problem is to determine when the universities will stop hiring new computer science faculty and, in particular, to answer the question, "Are the 6.042 TAs doomed?" Here are the details of the problem.

- There are a total of N faculty positions available worldwide. This number never changes due to budgetary constraints.
- Congress has passed a law forbidding forced retirement in universities, and no one will retire voluntarily. (This is true and a problem!) In our analysis, therefore, once a faculty position is filled, it never opens up again.
- Each year, every professor trains exactly 1 student who will go on to become a professor the following year. The only exception is that first year professors do not train students; they are too busy publishing, teaching, getting grants, and serving on committees.
- In year 0, there are no computer science professors in the world. In year 1, the first professor is hired.

13.2.2 Finding a Recurrence

Ideally, we could find a formula for the number of professors in the world in a given year. Then we could determine the year in which all N faculty positions are filled. Let $f(n)$ be the number of professors during year n. To develop some intuition about the problem, we can compute values of this function for small n by hand.

$f(0) = 0$	No CS professors; the dark ages.
$f(1) = 1$	1 new professor; too busy to train a student.
$f(2) = 1$	1 old professor; now training a student.
$f(3) = 2$	1 new prof, 1 old prof; new prof too busy, old prof training.
$f(4) = 3$	1 new prof, 2 old profs; new prof too busy, both old profs training
$f(5) = 5$	2 new profs, 3 old profs
$f(6) = 8$	3 new profs, 5 old profs

In general, the number of professors in year n is equal to the number of professors last year plus the number of new hires. The number of professors last year is $f(n-1)$. The number of new hires is equal to the number of professors two years ago, $f(n-2)$, since each of these professors trained a student last year. These observations give the following

recurrence equation for the number of professors:

$$\begin{aligned} f(0) &= 0 \\ f(1) &= 1 \\ f(n) &= f(n-1) + f(n-2) \qquad (n \geq 2) \end{aligned}$$

This is the familiar Fibonacci recurrence. Looking back, the values of $f(n)$ that we computed by hand are indeed the first few Fibonacci numbers. Fibonacci numbers arise in all sorts of applications. Fibonacci himself introduced the numbers in 1202 to study rabbit reproduction. Fibonacci numbers also appear, oddly enough, in the spiral patterns on the faces of sunflowers. And the input numbers that make Euclid's GCD algorithm require the greatest number of steps are consecutive Fibonacci numbers. So how big is $f(n)$ anyway? Of course, we could compute as many Fibonacci numbers as we like using the recurrence, but it would be much nicer to find a closed form.

13.2.3 Solving the Recurrence

Solving the Fibonacci recurrence is easy because the recurrence is linear. (Well, "easy" in the sense that you can learn the technique in one lecture; discovering it actually took six centuries.) A ***linear recurrence*** has the form:

$$\begin{aligned} f(n) &= a_1 f(n-1) + a_2 f(n-2) + \ldots + a_d f(n-d) \\ &= \sum_{i=1}^{d} b_i f(n-i) \end{aligned}$$

where $a_1, a_2, \ldots a_d$ are constants. The ***order*** of the recurrence is d. For example, the Fibonacci recurrence is order 2 and has coefficients $a_1 = a_2 = 1$. (Later on, we'll slightly expand the definition of a linear recurrence.)

For now, let's try to solve just the Fibonacci recurrence; we'll see how to solve general linear recurrences later in the lecture. Our rule of thumb is that guess-and-verify is the first method to apply to an unfamiliar recurrence. It turns out that for a linear recurrence, an exponential solution is a good guess. However, since we know nothing beyond this, our initial guess-and-erify attempt will really only be a "dry run"; that is, we will not make an exact guess and will not verify it with a complete proof. Rather, the goal of this first attempt is only to clarify the form of the solution.

Guess. $f(n) = cx^n$

Here c and x are parameters introduced to improve our odds of having a correct guess; in the verification step, we can pick values that make the proof work. To further improve our odds, let's neglect the boundary conditions, $f(0) = 0$ and $f(1) = 1$.

Verification. Plugging our guess into the recurrence $f(n) = f(n-1) + f(n-2)$ gives:

$$\begin{aligned} cx^n &= cx^{n-1} + cx^{n-2} \\ x^2 &= x + 1 \\ x^2 - x - 1 &= 0 \\ x &= \frac{1 \pm \sqrt{5}}{2} \end{aligned}$$

In the first step, we divide both sides of the equation by cx^{n-2}. Then we rearrange terms and find x with the quadratic formula.

This calculation suggests that the connstant c can be anything, but that x must be $(1 \pm \sqrt{5})/2$. Evidently, there are two solutions to the recurrence:

$$f(n) = c\left(\frac{1+\sqrt{5}}{2}\right)^n \qquad \text{or} \qquad f(n) = c\left(\frac{1-\sqrt{5}}{2}\right)^n$$

□

In fact, any linear combination of these two solutions is also a solution. The following theorem states that this is true in general for linear recurrences.

Theorem 78. *If $f(n)$ and $g(n)$ are solutions to a linear recurrence (without boundary conditions), then $cf(n) + dg(n)$ is also a solution.*

Proof. Since $f(n)$ and $g(n)$ are both solutions, then:

$$\begin{aligned} f(n) &= \sum_{i=1}^{d} a_i f(n-i) \\ g(n) &= \sum_{i=1}^{d} a_i g(n-i) \end{aligned}$$

Multiplying the first equation by c, the second by d, and summing gives:

$$\begin{aligned} cf(n) + dg(n) &= c \cdot \left(\sum_{i=1}^{d} a_i f(n-i)\right) + d \cdot \left(\sum_{i=1}^{d} a_i g(n-i)\right) \\ &= \sum_{i=1}^{d} a_i \big(cf(n-i) + dg(n-i)\big) \end{aligned}$$

Thus, $cf(n) + dg(n)$ is a solution as well. □

This same phenomenon— that a linear combination of solutions is another solution— also arises in differential equations and, consequently, many physical systems. In the present case, the theorem implies that

$$f(n) = c_1 \left(\frac{1+\sqrt{5}}{2}\right)^n + c_2 \left(\frac{1-\sqrt{5}}{2}\right)^n$$

is a solution to the Fibonacci recurrence without boundary conditions for all constants c_1 and c_2. All that remains is to choose these constants to obtain a solution consistent with the boundary conditions, $f(0) = 0$ and $f(1) = 1$. From the first condition, we know:

$$f(0) = c_1 \left(\frac{1+\sqrt{5}}{2}\right)^0 + c_2 \left(\frac{1-\sqrt{5}}{2}\right)^0 = c_1 + c_2 = 0$$

From the second boundary condition, we have:

$$f(1) = c_1 \left(\frac{1+\sqrt{5}}{2}\right)^1 + c_2 \left(\frac{1-\sqrt{5}}{2}\right)^1 = 1$$

We now have two linear equations in two unknowns. The system of equations is not degenerate, so there is a unique solution: $c_1 = 1/\sqrt{5}$ and $c_2 = -1/\sqrt{5}$. We're done! We have a complete solution to the Fibonacci recurrence *with* boundary conditions:

$$f(n) = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2}\right)^n$$

This *looks* completely wrong! All Fibonacci numbers are integers, but this expression is full of square roots of five! Amazingly, however, the square roots always cancel out. This expression really does give the Fibonacci numbers if we plug in $n = 0, 1, 2, \ldots$. It is easy to see why no one stumbled across this solution for six centuries!

13.2.4 Job Prospects

Let's return to the original question: how long until all N faculty positions are taken?

To answer this question, we must find the smallest n such that $f(n) \geq N$; that is, we must determine the year n in which there are as many potential professors as university positions. Graduates after year n will have to find other employment, e.g. shuffling golden disks in an obscure monastic community for the next 10^{19} years.

Because $f(n)$ has such a complicated form, it is hard to compute the right value of n exactly. However, we can find an excellent approximate answer. Note that in the closed

form for Fibonacci numbers, the second term rapidly goes to zero:

$$\begin{aligned} f(n) &= \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n \\ &= \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n + o(1) \end{aligned}$$

This is because $|(1-\sqrt{5})/2| = 0.618\ldots < 1$, and a big power of a number less than 1 is tiny.

From this approximation for $f(n)$, we can estimate the year in which all faculty positions will be filled. That happens when:

$$f(n) \approx \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n \geq N$$

Thus, all jobs are filled in about n years where:

$$\begin{aligned} n &= \frac{\log(\sqrt{5}\,N + o(1))}{\log(\frac{1+\sqrt{5}}{2})} \\ &= \Theta(\log N) \end{aligned}$$

This makes sense, since the number of professors is increasing exponentially. For example, $N = 10,000$ jobs would all be taken in about $n = 20.8$ years. Your TAs don't have a moment to lose!

The solution to the Fibonacci recurrence has an interesting corollary. The number:

$$\gamma = \left(\frac{1+\sqrt{5}}{2}\right)$$

is often called the "Golden Ratio". We can write the dominant term in the closed form for the n-th Fibonacci number in terms of the γ:

$$f(n) = \frac{\gamma^n}{\sqrt{5}} + o(1)$$

We've just shown that this expression involving irrational numbers is actually very close to an integer for all large n— namely, the n-th Fibonacci number. This is just one of many curious properties of the Golden Ratio.

13.3 General Linear Recurrences

The method we used to solve the Fibonacci recurrence can actually be used to solve any linear recurrence. Recall that a recurrence is linear if it has the form:

$$f(n) = a_1 f(n-1) + a_2 f(n-2) + \ldots + a_d f(n-d)$$

Substituting the guess $f(n) = x^n$, as with the Fibonacci recurrence, gives:

$$x^n = a_1 x^{n-1} + a_2 x^{n-2} + \ldots + a_d x^{n-d}$$
$$x^d = a_1 x^{d-1} + a_2 x^{d-2} + \ldots + a_{d-1} x + a_d$$

Dividing the first equation by x^{n-d} gives the second. This second equation is called the ***characteristic equation*** of the recurrence. The characteristic equation can be read off very quickly since the coefficients of the equation are the same as the coefficients of the recurrence.

The solutions to a linear recurrence are defined by the roots of the characteristic equation. Neglecting boundary conditions for the moment:

- If r is a nonrepeated root of the characteristic equation, then r^n is a solution to the recurrence.
- If r is a repeated root with multiplicity k, then

  $$r^n, \quad nr^n, \quad n^2 r^n, \quad \ldots, \quad n^{k-1} r^n$$

  are all solutions to the recurrence.

Futhermore, Theorem 78 implies that every linear combination of these solutions is also a solution.

For example, suppose that the characteristic equation of a recurrence has roots r_1, r_2, and r_3 twice. These four roots imply four distinct solutions:

$$f(n) = r_1^n \qquad f(n) = r_2^n \qquad f(n) = r_3^n \qquad f(n) = nr_3^n$$

Thus, every linear combination

$$f(n) = a \cdot r_1^n + b \cdot r_2^n + c \cdot r_3^n + d \cdot nr_3^n$$

is also a solution.

All that remains is to find a solution consistent with the boundary conditions by choosing the constants appropriately. Each boundary condition implies a linear equation involving these constants. So we can determine the constants by solving a system of linear equations. For example, suppoose our boundary conditions were $f(0) = 0$, $f(1) = 1$, $f(2) = 4$ and $f(3) = 9$. Then we would obtain four equations in four unknowns:

$$\begin{aligned}
f(0) = 0 \quad &\Rightarrow \quad a \cdot r_1^0 + b \cdot r_2^0 + c \cdot r_3^0 + d \cdot 0r_3^0 = 0 \\
f(1) = 1 \quad &\Rightarrow \quad a \cdot r_1^1 + b \cdot r_2^1 + c \cdot r_3^1 + d \cdot 1r_3^1 = 1 \\
f(2) = 4 \quad &\Rightarrow \quad a \cdot r_1^2 + b \cdot r_2^2 + c \cdot r_3^2 + d \cdot 2r_3^2 = 4 \\
f(3) = 9 \quad &\Rightarrow \quad a \cdot r_1^3 + b \cdot r_2^3 + c \cdot r_3^3 + d \cdot 3r_3^3 = 9
\end{aligned}$$

All the nasty r_i^j things are actually just constants. Solving this system gives values for a, b, c, and d that define a solution to the recurrence consistent with the boundary conditions.

13.3.1 An Example

Suppose that there is a type of plant that lives forever, but only reproduces during its first year of life. How fast will the plant population grow? Notice that this is just the reverse of the graduate student job problem where faculty "reproduce" in every year except the first.

Let $f(n)$ be the number of plants in year n. As boundary conditions, define $f(0) = 0$ and $f(1) = 1$. Now the plant population in year n is equal to the population from the year before plus the number of new plants. The population from the year before is $f(n-1)$. And the number of new plants this year is equal to the number of new plants last year, which is $f(n-1) - f(n-2)$. Putting these observations together, we can form a recurrence equation for the plant population:

$$\begin{aligned} f(n) &= f(n-1) + (f(n-1) - f(n-2)) \\ &= 2f(n-1) - f(n-2) \end{aligned}$$

The characteristic equation is $x^2 - 2x + 1 = 0$, which has the single root $x = 1$ with multiplicity 2. Therefore, the solution to the recurrence has the form:

$$\begin{aligned} f(n) &= c_1(1)^n + c_2 n(1)^n \\ &= c_1 + c_2 n \end{aligned}$$

The boundary conditions imply two linear equations in two unknowns:

$$\begin{aligned} f(0) = 0 \quad &\Rightarrow \quad c_1 + c_2(0) = 0 \\ f(1) = 1 \quad &\Rightarrow \quad c_1 + c_2(1) = 1 \end{aligned}$$

The solution to the linear system is $c_1 = 0$, $c_2 = 1$. Therefore, the solution to the recurrence is:

$$\begin{aligned} f(n) &= c_1 + c_2 n \\ &= 0 + (1)n \\ &= n \end{aligned}$$

The answer turns out to be very simple! In year n, there are exactly n plants. Of course, we probably could have solved this problem more easily with guess-and-verify. But, as the Fibonacci recurrence demonstrated, guessing is not always so easy.

13.4 Inhomogeneous Recurrences

We can now solve all recurrences of the form:

$$f(n) = a_1 f(n-1) + a_2 f(n-1) + \ldots + a_d f(n-d)$$

Strictly speaking, this is the family of ***homogeneous*** linear recurrences. Adding an extra, arbitrary function $g(n)$ on the right side gives the general form of an ***inhomogeneous linear recurrence***:

$$f(n) = a_1 f(n-1) + a_2 f(n-1) + \ldots + a_d f(n-d) + g(n)$$

For example, adding $+1$ to the Fibonacci recurrence gives an inhomogeneous linear recurrence:

$$f(n) = f(n-1) + f(n-2) + 1$$

Solving inhomogenous linear recurrences is neither very different nor very difficult. We can divide the whole job into three steps.

1. Replace $g(n)$ by 0 and solve the resulting homogeneous recurrence as before. (Ignore boundary conditions for now; that is, do not solve for constants $c_1, c_2, \ldots, c_d$ yet.) The solution to the homogeneous recurrence is called the ***homogeneous solution***.

2. Now restore $g(n)$ and find a single solution to the recurrence, again ignoring boundary conditions. This is called the ***particular solution***. There are general methods for finding particular solutions, but we advise you to use guess-and-verify. In a moment, we'll explain how to guess wisely.

3. Add the homogeneous and particular solutions together to obtain the ***general solution***. *Now* use the boundary conditions to determine constants by the usual method of generating and solving a system of linear equations.

If you've studied differential equations, then all this probably sounds quite familiar. If you haven't, then— when you *do* get around to studying differential equations— *they* should seem quite familiar.

13.4.1 An Example

Let's demonstrate the method for solving an inhomogenous linear recurrence on this example:

$$\begin{aligned} f(1) &= 1 \\ f(n) &= 4f(n-1) + 3^n \end{aligned}$$

Step 1: Solve the Homogeneous Recurrence

The homogeneous recurrence is $f(n) = 4f(n-1)$. The characteristic equation is $x - 4 = 0$. The only root is $x = 4$. Therefore, the homogeneous solution is $f(n) = c4^n$.

Step 2: Find a Particular Solution

Now we must find a single solution to the full recurrence $f(n) = 4f(n-1) + 3^n$. Let's guess that there is a solution of the form $d3^n$, where d is a constant. Substituting this guess into the recurrence gives:

$$\begin{aligned} d3^n &= 4d3^{n-1} + 3^n \\ 3d &= 4d + 3 \\ d &= -3 \end{aligned}$$

Evidently, $f(n) = -3 \cdot 3^n = -3^{n+1}$ is a particular solution.

Step 3: Add Solutions and Find Constants

We now add the homogeneous solution and the particular solution to obtain the general solution:

$$f(n) = c4^n - 3^{n+1}$$

The boundary condition gives the value of the constant c:

$$\begin{aligned} f(1) = 1 \qquad &\Rightarrow \qquad c4^1 - 3^{1+1} = 1 \\ &\Rightarrow \qquad c = \frac{5}{2} \end{aligned}$$

Therefore, the solution to the recurrence is $f(n) = \frac{5}{2}4^n - 3^{n+1}$. Piece of cake!

Since we could easily have made a mistake, let's check to make sure that our solution at least works for $n = 2$. From the recurrence, $f(2) = 4f(1) + 3^2 = 13$. From our closed form, $f(2) = \frac{5}{2}4^2 - 3^3 = 40 - 27 = 13$. It looks right!

13.4.2 How to Guess a Particular Solution

The hardest part of solving inhomogeneous recurrences is finding a particular solution. This involves guessing, and you might guess wrong. However, some rules of thumb make this job fairly easy most of the time.

- Generally, look for a particular solution with the same form as the inhomogeneous term $g(n)$.
- If $g(n)$ is a constant, then guess a particular solution $f(n) = c$. If this doesn't work, try $f(n) = bn + c$, then $f(n) = an^2 + bn + c$, etc.
- More generally, if $g(n)$ is a polynomial, try a polynomial of the same degree, then a polynomial of degree one higher, then two higher, etc. For example, if $g(n) = 6n+5$, then try $f(n) = bn + c$ and then $f(n) = an^2 + bn + c$.

- If $g(n)$ is an exponential, such as 3^n, then first guess that $f(n) = c3^n$. Failing that, try $f(n) = bn3^n + c3^n$ and then $an^2 3^n + bn3^n + c3^n$, etc.

Final advice: Solving linear and divide-and-conquer recurrences is almost brainless. You just follow the right recipe. However, many students stumble in translating a "word problem" involving professors or plants or Tetris pieces into a recurrence equation. This is a key step for which there is no general recipe. Make sure you think about *where the recurrences come from* in every example in lecture, recitation, and the homework!

Short Guide to Solving Linear Recurrences

A *linear recurrence* is an equation

$$\underbrace{f(n) = a_1 f(n-1) + a_2 f(n-2) + \ldots + a_d f(n-d)}_{\text{homogeneous part}} \quad \underbrace{+g(n)}_{\text{inhomogeneous part}}$$

together with boundary conditions such as $f(0) = b_0$, $f(1) = b_1$, etc.

1. Find the roots of the *characteristic equation*:

$$x^n = a_1 x^{n-1} + a_2 x^{n-2} + \ldots + a_k$$

2. Write down the *homogeneous solution*. Each root generates one term and the homogeneous solution is the sum of these terms. A nonrepeated root r generates the term $c_r r^n$, where c_r is a constant to be determined later. A root r with multiplicity k generates the terms:

$$c_{r_1} r^n, \quad c_{r_2} n r^n, \quad c_{r_3} n^2 r^n, \quad \ldots, \quad c_{r_k} n^{k-1} r^n$$

   where $c_{r_1}, \ldots, c_{r_k}$ are constants to be determined later.

3. Find a *particular solution*. This is a solution to the full recurrence that need not be consistent with the boundary conditions. Use guess and verify. If $g(n)$ is a polynomial, try a polynomial of the same degree, then a polynomial of degree one higher, then two higher, etc. For example, if $g(n) = n$, then try $f(n) = bn + c$ and then $f(n) = an^2 + bn + c$. If $g(n)$ is an exponential, such as 3^n, then first guess that $f(n) = c3^n$. Failing that, try $f(n) = bn3^n + c3^n$ and then $an^2 3^n + bn3^n + c3^n$, etc.

4. Form the *general solution*, which is the sum of the homogeneous solution and the particular solution. Here is a typical general solution:

$$f(n) = \underbrace{c2^n + d(-1)^n}_{\text{homogeneous solution}} + \underbrace{3n+1}_{\text{particular solution}}$$

5. Substitute the boundary conditions into the general solution. Each boundary condition gives a linear equation in the unknown constants. For example, substituting $f(1) = 2$ into the general solution above gives:

$$\begin{aligned} & 2 = c \cdot 2^1 + d \cdot (-1)^1 + 3 \cdot 1 + 1 \\ \Rightarrow \quad & -2 = 2c - d \end{aligned}$$

   Determine the values of these constants by solving the resulting system of linear equations.

Chapter 14

Counting I

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	35748833930586539237113 65	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	95119725587798802882529 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	99132374763417642998139 87
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? We can't, actually. But we'll prove that they 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 also 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.

14.1 Counting One Thing by Counting Another

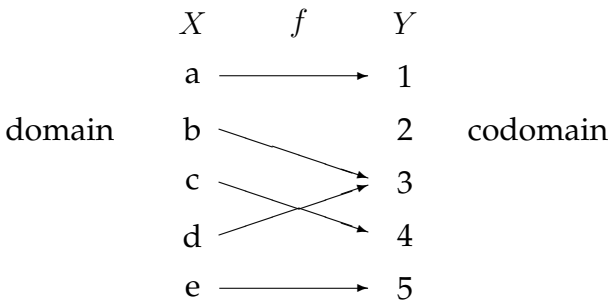
How do you count the number of people in a crowded room? We could count heads, since for each person there is exactly one head. Alternatively, we could count ears and divide by two. Of course, we might have to adjust the calculation if someone lost an ear in a pirate raid or someone was born with three ears. The point here is that we 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.

14.1.1 Functions

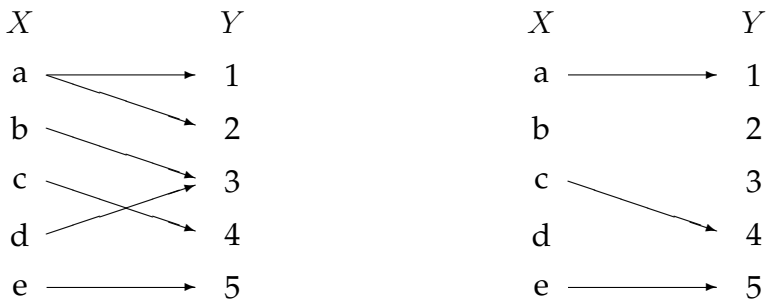
Functions like $f(x) = x^2 + 1$ and $g(x) = \tan^{-1}(x)$ are surely quite familiar from calculus. We're going to use functions for counting as well, but in a way that might not be familiar. Instead of using functions that map one real number to another, we'll use functions that relate elements of finite sets.

In order to count accurately, we need to carefully define the notion of a function. Formally, a ***function*** $f : X \to Y$ is a relation between two sets, X and Y, that relates *every* element of X to *exactly one* element of Y. The set X is called the ***domain*** of the function f, and Y is called the ***codomain***. Here is an illustration of a function:



In this example, the domain is the set $X = \{a, b, c, d, e\}$ and the range is the set $Y = \{1, 2, 3, 4, 5\}$. Related elements are joined by an arrow. This relation is a function because *every* element on the left is related to *exactly one* element on the right. In graph-theoretic terms, this is a function because every element on the left has degree *exactly 1*. If x is an element of the domain, then the related element in the codomain is denoted $f(x)$. We often say f ***maps*** x to $f(x)$. In this case, f maps a to 1, b to 3, c to 4, d to 3 and e to 5. Equivalently, $f(a) = 1$, $f(b) = 3$, $f(c) = 4$, $f(d) = 3$, and $f(e) = 5$.

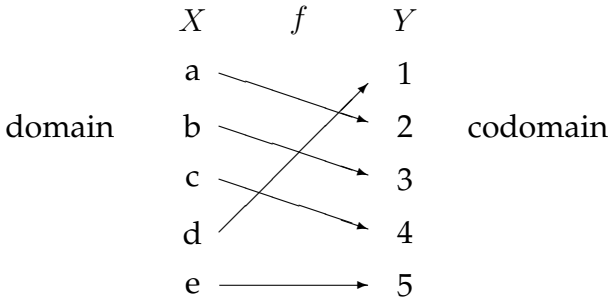
The relations shown below are *not* functions:



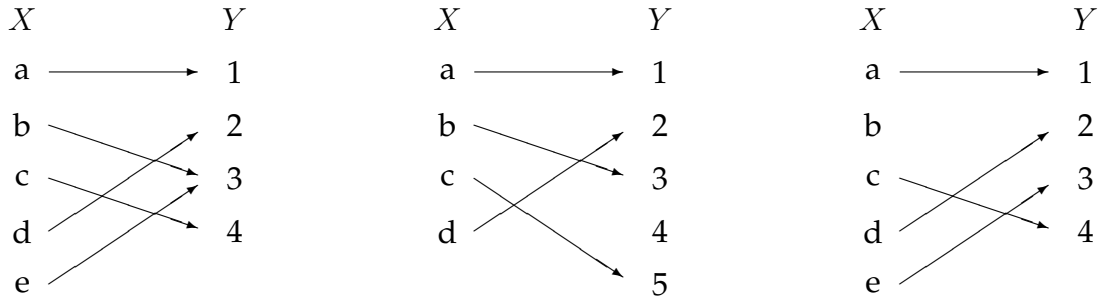
The relation on the left is not a function because a is mapped to *two* elements of the codomain. The relation on the right is not a function because b and d are mapped to *zero* elements of the codomain. Tsk-tsk!

14.1.2 Bijections

The definition of a function is rather asymmetric; there are restrictions on elements in the domain, but not on elements in the codomain. A ***bijection*** or ***bijective function*** is a function $f : X \to Y$ that maps exactly one element of the domain to each element of the codomain. In graph terms, *both* domain and codomain elements are required to have degree exactly 1 in a bijective function. Here is an example of a bijection:



In contrast, these relations that are *not* bijections:



The first function is not bijective because 3 is related to *two* elements of the domain. The second function is not bijective because 4 is related to *zero* elements of the domain. The last relation is not even a function, because b is associated with zero elements of the domain.

Bijective functions are also found in the more-familiar world of real-valued functions. For example, $f(x) = 6x+5$ is a bijective function with domain and codomain $\mathbb{R}$. For every real number y in the codomain, $f(x) = y$ for exactly one real number x in the domain. On the other hand, $f(x) = x^2$ is not a bijective function. The number 4 in the codomain is related to both 2 and -2 in the domain.

14.1.3 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 chocolate, l lemon-filled, s sugar, g glazed, and p plain

to the sequence:

$$\underbrace{0\ldots0}_{c}\;1\;\underbrace{0\ldots0}_{l}\;1\;\underbrace{0\ldots0}_{s}\;1\;\underbrace{0\ldots0}_{g}\;1\;\underbrace{0\ldots0}_{p}$$

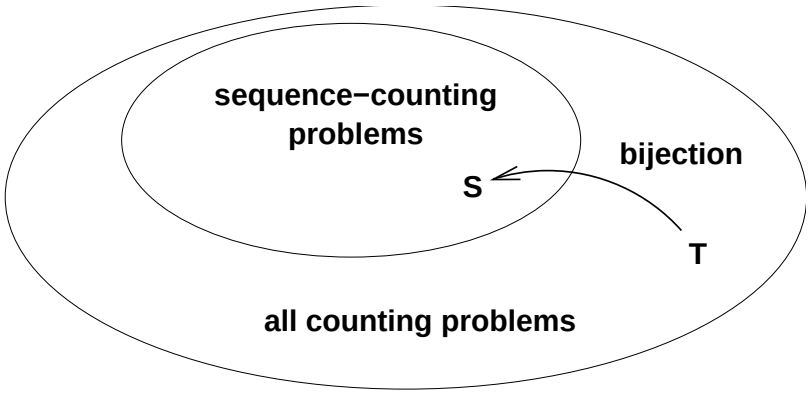
The resulting sequence always has 16 bits and exactly 4 ones, and thus is an element of B. Moreover, the mapping is a bijection; every such bit sequence is mapped to by exactly one order of a dozen doughnuts. Therefore, $|A| = |B|$ by the Bijection Rule!

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 over, and soon you'll consider it boringly routine.

14.1.4 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 precisely 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!

In order to pull this off, we need to clarify some issues concerning sequences and sets. Recall that a ***set*** is an unordered collection of distinct elements. A set is often represented by listing its elements inside curly-braces. For example, $\{a, b, c\}$ is a set, and $\{c, b, a\}$ is another way of writing the same set. On the other hand, $\{a, b, a\}$ is not a set, because element a appears twice.

On the other hand, a ***sequence*** is an ordered collection of elements (called ***components*** or ***terms***) that are not necessarily distinct. A sequence is often written by listing the terms inside parentheses. For example, (a, b, c) is a sequence, and (c, b, a) is a different sequence. Furthermore (a, b, a) is a perfectly valid three-term sequence.

The distinction between sets and sequences is crucial for everything that follows. If you don't keep the distinction clear in your mind, you're doomed!

14.2 Two Basic Counting Rules

We'll harvest our first crop of counting problems with two basic rules.

14.2.1 The Sum Rule

My mother used to have a Peanuts comic strip on her refrigerator. As I recall, Linus had decided to allocate his big sister, Lucy, a quota of 20 crabby days, 40 irritable days, and 60 generally surly days. She immedaiately smacks him in the head and says, "I still have 19 crabby days, all my irritables, and I haven't even touched the generally surly!" I'm not sure what my mother was trying to communicate, but I'd love to find the original comic strip; if anyone finds it, please email me at e_lehman@mit.edu!

Anyway, 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:

$$\begin{aligned} |C \cup I \cup S| &= |C| \cup |I| \cup |S| \\ &= 20 + 40 + 60 \\ &= 120 \text{ days} \end{aligned}$$

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.

14.2.2 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:

$$\begin{aligned} B &= \{\text{pancakes}, \text{bacon and eggs}, \text{bagel}, \text{Doritos}\} \\ L &= \{\text{burger and fries}, \text{garden salad}, \text{Doritos}\} \\ D &= \{\text{macaroni}, \text{pizza}, \text{frozen burrito}, \text{pasta}, \text{Doritos}\} \end{aligned}$$

Then $B \times L \times D$ is the set of all possible daily diets. Here are some sample elements:

$$\begin{gathered} (\text{pancakes}, \text{burger and fries}, \text{pizza}) \\ (\text{bacon and eggs}, \text{garden salad}, \text{pasta}) \\ (\text{Doritos}, \text{Doritos}, \text{frozen burrito}) \end{gathered}$$

The Product Rule tells us how many different daily diets are possible:

$$\begin{aligned} |B \times L \times D| &= |B| \cdot |L| \cdot |D| \\ &= 4 \cdot 3 \cdot 5 \\ &= 60 \end{aligned}$$

14.2.3 Putting Rules Together

Few counting problems can be solved with a single rule. More often, a solution is a flury 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:

$$\begin{array}{cccc} (0,0,0) & (0,0,1) & (0,1,0) & (0,1,1) \\ (1,0,0) & (1,0,1) & (1,1,0) & (1,1,1) \end{array}$$

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.

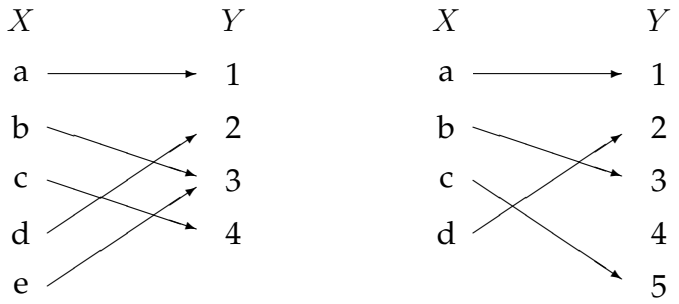
Incidentally, if you're at Brown and you're reading this... Hello! I hope the course is going well for you. Sorry for all the MIT-specific references in these notes. —Eric

14.3 More Functions: Injections and Surjections

Bijective functions are incredibly powerful counting tools. A few other kinds of functions are useful as well; we'll look at two now and one more next time. A function $f : X \to Y$ is:

- ***surjective*** if every element of Y is mapped to *at least once*
- ***injective*** if every element of Y is mapped to *at most once*
- ***bijective*** if every element of Y is mapped to *exactly once*

We've repeated the definition of a bijective function for comparison. Notice that these definitions immediately imply that a function is bijective if and only if it is both injective and surjective. Now the names "surjective" and "injective" are hopelessly unmemorable and nondescriptive. Some people prefer the terms ***onto*** and ***into***, respectively, perhaps on the grounds that these are hopelessly unmemorable and nondescriptive— but shorter.[1] Anyway, here are a couple examples:



The function on the left is surjective (every element on the right is mapped to at least once), but not injective (element 3 is mapped to twice). The function on the right is injective (every element is mapped to at most once), but not surjective (element 4 is mapped to zero times).

Earlier, we observed that two sets are the same size if there is a bijection between them. Similarly, surjections and injections imply certain size relationships between sets.

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 $|X| = |Y|$.*

14.3.1 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?

[1] Some people like the more-descriptive term "one-to-one". Unfortunately, this term describes both injective and bijective functions about equally well. So some people try to *further clarify* matters by saying that a "one-to-one function" is an injective function and a "one-to-one correspondence" is a bijective function. This is probably not going to go down as one of the great terminological triumphs of mathematical history. So, screw it: we're sticking with injective, surjective, and bijective and trusting that you'll put the definitions on your cheat sheet.

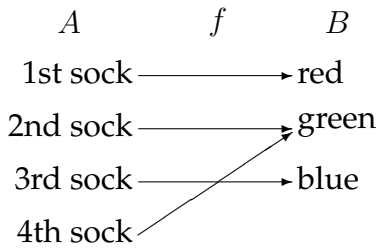
For example, picking out three socks is not enough; you might end up with one red, one green, and one blue. The solution relies on the Pigeonhole Principle, which is a friendly name for the contrapositive of part (2) of the Mapping Rule. Let's write it down:

> If $|X| > |Y|$, then no function $f : X \to Y$ is injective.

Now let's rewrite this a second time to eliminate the word "injective" since, by now, there's not a ghost of a chance that you remember what that means:

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 more than n pigeons fly into n pigeonholes, then at least two pigeons must fly into some hole. In this case, the pigeons form set A, the pigeonholes are set B, and f describes the assignment of pigeons to pigeonholes.

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 I'm talking about non-bald people.

Boston has about 500,000 non-bald peole, 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. I don't know who they are, but I 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 $\{0, 1, \ldots, 90 \cdot 10^{25}\}$, 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***.

> **CONTEST: There is a $100 prize for being the first 6.042 student to actually find two different subsets of the ninety 25-digit numbers on the first page that have the same sum. Send results to e_lehman@mit.edu. The contest expires at the end of the 6.042 final.**

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 $\Omega(2^n)$. He offered $500 to anyone who could prove or disprove his conjecture, but the problem remains unsolved.

Chapter 15

Counting II

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 the student who wrote "I worked alone" on quiz 1.

Olfactory Fixation Award A surprisingly competitive category this term, this goes to the student who comes up with the greatest number of odor-related mathematical examples.

We also considered some less flattering categories such as Proof Most Likely Readable from the Surface of the Moon, Solution Most Closely Resembling a Football Play Diagram with Good Yardage Potential, etc. But then we realized that you all might think up similar "awards" for the course staff and decided to turn the whole matter into a counting problem. In how many ways can, say, three different prizes be awarded to n people?

Remember our basic strategy for counting:

1. Learn to count sequences.
2. Translate everything else into a sequence-counting problem via bijections.

We'll flesh out this strategy considerably today, but the rough outline above is good enough for now.

So we first need to find a bijection that translates 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 $P \times P \times P$. 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). All that remains is to count these sequences. By the Product Rule, we have:

$$\begin{aligned} |P \times P \times P| &= |P| \cdot |P| \cdot |P| \\ &= n^3 \end{aligned}$$

Thus, there are n^3 ways to award the prizes to a class of n people.

15.1 The Generalized Product Rule

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 $(x, y, z) \in P \times P \times P$. 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:

$$S = \{(x, y, z) \in P \times P \times P \mid x, y, \text{and } z \text{ are different people}\}$$

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.

15.1.1 Defective Dollars

A dollar is *defective* 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 Generalized Product Rule. (Alternatively, you could conclude this using the ordinary Product Rule; however, the Generalized Product Rule is strictly more powerful. So you might as well forget the orignial Product Rule now and free up some brain space for 6.002.)

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 there are

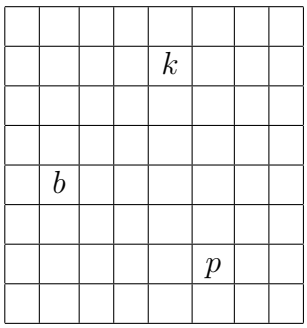
$$\begin{aligned} 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 &= \frac{10!}{2} \\ &= 1,814,400 \end{aligned}$$

serial numbers with all digits different. Plugging these results into the equation above, we find:

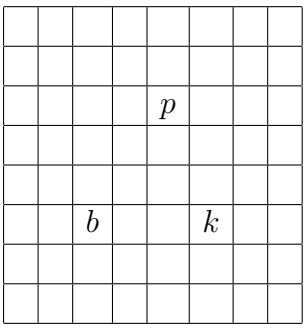
$$\begin{aligned} \text{fraction dollars that are nondefective} &= \frac{1,814,400}{100,000,000} \\ &= 1.8144\% \end{aligned}$$

15.1.2 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 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 $(8 \cdot 7 \cdot 6)^2$.

15.1.3 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.

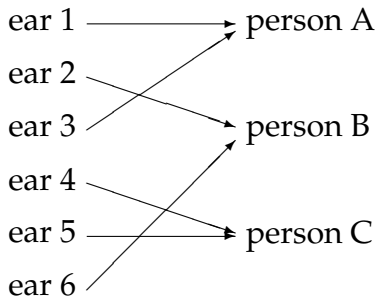
Permutations will come up again in this course approximately 1.6 bazillion times. In fact, permutations are the reason why factorial comes up so often and why we taught you Stirling's approximation:

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

15.2 The Division Rule

We can count the number of people in a room by counting ears and dividing by two. Or we could count the number of fingers and divide by 10. Or we could count the number of fingers and toes and divide by 20. (Someone is probably short a finger or has an extra ear, but let's not worry about that right now.) These observations lead to an important counting rule.

A k***-to-1 function*** maps exactly k elements of the domain to every element of the range. 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 maping each each finger or toe to its owner is 20-to-1. Now just as a bijection implies two sets are the same size, a k-to-1 function implies that the domain is k times larger than the domain:

Rule 8 (Division Rule). *If $f : A \to B$ is k-to-1, then $|A| = k \cdot |B|$.*

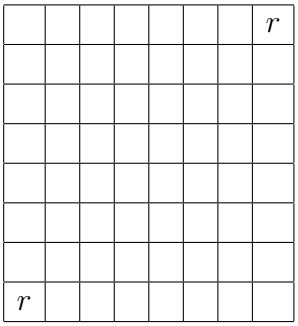
Suppose A is the set of ears in the room and B is the set of people. Since we know there is a 2-to-1 mapping from ears to people, the Division Rule says that $|A| = 2 \cdot |B|$ or, equivalently, $|B| = |A|/2$. Thus, the number of people is half the number of ears.

Now this might seem like a stupid way to count people. But, surprisingly, 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.

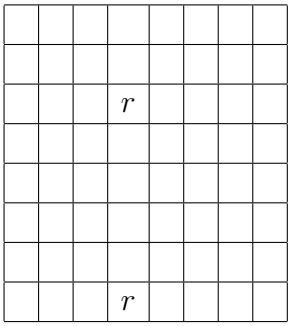
15.2.1 Another Chess Problem

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 map 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:

$$\begin{aligned} |B| &= \frac{|A|}{2} \\ &= \frac{(8 \cdot 7)^2}{2} \end{aligned}$$

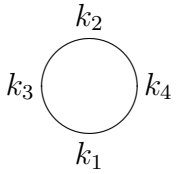
On the second line, we've computed the size of A using the General Product Rule just as in the earlier chess problem.

15.2.2 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:

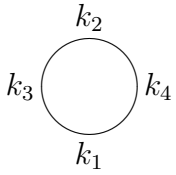
Let A be all the permutations of the knights, and let B be the set of all possible seating arrangements at the round table. We can map each permutaton in set A to a circular seating arrangement in set B by seating the first knight in the permutation anywhere, putting the second knight to his left, the third knight to the left of the second, and so forth all the way around the table. For example:

$$(k_2, k_4, k_1, k_3) \quad \Rightarrow$$



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 arranement:

$$\begin{array}{l}(k_2, k_4, k_1, k_3)\\(k_4, k_1, k_3, k_2)\\(k_1, k_3, k_2, k_4)\\(k_3, k_2, k_4, k_1)\end{array} \quad \Rightarrow$$



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.

15.3 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:

$$|M \cup E \cup P| = |M| + |E| + |P| \qquad \text{(if } M\text{, } E\text{, and } P \text{ are disjoint)}$$

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.

15.3.1 Union of Two Sets

For two sets, S_1 and S_2, the size of the union is given by the following equation:

$$|S_1 \cup S_2| = |S_1| + |S_2| - |S_1 \cap S_2| \tag{15.1}$$

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 prove equation (15.1) rigorously by applying the Sum Rule to some disjoint subsets of $S_1 \cup S_2$. As a first step, we observe that given any two sets, S, T, we can decompose S into the disjoint sets consisting of those elements in S but not T, and those elements in S and also in T. That is, S is the union of the disjoint sets $S - T$ and $S \cap T$. So by the Sum Rule we have

$$\begin{aligned} |S| &= |S - T| + |S \cap T|\,, && \text{and so} \\ |S - T| &= |S| - |S \cap T|\,. \end{aligned} \tag{15.2}$$

Now we decompose $S_1 \cup S_2$ into three disjoint sets:

$$S_1 \cup S_2 = (S_1 - S_2) \cup (S_2 - S_1) \cup (S_1 \cap S_2). \tag{15.3}$$

Now we have

$$\begin{aligned} |S_1 \cup S_2| &= |(S_1 - S_2) \cup (S_2 - S_1) \cup (S_1 \cap S_2)| && \text{(by (15.3))} \\ &= |S_1 - S_2| + |S_2 - S_1| + |S_1 \cap S_2| && \text{(Sum Rule)} \\ &= (|S_1| - |S_1 \cap S_2|) + (|S_2| - |S_1 \cap S_2|) + |S_1 \cap S_2| && \text{(by (15.2))} \\ &= |S_1| + |S_2| - |S_1 \cap S_2| && \text{(algebra)} \end{aligned}$$

15.3.2 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 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 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:

$$|P_{42} \cup P_{04} \cup P_{60}| = 9! + 9! + 9! - 8! - 8! - 8! + 7!$$

15.3.3 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.

15.4 The Grand Scheme for Counting

The rules and techniques we've covered to this point snap together into an overall scheme for solving elementary counting problems. Here it is:

Grand Scheme for Counting

1. Learn to count sequences using two techniques:
   - the General Product Rule
   - the BOOKKEEPER formula
2. Translate everything else to a sequence-counting problem via:
   - bijections
   - k-to-1 functions
3. But for unions of sets, use Inclusion-Exclusion.

Everything here should be familiar to you by now, except for the BOOKKEEPER formula, which you'll see in recitation tomorrow.

The Tao of BOOKKEEPPER

Problem 1. In this problem, we seek enlightenment through contemplation of the word $BOOKKEEPER$.

(a) In how many ways can you arrange the letters in the word $POKE$?

(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.

(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?

(e) In light of the Division Rule, how many arrangements are there of $BOOK$?

(f) Very good, young master! How many arrangements are there of the letters in $KE_1E_2PE_3R$?

(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.

(h) What kind of mapping is this?

(i) So how many arrangements are there of the letters in $KEEPER$?

(j) *Now you are ready to face the BOOKKEEPER!*

How many arrangements of $BO_1O_2K_1K_2E_1E_2PE_3R$ are there?

(k) How many arrangements of $BOOK_1K_2E_1E_2PE_3R$ are there?

(l) How many arrangements of $BOOKKE_1E_2PE_3R$ are there?

(m) How many arrangements of $BOOKKEEPER$ are there?

(n) How many arrangements of $VOODOODOLL$ are there?

(o) (IMPORTANT) How many n-bit sequences contain k zeros and $(n-k)$ ones?

This quantity is denoted $\binom{n}{k}$ and read "n choose k". You will see it almost every day in 6.042 from now until the end of the term.

Remember well what you have learned: subscripts on, subscripts off.

This is the Tao of Bookkeeper.

Chapter 16

Counting III

Today we'll briefly review some facts you dervied in recitation on Friday and then turn to some applications of counting.

16.1 The Bookkeeper Rule

In recitation you learned 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 is a special case of an exceptionally useful counting principle.

Rule 10 (Bookkeeper Rule). *The number of sequences with n_1 copies of l_1, n_2 copies of l_2, ..., and n_k copies of l_k is*

$$\frac{(n_1 + n_2 + \ldots + n_k)!}{n_1!\, n_2!\, \ldots\, n_k!}$$

provided $l_1, \ldots, l_k$ are distinct.

Let's review some applications and implications of the Bookkeeper Rule.

16.1.1 20-Mile Walks

I'm planning a 20 miles 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:

$$\frac{20!}{5!^4}$$

16.1.2 Bit Sequences

How many n-bit sequences contain exactly k ones?

Each such sequence also contains $n - k$ zeroes, so there are

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

by the Bookkeeper Rule.

16.1.3 *k*-element Subsets of an *n*-element Set

How many k-elements 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?

There is a natural bijection between k-element subsets of an n-element set and n-bit sequences with exactly k ones. For example, here is a 3-element subset of $\{x_1, x_2, \ldots, x_8\}$ and the associated 8-bit sequence with exactly 3 ones:

$$\begin{array}{cccccccccc} \{ & x_1, & & & x_4, & x_5 & & & & \} \\ (& 1, & 0, & 0, & 1, & 1, & 0, & 0, & 0 &) \end{array}$$

Therefore, the answer to this problem is the same as the answer to the earlier question about bit sequences.

Rule 11 (Subset Rule). *The number of k-element subsets of an n-element set is:*

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

The factorial expression in the Subset Rule comes up so often that there is a shorthand, $\binom{n}{k}$. This is read "n choose k" since it denotes the number of ways to choose k items from among n. We can immediately knock off all three questions above using the Sum Rule:

- 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.

The k-element subsets of an n-element set are sometimes called ***k-combinations***. There are a great many similar-sounding terms: permutations, r-permutations, permutations with repetition, combinations with repetition, permutations with indistinguishable objects, and so on. For example, the Bookkeeper rule is elsewhere called the "formula for permutations with indistinguishable objects". We won't drag you through all this terminology other than to say that, broadly speaking, all the terms mentioning *permutations* concern sequences and all terms mentioning *combinations* concern subsets.

16.1.4 An Alternative Derivation

Let's derive the Subset Rule another way to gain an alternative perspective. The number of *sequences* consisting of k distinct elements drawn from an n-element set is

$$n \cdot (n-1) \cdot (n-2) \cdots (n-k+1) = \frac{n!}{(n-k)!}$$

by the Generalized Product Rule. Now suppose we map each sequence to the set of elements it contains. For example:

$$(x_1, x_2, x_3) \qquad \to \qquad \{x_1, x_2, x_3\}$$

This is a $k!$-to-1 mapping since each k-element set is mapped to by all of its $k!$ permutations. Thus, by the Quotient Rule, the number of k-element subsets of an n-element set is:

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

16.1.5 Word of Caution

Someday you might refer to 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 just 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???*"

16.2 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 let's consider a postive, integral power of a binomial:

$$(a+b)^n$$

Suppose we multiply out this expression completely for, say, $n = 4$:

$$\begin{aligned}(a+b)^4 = {} & aaaa + aaab + aaba + aabb \\ & + abaa + abab + abba + abbb \\ & + baaa + baab + baba + babb \\ & + bbaa + bbab + bbba + bbbb\end{aligned}$$

Notice that there is one term for every sequence of a's and b's. Therefore, 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 79 (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.

16.3 Poker Hands

There are 52 cards in a deck. Each card has a *suit* and a *value*. There are four suits:

spades ♠ hearts ♡ clubs ♣ 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.

16.3.1 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 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:

$$\begin{array}{lcllllll} (8, Q, \heartsuit) & \leftrightarrow & \{ & 8\spadesuit, & 8\diamondsuit, & 8\heartsuit, & 8\clubsuit, & Q\heartsuit & \} \\ (2, A, \clubsuit) & \leftrightarrow & \{ & 2\clubsuit, & 2\heartsuit, & 2\diamondsuit, & 2\spadesuit, & A\clubsuit & \} \end{array}$$

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!

16.3.2 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:

$$\begin{array}{llllll} \{ & 2\spadesuit, & 2\clubsuit, & 2\diamondsuit, & J\clubsuit, & J\diamondsuit & \} \\ \{ & 5\diamondsuit, & 5\clubsuit, & 5\heartsuit, & 7\heartsuit, & 7\clubsuit & \} \end{array}$$

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.

16.3.3 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:

$$\begin{array}{lcl} (3, \{\diamondsuit, \spadesuit\}, Q, \{\diamondsuit, \heartsuit\}, A, \clubsuit) & \searrow & \\ & & \{\ 3\diamondsuit,\ \ 3\spadesuit,\ \ Q\diamondsuit,\ \ Q\heartsuit,\ \ A\clubsuit\ \} \\ (Q, \{\diamondsuit, \heartsuit\}, 3, \{\diamondsuit, \spadesuit\}, A, \clubsuit) & \nearrow & \\ & & \\ (9, \{\heartsuit, \diamondsuit\}, 5, \{\heartsuit, \clubsuit\}, K, \spadesuit) & \searrow & \\ & & \{\ 9\heartsuit,\ \ 9\diamondsuit,\ \ 5\heartsuit,\ \ 5\clubsuit,\ \ K\spadesuit\ \} \\ (5, \{\heartsuit, \clubsuit\}, 9, \{\heartsuit, \diamondsuit\}, K, \spadesuit) & \nearrow & \end{array}$$

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 and 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 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.

16.3.4 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:

$$(7, K, A, 2, \diamondsuit, 3) \ \leftrightarrow\ \{\ 7\diamondsuit,\ \ K\clubsuit,\ \ A\heartsuit,\ \ 2\spadesuit,\ \ 3\diamondsuit\ \}$$

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:

$$\frac{13^4 \cdot 4 \cdot 12}{2}$$

16.4 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 reveals four of them to the Magician, one at a time. The Magician concentrates for a short time and then correctly names the secret, fifth card!

16.4.1 The Secret

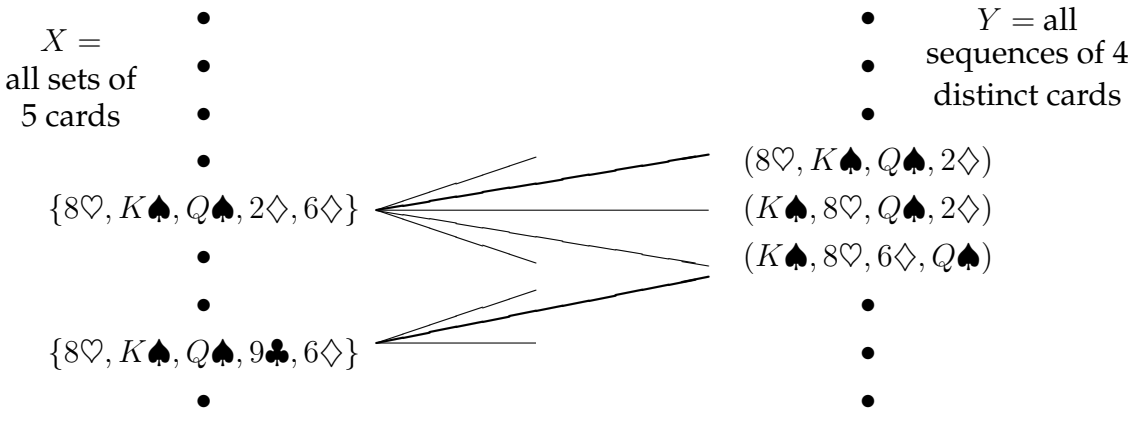
The Assistant somehow communicated the secret card to the Magician just by naming the other four cards. In particular, the Assistant has two ways to communicate:

1. He can announce the four cards in any order. The number of orderings of four cards is $4! = 24$, so this alone is insufficient to identify which of the remaining 48 cards is the secret one.

2. The Assistant can choose which four of the five cards to reveal. Of course, the Magician can not determine which of these five possibilities the Assistant selected since he does not know the secret card.

Nevertheless, these two forms of communication allow the Assistant to covertly reveal the secret card to the Magician.

Our counting tools give a lot of insight into the magic trick. 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.



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 $(K\spadesuit, 8\heartsuit, 6\diamondsuit, Q\spadesuit)$.

Let's think about this problem in terms of graphs. Regard the elements of X and Y as the vertices of a bipartite graph. Put 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 above.

What we need to perform the trick is a *matching* for the X vertices; that is, we need a subset of edges that join every vertex on the left to exactly one, distinct vertex on the right. If such a matching exists, then the Assistant and Magician can agree one in advance. Then, when the audience selects a set of 5 cards, the Assistant reveals the corresponding sequence of 4 cards. The Magician translates back to the correspoding set of 5 cards and names 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 $(K\spadesuit, 8\heartsuit, 6\diamondsuit, Q\spadesuit)$. The Magician names the one card in the corresponding set not already revealed, the $9\clubsuit$. Notice that the sets must be matched with *distinct* sequences; if the Assistant revealed the same sequence when the audience picked the set $\{8\heartsuit, K\spadesuit, Q\spadesuit, 2\diamondsuit, 6\diamondsuit\}$, then the Magician would be unable to determine whether the remaining card was the $9\clubsuit$ or $2\diamondsuit$!

The only remaining question is whether a matching for the X vertices exists. This is precisely the subject of Hall's Theorem. Regard the X vertices as girls, the Y vertices as boys, and each edge as an indication that a girl likes a boy. Then a matching for the girls exists if and only if the marriage condition is satisfied:

Every subset of girls likes at least as large a set of boys.

Let's prove that the marriage condition holds for the magic trick graph. We'll need a couple preliminary facts:

- 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 terms of the marriage metaphor, every girl like 120 boys.
- Each vertex on the right has degree 48, since there are 48 possibilities for the fifth card. Thus, every boy is liked by 48 girls.

Now let S be an arbitrary set of vertices on the left, which we're regarding as girls. There are $120\,|S|$ edges incident to vertices in this set. Since each boy is liked by at most 48 girls, this set of girls likes at least $120\,|S|\,/48 \geq |S|$ different boys. Thus, the marriage condition is satisfied, a matching exists by Hall's Theorem, and the trick can be done without magic!

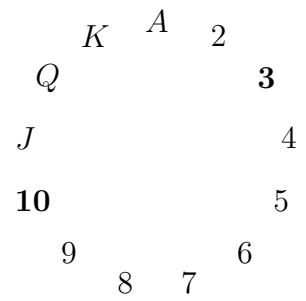
16.4.2 The Real Secret

You might not find the preceding answer very satisfying. After all, as a practical matter, the Assistant and the Magician can not memorize a matching containing $\binom{52}{5} = 2,598,960$ edges! The remaining challenge is to choose a matching that can be readily computed on the fly. We'll describe one approach. As an 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:

$$A\clubsuit\ 2\clubsuit\ \ldots\ K\clubsuit\ A\diamondsuit\ 2\diamondsuit\ \ldots\ Q\diamondsuit\ A\heartsuit\ 2\heartsuit\ \ldots\ Q\heartsuit\ A\spadesuit\ 2\spadesuit\ \ldots\ Q\spadesuit$$

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!

16.4.3 Same Trick with Four Cards?

Suppose that the audience selects only *four* cards and the Assistant reviews 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, one on 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 hears that sequence of three, then there are at least two possibilities for the fourth card which he cannot distinguish!

16.5 Combinatorial Proof

Suppose you have n different T-shirts 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.

16.5.1 Boxing

Flo, 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*?!" Flo figures that n people (including himself) are competing for spots on the team and only k will be selected. Thus, there are two cases to consider:

- Flo 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 be formed in this way is:
$$\binom{n-1}{k-1}$$
- Flo 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 Flo, 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:

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

Claire, equally-famed 6.042 TA, thinks Flo isn't so tough and so she might as well try out also. She reasons that n people (including herself) are trying out for k spots. Thus, the number of ways to select the team is simply:

$$\binom{n}{k}$$

Claire and Flo each correctly counted the number of possibly 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.

16.5.2 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 United States Olympic boxing teams. Flo computed $|S| = \binom{n-1}{k-1} + \binom{n-1}{k}$ by counting one way, and Claire 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, invented story. (You probably realized this was invention; after all, Flo is a French citizen and thus wouldn't be *eligible* for the US team.) Here is less-colorful example of a combinatorial argument.

Theorem 80.

$$\sum_{r=0}^{n} \binom{n}{r}\binom{2n}{n-r} = \binom{3n}{n}$$

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 80 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 80 is $\binom{3n}{n}$, which suggests choosing S to be all n-element subsets of some $3n$-element set.

Chapter 17

Generating Functions

Generating functions are one of the most surprising, useful, and clever inventions in discrete math. Roughly speaking, generating functions transform problems about *sequences* into problems about *real-valued functions*. This is great because we've got piles of mathematical machinery for manipulating real-valued 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 problem. There is a huge chunk of mathematics concerning generating functions, so we will only get a taste of the subject.

In this lecture, we'll put sequences in angle brackets to more clearly distinguish them from the many other mathemtical expressions floating around.

17.1 Generating Functions

The ***ordinary generating function (OGF)*** for the infinite sequence $\langle g_0, g_1, g_2, g_3 \dots \rangle$ is the formal power series:

$$G(x) = g_0 + g_1 x + g_2 x^2 + g_3 x^3 + \cdots$$

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 let x be a real number and actually evaluate a generating function, so we can largely forget about questions of convergence. Not all generating functions are ordinary, but those are the only kind we'll consider here.

Throughout the lecture, 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:

$$\begin{aligned}
\langle 0,0,0,0,\ldots\rangle &\longleftrightarrow 0+0x+0x^2+0x^3+\cdots=0\\
\langle 1,0,0,0,\ldots\rangle &\longleftrightarrow 1+0x+0x^2+0x^3+\cdots=1\\
\langle 3,2,1,0,\ldots\rangle &\longleftrightarrow 3+2x+1x^2+0x^3+\cdots=3+2x+x^2
\end{aligned}$$

The pattern here is simple: the i-th 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 once again we won't worry about convergence issues. This formula gives closed-form generating functions for a whole range of sequences. For example:

$$\begin{aligned}
\langle 1,1,1,1,\ldots\rangle &\longleftrightarrow 1+x+x^2+x^3+\cdots &&=\frac{1}{1-x}\\
\langle 1,-1,1,-1,\ldots\rangle &\longleftrightarrow 1-x+x^2-x^3+x^4-\cdots &&=\frac{1}{1+x}\\
\langle 1,a,a^2,a^3,\ldots\rangle &\longleftrightarrow 1+ax+a^2x^2+a^3x^3+\cdots &&=\frac{1}{1-ax}\\
\langle 1,0,1,0,1,0,\ldots\rangle &\longleftrightarrow 1+x^2+x^4+x^6+\cdots &&=\frac{1}{1-x^2}
\end{aligned}$$

17.2 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.

17.2.1 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 12 (Scaling Rule). *If*

$$\langle f_0, f_1, f_2, \ldots \rangle \longleftrightarrow F(x),$$

then

$$\langle cf_0,\ cf_1,\ cf_2,\ \ldots \rangle \longleftrightarrow c \cdot F(x).$$

Proof.

$$\begin{array}{rcl} \langle cf_0, cf_1, cf_2, \ldots \rangle & \longleftrightarrow & cf_0 + cf_1 x + cf_2 x^2 + \cdots \\ & = & c \cdot (f_0 + f_1 x + f_2 x^2 + \cdots) \\ & = & cF(x) \end{array}$$

□

17.2.2 Addition

Adding generating functions corresponds to adding the two sequences term by term. For example, adding two of our earlier examples gives:

$$\begin{array}{rccccccccccc} & \langle & 1, & 1, & 1, & 1, & 1, & 1, & \ldots & \rangle & \longleftrightarrow & \dfrac{1}{1-x} \\ + & \langle & 1, & -1, & 1, & -1, & 1, & -1, & \ldots & \rangle & \longleftrightarrow & \dfrac{1}{1+x} \\ \hline & \langle & 2, & 0, & 2, & 0, & 2, & 0, & \ldots & \rangle & \longleftrightarrow & \dfrac{1}{1-x} + \dfrac{1}{1+x} \end{array}$$

We've now derived two different expressions that both generate the sequence $\langle 2, 0, 2, 0, \ldots \rangle$. Not surprisingly, they turn out to be equal:

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

Rule 13 (Addition Rule). *If*

$$\begin{array}{ll} \langle f_0, f_1, f_2, \ldots \rangle \longleftrightarrow F(x), & \textit{and} \\ \langle g_0, g_1, g_2, \ldots \rangle \longleftrightarrow G(x), & \end{array}$$

then

$$\langle f_0 + g_0,\ f_1 + g_1,\ f_2 + g_2,\ \ldots \rangle \longleftrightarrow F(x) + G(x).$$

Proof.

$$\begin{aligned}\langle f_0+g_0,\ f_1+g_1,\ f_2+g_2,\ \dots\rangle &\longleftrightarrow \sum_{n=0}^{\infty}(f_n+g_n)x^n\\ &= \left(\sum_{n=0}^{\infty}f_nx^n\right)+\left(\sum_{n=0}^{\infty}g_nx^n\right)\\ &= F(x)+G(x)\end{aligned}$$

□

17.2.3 Right Shifting

Let's start over again with a simple sequence and its generating function:

$$\langle 1,1,1,1,\dots\rangle \longleftrightarrow \frac{1}{1-x}$$

Now let's *right-shift* the sequence by adding k leading zeros:

$$\begin{aligned}\langle\underbrace{0,0,\dots,0}_{k\text{ zeroes}},1,1,1,\dots\rangle &\longleftrightarrow x^k+x^{k+1}+x^{k+2}+x^{k+3}+\cdots\\ &= x^k\cdot(1+x+x^2+x^3+\cdots)\\ &= \frac{x^k}{1-x}\end{aligned}$$

Evidently, adding k leading zeros to the sequence corresponds to multiplying the generating function by x^k. This holds true in general.

Rule 14 (Right-Shift Rule). *If* $\langle f_0,f_1,f_2,\dots\rangle \longleftrightarrow F(x)$*, then:*

$$\langle\underbrace{0,0,\dots,0}_{k\text{ zeroes}},f_0,f_1,f_2,\dots\rangle \longleftrightarrow x^k\cdot F(x)$$

Proof.

$$\begin{aligned}\langle\overbrace{0,0,\dots,0}^{k\text{ zeroes}},f_0,f_1,f_2,\dots\rangle &\longleftrightarrow f_0x^k+f_1x^{k+1}+f_2x^{k+2}+\cdots\\ &= x^k\cdot(f_0+f_1x+f_2x^2+f_3x^3+\cdots)\\ &= x^k\cdot F(x)\end{aligned}$$

□

17.2.4 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.

$$\begin{aligned} \frac{d}{dx}\left(1 + x + x^2 + x^3 + x^4 + \cdots\right) &= \frac{d}{dx}\left(\frac{1}{1-x}\right) \\ 1 + 2x + 3x^2 + 4x^3 + \cdots &= \frac{1}{(1-x)^2} \\ \langle 1, 2, 3, 4, \ldots \rangle &\longleftrightarrow \frac{1}{(1-x)^2} \end{aligned}$$

We found a generating function for the sequence $\langle 1, 2, 3, 4, \ldots \rangle$!

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 15 (Derivative Rule). *If*

$$\langle f_0, f_1, f_2, f_3, \ldots \rangle \longleftrightarrow F(x),$$

then

$$\langle f_1, 2f_2, 3f_3, \ldots \rangle \longleftrightarrow F'(x).$$

Proof.

$$\begin{aligned} \langle f_1, 2f_2, 3f_3, \ldots \rangle &= f_1 + 2f_2 x + 3f_3 x^2 + \cdots \\ &= \frac{d}{dx}\left(f_0 + f_1 x + f_2 x^2 + f_3 x^3 + \cdots\right) \\ &= \frac{d}{dx} F(x) \end{aligned}$$

□

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, $\langle 0, 1, 4, 9, 16, \ldots \rangle$. If we could start with the sequence $\langle 1, 1, 1, 1, \ldots \rangle$ and multiply each term by its index two times, then we'd have the desired result:

$$\langle 0 \cdot 0,\ 1 \cdot 1,\ 2 \cdot 2,\ 3 \cdot 3,\ \ldots \rangle = \langle 0, 1, 4, 9, \ldots \rangle$$

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 14 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 $\langle 1, 1, 1, 1, \ldots \rangle$, differentiate, multiply by x, and then differentiate and multiply by x once more.

$$\begin{aligned}
\langle 1, 1, 1, 1, \ldots \rangle &\longleftrightarrow \frac{1}{1-x} \\
\langle 1, 2, 3, 4, \ldots \rangle &\longleftrightarrow \frac{d}{dx}\frac{1}{1-x} = \frac{1}{(1-x)^2} \\
\langle 0, 1, 2, 3, \ldots \rangle &\longleftrightarrow x \cdot \frac{1}{(1-x)^2} = \frac{x}{(1-x)^2} \\
\langle 1, 4, 9, 16, \ldots \rangle &\longleftrightarrow \frac{d}{dx}\frac{x}{(1-x)^2} = \frac{1+x}{(1-x)^3} \\
\langle 0, 1, 4, 9, \ldots \rangle &\longleftrightarrow x \cdot \frac{1+x}{(1-x)^3} = \frac{x(1+x)}{(1-x)^3}
\end{aligned}$$

Thus, the generating function for squares is:

$$\frac{x(1+x)}{(1-x)^3}$$

17.3 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 are 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 n-Fibonacci number. Of course, we already *have* a closed form for Fibonacci numbers, obtained from the cookbook procedure for solving linear recurrences. But there are a couple reasons to cover the same ground again. First, we'll gain some insight into why the cookbook method for linear recurrences works. And, second, the techniques we'll use are applicable to a large class of recurrence equations, including some that we have no other way to tackle.

17.3.1 Finding a Generating Function

Let's begin by recalling the definition of the Fibonacci numbers:

$$\begin{aligned}
f_0 &= 0 \\
f_1 &= 1 \\
f_n &= f_{n-1} + f_{n-2} \qquad \text{(for } n \geq 2\text{)}
\end{aligned}$$

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:

$$\begin{array}{ccccccccccccc} F(x) & = & f_0 + & & f_1 & x + & & f_2 & x^2 + & & f_3 & x^3 + & f_4 & x^4 + \ldots \\ \| & & \| & & \| & & & \| & & & \| & & \| & \\ x + xF(x) + x^2F(x) & = & 0 & + & (1+f_0)\,x & + & & (f_1+f_0)\,x^2 & + & & (f_2+f_1)\,x^3 & + & (f_3+f_2)\,x^4 & + \cdots \end{array}$$

Solving for $F(x)$ gives the generating function for the Fibonacci sequence:

$$\begin{aligned} & F(x) = x + xF(x) + x^2F(x) \\ \Rightarrow \qquad & F(x) = \frac{x}{1 - x - x^2} \end{aligned}$$

Sure enough, this is the simple generating function we claimed at the outset!

17.3.2 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 n-th 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 n-th 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 fractions 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:

$$\begin{aligned} F(x) &= \frac{1}{\sqrt{5}}\left(\frac{1}{1-\alpha_1 x} - \frac{1}{1-\alpha_2 x}\right) \\ &= \frac{1}{\sqrt{5}}\left((1+\alpha_1 x+\alpha_1^2 x^2+\cdots) - (1+\alpha_2 x+\alpha_2^2 x^2+\cdots)\right) \\ \Rightarrow f_n &= \frac{\alpha_1^n - \alpha_2^n}{\sqrt{5}} \\ &= \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right) \end{aligned}$$

This is the same scary formula for the n-th Fibonacci number that we found using the method for solving linear recurrences. And this alternate approach sheds some light on that method. In particular, the strange rules involving repeated roots of the characteristic equation are reflections of the rules for finding a partial fractions expansion!

17.4 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. When generating functions are used in this way, the coefficient of x^n is the number of ways to choose n items.

17.4.1 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 k-element set. For example, the coefficient of x^2 is $\binom{k}{2}$, the number of ways to choose 2 items from a k-element set. Similarly, the coefficient of x^{k+1} is the number of ways to choose $k+1$ items from a k-element set, which is zero.

17.4.2 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 subset 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 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 16 (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.)

Proof. Define:

$$A(x) = \sum_{n=0}^{\infty} a_n x^n, \qquad B(x) = \sum_{n=0}^{\infty} b_n x^n, \qquad C(x) = A(x) \cdot B(x) = \sum_{n=0}^{\infty} c_n x^n.$$

Let's first evaluate the product $A(x) \cdot B(x)$ and express the coefficient c_n in terms of the a and b coefficients. We can tabulate all of the terms in this product in a table:

	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:

$$c_n = a_0b_n + a_1b_{n-1} + a_2b_{n-2} + \cdots + a_nb_0$$

Now we must show that this is also the number of ways to select n items from $\mathcal{A} \cup \mathcal{B}$. In general, we can select a total of n items from $\mathcal{A} \cup \mathcal{B}$ 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_jb_{n-j} ways. Summing over all the possible values of j gives a total of

$$a_0b_n + a_1b_{n-1} + a_2b_{n-2} + \cdots + a_nb_0$$

ways to select n items from $\mathcal{A} \cup \mathcal{B}$. This is precisely the value of c_n computed above. □

The expression $c_n = a_0b_n + a_1b_{n-1} + a_2b_{n-2} + \cdots + a_nb_0$ may be familiar from a signal processing course; the sequence $\langle c_0, c_1, c_2, \ldots \rangle$ is the ***convolution*** of sequences $\langle a_0, a_1, a_2, \ldots \rangle$ and $\langle b_0, b_1, b_2, \ldots \rangle$.

17.4.3 Choosing Items with Repetition

The first counting problem we considered asked for the number of ways to select a dozen doughnuts when there were five varieties available. We can generalize this question as follows: in how many ways can we select k items from an n-element set if we're allowed

to pick the same item multiples times? In these terms, the doughnut problem asks in how many ways we can select a dozen doughnuts from the set:

$$\{\text{chocolate}, \text{lemon-filled}, \text{sugar}, \text{glazed}, \text{plain}\}$$

if we're allowed to pick several doughnuts of the same variety. 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:

$$\begin{aligned} \langle 1, 1, 1, 1, \ldots \rangle &\quad\longleftrightarrow\quad 1 + x + x^2 + x^3 + \cdots \\ &\quad=\quad \frac{1}{1-x} \end{aligned}$$

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:

$$\underbrace{\frac{1}{1-x}}_{\substack{\text{OGF for} \\ \{a_1\}}} \cdot \underbrace{\frac{1}{1-x}}_{\substack{\text{OGF for} \\ \{a_2\}}} \cdots \underbrace{\frac{1}{1-x}}_{\substack{\text{OGF for} \\ \{a_n\}}} = \underbrace{\frac{1}{(1-x)^n}}_{\substack{\text{OGF for} \\ \{a_1, a_2, \ldots, a_n\}}}$$

Therefore, the generating function for selecting items from a n-element set with repetition allowed is $1/(1-x)^n$.

Now we need to find the coefficients of this generating function. We could try to use partial fractions, but $(1-x)^n$ has a nasty repeated root at 1. An alternative is to use Taylor's Theorem:

Theorem 81 (Taylor's Theorem).

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots + \frac{f^{(k)}(0)}{k!}x^k + \cdots$$

This theorem says that the k-th coefficient of $1/(1-x)^n$ is equal to its k-th derivative evaluated at 0 and divided by $k!$. And computing the k-th derivative turns out not to be very difficult. Let

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

Then we have:

$$\begin{aligned} G'(x) &= n(1-x)^{-n-1} \\ G''(x) &= n(n+1)(1-x)^{-n-2} \\ G'''(x) &= n(n+1)(n+2)(1-x)^{-n-3} \\ G^{(k)}(x) &= n(n+1)\cdots(n+k-1)(1-x)^{-n-k} \end{aligned}$$

Thus, the coefficient of x^k in the generating function is:

$$\begin{aligned} G^{(k)}(0)/k! &= \frac{n(n+1)\cdots(n+k-1)}{k!} \\ &= \frac{(n+k-1)!}{(n-1)!\,k!} \\ &= \binom{n+k-1}{k} \end{aligned}$$

Therefore, the number of ways to select k items from an n-element set with repetition allowed is:

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

This makes sense, since there is a bijection between such selections and $(n+k-1)$-bit sequences with k zeroes (representing the items) and $n-1$ ones (separating the n different types of item).

17.5 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 apples in zero ways (since the number of apples must be

even), a set of 2 applies 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 ways, 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!*

Chapter 18

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.

18.1 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 embarassingly so.

The way to avoid errors is to distrust informal arguments and rely instead on a rigorous, systematic approach. In short: intuition *bad*, formalism *good*. If you insist on relying on intuition, then there are lots of compelling financial deals we'd love to offer you!

18.1.1 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!

18.1.2 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?"

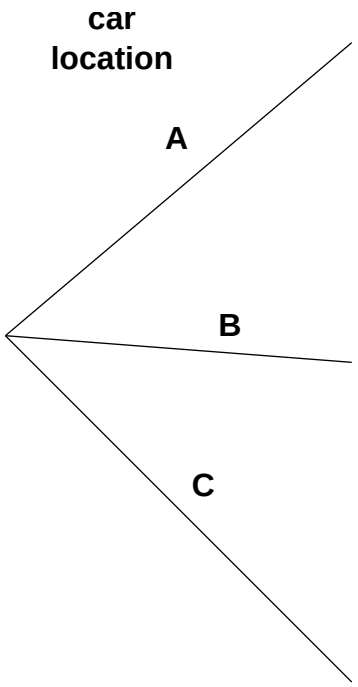
18.1.3 Step 1: Find the Sample Space

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:

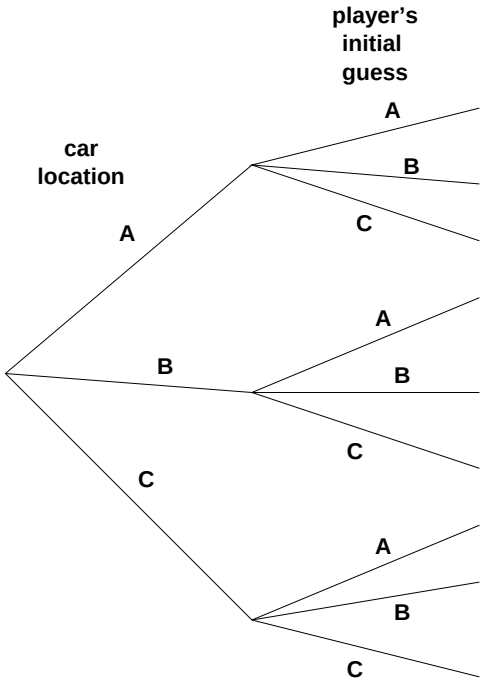
1. The door concealing the car.
2. The door initially chosen by the player.
3. The door that the host opens to reveal a goat.

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.

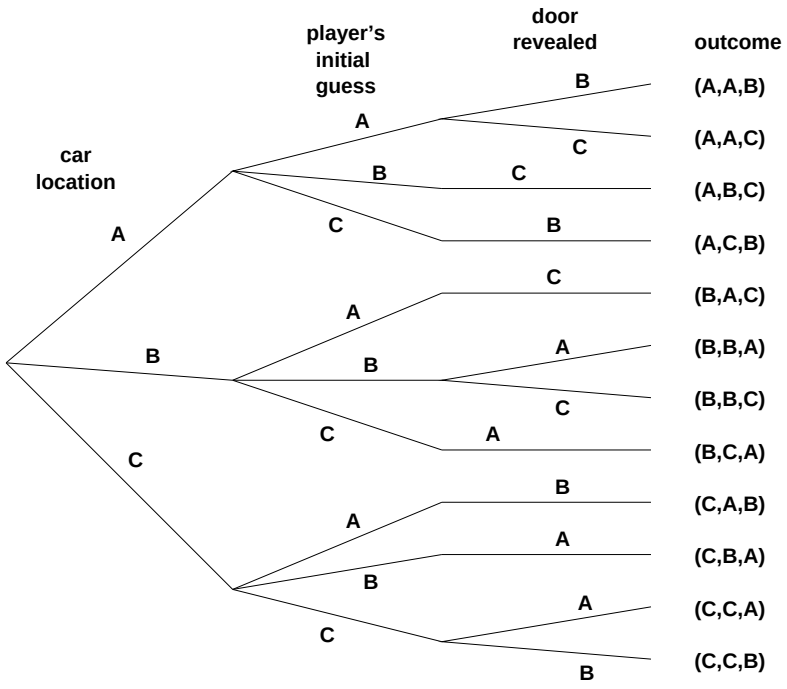
A ***tree diagram*** is a graphical tool that can help us work through the four-step approach when the number of outcomes is not too large or the problem is nicely structured. In particular, we can use a tree diagram to help understand the sample space of an experiment. The first randomly-determined 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 chose any of the three doors. We represent this by adding a second layer to the tree:



Finally, the host opens a door to reveal a goat. 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. All of these possibilities are worked out in a third layer of the tree:



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:

$$S = \left\{ \begin{array}{llllll} (A,A,B), & (A,A,C), & (A,B,C), & (A,C,B), & (B,A,C), & (B,B,A), \\ (B,B,C), & (B,C,A), & (C,A,B), & (C,B,A), & (C,C,A), & (C,C,B) \end{array} \right\}$$

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.

18.1.4 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:

$$\{(C, A, B), (C, B, A), (C, C, A), (C, C, B)\}$$

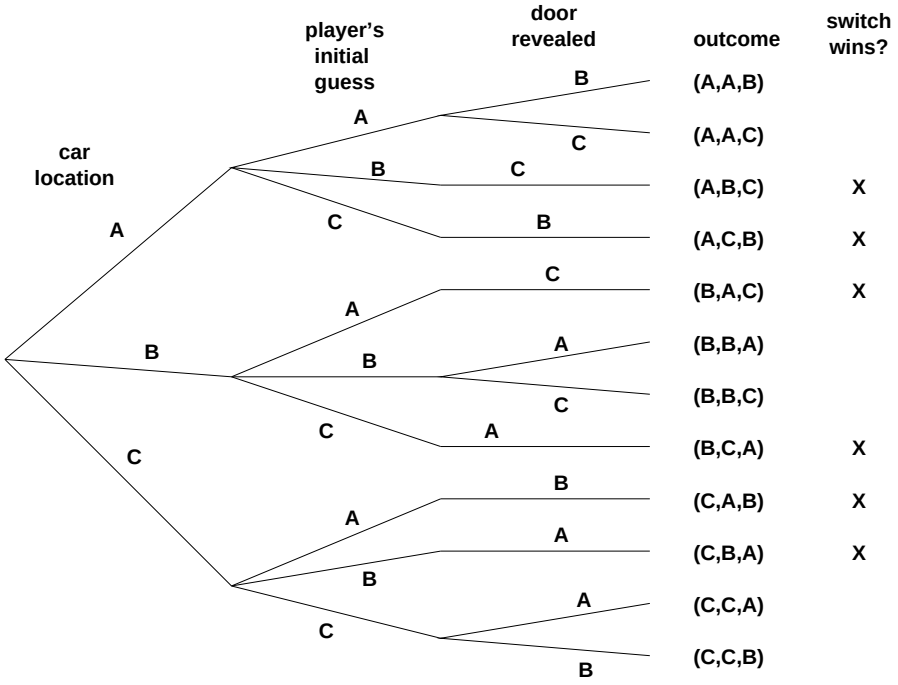
The event that the player initially picked the door concealing the prize is the set of outcomes:

$$\{(A, A, B), (A, A, C), (B, B, A), (B, B, C), (C, C, A), (C, C, B)\}$$

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 annonate 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 $\frac{1}{2}$. *This is wrong.* The reason is that these outcomes are not all equally likely, as we'll see shortly.

18.1.5 Step 3: Determine Outcome Probabilities

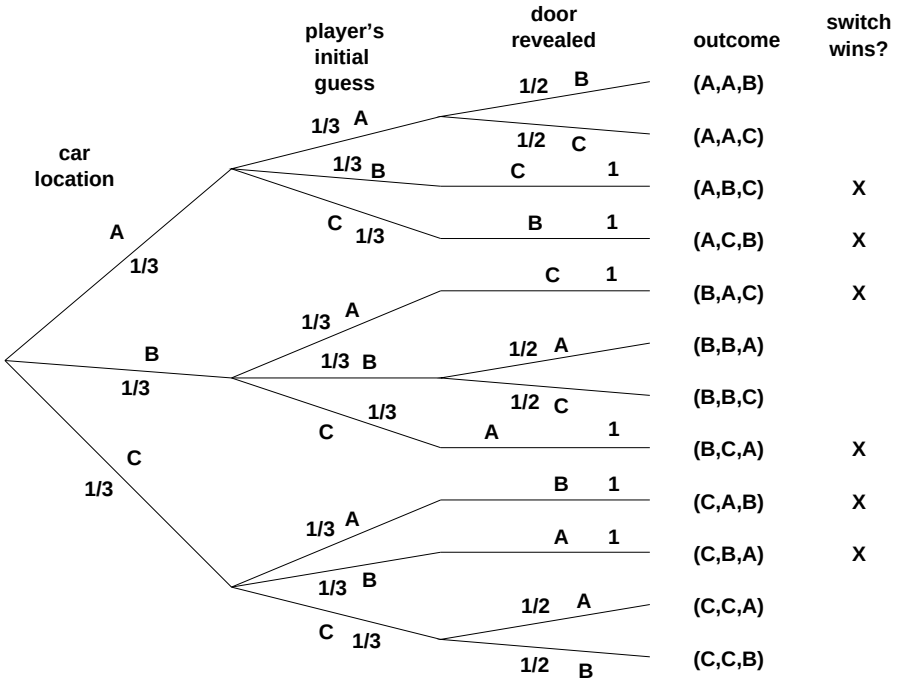
So far we've enumerated all the possible outcomes of the experiment. Now we must start assessing the likelihood of those outcomes. In particular, the goal of this step is to assign

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

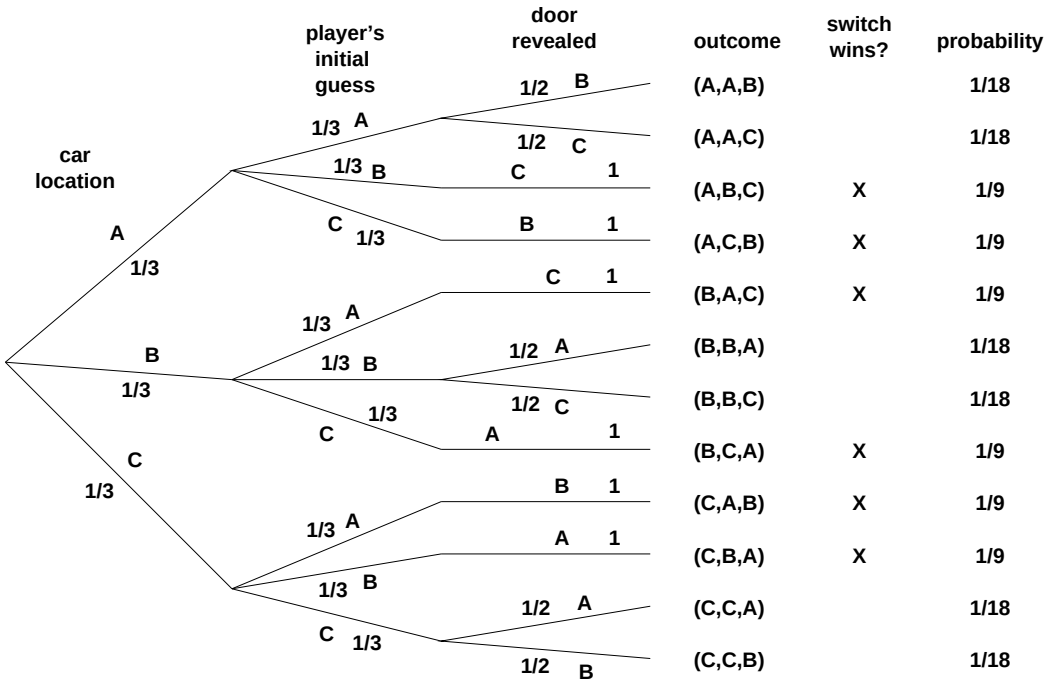
Our next job is to convert edge probabilities into outcome probabilities. This is a purely mechanical process: *the probability of an outcome is equal to the product of the edge-probabilities*

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 next time. In the meanwhile, 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:

$$\begin{aligned}\Pr(A, A, B) &= \frac{1}{18}\\ \Pr(A, A, C) &= \frac{1}{18}\\ \Pr(A, B, C) &= \frac{1}{9}\\ &\text{etc.}\end{aligned}$$

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:

$$\sum_{x \in S} \Pr(x) = 1$$

In this equation, 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 S and a probability function $\Pr : S \to [0, 1]$ 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.

18.1.6 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 $E \subseteq S$ 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:

$$\begin{aligned}\Pr(\text{switching wins}) &= \Pr(A, B, C) + \Pr(A, C, B) + \Pr(B, A, C) +\\ &\qquad \Pr(B, C, A) + \Pr(C, A, B) + \Pr(C, B, A)\\ &= \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9}\\ &= \frac{2}{3}\end{aligned}$$

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.

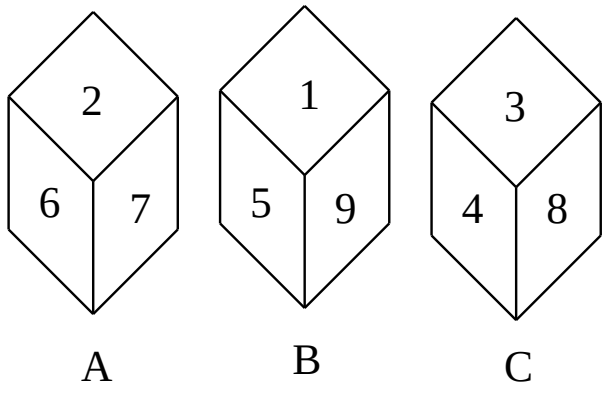
We're done with the problem! We didn't need any appeals to intuition or ingenious analogies. In fact, no mathematics more difficult than adding and multiplying fractions was required. The only hard part was resisting the temptation to leap to an "intuitively obvious" answer.

18.1.7 An Alternative Interpretation of the Monty Hall Problem

Was Marilyn really right? 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. If this case, switching never works!

18.2 Strange Dice

Let's play *Strange Dice*! The rules are simple. There are three dice, A, B, and C. Not surprisingly, the dice are numbered *strangely*, as shown below:



The number on each concealed face is the same as the number on the opposite, exposed face. The rules are simple. You pick one of the three dice, and then I pick one of the two remainders. We both roll and the player with the higher number wins.

Which of the dice should you choose to maximize your chances of winning? Die B is appealling, because it has a 9, the highest number overall. Then again, die A has two relatively large numbers, 6 and 7. But die C has an 8 and no very small numbers at all. Intuition gives no clear answer!

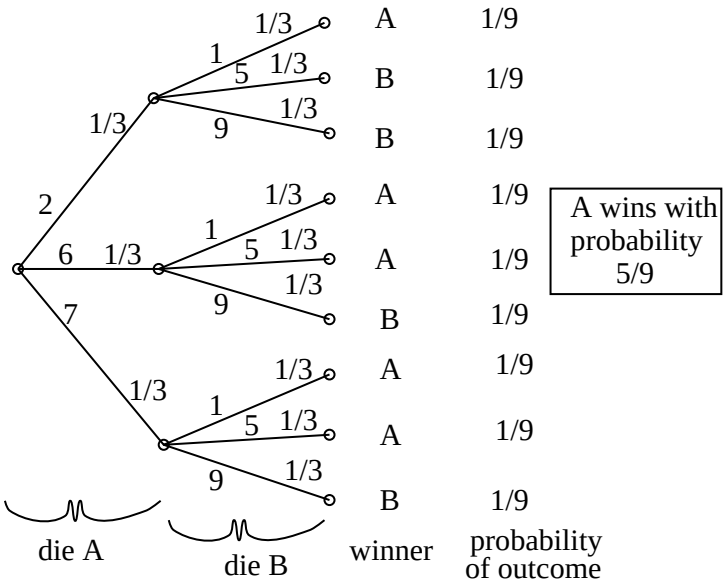
18.2.1 Analysis of Strange Dice

We can analyze Strange Dice using our standard, four-step method for solving probability problems. To fully understand the game, we need to consider three different experiments, corresponding to the three pairs of dice that could be pitted against one another.

Die *A* versus Die *B*

First, let's determine what happens when die A is played against die B.

Step 1: Find the sample space. The sample space for this experiment is worked out in the tree diagram show below. (Actually, the whole probability space is worked out in this one picture. But pretend that each component sort of fades in— nyyyrrroom!— as you read about the corresponding step below.)



For this experiment, the sample space is a set of nine outcomes:

$$S = \{\ (2,1),\ (2,5),\ (2,9),\ (6,1),\ (6,5),\ (6,9),\ (7,1),\ (7,5),\ (7,9)\ \}$$

Step 2: Define events of interest. We are interested in the event that the number on die A is greater than the number on die B. This event is a set of five outcomes:

$$\{\ (2,1),\ (6,1),\ (6,5),\ (7,1),\ (7,5)\ \}$$

These outcomes are marked A in the tree diagram above.

Step 3: Determine outcome probabilities. To find outcome probabilities, we first assign probabilities to edges in the tree diagram. Each number on each die comes up with probability $1/3$, regardless of the value of the other die. Therefore, we assign all edges probability $1/3$. The probability of an outcome is the product of probabilities on the corresponding root-to-leaf path, which means that every outcome has probability $1/9$. These probabilities are recorded on the right side of the tree diagram.

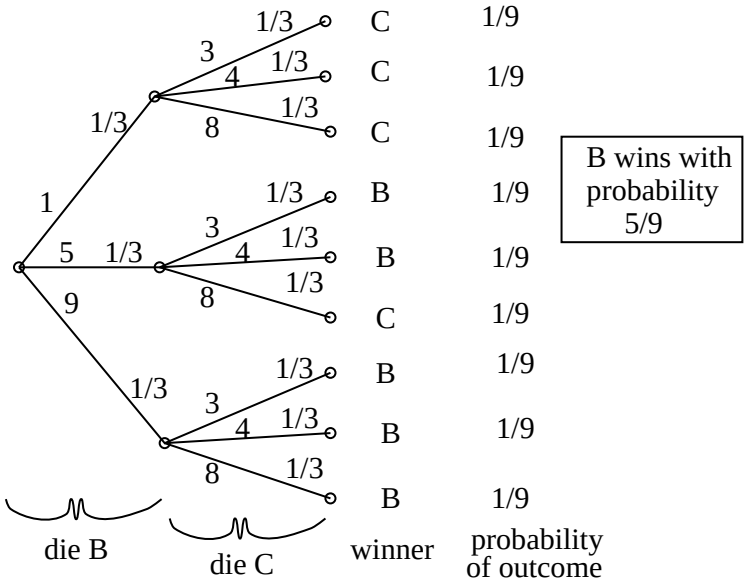
Step 4: Compute event probabilities. The probability of an event is the sum of the probabilities of the outcomes in that event. Therefore, the probability that die A comes up greater than die B is:

$$\begin{aligned}\Pr(A > B) &= \Pr(2,1) + \Pr(6,1) + \Pr(6,5) + \Pr(7,1) + \Pr(7,5) \\ &= \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} \\ &= \frac{5}{9}\end{aligned}$$

Therefore, die A beats die B more than half of the time. You had better not choose die B or else I'll pick die A and have a better-than-even chance of winning the game!

Die *B* versus Die *C*

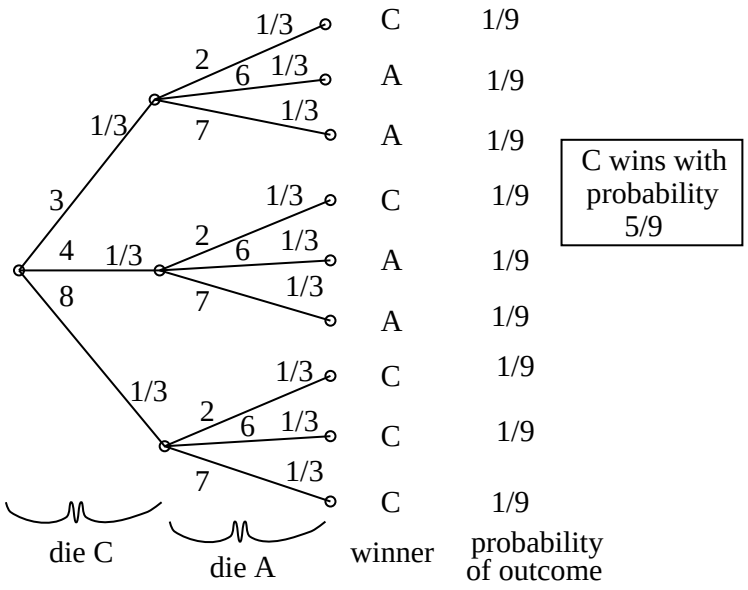
Now suppose that die B is played against die C. The tree diagram for this experiment is shown below.



The analysis is the same as before and leads to the conclusion that die B beats die C with probability $5/9$ as well. Therefore, you had beter not choose die C; if you do, I'll pick die B and most likely win!

Die *C* versus Die *A*

We've seen that A beats B and B beats C. Apparently, die A is the best and die C is the worst. The result of a confrontation between A and C seems a forgone conclusion. A tree diagram for this final experiment is worked out below.



Surprisingly, die C beats die A with probability $5/9$!

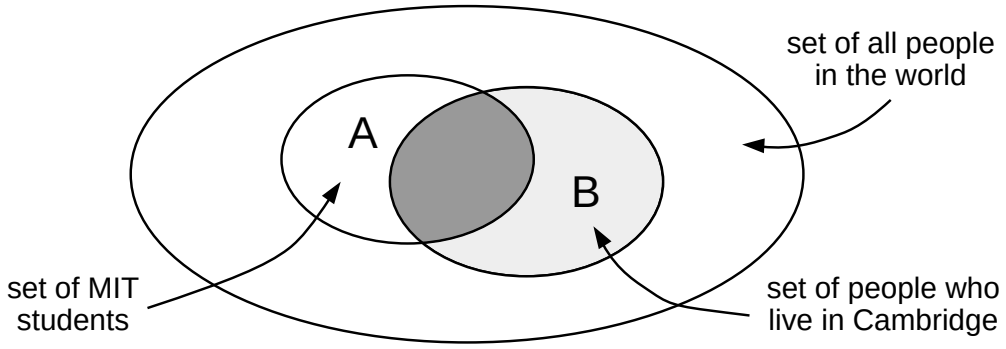
In summary, die A beats B, B beats C, and C beats A! Evidently, there is a relation between the dice that is *not transitive*! This means that no matter what die the first player chooses, the second player can choose a die that beats it with probability $5/9$. The player who picks first is always at a disadvantage!

Challenge: The dice can be renumbered so that A beats B and B beats C, each with probability $2/3$, and C still beats A with probability $5/9$. Can you find such a numbering?

Chapter 19

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 it is 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.

Probability is generally counterintuitive, but conditional probability is the worst! 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.

19.1 The Halting Problem

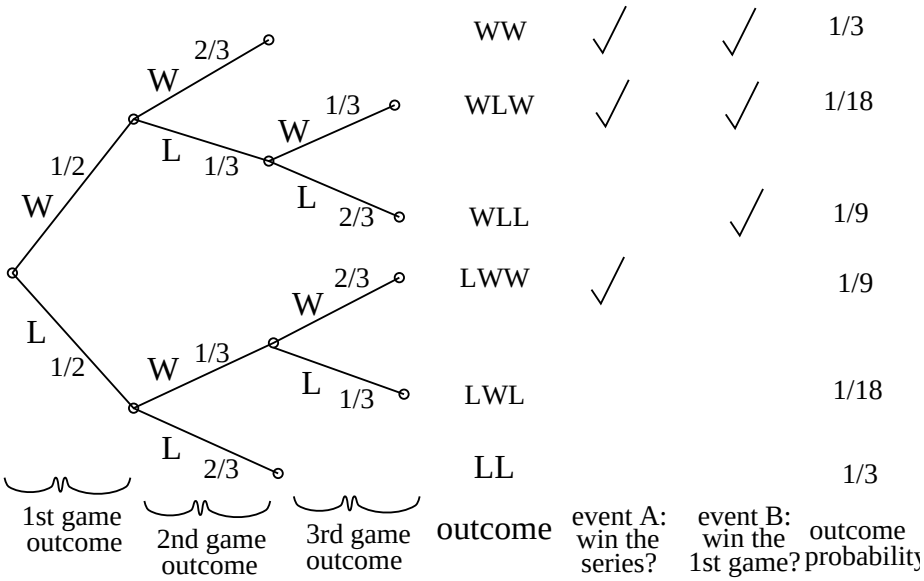
The *Halting Problem* is the canonical undecidable problem in computation theory that was first introduced by Alan Turing in his seminal 1936 paper. The problem is to determine whether a Turing machine halts on a given blah, blah, blah. Anyway, *much more importantly*, it is the name of the MIT EECS department's famed C-league hockey team.

In a best-of-three tournament, the Halting Problem wins the first game with probability $\frac{1}{2}$. 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 envigorated by victory and win the current game with probability $\frac{2}{3}$. If they lost the previous game, then they are demoralized by defeat and win the current game with probablity only $\frac{1}{3}$. What is the probability that the Halting Problem wins the tournament, given that they win the first game?

19.1.1 Solution to the Halting Problem

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:

$$S = \{WW,\ WLW,\ WLL,\ LWW,\ LWL,\ LL\}$$

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 $7/9$.

19.1.2 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 justficiation 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:

Rule (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 the Product Rule:

Rule (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 traverse 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.

19.2 *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 an ***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 $1/2$.

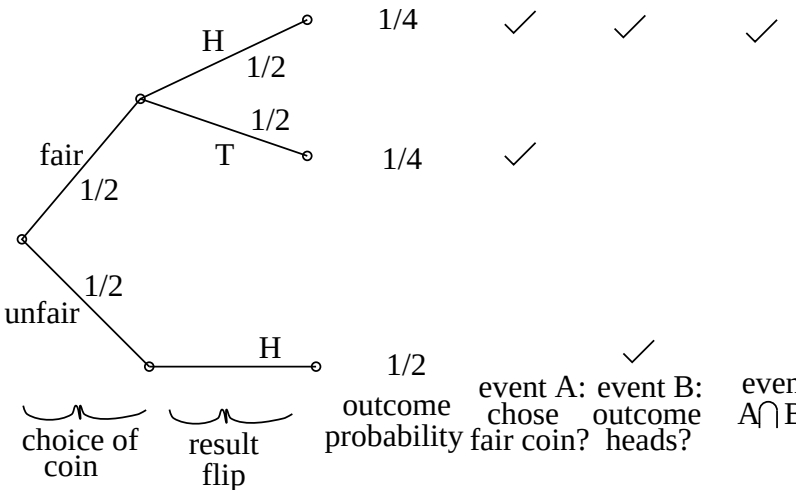
19.2.1 A Coin Problem

Suppose you have two coins. One coin is fair; that is, comes up heads with probability $1/2$ and tails with probability $1/2$. The other is a trick coin; it has heads on both sides, and so always comes up heads. Now you choose a coin at random so that you're equally likely to pick each one. If you flip the coin you select and get heads, then what is the probability that you flipped the fair coin?

This is another *a posteriori* problem since we want the probability of an event (that the fair coin was chosen) given the outcome of a later event (that heads came up). Intuition may fail us, but the standard four-step method works perfectly well.

Step 1: Find the Sample Space

The sample space is worked out in the tree diagram below.



Step 2: Define Events of Interest

Let A be the event that the fair coin was chosen. Let B the event that the result of the flip was heads. The outcomes in each event are marked in the figure. We want to compute $\Pr(A \mid B)$, the probability that the fair coin was chosen, given that the result of the flip was heads.

Step 2: Compute Outcome Probabilities

First, we assign probabilities to edges in the tree diagram. Each coin is chosen with probability $1/2$. If we choose the fair coin, then head and tails each come up with probability $1/2$. If we choose the trick coin, then heads comes up with probability 1. By the Product Rule, the probability of an outcome is the product of the probabilities on the corresponding root-to-leaf path. All of these probabilities are shown in the tree diagram.

Step 4: Compute Event Probabilities

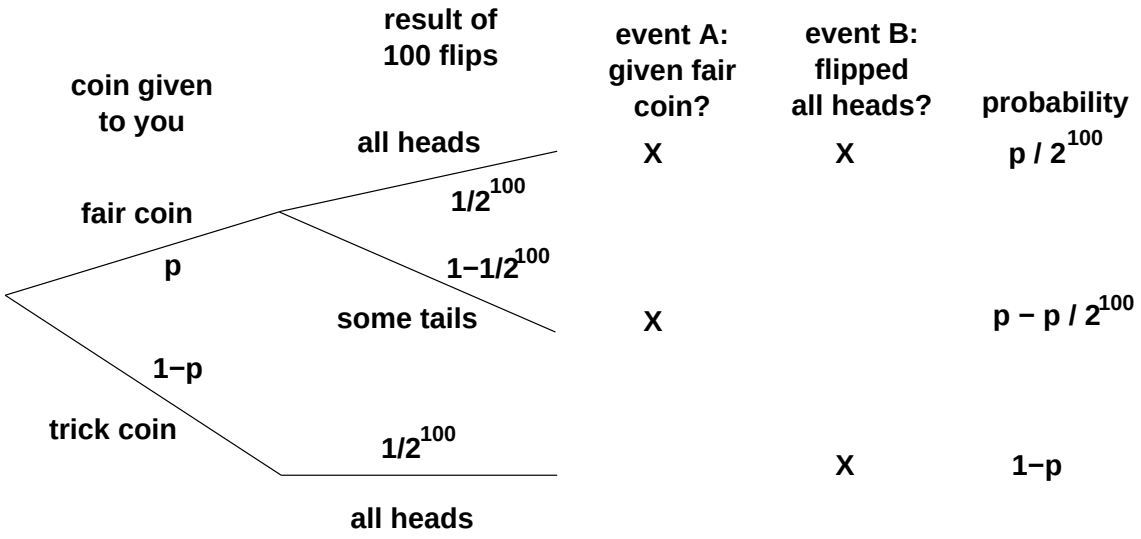
$$\begin{aligned}\Pr(A \mid B) &= \frac{\Pr(A \cap B)}{\Pr(B)} \\ &= \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{2}} \\ &= \frac{1}{3}\end{aligned}$$

The first equation uses the Product Rule. On the second line, we use the fact that the probability of an event is the sum of the probabilities of the outcomes it contains. The final line is simplification. The probability that the fair coin was chosen, given that the result of the flip was heads, is $1/3$.

19.2.2 A Variant of the Two Coins Problem

Let's consider a variant of the two coins problem. Someone hands you either a fair coin or a trick coin with heads on both sides. You flip the coin 100 times and see heads every time. What can you say about the probability that you flipped the fair coin? Remarkably—nothing!

In order to make sense out of this outrageous claim, let's formalize the problem. The sample space is worked out in the tree diagram below. We do not know the probability that you were handed the fair coin initially— you were just given one coin or the other—so let's call that p.



Let A be the event that you were handed the fair coin, and let B be the event that you flipped 100 heads. Now, we're looking for $\Pr(A \mid B)$, the probability that you were handed the fair coin, given that you flipped 100 heads. The outcome probabilities are worked out in the tree diagram. Plugging the results into the definition of conditional probability gives:

$$\begin{aligned}\Pr(A \mid B) &= \frac{\Pr(A \cap B)}{\Pr(B)} \\ &= \frac{p/2^{100}}{1-p+p/2^{100}} \\ &= \frac{p}{2^{100}(1-p)+p}\end{aligned}$$

This expression is very small for moderate values of p because of the 2^{100} term in the denominator. For example, if $p = 1/2$, then the probability that you were given the fair coin is essentially zero.

But we *do not know* the probability p that you were given the fair coin. And perhaps the value of p is *not* moderate; in fact, maybe $p = 1 - 2^{-100}$. Then there is nearly an even chance that you have the fair coin, given that you flipped 100 heads. In fact, maybe you were handed the fair coin with probability $p = 1$. Then the probability that you were given the fair coin is, well, 1!

A similar problem arises in polling before an election. A pollster picks a random American and asks his or her party affiliation. If this process is repeated many times, what can be said about the population as a whole? To clarify the analogy, suppose that the country contains only two people. There is either one Republican and one Democrat (like the fair coin), or there are two Republicans (like the trick coin). The pollster picks a random citizen 100 times, which is analogous to flipping the coin 100 times. Suppose that he picks a Republican every single time. We just showed that, even given this polling data, the probability that there is one citizen in each party could still be anywhere between 0 and 1!

What the pollster *can* say is that either:

1. Something earth-shatteringly unlikely happened during the poll.

2. There are two Republicans.

This is as far as probability theory can take us; from here, you must draw your own conclusions. Based on life experience, many people would consider the second possibility more plausible. However, if you are just *convinced* that the country isn't entirely Republican (say, because you're a citizen and a Democrat), then you might believe that the first possibility is actually more likely.

19.3 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:

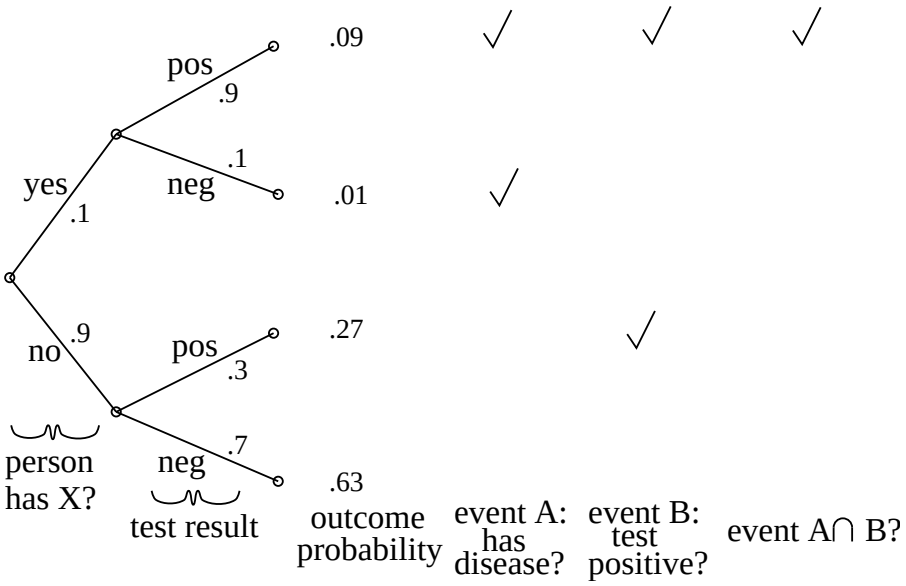
- If you have the disease, there is a 10% chance that the test will say you do not. (These are called "false negatives".)

- If you do not have 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!

19.4 Conditional Probability Pitfalls

The remaining sections illustrate some common blunders involving conditional probability.

19.4.1 Carnival Dice

There is a gambling game called Carnival Dice. A player picks a number between 1 and 6 and then rolls three fair dice. The player wins if his number comes up on at least one die. The player loses if his number does not appear on any of the dice. What is the probability that the player wins? This problem sounds simple enough that we might try an intuitive lunge for the solution.

False Claim 82. *The player wins with probability $\frac{1}{2}$.*

Proof. Let A_i be the event that the i-th die matches the player's guess.

$$\begin{aligned}
\Pr(\text{win}) &= \Pr(A_1 \cup A_2 \cup A_3) \\
&= \Pr(A_1) \cup \Pr(A_2) \cup \Pr(A_3) \\
&= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\
&= \frac{1}{2}
\end{aligned}$$

□

The only justification for the second equality is that it looks vaguely reasonable; in fact, equality does not hold. Let's examine the expression $\Pr(A_1 \cup A_2 \cup A_3)$ to see exactly what is happening. Recall that the probability of an event is the sum of the probabilities of the outcomes it contains. Therefore, we could argue as follows:

$$\begin{aligned}\Pr(A_1 \cup A_2 \cup A_3) &= \sum_{w \in A_1 \cup A_2 \cup A_3} \Pr(w) \\ &= \sum_{w \in A_1} \Pr(w) + \sum_{w \in A_2} \Pr(w) + \sum_{w \in A_3} \Pr(w) \\ &= \Pr(A_1) + \Pr(A_2) + \Pr(A_3)\end{aligned}$$

This argument is valid provided that the events A_1, A_2, and A_3 are *disjoint*; that is, there is no outcome in more than one event. If this were not true for some outcome, then a term would be duplicated when we split the one sum into three. Subject to this caveat, the argument generalizes to prove the following theorem:

Theorem 83. *Let $A_1, A_2, \ldots A_n$ be disjoint events. Then:*

$$\Pr(A_1 \cup A_2 \cup \ldots \cup A_n) = \Pr(A_1) + \Pr(A_2) + \ldots + \Pr(A_n)$$

We can evaluate the probability of a union of events that are not necessarily disjoint using a theorem analogous to Inclusion-Exclusion. Here is the special case for a union of three events.

Theorem 84. *Let A_1, A_2, and A_3 be events, not necessarily disjoint. Then:*

$$\begin{aligned}\Pr(A_1 \cup A_2 \cup A_3) = &\Pr(A_1) + \Pr(A_2) + \Pr(A_3) \\ &- \Pr(A_1 \cap A_2) - \Pr(A_1 \cap A_3) - \Pr(A_2 \cap A_3) \\ &+ \Pr(A_1 \cap A_2 \cap A_3)\end{aligned}$$

We can use this theorem to compute the real chance of winning at Carnival Dice. The probability that one die matches the player's guess is $1/6$. The probability that two dice both match the player's guess is $1/36$ by the Product Rule. Similarly, the probability that all three dice match is $1/216$. Plugging these numbers into the preceding theorem gives:

$$\begin{aligned}\Pr(\text{win}) &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\ &\quad - \frac{1}{36} - \frac{1}{36} - \frac{1}{36} \\ &\quad + \frac{1}{216} \\ &\approx 42\%\end{aligned}$$

These are terrible odds for a gambling game; you'd be much better off playing roulette, craps, or blackjack!

19.4.2 Other Identities

There is a close relationship between computing the size of a set and computing the probability of an event. Theorem 84 is one 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.

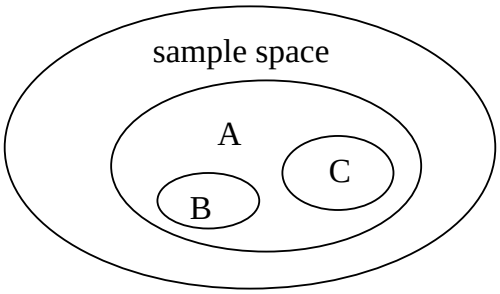
Many general probability identities still hold when all probabilities are conditioned on the same event. For example, the following identity is analogous to the Inclusion-Exclusion formula for two sets, except that all probabilities are conditioned on an event C.

$$\Pr(A \cup B \mid C) = \Pr(A \mid C) + \Pr(A \mid C) - \Pr(A \cap B \mid C)$$

Be careful not to mix up events before and after the conditioning bar! For example, the following is *not* a valid identity:

$$\Pr(A \mid B \cup C) = \Pr(A \mid B) + \Pr(A \mid C) \qquad (B \cap C = \phi)$$

A counterexample is shown below. In this case, $\Pr(A \mid B) = 1$, $\Pr(A \mid C) = 1$, and $\Pr(A \mid B \cup C) = 1$. However, since $1 \neq 1 + 1$, the equation above does not hold.



So you're convinced that this equation is false in general, right? Let's see if you *really* believe that.

19.4.3 Discrimination Lawsuit

Several years ago there was a sex discrimination lawsuit against Berkeley. A female professor was denied tenure, allegedly because she was a woman. She argued that in every one of Berkeley's 22 departments, the percentage of male applicants accepted was greater than the percentage of female applicants accepted. This sounds very suspicious!

However, Berkeley's lawyers argued that across the whole university the percentage of male tenure applicants accepted was actually *lower* than the percentage of female applicants accepted. This suggests that if there was any sex discrimination, then it was against men! Surely, at least one party in the dispute must be lying.

Let's simplify the problem and express both arguments in terms of conditional probabilities. Suppose that there are only two departments, EE and CS, and consider the experiment where we pick a random applicant. Define the following events:

- Let A be the event that the applicant is accepted.
- Let F_{EE} the event that the applicant is a female applying to EE.
- Let F_{CS} the event that the applicant is a female applying to CS.
- Let M_{EE} the event that the applicant is a male applying to EE.
- Let M_{CS} the event that the applicant is a male applying to CS.

Assume that all applicants are either male or female, and that no applicant applied to both departments. That is, the events F_{EE}, F_{CS}, M_{EE}, and M_{CS} are all disjoint.

In these terms, the plaintiff is make the following argument:

$$\begin{aligned}\Pr(A \mid F_{EE}) &< \Pr(A \mid M_{EE})\\ \Pr(A \mid F_{CS}) &< \Pr(A \mid M_{CS})\end{aligned}$$

That is, in both departments, the probability that a woman is accepted for tenure is less than the probability that a man is accepted. The university retorts that overall a woman applicant is *more* likely to be accepted than a man:

$$\Pr(A \mid F_{EE} \cup F_{CS}) > \Pr(A \mid M_{EE} \cup M_{CS})$$

It is easy to believe that these two positions are contradictory. In fact, we might even try to prove this by adding the plaintiff's two inequalities and then arguing as follows:

$$\begin{aligned}&\Pr(A \mid F_{EE}) + \Pr(A \mid F_{CS}) < \Pr(A \mid M_{EE}) + \Pr(A \mid M_{CS})\\ \Rightarrow\quad &\Pr(A \mid F_{EE} \cup F_{CS}) < \Pr(A \mid M_{EE} \cup M_{CS})\end{aligned}$$

The second line exactly contradicts the university's position! But there is a big problem with this argument; the second inequality follows from the first only if we accept the "false identity" from the preceding section. This argument is bogus! Maybe the two parties do not hold contradictory positions after all!

In fact, the table below shows a set of application statistics for which the assertions of both the plaintiff and the university hold:

CS	0 females accepted, 1 applied	0%
	50 males accepted, 100 applied	50%
EE	70 females accepted, 100 applied	70%
	1 male accepted, 1 applied	100%
Overall	70 females accepted, 101 applied	$\approx 70\%$
	51 males accepted, 101 applied	$\approx 51\%$

In this case, a higher percentage of males were accepted in both departments, but overall a higher percentage of females were accepted! Bizarre!

19.4.4 On-Time Airlines

Newspapers publish on-time statistics for airlines to help travelers choose the best carrier. The on-time rate for an airline is defined as follows:

$$\text{on-time rate} = \frac{\text{\# flights less than 15 minutes late}}{\text{\# flights total}}$$

This seems reasonable, but actually can be badly misleading! Here is some on-time data for two airlines in the late 80's.

	Alaska Air			America West		
Airport	#on-time	#flights	%	#on-time	#flights	%
Los Angeles	500	560	89	700	800	87
Phoenix	220	230	95	4900	5300	92
San Diego	210	230	92	400	450	89
San Francisco	500	600	83	320	450	71
Seattle	1900	2200	86	200	260	77
OVERALL	3820	3020	89	6520	7260	90

America West had a better overall on-time percentage, but Alaska Airlines did better at *every single airport!* This is the same paradox as in the Berkeley tenure lawsuit. The problem is that Alaska Airlines flew proportionally more of its flights to bad weather airports like Seattle, whereas America West was based in fair-weather, low-traffic Phoenix!

Chapter 20

Independence

20.1 Independent Events

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***. In particular, 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.

20.1.1 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 $1/10$.

20.1.2 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:

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B).$$

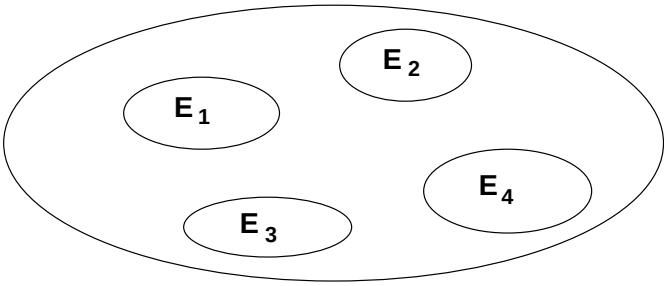
The equation on the left always holds if $\Pr(B) = 0$. Otherwise, we can divide both sides by $\Pr(B)$ and use the definition of conditional probability to obtain an alternative definition of independence:

$$\Pr(A \mid B) = \Pr(A) \qquad \text{or} \qquad \Pr(B) = 0$$

This equation 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.

20.1.3 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

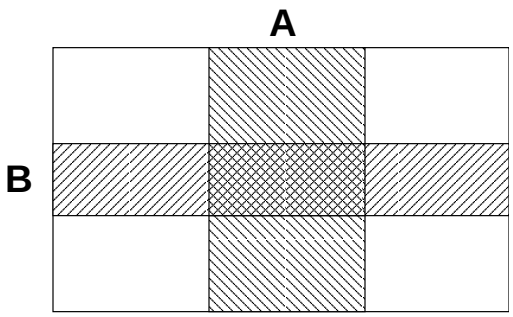
$$\Pr(A \cap B) = 0$$

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:

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

20.1.4 An Experiment with Two Coins

Suppose that we flip two independent, fair coins. Consider the following two events:

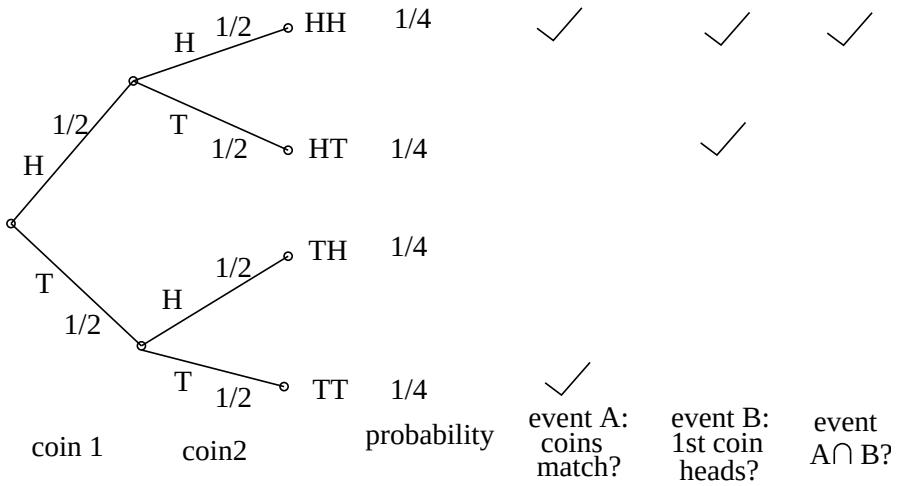
$A =$ the coins match (both are heads or both are tails)
$B =$ the first coin is heads

Are these independent events? Intuitively, the answer is "no". After all, whether or not the coins match *depends* on how the first coin comes up; if we toss HH, they match, but if we toss TH, then they do not. However, the mathematical definition of independence does not correspond perfectly to the intuitive notion of "unrelated" or "unconnected". These events actually are independent!

Claim 85. *Events A and B are independent.*

Proof. We must show that $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$.

Step 1: Find the Sample Space. As shown in the tree diagram below, there are four possible outcomes: HH, HT, TH, and TT.



Step 2: Define Events of Interest. The outcomes in event A (coins match) and event B (first coin is heads) are checked in the tree diagram above

Step 3: Compute Outcome Probabilities. Since the coins are independent and fair, all edge probabilities are $1/2$. We find outcome probabilities by multiplying edge probabilities along each root-to-leaf path. All outcomes have probability $1/4$.

Step 4: Compute Event Probabilities. Now we can verify that $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$:

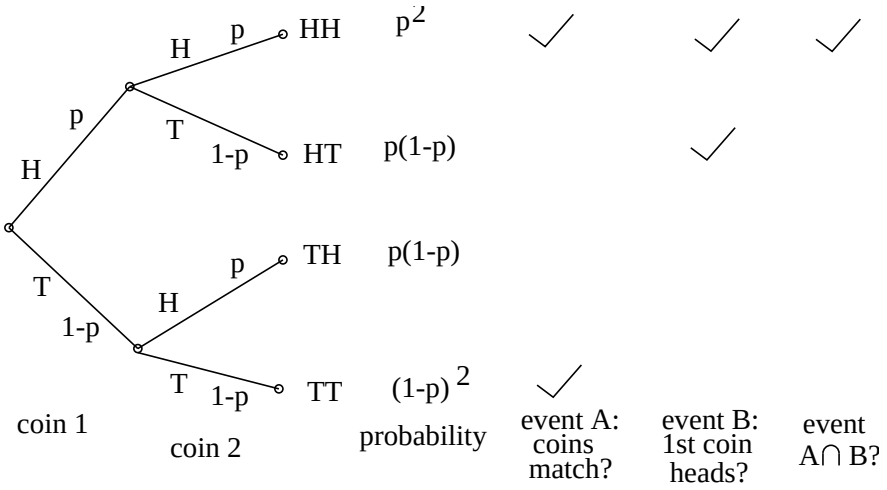
$$\Pr(A) = \Pr(HH) + \Pr(TT) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$
$$\Pr(B) = \Pr(HH) + \Pr(HT) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$
$$\Pr(A \cap B) = \Pr(HH) = \frac{1}{4}$$

Therefore, A and B are independent events as claimed. □

20.1.5 A Variation of the Two-Coin Experiment

Suppose that we alter the preceding experiment so that the coins are independent, but not fair. In particular, suppose each coin is heads with probability p and tails with probability $1 - p$ where p might not be $1/2$. As before, let A be the event that the coins match, and let B the event that the first coin is heads. Are events A and B independent for all values of p?

The problem is worked out in the tree diagram below.



We can read event probabilities off the tree diagram:

$$\begin{aligned}\Pr(A) &= \Pr(HH) + \Pr(TT) = p^2 + (1-p)^2 = 2p^2 - 2p + 1\\ \Pr(B) &= \Pr(HH) + \Pr(HT) = p^2 + p(1-p) = p\\ \Pr(A \cap B) &= \Pr(HH) = p^2\end{aligned}$$

Now events A and B are independent if and only if $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$ or, equivalently:

$$\left(2p^2 - 2p + 1\right) \cdot p = p^2$$

Since both sides are multiples of p, one solution is $p = 0$. Dividing both sides by p and simplifying leaves a quadratic equation:

$$2p^2 - 3p + 1 = 0$$

According to the quadratic formula, the remaining solutions are $p = 1$ and $p = 1/2$.

This analysis shows that events A and B are independent only if the coins are either *fair* or *completely biased* toward either heads or tails. Evidently, there was some dependence lurking at the fringes of the previous problem, but it was kept at bay because the coins were fair!

The Ultimate Application

Surprisingly, this has an application to Ultimate Frisbee. Here is an excerpt from the Tenth Edition rules:

A. Representatives of the two teams each flip a disc. The representative of one team calls ”same” or ”different” while the discs are in the air. The team winning the flip has the choice of:

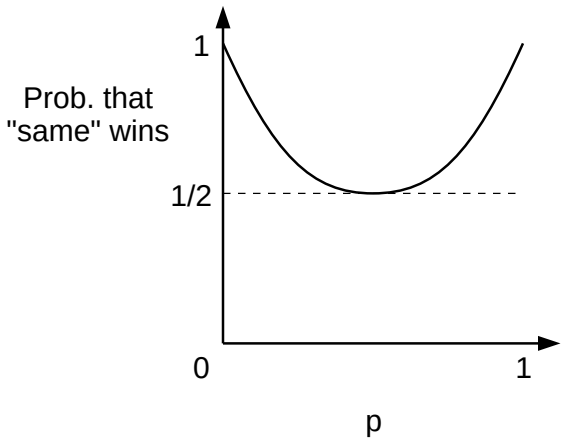
   1. Receiving or throwing the initial throw-off; or
   2. Selecting which goal they wish to defend initially.

B. The team losing the flip is given the remaining choice.

As we computed above, the probability that two flips match is:

$$\Pr(A) = p^2 + (1-p)^2$$

Below we've plotted this match probability as a function of p, the probability that one disc lands face-up.



Frisbees are asymmetric objects with strong aerodynamic properties, so p is not likely to be $1/2$. That plot shows that if there is any bias one way or the other, then saying “same” wins *more* than half the time. In fact, even if frisbees land face up exactly half the time ($p = 1/2$), then saying “same” still wins half the time. Therefore, might as well *always* say “same” during the opening flip!

20.2 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 than $E_1, \ldots, E_{100}$ are mutually independent.

20.2.1 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?

20.2.2 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 in ***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.

20.3 The Birthday Paradox

Suppose that there are 100 students in a lecture hall. There are 365 possible birthdays, ignoring February 29. What is the probability that two students have the same birthday? 50%? 90%? 99%? Let's make some modeling assumptions:

- For each student, all possible birthdays are equally likely. The idea underlying this assumption is that each student's birthday is determined by a random process involving parents, fate, and, um, some issues that we discussed earlier in the context of graph theory. Our assumption is not completely accurate, however; a disproportionate number of babies are born in August and September, for example. (Counting back nine months explains the reason why!)
- Birthdays are mutually independent. This isn't perfectly accurate either. For example, if there are twins in the lecture hall, then their birthdays are surely not independent.

We'll stick with these assumptions, despite their limitations. Part of the reason is to simplify the analysis. But the bigger reason is that our conclusions will apply to many situations in computer science where twins, leap days, and romantic holidays are not considerations. Also in pursuit of generality, let's switch from specific numbers to variables. Let m be the number of people in the room, and let N be the number of days in a year.

20.3.1 The Four-Step Method

We can solve this problem using the standard four-step method. However, a tree diagram will be of little value because the sample space is so enormous. This time we'll have to proceed without the visual aid!

Step 1: Find the Sample Space

Let's number the people in the room from 1 to m. An outcome of the experiment is a sequence $(b_1, \ldots, b_m)$ where b_i is the birthday of the ith person. The sample space is the set of all such sequences:

$$S = \{(b_1, \ldots, b_m) \mid b_i \in \{1, \ldots, N\}\}$$

Step 2: Define Events of Interest

Our goal is to determine the probability of the event A, in which some two people have the same birthday. This event is a little awkward to study directly, however. So we'll use a common trick, which is to analyze the *complementary* event $\overline{A}$, in which all m people have different birthdays:

$$\overline{A} = \{(b_1, \ldots, b_m) \in S \mid \text{all } b_i \text{ are distinct}\}$$

If we can compute $\Pr\left(\overline{A}\right)$, then we can compute what we really want, $\Pr(A)$, using the relation:

$$\Pr(A) + \Pr\left(\overline{A}\right) = 1$$

Step 3: Assign Outcome Probabilities

We need to compute the probability that m people have a particular combination of birthdays $(b_1, \ldots, b_m)$. There are N possible birthdays and all of them are equally likely for each student. Therefore, the probability that the ith person was born on day b_i is $1/N$. Since we're assuming that birthdays are mutually independent, we can multiply probabilities. Therefore, the probability that the first person was born on day b_1, the second on day b_2, and so forth is $(1/N)^m$. This is the probability of every outcome in the sample space.

Step 4: Compute Event Probabilities

Now we're interested in the probability of event $\overline{A}$ in which everyone has a different birthday:

$$\overline{A} = \{(b_1, \ldots, b_m) \in S \mid \text{all } b_i \text{ are distinct}\}$$

This is a gigantic set. In fact, there are N choices for b_1, $N-1$ choices for b_2, and so forth. Therefore, by the Generalized Product Rule:

$$\left|\overline{A}\right| = N(N-1)(N-2)\ldots(N-m+1)$$

The probability of the event $\overline{A}$ is the sum of the probabilities of all these outcomes. Happily, this sum is easy to compute, owing to the fact that every outcome has the same probability:

$$\begin{aligned}
\Pr\left(\overline{A}\right) &= \sum_{w \in \overline{A}} \Pr(w) \\
&= \frac{\left|\overline{A}\right|}{N^m} \\
&= \frac{N(N-1)(N-2)\ldots(N-m+1)}{N^m}
\end{aligned}$$

We're done!

20.3.2 An Alternative Approach

The probability theorems and formulas we've developed provide some other ways to solve probability problems. Let's demonstrate this by solving the birthday problem using a different approach— which had better give the same answer! As before, there are m people and N days in a year. Number the people from 1 to m, and let E_i be the event that the ith person has a birthday different from the preceding $i-1$ people. In these terms, we have:

$$\begin{aligned}&\Pr(\text{all } m \text{ birthdays different})\\ &\qquad = \Pr(E_1 \cap E_2 \cap \ldots \cap E_m)\\ &\qquad = \Pr(E_1) \cdot \Pr(E_2 \mid E_1) \cdot \Pr(E_3 \mid E_1 \cap E_2) \cdots \Pr(E_m \mid E_1 \cap \ldots \cap E_{m-1})\end{aligned}$$

On the second line, we're using the Product Rule for probabilities. The nasty-looking conditional probabilities aren't really so bad. The first person has a birthday different from all predecessors, because there are no predecessors:

$$\Pr(E_1) = 1$$

We're assuming that birthdates are equally probable and birthdays are independent, so the probability that the second person has the same birthday as the first is only $1/N$. Thus:

$$\Pr(E_2 \mid E_1) = 1 - \frac{1}{N}$$

Given that the first two people have different birthdays, the third person shares a birthday with one or the other with probability $2/N$, so:

$$\Pr(E_3 \mid E_1 \cap E_2) = 1 - \frac{2}{N}$$

Extending this reasoning gives:

$$\Pr(\text{all } m \text{ birthdays different}) = \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\cdots\left(1 - \frac{m-1}{N}\right)$$

We're done— again! This is our previous answer written in a different way.

20.3.3 An Upper Bound

One justification we offered for teaching approximation techniques was that approximate answers are often easier to work with and interpret than exact answers. Let's use the birthday problem as an illustration. We proved that m people all have different birthdays with probability

$$\Pr(\text{all } m \text{ birthdays different}) = \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\cdots\left(1 - \frac{m-1}{N}\right)$$

where N is the number of days in a year. This expression is exact, but inconvenient; evaluating it would require $\Omega(m)$ arithmetic operations. Furthermore, this expression is difficult to interpret; for example, how many people must be in a room to make the probability of a birthday match about $1/2$? Hard to say!

Let's look for a simpler, more meangingful approximate solution to the birthday problem. Every term in the product has the form $1 - x$ where x is relatively small, provided $m \ll N$. This is good news, because $1 - x$ figures prominently in one of the most useful of all approximation tricks:

$$1 - x \approx e^{-x} \qquad \text{for small } x$$

We'll use this trick several more times this term, so let's see where it comes from. Start with the Taylor series for $\ln(1 - x)$:

$$\ln(1 - x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \ldots$$

Now exponentiate both sides:

$$1 - x = e^{-x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \ldots}$$

Later we'll need this whole equation, but our immediate goal is to justify erasing most of the terms. Notice that if x is small, then $x^2/2$, $x^3/3$, $x^4/4$, etc. are *really* small, *shockingly* small, and *unbe-freakin'-lievably* small, respectively. Furthermore, if x is nonnegative, then:

$$1 - x \leq e^{-x}$$

The approximation $1 - x \approx e^{-x}$ is particularly helpful because it converts products to sums and vice-versa. For example, plugging this fact into the birthday problem gives:

$$\begin{aligned}
\Pr(\text{all } m \text{ birthdays different}) &= \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\cdots\left(1 - \frac{m-1}{N}\right) \\
&\leq e^{-1/N} \cdot e^{-2/N} \cdots e^{-(m-1)/N} \\
&= e^{-(1+2+\ldots(m-1))/N} \\
&= e^{-\frac{m(m-1)}{2N}} \qquad (20.1)
\end{aligned}$$

Notice how we began with a product, but ended up with a sum in the exponent. Applying a standard sum formula in the next step gives a closed-form (approximate) solution to the birthday problem!

Now let's get some concrete answers. If there are $m = 100$ people in a room and $N = 365$ days in a year, then the probability that no two have the same birthday is at most:

$$e^{-100 \cdot 99/(2 \cdot 365)} = e^{-13.56\ldots} < 0.0000013$$

So the odds everyone has a different birthday are around 1 in a million! In principle, there could be $m = 365$ people in a room, all with different birthdays. However, the probability of that happening by chance is at most:

$$e^{-365 \cdot 364/(2 \cdot 365)} = e^{-182} < 10^{-79}$$

Not gonna happen!

In fact, our upper bound implies that if there are only $m = 23$ people in a room, then the probability that all have different birthdays is *still less than half*. In other words, a room with only $m = 23$ people contains two people with the same birthday, more likely than not!

20.3.4 A Lower Bound

Like computer programs, approximation arguments are never done. You think you're finished, but then that seventh-order error term starts to nag at you. Soon you're waking up with a clenched jaw because that term is just *offensively* large. So— in the middle of the night— you're off again, trying to tune the approximation just a *little* more.[1] For example, for the birthday problem, we already have a good, approximate answer. Furthermore, it is an upper bound, so we even know in what direction the exact answer lies. Oh, but what about a lower bound? That would tell us how well our upper bound approximates the true answer.

There are many ways to obtain a lower bound. (In fact, an argument somewhat different from this one was presented in lecture.) The analysis given here demonstrates several techniques that you should understand individually, though the choice of this particular sequence of operations may seem mysterious. Let's start over with the exact answer:

$$\Pr(\text{all } m \text{ birthdays different}) = \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right) \cdots \left(1 - \frac{m-1}{N}\right)$$

Now let's rewrite each term using the series we derived above:

$$1 - x = e^{-x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \cdots}$$

This gives a hairy expression for the probability that all m birthdays are different:

$$e^{-\frac{1}{N} - \frac{1}{2N^2} - \frac{1}{3N^3} - \cdots} \cdot e^{-\frac{2}{N} - \frac{2}{2N^2} - \frac{2}{3N^3} - \cdots} \cdots e^{-\frac{m-1}{N} - \frac{m-1}{2N^2} - \frac{m-1}{3N^3} - \cdots}$$
$$= e^{-\frac{1+2+\cdots+(m-1)}{N} - \frac{1^2+2^2+\cdots+(m-1)^2}{2N^2} - \frac{1^3+2^3+\cdots+(m-1)^3}{3N^3} - \cdots}$$

[1] Okay, maybe this only happens to me.

On the second line, we've grouped terms with the same denominator. The numerators have a familiar form: they're sums of powers. These were among our "favorite" formulas to prove by induction back at the beginning of the term! The first sum, for example, is:

$$1 + 2 + \ldots + (m-1) = \frac{m(m-1)}{2}$$

We also established closed forms for the next couple sums. But since our overall goal is to find a lower bound on the whole expression, we can replace these sums with simple upper bounds. We can get such upper bounds via another blast from the past, the integration method:

$$1^k + 2^k + \ldots + (m-1)^k \leq \int_0^m m^k = \frac{m^{k+1}}{k+1}$$

Substituting in these results gives:

$$\begin{aligned}
\Pr(\text{all birthdays different}) &\leq e^{-\frac{m(m-1)}{2N} - \frac{m^3}{2\cdot 3N^2} - \frac{m^4}{3\cdot 4N^3} - \frac{m^5}{4\cdot 5N^4} - \cdots} \\
&= e^{-\frac{m(m-1)}{2N} - \frac{m^3}{N^2}\left(\frac{1}{2\cdot 3} + \frac{1}{3\cdot 4}\left(\frac{m}{N}\right) + \frac{1}{4\cdot 5}\left(\frac{m}{N}\right)^2 + \ldots\right)} \\
&\leq e^{-\frac{m(m-1)}{2N} - \frac{m^3}{N^2}\left(\frac{1}{2\cdot 3} + \frac{1}{3\cdot 4} + \frac{1}{4\cdot 5} + \ldots\right)} \\
&= e^{-\frac{m(m-1)}{2N} - \frac{m^3}{2N}} \qquad (20.2)
\end{aligned}$$

The last expression is the lower bound we were looking for. On the second line, we pulled out m^3/N^2. The third line follows from the fact that $m/N \leq 1$. The remaining sum is a famous "telescoping series" in which consecutive terms cancel:

$$\begin{aligned}
&\frac{1}{2\cdot 3} + \frac{1}{3\cdot 4} + \frac{1}{4\cdot 5} + \frac{1}{5\cdot 6} + \ldots \\
&\qquad = \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \left(\frac{1}{4} - \frac{1}{5}\right) + \left(\frac{1}{5} - \frac{1}{6}\right) + \ldots \\
&\qquad = \frac{1}{2}
\end{aligned}$$

20.3.5 The Birthday Principle

Let's put our lower bound (20.2) together with our upper bound (20.1):

$$e^{-\frac{m(m-1)}{2N} - \frac{m^3}{2N^2}} \leq \Pr(\text{all } m \text{ birthdays different}) \leq e^{-\frac{m(m-1)}{2N}}$$

The only difference is the $m^3/2N^2$ term in the lower bound. Thus, if m (the number of students) is not too large relative to N (the number of days in a year), then the upper and

lower bounds are really close. In particular, if $m = o\left(N^{2/3}\right)$, then the extra term goes to zero as N goes to infinity. Therefore, in the limit, the ratio of the upper bound to the lower bound is 1. Since the exact probability is sandwiched in between these two, we have an asymptotically tight solution to the birthday problem:

$$\Pr\left(\text{all } m \text{ birthdays different}\right) \sim e^{-\frac{m(m-1)}{2N}}$$

So how many people must be in a room so that there's a half chance that two have the same birthday? Letting the expression above equal $1/2$ and solving for m gives:

$$m \sim \sqrt{(2\ln 2)N} \approx 1.18\sqrt{N}.$$

This is called the ***birthday principle***:

> If there are N days in a year and about $\sqrt{(2\ln 2)N}$ people in a room, then there is an even chance that two have the same birthday.

An informal argument partly explains this phenomenon. Two people share a birthday with probability $1/N$. Therefore, we should expect to find matching birthdays when the number of *pairs* of people in the room is around N, which happens when $\binom{m}{2} = N$ or $m \approx \sqrt{2N}$, which roughly agrees with the Birthday Principle.

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 says a hash table with N buckets starts to experience collisions when around $\sqrt{(2\ln 2)N}$ items are inserted.

The Birthday Principle could even help you solve the subset-sum problem given earlier in the term. There, the challenge was to find two different subsets of 90 twenty-five digit numbers with the same sum. Suppose that we regard the sum of a subset as that subset's "birthday". Then the number of different birthdays is at most $90 \cdot 10^{25}$. This is an enormous number, but the Birthday Principle suggests that among

$$\sqrt{2\ln 2 \cdot 90 \cdot 10^{25}} \approx 3.5 \cdot 10^{13}$$

randomly-selected subsets of the 90 numbers, there is an even chance that two have the same sum. That's still a big number of subsets, but nothing wildly out of computational range. Of course, this calculation assumes that all possible sums are equally probable, which is not correct; many sums are not even possible. But the fact that some "birthdays" are more common than others should only *increase* the probability of finding two that match!

Chapter 21

Random Variables

We've used probablity to model a variety of experiments, games, and tests. Throughout, we have tried to compute probabilities of *events*. We asked, for example, what is the probability of the event that you win the Monty Hall game? What is the probability of the event that it rains, given that the weatherman carried his umbrella today? What is the probability of the event that you have a rare disease, given that you tested positive?

But one can ask more general questions about an experiment. *How hard* will it rain? *How long* will this illness last? *How much* will I lose playing 6.042 games all day? These questions are fundamentally different and not easily phrased in terms of events. The problem is that an event either does or does not happen: you win or lose, it rains or doesn't, you're sick or not. But these new questions are about matters of degree: how much, how hard, how long? To approach these questions, we need a new mathematical tool.

21.1 Random Variables

Let's begin with an example. 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:

$$S = \{HHH,\ HHT,\ HTH,\ HTT,\ THH,\ THT,\ TTH,\ TTT\}$$

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:

$$\begin{array}{rclrcl} M(HHH) & = & 1 & M(THH) & = & 0 \\ M(HHT) & = & 0 & M(THT) & = & 0 \\ M(HTH) & = & 0 & M(TTH) & = & 0 \\ M(HTT) & = & 0 & M(TTT) & = & 1 \end{array}$$

The functions C and M are examples of ***random variables***. In general, a random variable is a function whose domain is the sample space. (The codomain can be anything, but we'll usually use a subset of the real numbers.) Notice that the name "random variable" is a misnomer; random variables are actually functions!

21.1.1 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 partitions the sample space into those outcomes in the event and those outcomes not in the event. Therefore, each event is naturally associated with a certain indicator random variable and vice versa: an ***indicator for an event*** E is an indicator random variable that is 1 for all outcomes in E and 0 for all outcomes not in E. Thus, M is an indicator random variable for the event that all three coins match.

21.1.2 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 and inequalities involving random variables. For example:

$$\begin{aligned}\Pr(M = 1) &= \Pr(TTT) + \Pr(HHH) \\ &= \frac{1}{8} + \frac{1}{8} \\ &= \frac{1}{4}\end{aligned}$$

As another example:

$$\begin{aligned}\Pr(C \geq 2) &= \Pr(THH) + \Pr(HTH) + \Pr(HHT) + \Pr(HHH) \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \\ &= \frac{1}{2}\end{aligned}$$

This is pretty wild; one normally thinks of equations and inequalities as either true or false. But when variables are replaced by random variables, there is a *probability* that the relationship holds!

21.1.3 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, which *sets* of outcomes. So taking their intersection is perfectly valid!

21.1.4 Independence

The notion of independence carries over from events to random variables as well. Random variables R_1 and R_2 are ***independent*** if

$$\Pr(R_1 = x_1 \cap R_2 = x_2) = \Pr(R_1 = x_1) \cdot \Pr(R_2 = x_2)$$

for all x_1 in the codomain of R_1 and x_2 in the codomain of R_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 and only if

$$\Pr(R_1 = x_1 \mid R_2 = x_2) = \Pr(R_1 = x_1) \text{ or } \Pr(R_2 = x_2) = 0$$

for all x_1 in the codomain of R_1 and x_2 in the codomain of R_2. In words, the probability that R_1 takes on a particular value is unaffected by the value of R_2.

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 that case, we have:

$$\Pr(C = 2 \cap M = 1) = 0 \qquad \text{but} \qquad \Pr(C = 2) \cdot \Pr(M = 1) = \frac{3}{8} \cdot \frac{1}{4} \neq 0$$

The notion of independence generalizes to a set of random variables as follows. Random variables $R_1, R_2, \ldots, R_n$ are ***mutually independent*** if

$$\begin{aligned}\Pr(R_1 = x_1 \cap R_2 = x_2 \cap \cdots \cap R_n = x_n) \\ = \Pr(R_1 = x_1) \cdot \Pr(R_2 = x_2) \cdots \Pr(R_n = x_n)\end{aligned}$$

for all $x_1, \ldots, x_n$ in the codomains of $R_1, \ldots, R_n$.

A consequence of this definition of mutual independence is that the probability of an assignment to a *subset* of the variables is equal to the product of the probabilites of the individual assignments. Thus, for example, if $R_1, R_2, \ldots, R_{100}$ are mutually independent random variables with codomain $\mathbb{N}$, then it follows that:

$$\Pr(R_1 = 9 \cap R_7 = 84 \cap R_{23} = 13) = \Pr(R_1 = 9) \cdot \Pr(R_7 = 84) \cdot \Pr(R_{23} = 13)$$

(This follows by summing over all possible values of the other random variables; we omit the details.)

21.1.5 An Example with Dice

Suppose that we roll two fair, independent dice. The sample space for this experiment consists of all pairs (r_1, r_2) where $r_1, r_2 \in \{1, 2, 3, 4, 5, 6\}$. Thus, for example, the outcome $(3, 5)$ corresponds to rolling a 3 on the first die and a 5 on the second. The probability of each outcome in the sample space is $1/6 \cdot 1/6 = 1/36$ since the dice are fair and independent.

We can regard the numbers that come up on the individual dice as random variables D_1 and D_2. So $D_1(3, 5) = 3$ and $D_2(3, 5) = 5$. Then the expression $D_1 + D_2$ is another random variable; let's call it T for "total". More precisely, we've defined:

$$T(w) = D_1(w) + D_2(w) \qquad \text{for every outcome } w$$

Thus, $T(3, 5) = D_1(3, 5) + D_2(3, 5) = 3 + 5 = 8$. In general, any function of random variables is itself a random variable. For example, $\sqrt{D_1} + \cos(D_2)$ is a strange, but well-defined random variable.

Let's also define an indicator random variable S for the event that the total of the two dice is seven:

$$S(w) = \begin{cases} 1 & \text{if } T(w) = 7 \\ 0 & \text{if } T(w) \neq 7 \end{cases}$$

So S is equal to 1 when the sum is seven and is equal to 0 otherwise. For example, $S(4, 3) = 1$, but $S(5, 3) = 0$.

Now let's consider a couple questions about independence. First, are D_1 and T independent? Intuitively, the answer would seem to be "no" since the number that comes up on the first die strongly affects the total of the two dice. But to prove this, we must find integers x_1 and x_2 such that:

$$\Pr(D_1 = x_1 \cap T = x_2) \neq \Pr(D_1 = x_1) \cdot \Pr(T = x_2)$$

For example, we might choose $x_1 = 2$ and $x_2 = 3$. In this case, we have

$$\Pr(T = 2 \mid D_1 = 3) = 0$$

since the total can not be only 2 when one die alone is 3. On the other hand, we have:

$$\begin{aligned} \Pr(T = 2) \cdot \Pr(D_1 = 3) &= \Pr(\{1, 1\}) \cdot \Pr(\{(3, 1),\ (3, 2),\ \ldots,\ (3, 6)\}) \\ &= \frac{1}{36} \cdot \frac{6}{36} \neq 0 \end{aligned}$$

So, as we suspected, these random variables are not independent.

Are S and D_1 independent? Once again, intuition suggests that the answer is "no". The number on the first die ought to affect whether or not the sum is equal to seven. But this time intuition turns out to be wrong! These two random variables actually are independent.

Proving that two random variables are independent takes some work. (Fortunately, this is an uncommon task; usually independence is a modeling assumption. Only rarely do random variables unexpectedly turn out to be independent.) In this case, we must show that

$$\Pr(S = x_1 \cap D_1 = x_2) = \Pr(S = x_1) \cdot \Pr(D_1 = x_2) \tag{21.1}$$

for all $x_1 \in \{0, 1\}$ and all $x_2 \in \{1, 2, 3, 4, 5, 6\}$. We can work through all these possibilities in two batches:

- Suppose that $x_1 = 1$. Then for every value of x_2 we have:

$$\Pr(S = 1) = \Pr((1, 6),\ (2, 5),\ \ldots,\ (6, 1)) = \frac{1}{6}$$
$$\Pr(D_1 = x_2) = \Pr((x_2, 1),\ (x_2, 2),\ \ldots,\ (x_2, 6)) = \frac{1}{6}$$
$$\Pr(S = 1 \cap D_1 = x_2) = \Pr((x_2, 7 - x_2)) = \frac{1}{36}$$

  Since $1/6 \cdot 1/6 = 1/36$, the independence condition is satisfied.

- Otherwise, suppose that $x_1 = 0$. Then we have $\Pr(S = 0) = 1 - \Pr(S = 1) = 5/6$ and $\Pr(D_1 = x_2) = 1/6$ as before. Now the event

$$S = 0 \cap D_1 = x_2$$

  consists of 5 outcomes: all of $(x_2, 1),\ (x_2, 2),\ \ldots,\ (x_2, 6)$ except for $(x_2, 7 - x_2)$. Therefore, the probability of this event is $5/36$. Since $5/6 \cdot 1/6 = 5/36$, the independence condition is again satisfied.

Thus, the outcome of the first die roll is independent of the fact that the sum is 7. This is a strange, isolated result; for example, the first roll is *not* independent of the fact that the sum is 6 or 8 or any number other than 7. But this example shows that the mathematical notion of independent random variables— while closely related to the intutive notion of "unrelated quantities"— is not exactly the same thing.

21.2 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 variable 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 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:

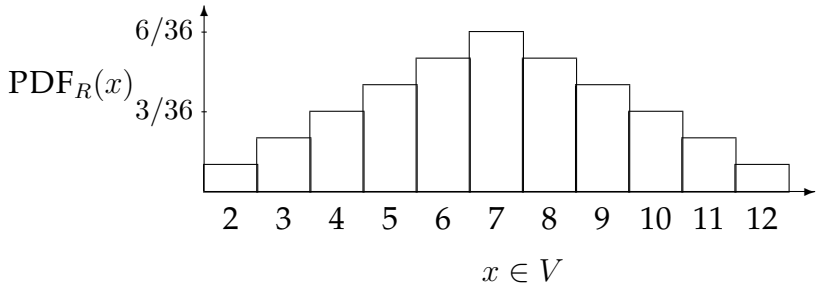
$$\text{PDF}_R(x) = \Pr(R = x)$$

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:

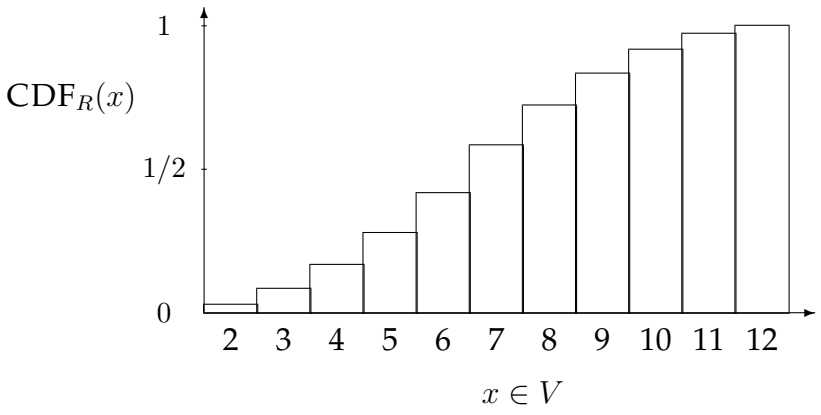


The lump in the middle indicates that sums close to 7 are the most likely. The total area of all the rectangles is 1 since the dice must take on exactly one of the sums in $V = \{2, 3, \ldots, 12\}$.

A closely-related idea is the ***cumulative distribution function (cdf)*** for a random variable R. This is a function $\text{CDF}_R : V \to [0, 1]$ 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.

For the remainder of today, we'll look at three important distributions and some applications.

21.2.1 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 ditribution function is:

$$\begin{aligned}\text{CDF}_B(0) &= p \\ \text{CDF}_B(1) &= 1\end{aligned}$$

This is called the ***Bernoulli distribution***. The number of heads flipped on a (possibly biased) coin has a Bernoulli distribution.

21.2.2 Uniform Distribution

A random variable that takes on each possible values 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 $\{1, 2, \ldots, 6\}$.

21.2.3 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 an 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:

$$\left\{\frac{1}{2},\ 1\frac{1}{2},\ 2\frac{1}{2},\ \ldots,\ n-\frac{1}{2}\right\}$$

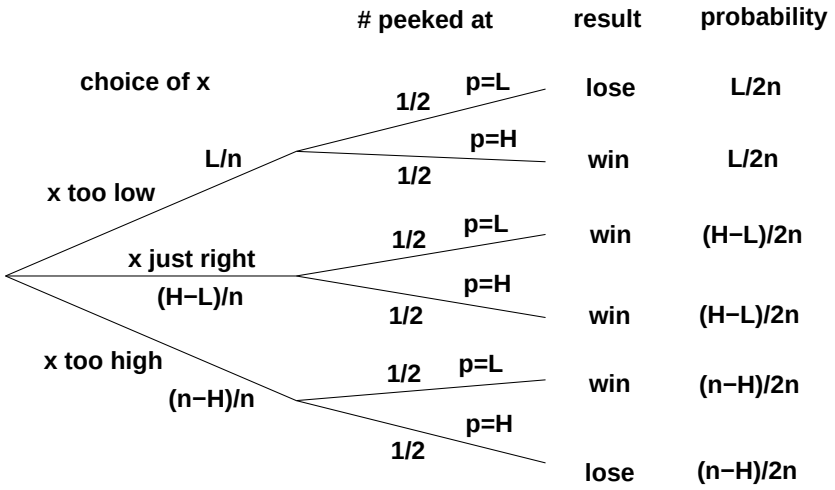
But what probability distribution should you use?

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 range $0, \ldots, 10$, then your probability of winning rises to 55%! By Las Vegas standards, those are great odds!

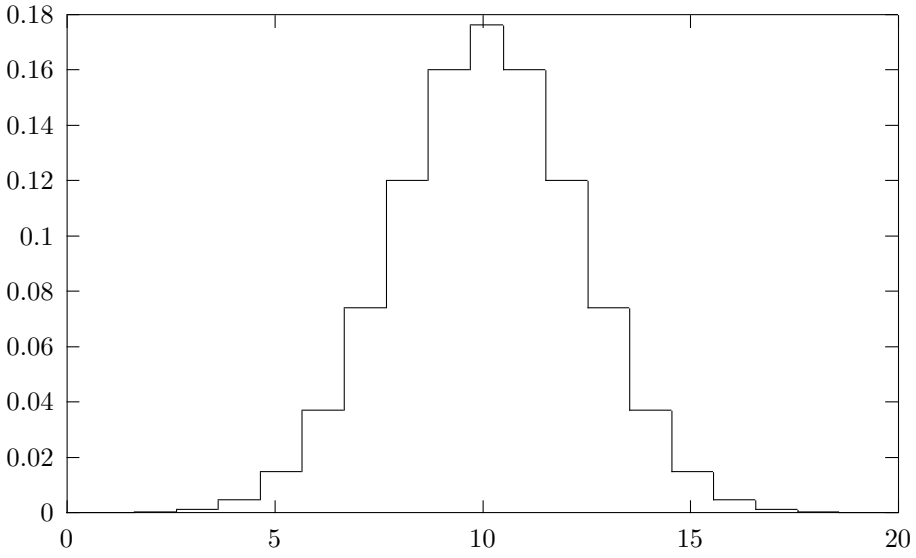
21.2.4 Binomial Distribution

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. If the coin is fair, then H has an *unbiased binomial density function*:

$$\text{PDF}_H(k) = \binom{n}{k} 2^{-n}$$

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 $\text{PDF}_H(k)$ 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.

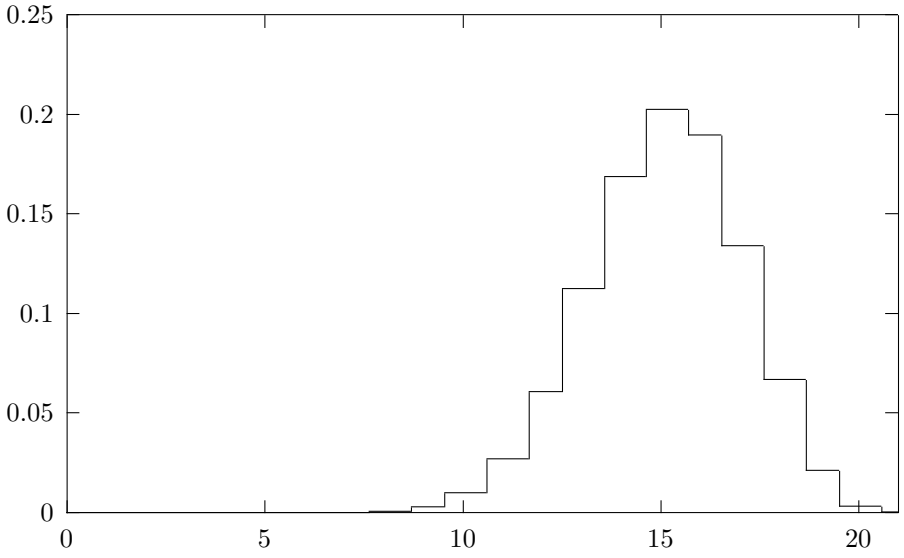
The General Binomial Distribution

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

There is an approximate closed-form formula for the general binomial density function, though it is a bit unwieldy. First, we need an approximation for a key term in the exact formula, $\binom{n}{k}$. 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***:

$$H(\alpha) = \alpha \log_2 \frac{1}{\alpha} + (1-\alpha) \log_2 \frac{1}{1-\alpha}$$

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:

$$\text{PDF}_J(\alpha n) \leq \frac{2^{nH(\alpha)}}{\sqrt{2\pi\alpha(1-\alpha)n}} \cdot p^{\alpha n}(1-p)^{(1-\alpha)n} \tag{21.2}$$

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:

$$\begin{aligned}\text{PDF}_J(\alpha n) &\leq \frac{2^{nH(1/2)}}{\sqrt{2\pi(1/2)(1-(1/2))n}} \cdot 2^{-n} \\ &= \sqrt{\frac{2}{\pi n}}\end{aligned}$$

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%.

21.2.5 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:

$$\begin{aligned}\text{CDF}_J(k) &= \Pr(J \le k) \\ &= \sum_{i=0}^{k} \text{PDF}_J(i) \\ &= \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}\end{aligned}$$

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 $\alpha < p$:

$$\begin{aligned}\text{CDF}_J(\alpha n) &\le \frac{1-\alpha}{1-\alpha/p} \cdot \text{PDF}_J(\alpha n) \\ &\le \frac{1-\alpha}{1-\alpha/p} \cdot \frac{2^{nH(\alpha)}}{\sqrt{2\pi\alpha(1-\alpha)n}} \cdot p^{\alpha n}(1-p)^{(1-\alpha)n}\end{aligned}$$

In the first step, we upper bound the summmation with a geometric sum and apply the formula for the sum of a geometric series. (The details are dull and omitted.) Then we insert the approximate formula (21.2) for $\text{PDF}_J(\alpha n)$ from the preceding section.

You have to press a lot of buttons on a calculator to evaluate this formula for a specific choice of α, p, and n. (Even computing $H(\alpha)$ is a fair amount of work!) But for large n, evaluating the cumulative distribution function exactly requires vastly *more* work! So don't look gift blessings in the mouth before they hatch. Or something.

As an example, the probability of fliping at most 25 heads in 100 tosses of a fair coin is obtained by setting $\alpha = 1/4$, $p = 1/2$ and $n = 100$:

$$\text{CDF}_J(n/4) \le \frac{1-(1/4)}{1-(1/4)/(1/2)} \cdot \text{PDF}_J(n/4) \qquad \le \frac{3}{2} \cdot 1.913 \cdot 10^{-7}$$

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.

21.3 Philosophy of Polling

On place where the binomial distribution comes up is in polling. Polling involves not only some tricky mathematics, but also some philosophical issues.

The difficulty is that polling tries to apply probabilty theory to resolve a question of fact. Let's first consider a slightly different problem where the issue is more stark. What is the probability that

$$N = 2^{6972607} - 1$$

is a prime number? One might guess $1/10$ or $1/100$. Or one might get sophisticated and point out that the Prime Number Theorem implies that only about 1 in 5 million numbers in this range are prime. But these answers are all wrong. There is no random process here. The number N is either prime or composite. You can conduct as many "repeated trials" as you like; the answer will always be the same. Thus, it seems probability does not touch upon this question.

However, there is a probabilistic primality test due to Rabin and Miller. If N is composite, there is at least a $3/4$ chance that the test will discover this. (In the remaining $1/4$ of the time, the test is inconclusive; it never produces a wrong answer.) Moreover, the test can be run again and again and the results are independent. So if N actually is composite, then the probability that $k = 100$ repetitions of the Rabin-Miller do not discover this is at most:

$$\left(\frac{1}{4}\right)^{100}$$

So 100 consecutive inconclusive answers would be extremely convincing evidence that N is prime! But we still couldn't say anything about the *probability* that N is prime: that is still either 0 or 1 and we don't know which.

A similar situation arises in the context of polling: we can make a convincing argument that a statement about public opinion is true, but can not actually say that the statement is true with any particular probability. Suppose we're conducting a yes/no poll on some question. Then we assume that some fraction p of the population would answer "yes" to the question and the remaining $1 - p$ fraction would answer "no". (Let's forget about the people who hang up on pollsters or launch into long stories about their little dog Fi-Fi— real pollsters have no such luxury!) Now, p is a fixed number, not a randomly-determined quantity. So trying to determine p by a random experiment is analogous to trying to determine whether N is prime or composite using a probabilistic primality test.

Probability slips into a poll since the pollster samples the opinions of a people selected uniformly and independently at random. The results are qualified by saying something like this:

> "One can say with 95% confidence that the maximum margin of sampling error is ± 3 percentage points."

This means that either the number reported in the poll is within 3% of the actual fraction p or else an unlucky 1-in-20 event happened during the polling process; specifically, the pollster's random sample was not representative of the population at large. This is *not* the same thing as saying that there is a 95% chance that the poll is correct; it either is or it isn't, just as N is either prime or composite regardless of the Rabin-Miller test results.

Chapter 22

Expected Value I

The ***expectation*** or ***expected value*** of a random variable is a single number that tells you a lot about the behavior of the variable. Roughly, the expectation is the average value, where each value is weighted according to the probability that it comes up. Formally, the expected value of a random variable R defined on a sample space S is:

$$\mathrm{Ex}\left(R\right) = \sum_{w \in S} R(w) \Pr\left(w\right)$$

To appreciate its signficance, suppose S is the set of students in a class, and we select a student uniformly at random. Let R be the selected student's exam score. Then $\mathrm{Ex}\left(R\right)$ is just the class average— the first thing everyone want to know after getting their test back! In the same way, expectation is usually the first thing one wants to determine about any random variable.

Let's work through an example. Let R be the number that comes up on a fair, six-sided die. Then the expected value of R is:

$$\begin{aligned}
\mathrm{Ex}\left(R\right) &= \sum_{k=1}^{6} k\left(\frac{1}{6}\right) \\
&= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\
&= \frac{7}{2}
\end{aligned}$$

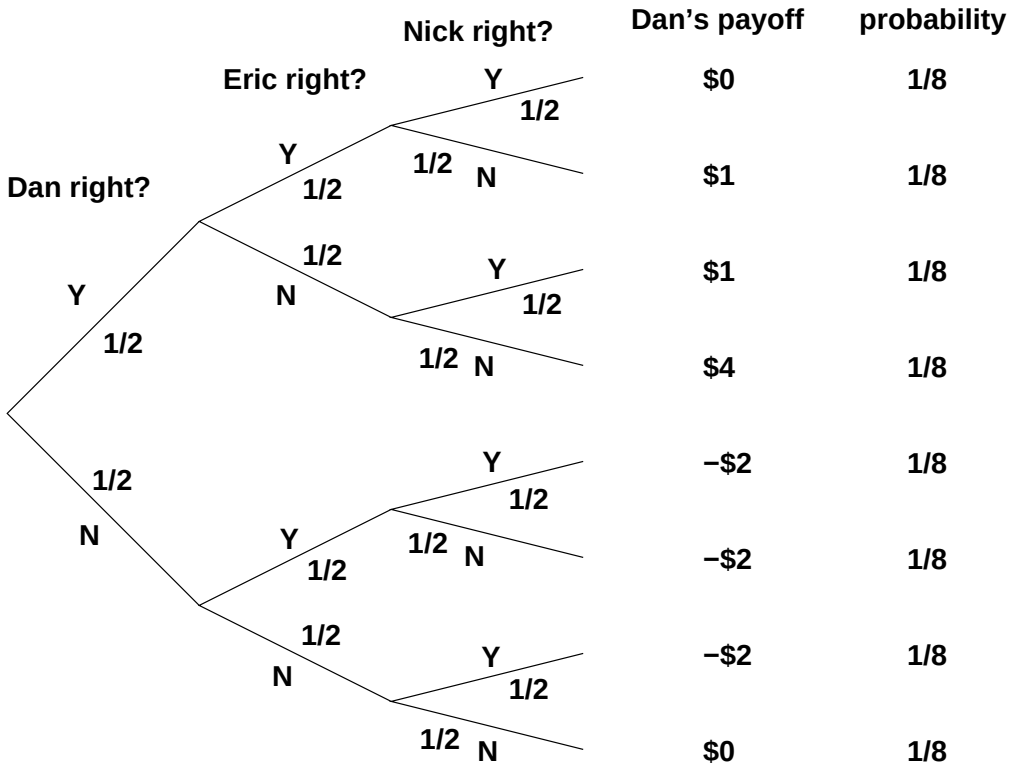
This calculation shows that the name "expected value" is a little misleading; the random variable might *never* actually take on that value. You can't roll a $3\frac{1}{2}$ on an ordinary die!

22.1 Betting on Coins

Dan, Eric, and Nick decide to play a fun game. Each player puts $2 on the table and secretly writes down either "heads" or "tails". Then one of them tosses a fair coin. The

$6 on the table is divided evenly among the players who correctly predicted the outcome of the coin toss. If everyone guessed incorrectly, then everyone takes their money back. After many repetitions of this game, Dan has lost a lot of money— more than can be explained by bad luck. What's going on?

A tree diagram for this problem is worked out below, under the assumptions that everyone guesses correctly with probability $1/2$ and everyone is correct independently.

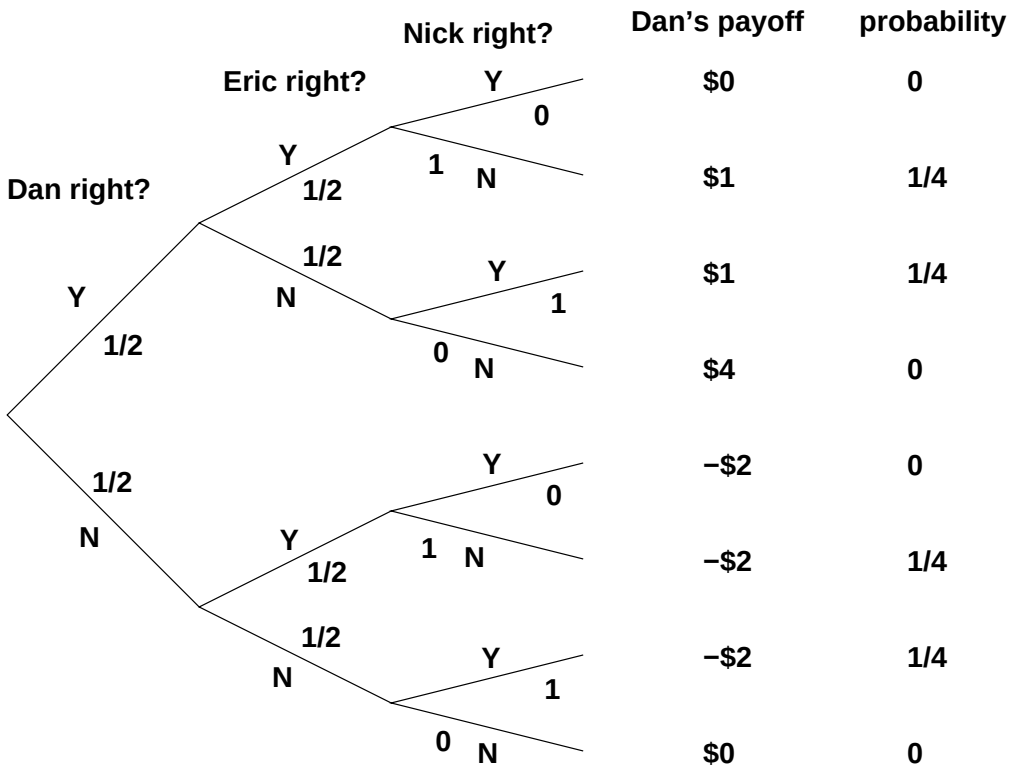


In the "payoff" column, we're accounting for the fact that Dan has to put in $2 just to play. So, for example, if he guesses correctly and Eric and Nick are wrong, then he takes all $6 on the table, but his net profit is only $4. Working from the tree diagram, Dan's expected payoff is:

$$\begin{aligned}\text{Ex}\,(\text{payoff}) &= 0\cdot\frac{1}{8}+1\cdot\frac{1}{8}+1\cdot\frac{1}{8}+4\cdot\frac{1}{8}+(-2)\cdot\frac{1}{8}+(-2)\cdot\frac{1}{8}+(-2)\cdot\frac{1}{8}+0\cdot\frac{1}{8}\\ &= 0\end{aligned}$$

So the game perfectly fair! Over time, he should neither win nor lose money.

The trick is that Nick and Eric are collaborating; in particular, they always make *opposite* guesses. So our assumption everyone is correct independently is wrong; actually the events that Nick is correct and Eric is correct are mutually exclusive! As a result, Dan can never win all the money on the table. When he guesses correctly, he always has to split his winnings with someone else. This lowers his overall expectation, as the corrected tree diagram below shows:



From this revised tree diagram, we can work out Dan's actual expected payoff:

$$\begin{aligned}\text{Ex}\,(\text{payoff}) &= 0\cdot 0 + 1\cdot\frac{1}{4} + 1\cdot\frac{1}{4} + 4\cdot 0 + (-2)\cdot 0 + (-2)\cdot\frac{1}{4} + (-2)\cdot\frac{1}{4} + 0\cdot 0\\ &= -\frac{1}{2}\end{aligned}$$

So he loses an average of a half-dollar per game!

Similar opportunities for subtle cheating come up in many betting games. For example, a group of friends once organized a football pool where each participant would guess the outcome of every game each week relative to the spread. This may mean nothing to you, but the upshot is that everyone was effectively betting on the outcomes of 12 or 13 coin tosses each week. The person who correctly predicts the most coin tosses won a lot of money. The organizer, thinking in terms of the first tree diagram, swore up and down that there was no way to get an unfair "edge". But actually the number of participants was small enough that just two players betting oppositely could gain a substantial advantage!

Another example involves a former MIT professor of statistics, Herman Chernoff. State lotteries are the worst gambling games around because the state pays out only a fraction of the money it takes in. But Chernoff figured out a way to win! Here are rules for a typical lottery:

- All players pay $1 to play and select 4 numbers from 1 to 36.
- The state draws 4 numbers from 1 to 36 uniformly at random.

- The state divides $1/2$ the money collected among the people who guessed correctly and spends the other half repairing the Big Dig.

This is a lot like our betting game, except that there are more players and more choices. Chernoff discovered that a small set of numbers was selected by a large fraction of the population— apparently many people think the same way. It was as if the players were collaborating to lose! If any one of them guessed correctly, then they'd have to split the pot with many other players. By selecting numbers uniformly at random, Chernoff was unlikely to get one of these favored sequences. So if he won, he'd likely get the whole pot! By analyzing actual state lottery data, he determined that he could win an average of 7 cents on the dollar this way!

22.2 Equivalent Definitions of Expectation

There are some other ways of writing the definition of expectation. Sometimes using one of these other formulations can make computing an expectation a lot easier. One option is to group together all outcomes on which the random variable takes on the same value.

Theorem 86.

$$\operatorname{Ex}(R) = \sum_{x \in \mathit{range}(R)} x \cdot \Pr(R = x)$$

Proof. We'll transform the left side into the right. Let $[R = x]$ be the event that $R = x$.

$$\begin{aligned}
\operatorname{Ex}(R) &= \sum_{w \in S} R(w) \Pr(w) \\
&= \sum_{x \in \text{ range}(R)} \sum_{w \in [R=x]} R(w) \Pr(w) \\
&= \sum_{x \in \text{ range}(R)} \sum_{w \in [R=x]} x \Pr(w) \\
&= \sum_{x \in \text{ range}(R)} \left(x \cdot \sum_{w \in [R=x]} \Pr(w) \right) \\
&= \sum_{x \in \text{ range}(R)} x \cdot \Pr(R = x)
\end{aligned}$$

On the second line, we break the single sum into two. The outer sum runs over all possible values x that the random variable takes on, and the inner sum runs over all outcomes taking on that value. Thus, we're still summing over every outcome in the sample space exactly once. On the last line, we use the definition of the probability of the event $[R = x]$. □

Corollary 87. *If R is a natural-valued random variable, then:*

$$\mathrm{Ex}\left(R\right) = \sum_{i=0}^{\infty} i \cdot \Pr\left(R = i\right)$$

There is another way to write the expected value of a random variable that takes on values only in the natural numbers, $\mathbb{N} = \{0, 1, 2, \ldots\}$.

Theorem 88. *If R is a natural-valued random variable, then:*

$$\mathrm{Ex}\left(R\right) = \sum_{i=0}^{\infty} \Pr\left(R > i\right)$$

Proof. Consider the sum:

$$\begin{array}{ccccccc} \Pr\left(R=1\right) & + & \Pr\left(R=2\right) & + & \Pr\left(R=3\right) & + & \cdots \\ & + & \Pr\left(R=2\right) & + & \Pr\left(R=3\right) & + & \cdots \\ & & & + & \Pr\left(R=3\right) & + & \cdots \\ & & & & & + & \cdots \end{array}$$

The columns sum to $1 \cdot \Pr\left(R = 1\right)$, $2 \cdot \Pr\left(R = 2\right)$, $3 \cdot \Pr\left(R = 3\right)$, etc. Thus, the whole sum is equal to:

$$\sum_{i=0}^{\infty} i \cdot \Pr\left(R = i\right) = \mathrm{Ex}\left(R\right)$$

Here, we're using Corollary 87. On the other hand, the rows sum to $\Pr\left(R > 0\right)$, $\Pr\left(R > 1\right)$, $\Pr\left(R > 2\right)$, etc. Thus, the whole sum is also equal to:

$$\sum_{i=0}^{\infty} \Pr\left(R > i\right)$$

These two expressions for the whole sum must be equal, which proves the theorem. □

22.2.1 Mean Time to Failure

Let's look at a problem where one of these alternative definitions of expected value is particularly helpful. 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 R be the number of hours until the crash, then the answer to our problem is $\mathrm{Ex}\left(R\right)$. This is a natural-valued variable, so we can use the formula:

$$\mathrm{Ex}\left(R\right) = \sum_{i=0}^{\infty} \Pr\left(R > i\right)$$

We have $R > i$ only if the system remains stable after i opportunities to crash, which happens with probability $(1-p)^i$. Plugging this into the formula above gives:

$$\begin{aligned}\operatorname{Ex}(R) &= \sum_{i=0}^{\infty}(1-p)^i \\ &= \frac{1}{1-(1-p)} \\ &= \frac{1}{p}\end{aligned}$$

The closed form on the second line comes from the formula for the sum of an infinite geometric series where the ratio of consecutive terms is $1-p$.

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 $1/p$.

22.2.2 Making a Baby Girl

A couple really wants to have a baby girl. 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 $p = \frac{1}{2}$. 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 1 boy.

22.3 An Expectation Paradox

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 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. The real answer is amazing: the expected number of picks that you need is *always infinite*, regardless of the distribution you choose!

This makes sense if you choose, say, the uniform distribution on $1, 2, \ldots, 6$. After all, there is a $\frac{1}{6}$ chance that I will pick 6. In this case, you must pick forever— you can never beat me!

In general, what is the probability that you need more than one pick? There are two cases to consider. If our numbers are different, then by symmetry there is a $\frac{1}{2}$ chance that mine is the larger, and you have to pick again. Otherwise, if our numbers are the same, then you lose and have to pick again. In either case, you need more than one pick with probability at least $\frac{1}{2}$.

What is the probability that you need more than two picks? Here is an erroneous argument. On the first pick, you beat me with probability about $\frac{1}{2}$. On the second pick, you beat me with probability about $\frac{1}{2}$ again. Thus the probability that you fail to beat me on both picks is only $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Therefore, the probability that you need more than two picks is around $\frac{1}{4}$. The error here is that beating me on your first pick is not independent of beating me on your second pick; multiplying the probabilities of these two events is therefore invalid!

Here is a correct argument for the probability that you need more than two picks. Suppose I pick z and then you pick y_1 and y_2. There are two cases. If there is a unique largest number among these three, then there is a $\frac{1}{3}$ chance that my number z is it, and you must pick again. After all, the largest number is equally likely to be chosen first, second, or third, regardless of the distribution. Otherwise, two or three of the numbers are tied for largest. My number is as likely to be among the largest as either of yours, so there is a better than $\frac{1}{3}$ chance that my number is as large as all of yours, and you must pick again. In both cases, you need more than two picks with probability at least $\frac{1}{3}$.

By the same argument, the probability that you need more than i picks is at least $1/(i+1)$. Suppose I pick z and you pick $y_1, y_2, \ldots, y_i$. Again, there are two cases. 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 $1/(i+1)$ 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 $1/(i+1)$ you must pick again. In both cases, with probability at least $1/(i+1)$, you need more than i picks to beat me.

These arguments suggest that you should choose a distribution such that ties are very rare. For example, you might choose the uniform distribution on $\{1, 2, \ldots, 10^{100}\}$. In this case, the probability that you need more than i picks to beat me is very close to $1/(i+1)$

for moderate values of i. 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.

Theorem 89. *Let T be the number picks you need to beat me. Then $\mathrm{Ex}(T) = \infty$.*

Proof.

$$\begin{aligned}\mathrm{Ex}(T) &= \sum_{i=0}^{\infty} \Pr(T > i) \\ &\geq \sum_{i=0}^{\infty} \frac{1}{i+1} \\ &= \infty\end{aligned}$$

Since T is a natural-valued random variable, we can use the special formulation of expectation on the first line. In the second step, we use the observation from above that you need more than i picks with probability at least $1/(i+1)$. The first n terms of this sum total to H_n, the n-th harmonic number, which is at least $\ln n$. Since $\ln n$ goes to infinity as n goes to infinity, the expectation is infinite! □

This phenomenon can cause all sorts of confusion. For example, suppose you have a communication network where each packet has a $1/i$ chance of being delayed by i or more steps. This sounds good; there is only a 1% chance of being delayed by 100 or more steps. But, by the argument above, 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:

$$\{1, -2, 3, -4, 5, -6, \ldots\}$$

Strictly speaking, we should qualify all theorems involving expectations with phrases such as "...provided all expectations exist." But we're going to leave such assumptions implicit. Fortunately, random variables without expectations rarely arise in practice.

22.4 Linearity of Expectation

Expected values obey a wonderful rule called ***linearity of expectation***. This says that the expectation of a sum is the sum of the expectationas.

Theorem 90 (Linearity of Expectation). *For every pair of random variables R_1 and R_2:*

$$\operatorname{Ex}(R_1 + R_2) = \operatorname{Ex}(R_1) + \operatorname{Ex}(R_2)$$

Proof. Let S be the sample space.

$$\begin{aligned}\operatorname{Ex}(R_1 + R_2) &= \sum_{w \in S} (R_1(w) + R_2(w)) \cdot \Pr(w) \\ &= \sum_{w \in S} R_1(w) \cdot \Pr(w) + \sum_{w \in S} R_2(w) \cdot \Pr(w) \\ &= \operatorname{Ex}(R_1) + \operatorname{Ex}(R_2)\end{aligned}$$

□

Linearity of expectation generalizes to any finite collection of random variables by induction:

Corollary 91. *For any random variables $R_1, R_2, \ldots, R_k$,*

$$\operatorname{Ex}(R_1 + R_2 + \cdots + R_k) = \operatorname{Ex}(R_1) + \operatorname{Ex}(R_2) + \cdots + \operatorname{Ex}(R_k)$$

The reason linearity of expectation is so wonderful is that, unlike many other probability rules, *the random variables are not required to be independent*. This probably sounds like a "yeah, whatever" technicality right now. But when you give an analysis using linearity of expectation, someone will almost invariably say, "No, you're wrong. There are all sorts of complicated dependencies here that you're ignoring." But that's the magic of linearity of expectation: you *can* ignore such dependencies!

22.4.1 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. At the start of these Notes, we showed that the expected value of one die is $3\frac{1}{2}$. We can find the expected value of the sum using linearity of expectation:

$$\begin{aligned}\operatorname{Ex}(R_1 + R_2) &= \operatorname{Ex}(R_1) + \operatorname{Ex}(R_2) \\ &= 3\frac{1}{2} + 3\frac{1}{2} \\ &= 7\end{aligned}$$

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! (This is provided that gluing somehow does not change weights to make the individual dice unfair.)

Proving that the expected sum is 7 with a tree diagram would be hard; there are 36 cases. And if we did not assume that the dice were independent, the job would be a nightmare!

22.4.2 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?

Without linearity of expectation, this would be a very difficult question to answer. We might try the following. Let the random variable R be the number of men that get their own hat. We want to compute $\mathrm{Ex}(R)$. By the definition of expectation, we have:

$$\mathrm{Ex}(R) = \sum_{k=0}^{\infty} k \cdot \Pr(R = k)$$

Now we're in trouble, because evaluating $\Pr(R = k)$ is a mess and we then need to substitute this mess into a summation. Furthermore, to have any hope, we would need to fix the probability of each permutation of the hats. For example, we might assume that all permutations of hats are equally likely.

Now let's try to use linearity of expectation. As before, let the random variable R be the number of men that get their own hat. The trick is to express R as a sum of indicator variables. In particular, let R_i be an indicator for the event that the ith man gets his own hat. That is, $R_i = 1$ is the event that he gets his own hat, and $R_i = 0$ is the event that he gets the wrong hat. The number of men that get their own hat is the sum of these indicators:

$$R = R_1 + R_2 + \cdots + R_n$$

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!

Let's take the expected value of both sides of the equation above and apply linearity of expectation:

$$\begin{aligned}\mathrm{Ex}(R) &= \mathrm{Ex}(R_1 + R_2 + \cdots + R_n) \\ &= \mathrm{Ex}(R_1) + \mathrm{Ex}(R_2) + \cdots + \mathrm{Ex}(R_n)\end{aligned}$$

All that remains is to compute the expected value of the indicator variables R_i. We'll use an elementary fact that is worth remembering in its own right:

Fact 3. *The expected value of an indicator random variable is the probability that the indicator is 1. In symbols:*

$$\mathrm{Ex}(I) = \Pr(I = 1)$$

Proof.

$$\begin{aligned}\mathrm{Ex}(I) &= 1 \cdot \Pr(I = 1) + 0 \cdot \Pr(I = 0) \\ &= \Pr(I = 1)\end{aligned}$$

□

So now we need only compute $\Pr(R_i = 1)$, which is the probability that the ith man gets his own hat. Since every man is as likely to get one hat as another, this is just $1/n$. Putting all this together, we have:

$$\begin{aligned}\operatorname{Ex}(R) &= \operatorname{Ex}(R_1) + \operatorname{Ex}(R_2) + \cdots + \operatorname{Ex}(R_n) \\ &= \Pr(R_1 = 1) + \Pr(R_2 = 1) + \cdots + \Pr(R_n = 1) \\ &= n \cdot \frac{1}{n} = 1.\end{aligned}$$

So we should expect 1 man to get his own hat back on average!

Notice that we did not assume that all permutations of hats are equally likely or even that all permutations are possible. We only needed to know that each man received his own hat with probability $1/n$. This makes our solution very general, as the next example shows.

22.4.3 The Chinese Appetizer Problem

There are n people at a circular table in a Chinese restaurant. On the table, there are n different appetizers arranged on a big Lazy Susan. Each person starts munching on the appetizer directly in front of him or her. Then someone spins the Lazy Susan so that everyone is faced with a random appetizer. What is the expected number of people that end up with the appetizer that they had originally?

This is just a special case of the hat-check problem, with appetizers in place of hats. In the hat-check problem, we assumed only that each man received his own hat with probability $1/n$. Beyond that, we made no assumptions about how the hats could be permuted. This problem is a special case because we happen to know that appetizers are cyclically shifted relative to their initial position. This means that either everyone gets their original appetizer back, or no one does. But our previous analysis still holds: the *expected* number of people that get their own appetizer back is 1.

Chapter 23

Expected Value II

23.1 The Expected Number of Events that Happen

Last time, we looked at linearity of expectation, the wonderful rule that say that the expectation of a sum is the sum of the expectations:

$$\mathrm{Ex}\left(R_1 + R_2 + \cdots + R_k\right) = \mathrm{Ex}\left(R_1\right) + \mathrm{Ex}\left(R_2\right) + \cdots + \mathrm{Ex}\left(R_k\right)$$

This gave us easy solutions to both the hat-check and Chinese appetizer problems.

More generally, suppose that $E_1, \ldots, E_n$ are arbitrary events that may or may not occur. For example, E_i might be the event that the i-th man gets his own hat back or the event that the i-th person gets her own appetizer after the lazy susan in spun. Then we can ask, "What is the expected number of these events that happen?" In the hat check problem, this amounts to asking how many men get their own hat back. The following theorem, based on linearity of expectation, provides a simple answer to the general question.

Theorem 92. *Let $E_1, \ldots, E_n$ be events. Then the expected number of events that occur is:*

$$\Pr\left(E_1\right) + \Pr\left(E_2\right) + \ldots + \Pr\left(E_n\right)$$

Proof. Let R_i be an indicator random variable for the event E_i:

$$R_i(w) = \begin{cases} 1 & \text{if } w \in E_i \\ 0 & \text{if } w \notin E_i \end{cases}$$

The number of events that happen is then the sum of these indicators:

$$T = R_1 + R_2 + \ldots + R_n$$

We can evaluate this sum using linearity of expectation:

$$\begin{aligned}\operatorname{Ex}(T) &= \sum_{i=1}^{n} \operatorname{Ex}(R_i) \\ &= \sum_{i=1}^{n} \Pr(R_i = 1) \\ &= \sum_{i=1}^{n} \Pr(E_i)\end{aligned}$$

On the second line, we're using the fact that the expected value of an indicator is equal to the probability that the indicator is 1. The last line uses the fact that R_i is an indicator for event E_i; thus, $R_i = 1$ if and only if event E_i happens. □

23.1.1 A Coin Problem— the Easy Way

Whenever you're confronted by a complicated expected value problem, your first question should be, "Can I use linearity of expectation?" Sometimes you can, sometimes you can't. But when using linearity is an option, it is often the best option by far. For example, suppose that we flip n fair coins. What is the expected number of heads that come up?

Let E_i be the event that the i-th coin is heads. Since each of the n coins is heads with probabilty $1/2$, the expected number of these events that occur is:

$$\Pr(E_1) + \Pr(E_2) + \ldots + \Pr(E_n) = n \cdot \frac{1}{2}$$

That was easy! Furthermore, we did *not* assume that the results of the coin tosses were mutually independent. Our solution is valid even if, say, pairs of coins are taped together, provided that each coin individually is fair.

23.1.2 The Hard Way

To better appreciate linearity of expectation, let's try to determine the expected number of heads on n coin tosses directly. Let the random variable R be the number of heads that come up in n tosses. Now we need to compute:

$$\operatorname{Ex}(R) = \sum_{k=0}^{n} k \cdot \Pr(R = k)$$

The probability that $R = k$ is the probability of flipping exactly k heads, which is

$$\Pr(R = k) = \binom{n}{k} 2^{-n}$$

since there are $\binom{n}{k}$ different heads/tails sequences with exactly k heads and each of these sequences has probability 2^{-n}. (Here we're assuming that the results of the coin tosses are mutually independent; without that assumption, we'd be really stuck!) Plugging this into expectation formula gives:

$$\mathrm{Ex}\,(R) = \sum_{k=0}^{n} k \cdot \binom{n}{k} 2^{-n}$$

It isn't obvious, but this nasty sum is actually equal $n/2$. (In fact, since we already know that $\mathrm{Ex}\,(R) = n/2$ via linearity of expectation, we have a "probabilistic proof" of this fact.)

That wasn't so easy *and* this approach is only valid if the coins are mutually independent.

23.2 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 kids 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.

23.2.1 A Solution Using Linearity of Expectation

Linearity of expectation is somewhat like induction and the pigeonhole principle; it's a simple idea that can be used in all sorts of ingenious ways. For example, we can use linearity of expecatation in a clever way to solve 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 k-th 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:

$$T = X_0 + X_1 + \ldots + X_{n-1}$$

Now let's focus our attention of the X_k, the length of the k-th 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 that we worked out last time. So we have:

$$\text{Ex}\left(X_k\right) = \frac{n}{n-k}$$

Linearity of expecatation, together with this observation, solves the coupon collector problem:

$$\begin{aligned}
\text{Ex}\left(T\right) &= \text{Ex}\left(X_0 + X_1 + \ldots + X_{n-1}\right) \\
&= \text{Ex}\left(X_0\right) + \text{Ex}\left(X_1\right) + \ldots + \text{Ex}\left(X_{n-1}\right) \\
&= \frac{n}{n-0} + \frac{n}{n-1} + \ldots + \frac{n}{3} + \frac{n}{2} + \frac{n}{1} \\
&= n\left(\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{3} + \frac{1}{2} + \frac{1}{1}\right) \\
&= nH_n
\end{aligned}$$

The summation on the next-to-last line is the n-th harmonic sum with the terms in reverse order. As you may recall, this sum is denoted H_n and is approximately $\ln n$.

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:

$$365H_{365} = 2364.6\ldots$$

23.3 Expected Value of a Product

Enough with sums! What about the expected value of a *product* of random variables? If R_1 and R_2 are independent, then the expected value of their product is the product of their expected values.

Theorem 93. *For* **independent** *random variables R_1 and R_2:*

$$\operatorname{Ex}(R_1 \cdot R_2) = \operatorname{Ex}(R_1) \cdot \operatorname{Ex}(R_2)$$

Proof. We'll transform the right side into the left side:

$$\begin{aligned}
\operatorname{Ex}(R_1) \cdot \operatorname{Ex}(R_2) &= \left(\sum_{x \in \text{Range}(R_1)} x \cdot \Pr(R_1 = x)\right) \cdot \left(\sum_{x \in \text{Range}(R_1)} y \cdot \Pr(R_2 = y)\right) \\
&= \sum_{x \in \text{Range}(R_1)} \sum_{y \in \text{Range}(R_2)} xy \Pr(R_1 = x) \Pr(R_1 = y) \\
&= \sum_{x \in \text{Range}(R_1)} \sum_{y \in \text{Range}(R_2)} xy \Pr(R_1 = x \cap R_1 = y)
\end{aligned}$$

The second line comes from multiplying out the product of sums. Then we used the fact that R_1 and R_2 are independent. Now let's group terms for which the product xy is the same:

$$\begin{aligned}
&= \sum_{z \in \text{Range}(R_1 \cdot R_2)} \sum_{x,y:\ xy=z} xy \Pr(R_1 = x \cap R_1 = y) \\
&= \sum_{z \in \text{Range}(R_1 \cdot R_2)} \left(z \sum_{x,y:\ xy=z} \Pr(R_1 = x \cap R_1 = y)\right) \\
&= \sum_{z \in \text{Range}(R_1 \cdot R_2)} z \cdot \Pr(R_1 \cdot R_2 = z) \\
&= \operatorname{Ex}(R_1 \cdot R_2)
\end{aligned}$$

□

23.3.1 The Product of Two Independent Dice

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:

$$\begin{aligned}\operatorname{Ex}(R_1 \cdot R_2) &= \operatorname{Ex}(R_1) \cdot \operatorname{Ex}(R_2) \\ &= 3\frac{1}{2} \cdot 3\frac{1}{2} \\ &= 12\frac{1}{4}\end{aligned}$$

On the first line, we're using Theorem 93. Then we use the result from last lecture that the expected value of one die is $3\frac{1}{2}$.

23.3.2 The Product of Two Dependent Dice

Suppose that the two dice are not independent; in fact, suppose that the second die is always the same as the first. Does this change the expected value of the product? Is the independence condition in Theorem 93 *really* necessary?

As before, 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 directly as follows:

$$\begin{aligned}\operatorname{Ex}(R_1 \cdot R_2) &= \operatorname{Ex}\left(R_1^2\right) \\ &= \sum_{i=1}^{6} i^2 \cdot \Pr(R_1 = i) \\ &= \frac{1^2}{6} + \frac{2^2}{6} + \frac{3^2}{6} + \frac{4^2}{6} + \frac{5^2}{6} + \frac{6^2}{6} \\ &= 15\frac{1}{6}\end{aligned}$$

The first step uses the fact that the outcome of the second die is always the same as the first. Then we expand $\operatorname{Ex}(R_1^2)$ using one of our formulations of expectation. Now that the dice are no longer independent, the expected value of the product has changed to $15\frac{1}{6}$. So the expectation of a product of dependent random variables need not equal the product of their expectations.

23.3.3 Corollaries

Theorem 93 extends to a collection of mutually independent variables.

Corollary 94. *If random variables $R_1, R_2, \ldots, R_n$ are mutually independent, then*

$$\operatorname{Ex}(R_1 \cdot R_2 \cdots R_n) = \operatorname{Ex}(R_1) \cdot \operatorname{Ex}(R_2) \cdots \operatorname{Ex}(R_n)$$

The proof uses induction, Theorem 93, and the definition of mutual independence. We'll omit the details.

Adjusting a random variable by an additive or multiplicative constant adjusts the expected value in the same way.

Corollary 95. *If R is a random variable and a and b are constants, then*

$$\operatorname{Ex}(aR + b) = a \operatorname{Ex}(R) + b$$

This corollary follows if we regard a and b as random variables that each take on one particular value with probability 1. Constants are always independent of other random variables, so the equation holds by linearity of expectation and Theorem 93.

We now know the expected value of a sum or product of random variables. Unfortunately, the expected value of a reciprocal is not so easy to characterize. Here is a flawed attempt.

False Corollary 96. *If R is a random variable, then*

$$\operatorname{Ex}\left(\frac{1}{R}\right) = \frac{1}{\operatorname{Ex}(R)}$$

As a counterexample, suppose the random variable R is 1 with probability $\frac{1}{2}$ and is 2 with probability $\frac{1}{2}$. Then we have:

$$\begin{aligned}
\frac{1}{\operatorname{Ex}(R)} &= \frac{1}{1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2}} \\
&= \frac{2}{3} \\
\operatorname{Ex}\left(\frac{1}{R}\right) &= \frac{1}{1} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \\
&= \frac{3}{4}
\end{aligned}$$

The two quantities are not equal, so the corollary must be false. But here is another false corollary, which we can actually "prove"!

False Corollary 97. *If $\operatorname{Ex}(R/T) > 1$, then $\operatorname{Ex}(R) > \operatorname{Ex}(T)$.*

"Proof". We begin with the if-part, multiply both sides by $\operatorname{Ex}(T)$, and then apply Theorem 93:

$$\begin{aligned}
\operatorname{Ex}(R/T) &> 1 \\
\operatorname{Ex}(R/T) \cdot \operatorname{Ex}(T) &> \operatorname{Ex}(T) \\
\operatorname{Ex}(R) &> \operatorname{Ex}(T)
\end{aligned}$$

□

This "proof" is bogus! The first step is valid only if $\mathrm{Ex}(T) > 0$. More importantly, we can't apply Theorem 93 in the second step because R/T and T are not necessarily independent. Unfortunately, the fact that Corollary 97 is false does not mean it is never used!

A RISC Paradox

The following data is taken from a paper by some famous professors. They wanted to show that programs on a RISC processor are generally shorter than programs on a CISC processor. For this purpose, they made a table of program lengths for some benchmark problems, which looked like this:

Benchmark	RISC	CISC	CISC / RISC
E-string search	150	120	0.8
F-bit test	120	180	1.5
Ackerman	150	300	2.0
Rec 2-sort	2800	1400	0.5
Average			1.2

Each row contains the data for one benchmark. The numbers in the first two columns are program lengths for each type of processor. The third column contains the ratio of the CISC program length to the RISC program length. Averaging this ratio over all benchmarks gives the value 1.2 in the lower right. The authors conclude that "CISC programs are 20% longer on average".

But there's a pretty serious problem here. Suppose we redo the final column, taking the inverse ratio, RISC / CISC instead of CISC / RISC.

Benchmark	RISC	CISC	RISC / CISC
E-string search	150	120	1.25
F-bit test	120	180	0.67
Ackerman	150	300	0.5
Rec 2-sort	2800	1400	2.0
Average			1.1

By exactly the same reasoning used by the authors, we could conclude that RISC programs are 10% longer on average than CISC programs! What's going on?

A Probabilistic Interpretation

To shed some light on this paradox, we can model the RISC vs. CISC debate with the machinery of probability theory.

Let the sample space be the set of benchmark programs. Let the random variable R be the length of the RISC program, and let the random variable C be the length of the CISC

program. We would like to compare the average length of a RISC program, $\mathrm{Ex}(R)$, to the average length of a CISC program, $\mathrm{Ex}(C)$.

To compare average program lengths, we must assign a probability to each sample point; in effect, this assigns a "weight" to each benchmark. One might like to weigh benchmarks based on how frequently similar programs arise in practice. But let's follow the original authors' lead. They assign each ratio equal weight in their average, so they're implicitly assuming that similar programs arise with equal probability. Let's do that same and make the sample space uniform. We can now compute $\mathrm{Ex}(R)$ and $\mathrm{Ex}(C)$ as follows:

$$\begin{aligned}\mathrm{Ex}(R) &= \frac{150}{4} + \frac{120}{4} + \frac{150}{4} + \frac{2800}{4} \\ &= 805 \\ \mathrm{Ex}(C) &= \frac{120}{4} + \frac{180}{4} + \frac{300}{4} + \frac{1400}{4} \\ &= 500\end{aligned}$$

So the average length of a RISC program is actually $\mathrm{Ex}(R) / \mathrm{Ex}(C) = 1.61$ times greater than the average length of a CISC program. RISC is even worse than either of the two previous answers would suggest!

In terms of our probability model, the authors computed C/R for each sample point and then averaged to obtain $\mathrm{Ex}(C/R) = 1.2$. This much is correct. However, they interpret this to mean that CISC programs are longer than RISC programs on average. Thus, the key conclusion of this milestone paper rests on Corollary 97, *which we know to be false!*

A Simpler Example

The root of the problem is more clear in the following, simpler example. Suppose the data were as follows.

Benchmark	Processor A	Processor B	B/A	A/B
Problem 1	2	1	1/2	2
Problem 2	1	2	2	1/2
Average			1.25	1.25

Now the statistics for processors A and B are exactly symmetric. Yet, from the third column we would conclude that Processor B programs are 25% longer on average, and from the fourth column we would conclude that Processor A programs are 25% longer on average. Both conclusions are obviously wrong. The moral is that *averages of ratios can be very misleading*. More generally, if you're computing the expectation of a quotient, think twice; you're going to get a value ripe for misuse and misinterpretation.

Chapter 24

Weird Happenings

> **Administrative note:** We've decided to provide an extra incentive on the final exam: if more than 80% of the class scores at least 1.25 times the class average (and the average is nonzero), then **everyone gets an *A* for the course!** We hope that this will encourage you all to study together so that you all succeed together.

Earlier this term there were 46 people in class, yet no two had the same birthday, which should happen only about 1 time in 17. Another term, students won the Monty Hall game 10 times out of 10. If everyone used the optimal "switch" strategy, this should happen only about 1 time in 57. But some students used the suboptimal "stay" strategy and they still won! This year the Boston Red Sox finally won the world series after managing to lose for 86 consecutive years. And in the recent presidential election, exit polls based on random sampling showed a decisive Kerry victory, though Bush got far more votes in the end! Weird things happen sometimes.

Yet many computer systems and algorithms are designed assuming that *weird things won't happen*. Many web sites are built assuming that many people will visit occasionally. So if everyone happened to visit such a site at the same time, by some weird coincidence, the system would collapse under the load. The Quicksort algorithm usually sorts a list of n items by comparing $O(n \log n)$ pairs of items to one another. If n is a million, then this is only a few million operations. But the algorithm relies on randomization; so if weird things happen, Quicksort could take a *half-trillion* operations instead! Hash tables are a standard data structure for rapidly storing and retreiving records. But, with sufficient bad luck, accesses can slow to a crawl. (We'll look at this example more closely later.) And the assumption that weird things won't happen is not only built into computer systems. What would happen to the phone system if everyone in America tried to call Topeka at the same time? Or what if everyone in Cambridge decided to go for ice cream at 7 PM next Friday? Our whole society is built around bets that weird things won't happen!

So to avoid catastrophe, we need mathematical tools to figure out just how unlikely weird things really are. That's today's topic.

24.1 The New Grading Policy

Let's return to the special grading policy introduced at the beginning: if more than 80% of the class scores at least 1.25 times the class average and the average is nonzero, then everyone gets an A for the course.

Suppose there are n students and the class average is $m > 0$. Let's look at the implications of these conditions, starting with the definition of class average:

$$\begin{aligned}\text{class average} &= \frac{\text{sum of all scores}}{\text{number of students}} \\ &> \frac{(0.80n)\cdot(1.25m)}{n} \\ &= m\end{aligned}$$

Thus, the class average must be greater than m— which was defined to *be* the class average. This is a contradiction! In the same way that not everyone can score above the average, there is no way more than 80% can score at least 1.25 times the average. In other words, the conditions of the new grading policy can never be satisfied! (Sorry.)

24.1.1 Markov's Inequality

Let's recast the analysis of the grading policy in probabilistic terms. Suppose that we select a student uniformly at random. Let the random variable X denote that student's final exam score. Then the class average is just $\operatorname{Ex}(X)$, and the conclusion reached above is a special case of an important general theorem:

Theorem 98 (Markov's Inequality). *Let X be a nonnegative random variable. If $c > 0$, then:*

$$\Pr(X \geq c) \leq \frac{\operatorname{Ex}(X)}{c}$$

Proof. If $\Pr(X \geq c) = 0$ or $\Pr(X < c) = 0$, then the theorem is trivially true. Otherwise, the Total Expectation theorem says:

$$\begin{aligned}\operatorname{Ex}(X) &= \Pr(X \geq c)\cdot\underbrace{\operatorname{Ex}(X \mid X \geq c)}_{\geq c} + \Pr(X < c)\cdot\underbrace{\operatorname{Ex}(X \mid X < c)}_{\geq 0} \\ &\geq \Pr(X \geq c)\cdot c\end{aligned}$$

Dividing the first and last expressions by c proves the theorem. □

For example, if we set $c = (5/4)\operatorname{Ex}(X)$ and $\operatorname{Ex}(X) > 0$, then the Markov Inequality says:

$$\Pr(X \geq (5/4)\operatorname{Ex}(X)) \leq \frac{\operatorname{Ex}(X)}{(5/4)\operatorname{Ex}(X)} = \frac{4}{5}$$

In words, the probability that a random student scores 1.25 times the class average or better can be at most 80%, provided the class average is greater than zero.

The Markov Inequality puts a limit on weird happenings; in particular, a nonnegative random variable can deviate far, above its expection only very rarely.

24.1.2 Limitations of the Markov Inequality

Let's apply the Markov Inequality to a couple more problems:

- Marilyn vos Savant's IQ is reportedly 228. How probable is such an IQ, given that the average is 100? Let Q be the IQ of a person selected uniformly at random. The Markov Inequality says:

  $$\Pr(Q \geq 228) \leq \frac{\mathrm{Ex}(Q)}{228} = \frac{100}{228}$$

  So less than half the population is as smart as Marilyn.

- Let D be the number rolled on a fair die. How often is a number greater than or equal to 4 rolled? Markov's Inequality says:

  $$\Pr(D \geq 4) \leq \frac{\mathrm{Ex}(D)}{4} = \frac{7/2}{4} = \frac{7}{8}$$

  Therefore, there is at most a $7/8$ chance of rolling a 5 or 6.

What's going on here?! These two conclusions are correct, but ridiculously weak. Far less than half the population has a 288 IQ, and rolling a 4 or more on a fair die has probabilty $1/2$— which is much less than $7/8$!

The difficulty is that the Markov Inequality is fed very little information, just a random variable's expectation and the fact that it's nonnegative. Based on this scanty information, the Markov Inequality gives the best possible bounds. Sometimes we don't know much about a random variable and the Markov Inequality is the only tool available. Other times, we can supercharge the Markov Inequality by incorporating additional data. We'll return to that approach in a little while.

24.2 The Tip of the Tail

A spaceship has n critical parts. Let E_k be the event that the k-th part fails on the next flight. If any critical part fails, then the spaceship is lost. This happens with probability:

$$\Pr(E_1 \cup \ldots \cup E_n)$$

What can be said about this quantity?

This sort of analysis comes up in the design of any critical system, where *weird* things can be very *bad* things. We define a set of events representing things that can go catastrophically wrong, and then try to compute the probability that something does.

24.2.1 Upper Bound: The Union Bound

We can *upper bound* the probability that some critical part fails using the Union Bound, which follows from the Markov Inequality:

Theorem 99 (Union Bound). *For events $E_1, \ldots, E_n$:*

$$\Pr(E_1 \cup \ldots \cup E_n) \leq \Pr(E_1) + \ldots + \Pr(E_n)$$

Proof. Let X be the number of the events $E_1, \ldots, E_n$ that occur. Then:

$$\begin{aligned}\Pr(E_1 \cup \ldots \cup E_n) &= \Pr(X \geq 1) \\ &\leq \frac{\mathrm{Ex}(X)}{1} \\ &= \Pr(E_1) + \ldots + \Pr(E_n)\end{aligned}$$

□

For example, suppose that the spaceship has 100,000 critical components and each has a 1-in-a-million probability of failure. Then the Union Bound says that the probability that *some* part fails is at most sum of the failure probabilities of the individual parts:

$$\begin{aligned}\Pr(E_1 \cup \ldots \cup E_{100{,}000}) &\leq \Pr(E_1) + \ldots + \Pr(E_{100{,}000}) \\ &= 100{,}000 \cdot \frac{1}{1{,}000{,}000} \\ &= \frac{1}{10}\end{aligned}$$

So the flight has at least a 90% chance of success.

Notice that the Union Bound makes no assumptions about whether the events E_i are independent or not. Thus, the Union Bound is great for conservative risk assessments; if we regard $E_1, \ldots, E_n$ as "bad events", then it gives an absolute upper bound on the probability that some "bad event" happens.

24.2.2 Lower Bound: "Murphy's Law"

Suppose that our spacecraft is a bit more cutting-edge. Now the critical components have the following characteristics:

- 10 components each which fail with probability $1/5$.
- 100 components each which fail with probability $1/40$.
- 1000 components each which fail with probability $1/200$.

In this design, components are carefully isolated so that they fail mutually independently. Suppose we just put our spaceship on the launch pad, "light the candle", and hope for the best. What is the probability that some component fails? We could crank out the exact answer, but there's a handy approximation available.

Theorem 100 ("Murphy's Law"). *If events $E_1, \ldots E_n$ are mutually independent and X is the number of these events that occur, then:*

$$\Pr(E_1 \cup \ldots \cup E_n) \geq 1 - e^{-\operatorname{Ex}(X)}$$

Proof.

$$\begin{aligned}\Pr(E_1 \cup \ldots \cup E_n) &= 1 - \Pr\left(\overline{E_1 \cup \ldots \cup E_n}\right) \\ &= 1 - \Pr\left(\overline{E_1} \cap \ldots \cap \overline{E_n}\right)\end{aligned}$$

Now we use then fact that $E_1, \ldots, E_n$ are mutually independent.

$$\begin{aligned}&= 1 - \Pr\left(\overline{E_1}\right) \cdots \Pr\left(\overline{E_n}\right) \\ &= 1 - (1 - \Pr(E_1)) \cdots (1 - \Pr(E_n))\end{aligned}$$

Next, we pull out the trusty inequality $1 - x \leq e^{-x}$, which holds for all x.

$$\begin{aligned}&\geq 1 - e^{-\Pr(E_1)} \cdots e^{-\Pr(E_n)} \\ &= 1 - e^{-(\Pr(E_1)+\ldots+\Pr(E_n))} \\ &= 1 - e^{-\operatorname{Ex}(X)}\end{aligned}$$

□

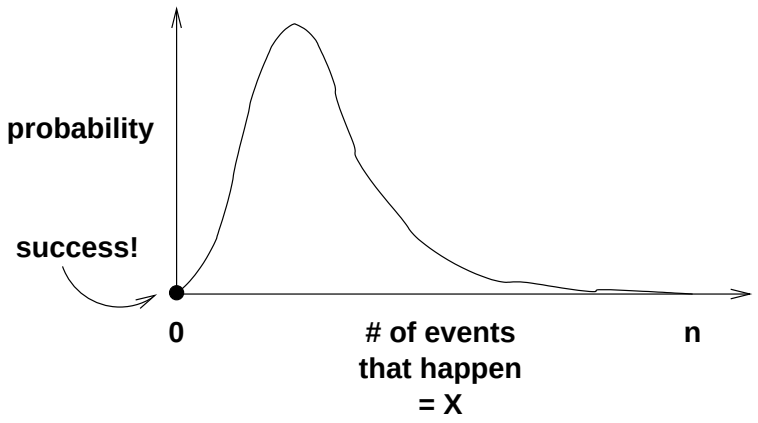
Theorem 100 can be regarded as a probabilistic version of Murphy's Law: *if you expect several things to go wrong, then something almost certainly will.* For the spaceship problem, the expected number component failures is:

$$\begin{aligned}\operatorname{Ex}(X) &= \Pr(E_1) + \ldots + \Pr(E_n) \\ &= 10 \cdot \frac{1}{5} + 100 \cdot \frac{1}{40} + 1000 \cdot \frac{1}{200} \\ &= 9.5\end{aligned}$$

So the probability of a successful flight is at most $e^{-9.5} \approx 0.000075$. Not a good gamble!

24.2.3 The Big Picture

Let's set the spaceship problem in a broader context. We have a sequence of events $E_1, \ldots, E_n$ and X is the number of these events that happen. For the second design spaceship design, the probability density function of X looks something like this:



The spaceship flies successfully only if no critical parts fail; that is, if $X = 0$. In terms of the picture, the flight is successful only at the absolute leftmost point of the distribution. So in analyzing the probability that the flight fails, we worked out general bounds on the probability that we're *not* at the tip of the tail of the distribution:

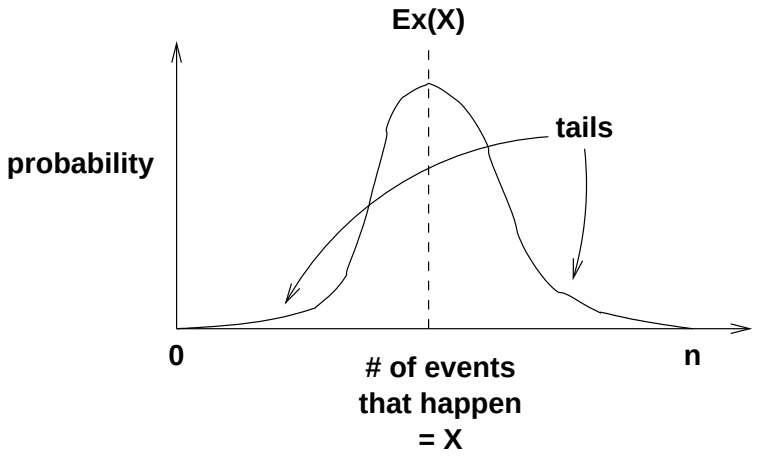
$$\underbrace{1 - e^{-\operatorname{Ex}(X)}}_{\substack{\text{"Murphy's Law"} \\ \text{(if } E_1, \ldots E_n \text{ are independent)}}} \leq \Pr(E_1 \cup \ldots \cup E_n) \leq \underbrace{\Pr(E_1) + \ldots + \Pr(E_n)}_{\substack{\text{Union Bound} \\ \text{(always holds)}}}$$

In particular, Murphy's Law says that if many independent events are expected to happen, then there's an extremely remote chance that none of them will. Thus, being out at the very tip of the tail is *extremely* weird. In fact, we're next going to show than being *anywhere* in either tail of the distribution is pretty unlikely.

24.3 Chernoff Bounds

MIT is admitting a new crop of students. The Institvte has offered admission to a few thousand applicants and carefully estimated the probability that each will accept, based on his or her interests, enthusiasm, and other likely offers. This calculation indicates that the expected number of new students is 1000, which is the ideal number. However, MIT must be wary of weird happenings. If the new class is too small, then expenses must be divided among fewer students, forcing tuition up. If the new class is too large, then living conditions and classes will be crowded. What is the probability that MIT must cope with signficantly fewer or more students?

Similar problems arise again and again in the analysis of computer systems and algorithms. The general theme is that there are many events that *can* occur, and we need to prove that the number that actually *do* occur is unlikely to be much greater or much less than the expected number. In terms of the probability density function, we're trying to show that the tails are small:



If the events are mutually independent, then we can get quick results to this effect from a powerful set of tools called ***Chernoff bounds***.

Theorem 101 (Chernoff Bounds). *Let $E_1, \ldots, E_n$ be a collection of mutually independent events, and let X be the number of these events that occur. Then:*

$$\Pr\left(X \leq (1-\delta)\operatorname{Ex}(X)\right) \leq e^{-\delta^2 \operatorname{Ex}(X)/2} \qquad \textit{when } 0 \leq \delta \leq 1$$

$$\Pr\left(X \geq (1+\delta)\operatorname{Ex}(X)\right) \leq e^{-\delta^2 \operatorname{Ex}(X)/3} \qquad \textit{when } 0 \leq \delta \leq 1$$

$$\Pr\left(X \geq c\operatorname{Ex}(X)\right) \leq e^{-(c\ln c - c + 1)\operatorname{Ex}(X)} \qquad \textit{when } c \geq 1$$

These are the supercharged Markov Inequalities that we mentioned earlier. The proof of this theorem is a bit intricate, so let's first apply it to the admissions problem.

24.3.1 MIT Admissions

Let E_k be the event that the k-th student accepts MIT's admission offer. Assume that all such events are mutually independent. Let X be the number of these events that occur; that is, X is the size of the incoming freshman class. The all-knowing admissions office has determined that $\operatorname{Ex}(X) = 1000$.

We can upper bound the probability that the new class contains 900 or fewer students using the first Chernoff inequality:

$$\begin{aligned}\Pr\left(X \geq 900\right) &= \Pr\left(X < \left(1 - \frac{1}{10}\right)\operatorname{Ex}(X)\right) \\ &\leq e^{-(1/10)^2 \cdot 1000/2} \\ &= e^{-5} \approx 0.0067\end{aligned}$$

On the other hand, we can upper bound the probability that 1200 or more new students come to MIT using the second inequality:

$$\begin{aligned}\Pr(X \geq 1200) &= \Pr\left(X > \left(1 + \frac{1}{5}\right)\operatorname{Ex}(X)\right) \\ &\leq e^{-(1/5)^2 \cdot 1000/3} \\ &= e^{-40/3} \approx 0.0000016\end{aligned}$$

If we want to estimate the probability of a complete disaster— say, 3000 or more students accept— then we can no longer use the second inequality; that holds only for deviations up to twice the expectation. We must use the third inequality instead. (Actually, the third Chernoff inequality always give an answer at least as good as the second; however, the second is often more convenient.)

$$\begin{aligned}\Pr(X \geq 3000) &= \Pr(X > 3 \cdot \operatorname{Ex}(X)) \\ &\leq e^{-(3\ln 3 - 3 + 1)\cdot 1000} \\ &< e^{-1295}\end{aligned}$$

That's pretty unlikely!

Like the Markov Inequality, a Chernoff bound may not yield the strongest possible conclusion because it is supplied with very little information about the random variable X. However, Chernoff bounds usually give *good* results and they're very easy to apply. So Chernoff bounds should among the first tools you reach for when you need to prove that weird things probably won't happen.

24.3.2 Proving Chernoff Bounds

Proving Chernoff bounds takes a good deal of work. To demonstrate the techniques involves, we'll prove the third inequality:

$$\Pr(X \geq c\operatorname{Ex}(X)) \leq e^{-(c\ln c - c + 1)\operatorname{Ex}(X)} \qquad \text{when } c \geq 1$$

The argument begins as follows:

$$\begin{aligned}\Pr(X \geq c\operatorname{Ex}(X)) &= \Pr\left(c^X \geq c^{c\operatorname{Ex}(X)}\right) \\ &\leq \frac{\operatorname{Ex}\left(c^X\right)}{c^{c\operatorname{Ex}(X)}} \qquad (24.1)\end{aligned}$$

In the first step, we exponentiate both sides of the inequality with base c. The probability remains unchanged because both inequalities describe the same event. The second step uses Markov's Inequality.

These two steps illustrate the key idea behind Chernoff bounds. Remember that Markov's Inequality upper bounds the probability that a random variable deviates above the mean.

For some probability distributions, Markov's Inequality gives a tight bound and for others it doesn't. Exponentiating before applying Markov's Inequality moves us to the sweet spot of the Markov Inequality, ensuring that we get good results. This isn't the sort of trick you'd immediately think up, but it works like a charm.

The next task is to find a convenient upper bound on the numerator in (24.1). There are roughly three steps: break the expression into little pieces, analyze each little piece, and then assemble them back together again. Let $I_1, \ldots, I_n$ be indicators for the events $E_1, \ldots, E_n$. In these terms, the number of events that happen is:

$$X = I_1 + \ldots + I_n$$

We'll use this as our starting point:

$$\begin{aligned}
\operatorname{Ex}\left(c^X\right) &= \operatorname{Ex}\left(c^{\sum_{k=1}^n I_k}\right) \\
&= \operatorname{Ex}\left(\prod_{k=1}^n c^{I_k}\right) \\
&= \prod_{k=1}^n \operatorname{Ex}\left(c^{I_k}\right)
\end{aligned}$$

The last step uses the fact that the indicators I_k are independent and the fact that functions of independent random variables are themselves independent. We've now decomposed the original expression into a product of little pieces, each involving a single indicator random variable. The next step is to compute the expected value of c^{I_k} using the definition of expectation:

$$\begin{aligned}
&= \prod_{k=1}^n \Pr\left(E_k\right) \cdot c^1 + \left(1 - \Pr\left(E_k\right)\right) \cdot c^0 \\
&= \prod_{k=1}^n 1 + (c-1)\Pr\left(E_k\right) \\
&\leq \prod_{k=1}^n e^{(c-1)\Pr\left(E_k\right)}
\end{aligned}$$

On the last line we're using the inequality $1 + x \leq e^x$. Now we put all the pieces back together again:

$$\begin{aligned}
&= e^{\sum_{k=1}^n (c-1)\Pr\left(E_k\right)} \\
&= e^{(c-1)\operatorname{Ex}(X)}
\end{aligned}$$

Plugging this upper bound into (24.1) gives:

$$\begin{aligned}\Pr\left(X \geq c\,\mathrm{Ex}\left(X\right)\right) &\leq \frac{e^{(c-1)\,\mathrm{Ex}\left(X\right)}}{c^{c\,\mathrm{Ex}\left(X\right)}} \\ &= e^{-(c\ln c - c + 1)\,\mathrm{Ex}\left(X\right)}\end{aligned}$$

This is the second Chernoff inequality. The third inequality follows by setting $c = 1 + \delta$ and using an approximation based on the Taylor series of the exponent. The proof of the first inequality has a similar structure, but differs in a few details.

A small corollary extends the usefulness of the Chernoff bounds in further. Sometimes we don't know $\mathrm{Ex}\left(X\right)$ exactly, but we at least know an upper bound. Fortunately, the second and third Chernoff inequalities still hold if we use this upper bound instead of the exact value of $\mathrm{Ex}\left(X\right)$.

Corollary 102. *The second and third bounds in Theorem 101 remain valid when all instances of* $\mathrm{Ex}\left(X\right)$ *are replaced by an upper bound on the expectation of* X*.*

The proof is a bunch of unenlightening algebra, which we'll omit.

24.4 Hashing

Suppose that we need to store credit histories for a great many people. We could create $n = 26^2$ bins labeled $AA, AB, AC, \ldots, ZZ$. Then we would store a person's record based on the first two letters of their name. For example, the record for "Lee, Edmond" would be stored in the bin labeled LE. Then, when we needed to look up Edmond's credit history, we would only need to search through the records in bin LE, rather than all the records.

In computer science, this trick for rapidly storing and retrieving records is called *hashing*. Each record consists of a *key* and *value*. A *hash function* maps each record's key to a bin. In our example, the keys were names, the values were credit histories, and the hash function mapped each name to its first two letters.

The fear in hashing is that one bin somehow ends up with too many records. In that case, retrieving any record in the overloaded bin is time-consuming, since there are so many to search through. This sort of imbalance is inevitable if the hash function is chosen poorly. For example, the hash function that maps each name to its first two letters is actually a horribe choice because some two letter prefixes are quite common (LEe, LEhman, LEighton) and others extremely uncommon (QZ, VZ, RR).

An ideal hash function would assign records to bins uniformly and independently at random. We can not achieve this goal in a rigorous sense— there is really no randomization involved— but this is still a decent practical model of a good hash function on typical data.

So let's assume that R records are hashed to N bins uniformly and independently at random. Let's see what our various probability tools say about the structure of the hash table.

24.4.1 The First Collision

When must there be a bin containing at least two records?

We can answer this question in two ways. In an absolute sense, the Pigeonhole Principle says that if there are $R > N$ records, then at least one of the N bins *must* contains two or more records.

Alternatively, we could regard the records as people and the bins as possible birthdays. Then the Birthday Principle says that there is an even chance that some bin contains two records when:

$$R \approx \sqrt{(2 \ln 2)N}$$

Thus, the first collision is likely to occur when the hash table still contains very few records. This can be frustrating. For example, if we create a hash table with a million bins, the probability that some bin contains two records is $1/2$ when the table contains only about 1177 records!

24.4.2 *N* Records in *N* Bins

Suppose that the number of records in our hash table is equal to the number of bins. So, for example, we might be storing a million records in a hash table with a million bins. What does the table look like?

Let's first consider a particular bin B. Let E_k be the event that the k-record is hashed to bin B. Since records are hashed uniformly, $\Pr(E_k) = 1/N$. And these events are independent because records are hashed to bins independently.

Now let X be the number of these events that happen; that is, X is the number of records hashed to bin B. The expected value of X is 1 since:

$$\begin{aligned} \mathrm{Ex}(X) &= \Pr(E_1) + \ldots + \Pr(E_N) \\ &= N \cdot 1/N \\ &= 1 \end{aligned}$$

We can use Murphy's Law to upper bound the probability that one or more records are hashed to bin B:

$$\begin{aligned} \Pr(E_1 \cup \ldots \cup E_N) &\geq 1 - e^{\mathrm{Ex}(X)} \\ &= 1 - 1/e \end{aligned}$$

So B is empty with probability at most $1/e$. Thus, the expected number of empty bins in the whole table is at most $N/e \approx 0.367N$ and this bound is asymptotically tight.

We can upper bound the probability that bin B gets more than c records using the third Chernoff inequality. Since $\mathrm{Ex}(X) = 1$, this has a simple form:

$$\Pr(X \geq c) \leq e^{-(c \ln c - c + 1)}$$

How high must we set the threshold c so that $\Pr(X > c)$, the probability that c or more records are stored in bin B, is still small? Let's try $c = \ln N$:

$$\begin{aligned}\Pr(X \geq \ln N) &\leq e^{-(\ln N \ln\ln N - \ln N + 1)} \\ &= \frac{1}{N^{\ln\ln N - 1 + 1/\ln N}}\end{aligned}$$

The dominant term in the exponent is $\ln\ln N$, which tends to infinity for large N. So this probability goes to zero faster than the inverse of any polynomial in N. So, asymptotically, it is very unlikely that any bin contains $\ln n$ or more records.

In fact, the probability that bin B contains more than c records is still less than $1/N^2$ when $c = e \ln N / \ln\ln N$. (This "log over log-log" function comes up pretty often. Say "nice function" and let it sniff you. Then give it a pat, and you'll be friends for life.) By the Union Bound, the probability that there exists *some* bin containing more than c records is at most:

$$\begin{aligned}\Pr\left(\text{some bin has } \geq \frac{e \ln N}{\ln\ln N}\right) &\leq \Pr(\text{bin 1 does}) + \ldots + \Pr(\text{bin } N \text{ does}) \\ &\leq N \cdot \frac{1}{N^2} \\ &= \frac{1}{N}\end{aligned}$$

So, for example, if we put a million records into a million-bin hash table, then there is less than a 1-in-a-million chance that any bin contains $15 > e \ln 10^6 / \ln\ln 10^6$ or more records.

24.4.3 All Bins Full

A final question: what is the expected number of records that we must add to a hash table in order for every bin to contain at least 1 record?

This is a restatement of the Coupon Collector problem, which we covered last time. The solution is $R = NH_n \approx N \ln N$. For example, if the hash table contains $N = 1,000,000$ bins, then we must add about 13.8 million records to get at least one record in every bin.

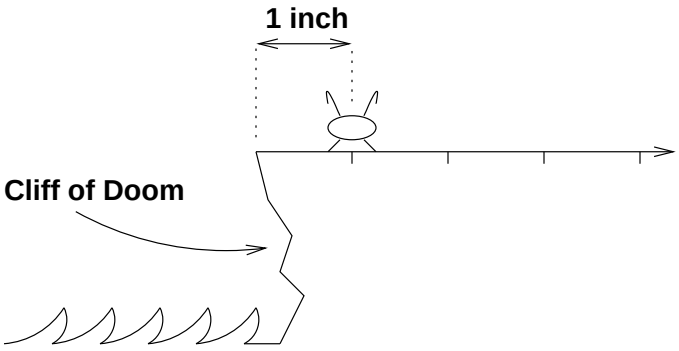
Unless something weird happens.

Chapter 25

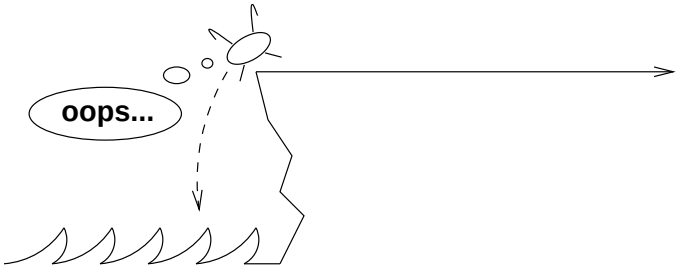
Random Walks

25.1 A Bug's Life

There is a small flea named Stencil. To his right, there is an endless flat plateau. One inch to his left is the Cliff of Doom, which drops to a raging sea filled with flea-eating monsters.



Each second, Stencil hops 1 inch to the right or 1 inch to the left with equal probability, independent of the direction of all previous hops. If he ever lands on the very edge of the cliff, then he teeters over and falls into the sea.



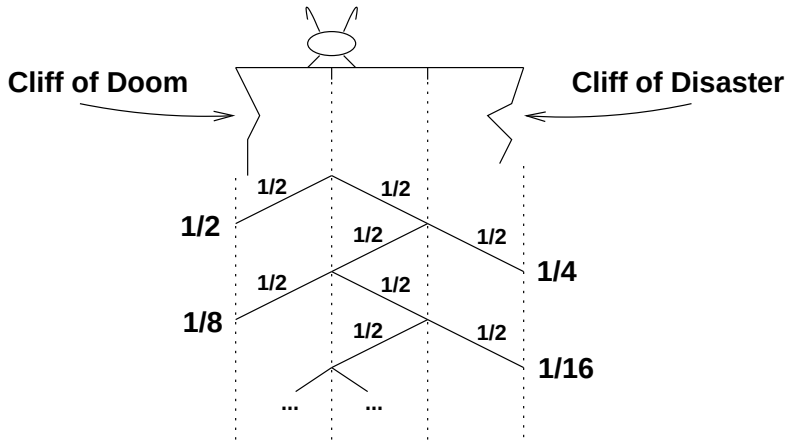
So, for example, if Stencil's first hop is to the left, he's fishbait. On the other hand, if his first few hops are to the right, then he may bounce around happily on the plateau for quite some time.

Our job is to analyze the life of Stencil. Does he have any chance of avoiding a fatal plunge? If not, how long will he hop around before he takes the plunge?

Stencil's movement is an example of a ***random walk***. A typical random walk involves some value that randomly wavers up and down over time. Many natural phenomona are nicely modeled by random walks. However, for some reason, they are traditionally discussed in the context of some social vice. For example, the value is often regarded as the position of a drunkard who randomly staggers left, staggers right, or just wobbles in place during each time step. Or the value is the wealth of a gambler who is continually winning and losing bets. So discussing random walks in terms of fleas is actually sort of elevating the discourse.

25.1.1 A Simpler Problem

Let's begin with a simpler problem. Suppose that Stencil is on a small island; now, not only is the Cliff of Doom 1 inch to his left, but also there is a Cliff of Disaster 2 inches to his right!



Below the figure, we've worked out a tree diagram for his possible fates. In particular, he falls off the Cliff of Doom on the left side with probability:

$$\begin{aligned}\frac{1}{2}+\frac{1}{8}+\frac{1}{32}+\ldots &= \frac{1}{2}\left(1+\frac{1}{4}+\frac{1}{16}+\ldots\right)\\ &= \frac{1}{2}\cdot\frac{1}{1-1/4}\\ &= \frac{2}{3}\end{aligned}$$

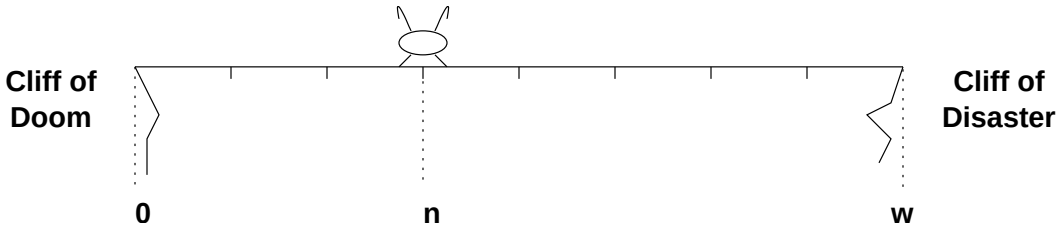
Similarly, he falls off the Cliff of Disaster on the right side with probability:

$$\frac{1}{4}+\frac{1}{16}+\frac{1}{64}+\ldots=\frac{1}{3}$$

There is a remaining possibility: he *could* hop back and forth in the middle of the table forever. However, we've already identified two disjoint events with probabilities $2/3$ and $1/3$, so this happy alternative must have probability 0.

25.1.2 A Big Island

Putting Stencil on such a tiny island was sort of cruel. Sure, he's probably carrying bubonic plague, but there's no reason to pick on the little fella. So suppose that we instead place him n inches from the left side of an island w inches across:



In other words, Stencil starts at position n and there are cliffs at positions 0 and w.

Now he has three possible fates: he could fall off the Cliff of Doom, fall off the Cliff of Disaster, or hop around on the island forever. We could compute the probabilities of these three events with a horrific summation, but fortunately there's a far easier method: we can use a linear recurrence.

Let R_n be the probability that Stencil falls to the right off the Cliff of Disaster, given that he starts at position n. In a couple special cases, the value of R_n is easy to determine. If he starts at position w, he falls from the Cliff of Disaster immediately, so $R_w = 1$. On the other hand, if he starts at position 0, then he falls from the Cliff of Doom immediately, so $R_0 = 0$.

Now suppose that our frolicking friend starts somewhere in the middle of the island; that is, $0 < n < w$. Then we can break the analysis of his fate into two cases based on the direction of his first hop:

- If his first hop is to the left, then he lands at position $n - 1$ and eventually falls off the Cliff of Disaster with probability R_{n-1}.
- On the other hand, if his first hop is to the right, then he lands at position $n + 1$ and eventually falls off the Cliff of Disaster with probability R_{n+1}.

Therefore, by the Total Probability Theorem, we have:

$$R_n = \frac{1}{2}R_{n-1} + \frac{1}{2}R_{n+1}$$

A Recurrence Solution

Let's assemble all our observations about R_n, the probability that Stencil falls from the Cliff of Disaster if he starts at position n:

$$\begin{aligned} R_0 &= 1 \\ R_w &= 0 \\ R_n &= \tfrac{1}{2}R_{n-1} + \tfrac{1}{2}R_{n+1} \qquad (0 < n < w) \end{aligned}$$

This is just a linear recurrence— and we know how to solve those! Uh, right? (We've attached a quick reference guide to be on the safe side.)

There is one unusual complication: in a normal recurrence, R_n is written a function of preceding terms. In this recurrence equation, however, R_n is a function of both a preceding term (R_{n-1}) and a *following* term (R_{n+1}). This is no big deal, however, since we can just rearrange the terms in the recurrence equation:

$$R_{n+1} = 2R_n - R_{n-1}$$

Now we're back on familiar territory.

Let's solve the recurrence. The characteristic equation is:

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

This equation has a double root at $x = 1$. There is no inhomogenous part, so the general solution has the form:

$$R_n = a \cdot 1^n + b \cdot n1^n = a + bn$$

Substituting in the boundary conditions $R_0 = 0$ and $R_w = 1$ gives two linear equations:

$$\begin{aligned} 0 &= a \\ 1 &= a + bw \end{aligned}$$

The solution to this system is $a = 0$, $b = 1/w$. Therefore, the solution to the recurrence is:

$$R_n = n/w$$

Interpreting the Answer

Our analysis shows that if we place Stencil n inches from the left side of an island w inches across, then he falls off the right side with probability n/w. For example, if Stencil is $n = 4$ inches from the left side of an island $w = 12$ inches across, then he falls off the right side with probability $n/w = 4/12 = 1/3$.

We can compute the probability that he falls off the *left* side by exploiting the symmetry of the problem: the probability the falls off the *left* side starting at position n is the same as the probability that he falls of the *right* side starting at position $w - n$, which is $(w - n)/n$.

Short Guide to Solving Linear Recurrences

A *linear recurrence* is an equation

$$\underbrace{f(n) = a_1 f(n-1) + a_2 f(n-2) + \ldots + a_d f(n-d)}_{\text{homogeneous part}} \quad \underbrace{+g(n)}_{\text{inhomogeneous part}}$$

together with boundary conditions such as $f(0) = b_0$, $f(1) = b_1$, etc.

1. Find the roots of the *characteristic equation*:

$$x^n = a_1 x^{n-1} + a_2 x^{n-2} + \ldots + a_k$$

2. Write down the *homogeneous solution*. Each root generates one term and the homogeneous solution is the sum of these terms. A nonrepeated root r generates the term $c_r r^n$, where c_r is a constant to be determined later. A root r with multiplicity k generates the terms:

$$c_{r_1} r^n, \quad c_{r_2} n r^n, \quad c_{r_3} n^2 r^n, \quad \ldots, \quad c_{r_k} n^{k-1} r^n$$

   where $c_{r_1}, \ldots, c_{r_k}$ are constants to be determined later.

3. Find a *particular solution*. This is a solution to the full recurrence that need not be consistent with the boundary conditions. Use guess and verify. If $g(n)$ is a polynomial, try a polynomial of the same degree, then a polynomial of degree one higher, then two higher, etc. For example, if $g(n) = n$, then try $f(n) = bn + c$ and then $f(n) = an^2 + bn + c$. If $g(n)$ is an exponential, such as 3^n, then first guess that $f(n) = c3^n$. Failing that, try $f(n) = bn3^n + c3^n$ and then $an^2 3^n + bn3^n + c3^n$, etc.

4. Form the *general solution*, which is the sum of the homogeneous solution and the particular solution. Here is a typical general solution:

$$f(n) = \underbrace{c2^n + d(-1)^n}_{\text{homogeneous solution}} + \underbrace{3n + 1}_{\text{particular solution}}$$

5. Substitute the boundary conditions into the general solution. Each boundary condition gives a linear equation in the unknown constants. For example, substituting $f(1) = 2$ into the general solution above gives:

$$\begin{aligned} & 2 = c \cdot 2^1 + d \cdot (-1)^1 + 3 \cdot 1 + 1 \\ \Rightarrow \quad & -2 = 2c - d \end{aligned}$$

   Determine the values of these constants by solving the resulting system of linear equations.

This is bad news. The probability that Stencil eventually falls off one cliff or the other is:

$$\frac{n}{w} + \frac{w-n}{w} = 1$$

There's no hope! The probability that he hops around on the island forever is zero. And there's even worse news. Let's go back to the original problem where Stencil is 1 inch from the left edge of an infinite plateau. In this case, the probability that he eventually falls into the sea is:

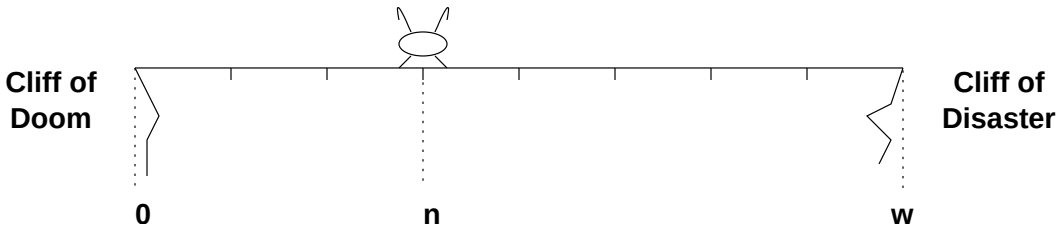
$$\lim_{w\to\infty} \frac{w-1}{w} = 1$$

Our little friend is doomed!

Hey, you know how in the movies they often make it look like the hero dies, but then he comes back in the end and everything turns out okay? Well, I'm not sayin' anything, just pointing that out.

25.1.3 Life Expectancy

On the bright side, Stencil may get to hop around for a while before he sinks beneath the waves. Let's use the same setup as before, where he starts out n inches from the left side of an island w inches across:



What is the expected number of hops he takes before falling off a cliff?

Let X_n be his expected lifespan, measured in hops. If he starts at either edge of the island, then he dies immediately:

$$X_0 = 0$$
$$X_w = 0$$

If he starts somewhere in the middle of the island ($0 < n < w$), then we can again break down the analysis into two cases based on his first hop:

- If his first hop is to the left, then he lands at position $n-1$ and can expect to live for another X_{n-1} steps.
- If his first hop is to the right, then he lands at position $n+1$ and his expected lifespan is X_{n+1}.

Thus, by the Total Expectation Theorem and linearity, his expected lifespan is:

$$X_n = 1 + \frac{1}{2}X_{n-1} + \frac{1}{2}X_{n+1}$$

The leading 1 accounts for his first hop.

Solving the Recurrence

Once again, Stencil's fate hinges on a recurrence equation:

$$\begin{aligned} X_0 &= 0 \\ X_w &= 0 \\ X_n &= 1 + \tfrac{1}{2}X_{n-1} + \tfrac{1}{2}X_{n+1} \qquad (0 < n < w) \end{aligned}$$

We can rewrite the last line as:

$$X_{n+1} = 2X_n - X_{n-1} - 2$$

As before, the characteristic equation is:

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

There is a double-root at 1, so the homogenous solution has the form:

$$X_n = a + bn$$

There's an inhomogenous term, so we also need to find a particular solution. Since this term is a constant, we should try a particular solution of the form $X_n = c$ and then try $X_n = c + dn$ and then $X_n = c + dn + en^2$ and so forth. As it turns out, the first two possibilities don't work, but the third does. Substituting in this guess gives:

$$\begin{aligned} X_{n+1} &= 2X_n - X_{n-1} - 2 \\ c + d(n+1) + e(n+1)^2 &= 2(c + dn + en^2) - (c + d(n-1) + e(n-1)^2 - 2 \\ e &= -1 \end{aligned}$$

All the c and d terms cancel, so $X_n = c + dn - n^2$ is a particular solution for all c and d. For simplicity, let's take $c = d = 0$. Thus, our particular solution is $X_n = -n^2$.

Adding the homogenous and particular solutions gives the general form of the solution:

$$X_n = a + bn - n^2$$

Substituting in the boundary conditions $X_0 = 0$ and $X_w = 0$ gives two linear equations:

$$\begin{aligned} 0 &= a \\ 0 &= a + bw - w^2 \end{aligned}$$

The solution to this system is $a = 0$ and $b = w$. Therefore, the solution to the recurrence equation is:

$$X_n = wn - n^2 = n(w - n)$$

Interpreting the Solution

Stencil's expected lifespan is $X_n = n(w-n)$, which is the *product* of the distances to the two cliffs. Thus, for example, if he's 4 inches from the left cliff and 8 inches from the right cliff, then his expected lifespan is $4 \cdot 8 = 32$.

Let's return to the original problem where Stencil has the Cliff of Doom 1 inch to his left and an infinite plateau to this right. (Also, cue the "hero returns" theme music.) In this case, his expected lifespan is:

$$\lim_{w\to\infty} 1(w-1) = \infty$$

Yes, Stencil is certain to eventually fall off the cliff into the sea— but his expected lifespan is infinite! This sounds almost like a contradiction, but both answers are correct!

Here's an informal explanation. The probability that Stencil falls from the Cliff of Doom on the k-th step is approximately $1/k^{3/2}$. Thus, the probability that he falls eventually is:

$$\Pr(\text{falls off cliff}) \approx \sum_{k=1}^{\infty} \frac{1}{k^{3/2}}$$

You can verify by integration that this sum converges. The exact sum actually converges to 1. On the other hand, the expected time until he falls is:

$$\mathrm{Ex}(\text{hops until fall}) \approx \sum_{k=1}^{\infty} k \cdot \frac{1}{k^{3/2}} = \sum_{k=1}^{\infty} \frac{1}{\sqrt{k}}$$

And you can verify by integration that this sum diverges. So our answers are compatible!

25.2 The Gambler's Ruin

We took the high road for a while, but now let's discuss random walks in more conventional terms. A gambler goes to Las Vegas with n dollars in her pocket. Her plan is to make only \$1 bets on red or black in roulette, each of which she'll win with probability $9/19 \approx 0.473$. She'll play until she's either broke or up \$100. What's the probability that she goes home a winner?

This is similar to the flea problem. The gambler's wealth goes up and down randomly, just like the Stencil's position. Going broke is analogous to falling off the Cliff of Doom and winning \$100 corresponds to falling off the Cliff of Disaster. In fact, the only substantive difference is that the gambler's wealth is slightly more likely to go down than up, whereas Stencil was equally likely to hop left or right.

We determined the flea usually falls off the nearest cliff. So we might expect that the gambler can improve her odds of going up \$100 before going bankrupt by bringing more

money to Vegas. But here's some actual data:

$n =$ starting wealth	probability she reaches $n + \$100$ before \$0
$\$100$	1 in 37649.619496...
$\$1000$	1 in 37648.619496...
$\$1,000,000,000$	1 in 37648.619496...

Except on the very low end, the amount of money she brings makes almost no difference! (The fact that only one digit changes from the first case to the second is a peripheral bit of bizarreness that we'll leave in your hands.)

25.2.1 Finding a Recurrence

We can approach the gambling problem the same way we studied the life of Stencil. Supose that the gambler starts with n dollars. She wins each bet with probability p and plays until she either goes bankrupt or has $w \geq n$ dollars in her pocket. (To be clear, w is the total amount of money she wants to end up with, not the amount by which she wants to increase her wealth.) Our objective is to compute R_n, the probability that she goes home a winner.

As usual, we begin by identifying some boundary conditions. If she starts with no money, then she's bankrupt immediately so $R_0 = 0$. On the other hand, if she starts with w dollars, then she's an instant winner, so $R_w = 1$.

Now we divide the analysis of the general situation into two cases based on the outcome of her first bet:

- She wins her first bet with probabilty p. She then has $n + 1$ dollars and probability R_{n+1} of reaching her goal of w dollars.
- She loses her first bet with probability $1 - p$. This leaves her with $n - 1$ dollars and probability R_{n-1} of reaching her goal.

Plugging these facts into the Total Probability Theorem gives the equation:

$$R_n = pR_{n+1} + (1-p)R_{n-1}$$

25.2.2 Solving the Recurrence

We now have a recurrence for R_n, the probability that the gambler reaches her goal of w dollars if she starts with n:

$$\begin{aligned} R_0 &= 0 \\ R_w &= 1 \\ R_n &= pR_{n+1} + (1-p)R_{n-1} \qquad (0 < n < w) \end{aligned}$$

The characteristic equation is:

$$px^2 - x + (1-p) = 0$$

The quadratic formula gives the roots:

$$\begin{aligned} x &= \frac{1 \pm \sqrt{1 - 4p(1-p)}}{2p} \\ &= \frac{1 \pm \sqrt{(1-2p)^2}}{2p} \\ &= \frac{1 \pm (1-2p)}{2p} \\ &= \frac{1-p}{p} \text{ or } 1 \end{aligned}$$

There's an important point lurking here. If the gambler is equally likely to win or lose each bet, then $p = 1/2$, and the characteristic equation has a double root at $x = 1$. This is the situation we considered in the flea problem. The double root led to a general solution of the form:

$$R_n = a + bn$$

Now suppose that the gambler is *not* equally likely to win or lose each bet; that is, $p \neq 1/2$. Then the two roots of the characteristic equation are different, which means that the solution has a completely different form:

$$R_n = a \cdot \left(\frac{1-p}{p}\right)^n + b \cdot 1^n$$

In mathematical terms, this is where the flea problem and the gambler problem take off in completely different directions: in one case we get a linear solution and in the other we get an exponenetial solution!

Anyway, substituting the boundary conditions into the general form of the solution gives a system of linear equations:

$$\begin{aligned} 0 &= a + b \\ 1 &= a \cdot \left(\frac{1-p}{p}\right)^w + b \end{aligned}$$

Solving this system, gives:

$$a = \frac{1}{\left(\frac{1-p}{p}\right)^w - 1} \qquad b = -\frac{1}{\left(\frac{1-p}{p}\right)^w - 1}$$

Substituting these values back into the general solution gives:

$$\begin{aligned} R_n &= \left(\frac{1}{\left(\frac{1-p}{p}\right)^w - 1}\right) \cdot \left(\frac{1-p}{p}\right)^n - \frac{1}{\left(\frac{1-p}{p}\right)^w - 1} \\ &= \frac{\left(\frac{1-p}{p}\right)^n - 1}{\left(\frac{1-p}{p}\right)^w - 1} \end{aligned}$$

(Suddenly, Stencil's life doesn't seem so bad, huh?)

25.2.3 Interpreting the Solution

We have an answer! If the gambler starts with n dollars and wins each bet with probability p, then the probability she reaches w dollars before going broke is:

$$\frac{\left(\frac{1-p}{p}\right)^n - 1}{\left(\frac{1-p}{p}\right)^w - 1}$$

Let's try to make sense of this expression. If the game is biased against her, as with roulette, then $1-p$ (the probability she loses) is greater than p (the probability she wins). If n, her starting wealth, is also reasonably large, then both exponentiated fractions are big numbers and the -1's don't make much difference. Thus, her probability of reaching w dollars is very close to:

$$\left(\frac{1-p}{p}\right)^{n-w}$$

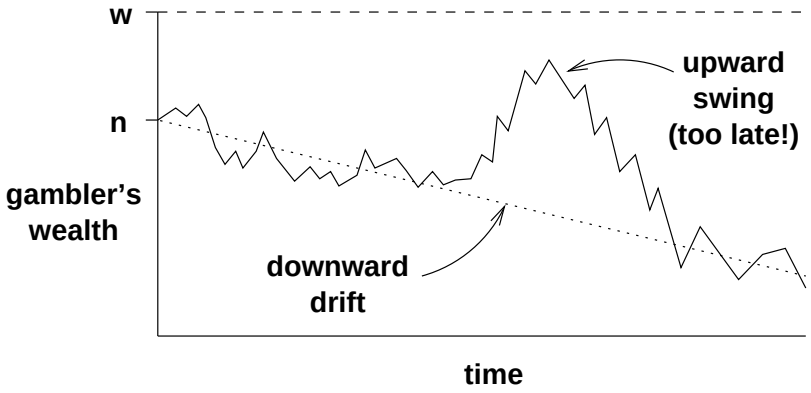
In particular, if she is hoping to come out \$100 ahead in roulette, then $p = 9/19$ and $w = n + 100$, so her probability of success is:

$$\left(\frac{10}{9}\right)^{-100} = 1 \text{ in } 37648.619496$$

This explains the strange number we arrived at earlier!

25.2.4 Some Intuition

Why does the gambler's starting wealth have so little impact on her probability of coming out ahead? Intuitively, there are two forces at work. First, the gambler's wealth has random upward and downward *swings* due to runs of good and bad luck. Second, her wealth has a steady, downward *drift* because she has a small expected loss on every bet. The situation is illustrated below:

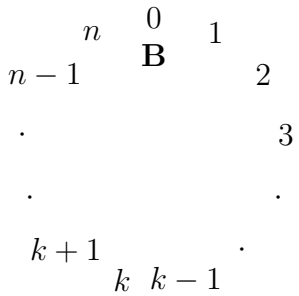


For example, in roulette the gambler wins a dollar with probability $9/19$ and loses a dollar with probability $10/19$. Therefore, her expected return on each bet is $1 \cdot 9/10 + (-1) \cdot 10/19 = -1/19$. Thus, her expected wealth drifts downward by a little over 5 cents per bet.

One might think that if the gambler starts with a billion dollars, then she will play for a long time, so at some point she should have a lucky, upward swing that puts her \$100 ahead. The problem is that her capital is steadily drifting downward. And after her capital drifts down a few hundred dollars, she needs a huge upward swing to save herself. And such a huge swing is extremely improbable. So if she does not have a lucky, upward swing early on, she's doomed forever. As a rule of thumb, *drift* dominates *swings* over the long term.

25.3 Pass the Broccoli

Here's a game that involves a random walk. There are $n+1$ people, numbered $0, 1, \ldots, n$, sitting in a circle:



The B indicates that person 0 has a big stalk of nutritious broccoli, which provides 250% of the US recommended daily allowance of vitamin C and is also a good source of vitamin A and iron. (Typical for a random walk problem, this game orginally involved a pitcher of beer instead of a broccoli. We're taking the high road again.)

Person 0 passes the broccoli either to the person on his left or the person on his right with equal probability. Then, that person also passes the broccoli left or right at random and so on. After a while, everyone in an arc of the circle has touched the broccoli and everyone outside that arc has not. Eventually, the arc grows until all but one person has touched the broccoli. That final person is declared the winner and gets to keep the broccoli!

Suppose that you allowed to position yourself anywhere in the circle. Where should you stand in order to maxmize the probability that you win? You shouldn't be person 0; you can't win in that position. The answer is "intuitively obvious": you should stand as far as possible from person 0 at position $n/2$.

Let's verify this intuition. Suppose that you stand at position $k \neq 0$. At some point, the broccoli is going to end up in the hands of one of your neighbors. This has to happen eventually; the game can't end until at least one of them touches it. Let's say that person $k-1$ gets the broccoli first. Now let's cut the circle between yourself and your other neighbor, person $k+1$:

$$\underset{\mathbf{B}}{k} \quad (k-1) \quad \ldots \quad 3 \quad 2 \quad 1 \quad 0 \quad n \quad (n-1) \quad \ldots \quad (k+1)$$

Now there are two possibilities. If the broccoli reaches you before it reaches person $k+1$, then you lose. But if the broccoli reaches person $k+1$ before it reaches you, then every other person has touched the broccoli and you win! This is just the flea problem all over again: the probability that the broccoli hops $n-1$ people to the right (reaching person $k+1$) before it hops 1 person to the left (reaching you) is $1/n$. Therefore, our intution was compeletely wrong: *your probability of winning is* $1/n$ *regardless of where you're standing!*