CS70: Discrete Mathematics and Probability Theory
Course Reader

Amir Kamil

Summer 2011

Contents

CS 70 Discrete Mathematics and Probability Theory
Summer 2011 Kamil Note 1

Course Outline

CS70 is a course on discrete mathematics and probability theory, especially tailored for EECS students. The purpose of the course is to teach you about:

- **Fundamental ideas in computer science and electrical engineering:**
  - Boolean logic
  - Modular arithmetic, public-key cryptography, error-correcting codes, secret sharing protocols
  - The power of randomization ("flipping coins") in computation: load balancing, hashing, inference, overcoming noise in communication channels
  - Uncomputability and the halting problem

  Many of these concepts form a foundation for more advanced courses in EECS. science.
- **Precise, reliable, powerful thinking:**
  - Proofs of correctness. These are essential to analyzing algorithms and programs
  - Induction and recursion
  - Probability theory
- **Problem solving skills:**
  - These are emphasized in the discussion sections and homeworks.

Course outline (abbreviated):

- Propositions, propositional logic and proofs
- Mathematical induction, recursion
- The stable marriage problem
- Modular arithmetic, the RSA cryptosystem
- Polynomials over finite fields and their applications: error-correcting codes, secret sharing
- Probability and probabilistic algorithms: load balancing, hashing, expectation, variance, Chebyshev bounds, conditional probability, Bayesian inference, law of large numbers, Central Limit Theorem.
- Diagonalization, self-reference, and uncomputability

Getting Started

In order to be fluent in mathematical statements, you need to understand the basic framework of the language of mathematics. This first week, we will start by learning about what logical forms mathematical theorems may take, and how to manipulate those forms to make them easier to prove. In the next few lectures, we will learn several different methods of proving things.

Propositions

A **proposition** is a statement which is either true or false.

These statements are all propositions:

(1) $\sqrt{3}$ is irrational.
(2) $1+1=5$.
(3) Julius Caesar had 2 eggs for breakfast on his 10^{th} birthday.

These statements are clearly not propositions:

(4) $2+2$.
(5) $x^2+3x=5$.

These statements aren't propositions either (although some books say they are). Propositions should not include fuzzy terms.

(6) Arnold Schwarzenegger often eats broccoli. (What is "often?")
(7) Barack Obama is popular. (What is "popular?")

Propositions may be joined together to form more complex statements. Let P, Q, and R be variables representing propositions (for example, P could stand for "3 is odd"). The simplest way of joining these propositions together is to use the connectives "and", "or" and "not."

(1) **Conjunction**: $P \wedge Q$ ("P and Q"). True only when both P and Q are true.

(2) **Disjunction**: $P \vee Q$ ("P or Q"). True when at least one of P and Q is true.

(3) **Negation**: $\neg P$ ("not P"). True when P is false.

Statements like these, with variables, are called *propositional forms.* If we let P stand for the proposition "3 is odd," Q stand for "4 is odd," and R for "5 is even," then the propositional forms $P \wedge R$, $P \vee R$ and $\neg Q$ are false, true, and true, respectively. Note that $P \vee \neg P$ is always true, regardless of the truth value of P. A propositional form that is always true regardless of the truth values of its variables is called a *tautology*. A statement such as $P \wedge \neg P$, which is always false, is called a *contradiction*.

A useful tool for describing the possible values of a propositional form is a **truth table**. Truth tables are the same as function tables. You list all possible input values for the variables, and then list the outputs given those inputs. (The order does not matter.)

Here are truth tables for conjunction, disjunction and negation:

P	Q	$P \wedge Q$
T	T	T
T	F	F
F	T	F
F	F	F

P	Q	$P \vee Q$
T	T	T
T	F	T
F	T	T
F	F	F

P	$\neg P$
T	F
F	T

The most important and subtle propositional form is an **implication**:

> (4) **Implication**: $P \Longrightarrow Q$ ("P implies Q"). This is the same interpreted as "If P, then Q." It is logically equivalent to $(\neg P) \vee Q$. In other words, $P \Longrightarrow Q$ is a more concise (and possibly more intuitive) way of writing $(\neg P) \vee Q$, but it has exactly the same meaning.

In the expression $P \Longrightarrow Q$, P is called the *hypothesis* of the implication, and Q is the *conclusion*. [1]

Examples of implications:

> If you stand in the rain, then you'll get wet.
> If you passed the class, you received a certificate.

An implication $P \Longrightarrow Q$ is false only when P is true and Q is false. For example, the first statement would be false only if you stood in the rain but didn't get wet. The second statement above would be false only if you passed the class yet you didn't receive a certificate.

Here is the truth table for $P \Longrightarrow Q$:

P	Q	$P \Longrightarrow Q$	$\neg P \vee Q$
T	T	T	T
T	F	F	F
F	T	T	T
F	F	T	T

Note that $P \Longrightarrow Q$ is always true when P is false. This means that many statements that sound nonsensical in English are true, mathematically speaking. Examples are statements like: "If pigs can fly, then horses can read" or "If 14 is odd then $1+2=18$." When an implication is stupidly true because the hypothesis is false, we say that it is *vacuously true*. Note also that $P \Longrightarrow Q$ is logically equivalent to $\neg P \vee Q$, as can be seen from the above truth table.

[1] P is also called the *antecedent* and Q the *consequent*.

$P \implies Q$ is the most common form mathematical theorems take. Here are some of the different ways of saying it:

(1) If P, then Q.
(2) Q if P.
(3) P only if Q.
(4) P is sufficient for Q.
(5) Q is necessary for P.

If both $P \implies Q$ and $Q \implies P$ are true, then we say "P if and only if Q" (abbreviated P iff Q). Formally, we write $P \iff Q$. $P \iff Q$ is true only when P and Q have the same truth values.

For example, if we let P be "3 is odd," Q be "4 is odd," and R be "6 is even," then $P \implies R$, $Q \implies P$ (vacuously), and $R \implies P$. Because $P \implies R$ and $R \implies P$, we have $P \iff R$ (i.e., P if and only if R).

Given an implication $P \implies Q$, we can also define its

(a) **Contrapositive**: $\neg Q \implies \neg P$
(b) **Converse**: $Q \implies P$

The contrapositive of "If you passed the class, you received a certificate" is "If you did not get a certificate, you did not pass the class." The converse is "If you got a certificate, you passed the class." Does the contrapositive say the same thing as the original statement? Does the converse?

Let's look at the truth tables:

P	Q	$\neg P$	$\neg Q$	$P \implies Q$	$Q \implies P$	$\neg Q \implies \neg P$	$P \iff Q$
T	T	F	F	T	T	T	T
T	F	F	T	F	T	F	F
F	T	T	F	T	F	T	F
F	F	T	T	T	T	T	T

Note that $P \implies Q$ and its contrapositive have the *same* truth values everywhere in their truth tables; propositional forms having the same truth values are said to be **logically equivalent**, written "$\equiv$". Thus we may write $(P \implies Q) \equiv (\neg Q \implies \neg P)$. Many students confuse the contrapositive with the converse: note that $P \implies Q$ and $\neg Q \implies \neg P$ are logically equivalent, but $P \implies Q$ and $Q \implies P$ are not!

When two propositional forms are logically equivalent, we can think of them as "meaning the same thing." We will see next time how useful this can be for proving theorems.

Quantifiers

The mathematical statements you'll see in practice will not be made up of simple propositions like "3 is odd." Instead you'll see statements like:

(1) For all natural numbers n, $n^2 + n + 41$ is prime.
(2) If n is an odd integer, so is n^3.
(3) There is an integer k that is both even and odd.

In essence, these statements assert something about lots of simple propositions all at once. For instance, the first statement is asserting that 0^2+0+41 is prime, 1^2+1+41 is prime, and so on. The last statement is saying that as k ranges over every possible integer, we will find at least one value for which the statement is satisfied.

Why are the above three examples considered to be propositions, while earlier we claimed that "$x^2+3x=5$" was not? The reason is these three examples are in some sense "complete", where "$x^2+3x=5$" is "incomplete" because whether it is true or false depends upon the value of x.

To explain it another way, in these three examples, there is an underlying "universe" that we are working in. The statements are then *quantified* over that universe. To express these statements mathematically we need two **quantifiers**: The *universal quantifier* $\forall$ ("for all") and the *existential quantifer* $\exists$ ("there exists"). The three statements above can then be expressed using quantifiers as follows[2]

(1) $(\forall n \in \mathbb{N})(n^2+n+41 \text{ is prime})$.
(2) $(\forall n \in \mathbb{Z})(n \text{ is odd} \implies n^3 \text{ is odd})$.
(3) $(\exists n \in \mathbb{Z})(n \text{ is odd} \wedge n \text{ is even})$.

Here $\mathbb{N} = \{0,1,2,\dots\}$ is the set of natural numbers and $\mathbb{Z} = \{\dots,-2,-1,0,1,2,\dots\}$ is the set of integers.

As another example, consider the statement "some mammals lay eggs." Mathematically, "some" means "at least one," so the statement is saying "There exists a mammal x such that x lays eggs." If we let our universe U be the set of mammals, then we can write: $(\exists x \in U)(x \text{ lays eggs})$. (Sometimes, when the universe is clear, we omit U and simply write $\exists x(x \text{ lays eggs})$.)

Some statements can have multiple quantifiers. As we will see, however, quantifiers do not commute. You can see this just by thinking about English statements. Consider the following (rather gory) example:

Example:

> "Every time I ride the subway in New York, somebody gets stabbed."
> "There is someone, such that every time I ride the subway in New York, that someone gets stabbed."

The first statement is saying that every time I ride the subway someone gets stabbed, but it could be a different person each time. The second statement is saying something truly horrible: that there is some poor guy Joe with the misfortune that every time I get on the New York subway, there is Joe, getting stabbed again. (Poor Joe will run for his life the second he sees me.)

Mathematically, we are quantifying over two universes: $T = \{\text{times when I ride on the subway}\}$ and $P = \{\text{people}\}$. The first statement can be written: $(\forall t \in T)(\exists p \in P)(p \text{ gets stabbed at time } t)$. The second statement says: $(\exists p \in P)(\forall t \in T)(p \text{ gets stabbed at time } t)$.

Let's look at a more mathematical example:

Consider

1. $(\forall x \in \mathbb{Z})(\exists y \in \mathbb{Z})(x < y)$
2. $(\exists y \in \mathbb{Z})(\forall x \in \mathbb{Z})(x < y)$

[2]Note that in a *finite* universe, we can express existentially and universally quantified propositions without quantifiers, using disjunctions and conjunctions respectively. For example, if our universe U is {1, 2, 3, 4}, then $(\exists x P(x))$ is logically equivalent to $P(1) \vee P(2) \vee P(3) \vee P(4)$, and $(\forall x P(x))$ is logically equivalent to $P(1) \wedge P(2) \wedge P(3) \wedge P(4)$. However, in an infinite universe, such as the natural numbers, this is not possible.

The first statement says that, given an integer, I can find a larger one. The second statement says something very different: that there is a largest integer! The first statement is true, the second is not.

A word on notation: some mathematicians write quantified statements using a slightly different notation. Instead of something like $(\forall n \in \mathbb{N})(n^2 \geq 0)$, you might see it written as $\forall n \in \mathbb{N} \,.\, n^2 \geq 0$. Either of these gets the job done—it's simply a matter of stylistic preferences. Finally, sometimes you might see people write $\forall n\, n^2 \geq 0$ or $(\forall n)(n^2 \geq 0)$. This way of writing it is potentially ambiguous, because it does not make the universe explicit; it should only be used in cases where there is no possibility of confusion about the universe being quantified over.

Quantifiers and Negation

What does it mean for a proposition P to be false? It means that its negation $\neg P$ is true. Often, we will need to negate a quantified proposition (the motivation for this will become more obvious next lecture when we look at proofs). For now, let's look at an example of how to go about this.

Example:

Assume that the universe is $U = \{1,2,3,4\}$ and let $P(x)$ denote the propositional formula "$x^2 > 10$." Check that $\exists x P(x)$ is true but $\forall x P(x)$ is false. Observe that both $\neg(\forall x P(x))$ and $\exists x \neg P(x)$ are true because $P(1)$ is false. Also note that both $\forall x \neg P(x)$ and $\neg(\exists x P(x))$ are false, since $P(4)$ is true. The fact that each pair of statements had the same truth value is no accident, as the formulae

$$\neg(\forall x P(x)) \equiv \exists x \neg P(x)$$

$$\neg(\exists x P(x)) \equiv \forall x \neg P(x)$$

are laws that hold for any proposition P quantified over any universe. (Recall that "$\equiv$" means logically equivalent.) It is helpful to think of English sentences to convince yourself (informally) that these laws are true. For example, assume that we are working within the universe $\mathbb{Z}$ (the set of all integers), and that $P(x)$ is the proposition "x is odd." We know that the statement $(\forall x P(x))$ is false, since not every integer is odd. Therefore, we expect its negation, $\neg(\forall x P(x))$, to be true. But how would you say the negation in English? Well, if it is not true that every integer is odd, then that must mean there is some integer which is not odd (i.e., even). How would this be written in propositional form? That's easy, it's just: $(\exists x \neg P(x))$.

To see a more complex example, fix some universe and propositional formula $P(x,y)$. Assume we have the proposition $\neg(\forall x \exists y P(x,y))$ and we want to push the negation operator inside the quantifiers. By the above laws, we can do it as follows:

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

Notice that we broke the complex negation into a smaller, easier problem as the negation propagated itself through the quantifiers. Note also that the quantifiers "flip" as we go.

Let's look at a trickier example:

Write the sentence "there are at least three distinct integers x that satisfy $P(x)$" as a proposition using quantifiers! One way to do it is

$$\exists x \exists y \exists z (x \neq y \wedge y \neq z \wedge z \neq x \wedge P(x) \wedge P(y) \wedge P(z)).$$

(Here all quantifiers are over the universe $\mathbb{Z}$ of integers.) Now write the sentence "there are **at most** three distinct integers x that satisfy $P(x)$" as a proposition using quantifiers. One way to do it is

$$\exists x \exists y \exists z \forall d (P(d) \Longrightarrow d = x \vee d = y \vee d = z).$$

Here is an equivalent way to do it:

$$\forall x \forall y \forall v \forall z((x \neq y \wedge y \neq v \wedge v \neq x \wedge x \neq z \wedge y \neq z \wedge v \neq z) \Longrightarrow \neg(P(x) \wedge P(y) \wedge P(v) \wedge P(z))).$$

[Check that you understand both of the above alternatives.] Finally, what if we want to express the sentence "there are **exactly** three distinct integers x that satisfy $P(x)$"? This is now easy: we can just use the *conjunction* of the two propositions above.

Proofs

Intuitively, the concept of proof should already be familiar. We all like to assert things, and few of us like to say things that turn out to be false. A proof provides a means for *guaranteeing* such claims.

Proofs in mathematics and computer science require a precisely stated proposition to be proved. But what exactly is a proof? How do you show that a proposition is true? Recall that there are certain propositions called axioms or postulates, that we accept without proof (we have to start somewhere). A formal proof is a sequence of statements, ending with the proposition being proved, with the property that each statement is either an axiom or its truth follows easily from the fact that the previous statements are true. For example, in high school geometry you may have written two-column proofs where one column lists the statements and the other column lists the justifications for each statement. The justifications invoke certain very simple rules of inference which we trust (such as if P is true and Q is true, then $P \wedge Q$ is true). Every proof has these elements, though it does not have to be written in a tabular format. And most importantly, the fact that each step follows from the previous step is so straightforward, it can be checked by a computer program.

A formal proof for all but the simplest propositions is too cumbersome to be useful. In practice, mathematicians routinely skip steps to give proofs of reasonable length. How do they decide which steps to include in the proof? The answer is sufficiently many steps to convince themselves and the reader that the details can easily be filled in if desired. This of course depends upon the knowledge and skill of the audience. So in practice proofs are socially negotiated.[1]

During the first few weeks of the semester, the proofs we will write will be quite formal. Once you get more comfortable with the notion of a proof, we will relax a bit. We will start emphasizing the main ideas in our proofs and sketching some of the routine steps. This will help increase clarity and understanding and reduce clutter. A proof, for the purposes of this class, is essentially a convincing argument. Convincing to whom? First, to you, the author, second, to your classmates, third, to your professor and your TA.

In this lecture you will see some examples of proofs. The proofs chosen are particularly interesting and elegant, and some are of great historical importance. But the purpose of this lecture is not to teach you about these particular proofs (and certainly not for you to attempt to memorize any of them!). Instead, you should see these as good illustrations of various basic proof techniques. You will notice that sometimes when it is hard to even get started proving a certain proposition using one proof technique, it is easy using a different technique. This will come in handy later in the course when you work on homework problems or try to prove a statement on your own. If you find yourself completely stuck, rather than getting discouraged you might find that using a different proof technique opens doors that were previously closed.

We now begin with a few definitions pertaining to proofs.

A **theorem**, informally speaking, is a true proposition that is guaranteed by a proof. If you believe that a statement is true but can't prove it, call it a **conjecture**, essentially an educated guess.

[1] Those interested in exploring this issue in more detail may like to read the influential paper "Social Processes and Proofs of Theorems and Programs" by DeMillo, Lipton and Perlis, *Communications of the ACM* **22** (1979) pages 271–280.

A concept useful for writing up complicated proofs is that of a **lemma**, which is a little theorem that you use in the proof of a bigger theorem. A lemma is to proofs what a subroutine is to programming.

An **axiom** is a statement we accept as true without proof.

There are many different types of proofs, as we shall see. The basic structure of these different types of proofs is best expressed in terms of propositional logic.

Direct Proof

Let us start with a very simple example.

Theorem: If x is an odd integer, then $x+1$ is even.

Following the notation introduced in the previous Note, the statement of the theorem is equivalent to

$$(\forall x \in \mathbb{Z})(x \text{ is odd} \Longrightarrow x+1 \text{ is even}).$$

(Here $\mathbb{Z}$ denotes the set of all integers.) For each x, the proposition that we are trying to prove is of the form $P(x) \Longrightarrow Q(x)$. A direct proof of this starts by assuming $P(x)$ for a generic value of x and eventually concludes $Q(x)$ through a chain of implications:

Direct Proof of $P \Longrightarrow Q$
Assume P
$\vdots$
Therefore Q

Let us proceed with a direct proof of the simple example given above:

Proof: Assume x is odd. Then by definition, $x = 2k+1$ for some $k \in \mathbb{Z}$. Adding one to both sides, we get $x+1 = 2k+2 = 2(k+1)$. Therefore, by definition, $x+1$ is an even number. □

Before turning to our next example, we recall that the integer d divides n (denoted $d|n$) if and only if there exists some integer q such that $n = dq$.

For the following, let n be a positive integer less than 1000.

Theorem: Let n be a positive integer. If the sum of the digits of n is divisible by 9, then n is divisible by 9.

Comment: The theorem is true for arbitrary n. We're just doing the three digit case here so the notation does not distract from the structure of the argument.

This theorem's statement is equivalent to

$$(\forall n \in \mathbb{Z}^+)(\text{sum of } n\text{'s digits divisible by } 9 \Longrightarrow n \text{ divisible by } 9).$$

(Here $\mathbb{Z}^+$ denotes the set of positive integers, $\{1,2,\ldots\}$.) So once again we start by assuming, for a generic value of n, that the sum of n's digits is divisible by 9. Then we perform a sequence of steps to conclude that n itself is divisible by 9. Here is the proof:

Proof: Suppose we have n such that the sum of the digits of n is divisible by 9. Let a be the hundred's digit of n, b the ten's digit, and c the one's digit. Then $n = 100a+10b+c$. Now suppose that the sum of the digits of n is divisible by 9. This amounts to supposing that

$$a+b+c = 9k$$

for some $k \in \mathbb{Z}$. Adding $99a+9b$ to both sides of the equation, we get

$$100a+10b+c = 9k+99a+9b = 9(k+11a+b),$$

which is certainly divisible by 9 (since $k+11a+b$ is an integer). And finally, since $n = 100a+10b+c$, we see that n is divisible by 9. □

Exercise: Generalize the above proof so that it works for *any* positive integer n. [HINT: Suppose n has k digits, and write a_i for the ith digit of n, so that $n = \sum_{i=0}^{k-1} a_i \times 10^i$.]

In this case the converse of the theorem is also true: If n is divisible by 9, the sum of its digits is divisible by 9, too. In other words, the sum of the digits of n is divisible by 9 *if and only if*[2] n is divisible by 9. In general, to prove $P \iff Q$ you have to do two proofs: You must show that $P \implies Q$ and then, separately, you must also show that $Q \implies P$.

Theorem: n is divisible by 9 if and only if the sum of the digits of n is divisible by 9.

Proof: We already proved above that if the sum of the digits of n is divisible by 9 then n is divisible by 9. So we only need to prove the converse. We use the same notation for the digits of n as we used in the previous proof:

n is divisible by 9
$\implies n = 9\ell$, for some $\ell \in \mathbb{Z}$
$\implies 100a+10b+c = 9\ell$
$\implies 99a+9b+(a+b+c) = 9\ell$
$\implies a+b+c = 9\ell-99a-9b$
$\implies a+b+c = 9(\ell-11a-b)$
$\implies a+b+c = 9k$, where $k = \ell-11a-b \in \mathbf{Z}$
$\implies a+b+c$ is divisible by 9. □

Note that, in this simple example, the proof of $Q \implies P$ is essentially the same as the proof of $Q \implies P$ "run backwards." In such a case, it's tempting to try to get away with proving both of the implications at the same time (using the symbol $\iff$ at every step of the proof). However, I do not recommend this approach. Doing so requires *great caution*: for the proof to be legitimate, the steps have to make just as much sense backwards as forwards! (Go back and read the last proof again, starting with the last line and ending with the first, and convince yourself that it also works backwards.) To avoid potential pitfalls, it is recommended that you always prove a statement of the form $P \iff Q$ using two *separate* proofs. This will in any case be necessary in more interesting examples, where the proofs of $P \implies Q$ and of $Q \implies P$ might look very different.

Proof by Contraposition

In the last lecture, we learned that a statement of the form $P \implies Q$ is logically equivalent to its contrapositive: $\neg Q \implies \neg P$. This means that proving an implication is equivalent to proving the contrapositive. A proof by contraposition of $P \implies Q$ is just a direct proof of its contrapositive $\neg Q \implies \neg P$:

[2]The phrase "if and only if" is often abbreviated to "iff".

Proof by Contraposition of $P \implies Q$
Assume $\neg Q$
$\vdots$
Therefore $\neg P$
So $\neg Q \implies \neg P$, or equivalently $P \implies Q$

Sometimes proving the contrapositive of a statement is easier than proving the statement directly. Here is an illustrative example.

Theorem: Let n be an integer and let d divide n. Then, if n is odd then d is odd.

Proving this directly would be difficult. We would assume n is odd but what then? Proving the contrapositive of the statement, however, is very straightforward. The contrapositive is: If d is even then n is even.

Proof: Suppose d is even, then (by definition) $d = 2k$ for some $k \in \mathbb{Z}$.
Because $d|n$, we have $n = dq$ for some $q \in \mathbb{Z}$.
Combining these two statements, we have $n = dq = (2k)q = 2(kq)$.
So n is even. So if d is even then n is even. Therefore if n is odd then d is odd. □

Proof by contraposition is a very common technique. When proving implications ($P \implies Q$) the contrapositive gives us a second option for how to approach the problem. As a warning, do not confuse the contrapositive with the converse! To give some intuition using English, consider the statement "If it is sunny, then it is daytime." The contrapositive is "If it is nighttime, then it is not sunny," and the converse is "If it is daytime, then it is sunny." We know the original statement is true, and its contrapositive is also true. However the converse is simply false (for example, a summer afternoon in San Francisco!).

Proof by Contradiction

Proof by contradiction is also called *reductio ad absurdum* (reduction to an absurdity). The idea is to assume the opposite of what one is trying to prove and then show that this leads to something that is clearly nonsensical: a contradiction.

Proof by Contradiction of P
Assume $\neg P$
$\vdots$
R
$\vdots$
$\neg R$
Contradiction
Therefore P

Before proceeding to an example, let us try to understand the logic behind a proof by contradiction. We assume $\neg P$, and then prove both R and $\neg R$. But for any proposition R, $R \wedge \neg R \equiv False$. So we have shown that $\neg P \implies False$. The only way this implication can be true is if $\neg P$ is false. i.e., P is true.

Our first example of a proof by contradiction dates back more than 2000 years—to Euclid.

Theorem: There are infinitely many prime numbers.

Proving this directly would be difficult. How do we construct infinitely many prime numbers? But, as we will see, bad things happen when we assume that this statement is false: that there are only finitely many primes. Before we prove the theorem, we will state a simple lemma that we'll use without proof. We will prove it next week when we learn about induction.

Lemma: Every natural number greater than one is either prime or has a prime divisor (greater than one).

Now for the proof of the theorem.

Proof: Suppose (in order to get a contradiction) that there are only finitely many primes. Then, we can enumerate them: $p_1, p_2, p_3, \ldots, p_k$. (Here k is the total number of primes.)

Consider the number $q = p_1 p_2 p_3 \ldots p_k + 1$, the product of all the primes plus one. Note that q cannot be prime because it is strictly larger than all the primes. Thus, by the lemma, it has a prime divisor, p. (This will be our statement R. More precisely, R is the assertion that $p > 1$.) Because $p_1, p_2, p_3, \ldots, p_k$ are all the primes, p must be equal to one of them, so p is a divisor of their product.

So we have that p divides $p_1 p_2 p_3 \ldots p_k$, and p divides q, but that means p divides their difference, which is 1. Therefore, $p \leq 1$ (this is $\neg R$). Contradiction. If we start with the assumption that there are finitely many primes, we derive a contradiction. The only remaining possibility is that our original assumption (finitely many primes) was wrong. Therefore there are infinitely many primes. □

Note that in the proof, q need not be prime, tempting as it might be to say so. It's certainly not the case that a product of primes plus one must always be prime (think of $7 \times 2 + 1$). Nor is it the case that the product of the first k primes plus one must necessarily be prime (e.g., $2 \times 3 \times 5 \times 7 \times 11 \times 13 + 1 = 30031 = 59 \times 509$). When writing a proof, it is important to carefully think through each step, ensuring that it's logically justified. The most important part of learning mathematics is learning a habit of thinking clearly and precisely.

Let's look at another classic proof by contradiction. A **rational number** is a number that can be expressed as the ratio of two integers. For example, $\frac{2}{3}$, $\frac{3}{5}$, and $\frac{9}{16}$ are all rational numbers. In fact, any number with a finite or recurring decimal representation is a rational. [Exercise: Can you prove this?] Numbers that cannot be expressed as fractions are called **irrational**.

Theorem: $\sqrt{2}$ is irrational.

Proof: Assume (for the sake of contradiction) that $\sqrt{2}$ is rational. By the definition of rational numbers, this means that there exist integers a and b with no common factor other than 1, such that $\sqrt{2} = a/b$. (This will be our assertion R.)

For any numbers x and y, we know that $x = y \implies x^2 = y^2$. Hence $2 = a^2/b^2$.

Multiplying both sides by b^2, we have $a^2 = 2b^2$.

b is an integer, hence b^2 is an integer, hence a^2 is even (by the definition of evenness).

Hence, a is even (by the lemma below).

Therefore, by the definition of evenness, there is an integer c such that $a = 2c$.

Hence $2b^2 = (2c)^2 = 4c^2$, hence $b^2 = 2c^2$.

Since c is an integer, c^2 is an integer, hence b^2 is even.

Thus, b is even (by the lemma below).

Thus a and b have a common factor 2, contradicting the assertion that a and b have no common factor other than 1. This shows that the original assumption that $\sqrt{2}$ is rational is false, and hence that $\sqrt{2}$ must be irrational. □

Lemma: If a^2 is even, then a is even.

Can you prove this lemma? First try a direct proof. How would you proceed? Now try a proof by contraposition.

Proof by contradiction can seem mysterious. A proof by contradiction of P starts by assuming $\neg P$, and then it explores the consequences of this assumption. But you might wonder: why is it OK to assume $\neg P$ is true? Proofs aren't allowed to make assumptions willy-nilly without justification, so why is it fair game to begin the proof by assuming $\neg P$? The answer: certainly P is either true or false. If P is true, then we're done: we wanted to prove P is true, and it is. So the only thing left to consider is the possibility that P is false (i.e., $\neg P$ is true)—and a proof by contradiction demonstrates that this is in fact impossible, from which it follows that P must be true. Once we conclusively rule out the possibility that P might be false (by deriving a contradiction), we're entitled to conclude that P must be true.

Here is one more proof by contradiction, if you'd like to see another example.

Theorem: $x^5 - x + 1 = 0$ has no solution in the rational numbers.

To prove this theorem, we will first state and prove a useful lemma.

Lemma 1: If x is a real number satisfying $x^5 - x + 1 = 0$, and if $x = a/b$ for some $a, b \in \mathbb{Z}$ with $b \neq 0$, then both a and b are even.

Proof: Plugging in $x = a/b$, we have

$$(a/b)^5 - (a/b) + 1 = 0.$$

Multiplying both sides by b^5 yields

$$a^5 - ab^4 + b^5 = 0.$$

Now we perform a case analysis, looking at the parity of a and b:

- Case 0 (a is odd and b is odd): In this case, a^5, ab^4, and b^5 are all odd. Thus the left-hand side (LHS) of the second equation above has the form odd $-$ odd $+$ odd, and so the LHS is odd. However, the right-hand side (RHS) is even, which is impossible.
- Case 1 (a is odd and b is even): In this case, the LHS has the form odd $-$ even $+$ even, so the LHS is odd. As before, this is impossible.
- Case 2 (a is even and b is odd): In this case, the LHS has the form even $-$ even $+$ odd, so the LHS is odd. This too is impossible.

We have eliminated the three cases above as impossible, so the only remaining possibility is that a must be even and b must be even. □

We're now ready to prove the theorem.

Proof: Suppose (for the sake of a contradiction) that there exists some rational number, call it x, such that $x^5 - x + 1 = 0$. Since x is rational, we can find $a, b \in \mathbb{Z}$ such that $x = a/b$ and $b > 0$. Actually, there might be many ways to express x in this way (i.e., many pairs of $a, b \in \mathbb{Z}$ such that $x = a/b$ and $b > 0$); among all of these ways, let a, b be the choice that makes b minimal, i.e., that makes b as small as possible.

Now define $\alpha = a/2$ and $\beta = b/2$. By Lemma 1, both a and b must be even, so both α and β must be integers. Also, $x = a/b = (2\alpha/2\beta) = \alpha/\beta$, so $x = \alpha/\beta$. in particular, we have $x = \alpha/\beta$, and also $\alpha, \beta \in \mathbb{Z}$ and $\beta > 0$. In other words, α, β provide another way to express x. But $\beta = b/2$, so $\beta < b$. So b must not have been minimal after all. This is a contradiction. So our assumption must have been wrong—we've proven there does not exist any rational number x satisfying $x^5 - x + 1 = 0$. □

Proof by Cases

Sometimes we don't know which of a set of possible cases is true, but we know that at least one of the cases is true. If we can prove our result in each of the cases, then we have a proof. The English phrase "damned if you do and damned if you don't" sums up this proof method. Here's a nice example:

Theorem: There exists some pair of irrational numbers x and y such that x^y is rational.

Comment: This is our first example in this Note of a theorem that is *existentially* quantified ("there exists"). In other words, the statement may be written as

$$(\exists x)(\exists y)(x \text{ is irrational} \wedge y \text{ is irrational} \wedge x^y \text{ is rational}).$$

Thus to prove the theorem we only need to prove the existence of at least one *example* of values x, y that satisfy the claim. (For this reason, proofs of existentially quantified statements are often—but not always—a little easier than proofs of universally quantified ones.)

Proof of the theorem: Consider the case $x = \sqrt{2}$ and $y = \sqrt{2}$. Clearly, either

(a) $\sqrt{2}^{\sqrt{2}}$ is rational; or

(b) $\sqrt{2}^{\sqrt{2}}$ is irrational.

In case (a), we have shown irrational numbers x and y such that x^y is rational, so we are done.
In case (b), consider the new values $x = \sqrt{2}^{\sqrt{2}}$ and $y = \sqrt{2}$. We have

$$\begin{aligned} x^y &= (\sqrt{2}^{\sqrt{2}})^{\sqrt{2}} \\ &= \sqrt{2}^{\sqrt{2}\sqrt{2}} \text{ by the axiom } (x^y)^z = x^{yz} \\ &= \sqrt{2}^2 = 2 \end{aligned}$$

Hence we have again shown irrational numbers x and y such that x^y is rational.
Since one of cases (a) and (b) must be true, and since in both cases we have exhibited irrational numbers x and y such that x^y is rational, we can conclude that such numbers must always exist. □

Notice that even after the proof, we still don't know which of the two cases is true, so we can't actually exhibit any irrational numbers satisfying the theorem. This is an example of a **nonconstructive** proof: one in which an existential theorem is proved without constructing an explicit example.

Non-proof

Failure to logically structure a proof or note the justification for each step can lead easily to "non-proofs." Consider the following examples.

Theorem: (not!) $-2 = 2$.

Proof: Assume $-2 = 2$. Squaring both sides, we get $(-2)^2 = 2^2$, or $4 = 4$, which is true. Therefore, $-2 = 2$. □

The theorem is obviously false, so what did we do wrong? Our arithmetic is correct, and it seems like each step follows from the previous step. The problem with this proof does not lie in the arithmetic, but rather the logic. We assumed the very theorem we were trying to prove was true! As you can see, logical soundness and structure are extremely important when proving propositions.

The next proof is incorrect for a different reason.

Theorem: (not!) $1 = -1$

Proof: $1 = \sqrt{1} = \sqrt{(-1)(-1)} = \sqrt{-1}\sqrt{-1} = \sqrt{-1}^2 = -1$. □

This proof appears to be logically sound, so the error lies elsewhere. Since we have concluded a falsehood, at least one of these steps must be false. Indeed, it is simply untrue that $\sqrt{xy} = \sqrt{x}\sqrt{y}$. If you think carefully through each step of your proofs, you can avoid such missteps.

Other classic errors:

- Dividing both sides of an equation by a variable. For example, suppose you see the following:

  $$ax = bx \text{ hence } a = b.$$

  The "axiom" to which this step implicitly appeals is false, because if $x = 0$, the claim $a = b$ is not necessarily true. So in this case, all we can conclude is that either $x = 0$ or $a = b$ (this can also be written as $x(a-b) = 0$). Some extra work may be needed to prove $x \neq 0$.

- Dividing both sides of an inequality by a variable. This is even worse! For example:

  $$ax < bx \text{ hence } a < b.$$

  Here the claim $a < b$ is false if $x < 0$, and unproven if $x = 0$.

- More generally, forgetting about 0. Forgetting to account for the possibility of variables being zero causes lots of headaches (including the above).

- "Working backwards." If you ever played with one of those puzzles where you solve a maze on paper, you may have learned the trick of starting at the exit of the maze and working backwards to the entrance. It's tempting to apply the same tricks to proving theorems: start with what you are trying to prove, and manipulate the claim until you get to one of the assumptions. However, this style of reasoning is erroneous; it amounts to a "converse error." You can't start by assuming what you are trying to prove. All reasoning should go "forwards": start with what you are given, and then work out what you can conclude from those givens, and so on.

Style and substance in proofs

We conclude with some general words of advice. First, get in the habit of thinking carefully before you write down the next sentence of your proof. If you cannot explain clearly why the step is justified, you are making a leap and you need to go back and think some more. In theory, each step in a proof must be justified by appealing to a definition or general axiom. In practice the depth to which one must do this is a matter of taste. For example, we could break down the step, "Since a is an integer, $(2a^2 + 2a)$ is an integer," into several more steps. [Exercise: what are they?] A justification can be stated without proof only if you are absolutely confident that (1) it is correct and (2) the reader will automatically agree that it is correct.

Notice that in the proof that $\sqrt{2}$ is irrational, we used the result, "For any integer n, if n^2 is even then n is even," twice. This suggests that it may be a useful fact in many proofs. A subsidiary result that is useful in a more complex proof is called a *lemma*. It is often a good idea to break down a long proof into several lemmas. This is similar to the way in which large programming tasks should be divided up into smaller subroutines. Furthermore, make each lemma (like each subroutine) as general as possible so it can be reused elsewhere.

The dividing line between lemmas and theorems is not clear-cut. Usually, when writing a paper, the theorems are those propositions that you want to "export" from the paper to the rest of the world, whereas the lemmas are propositions used locally in the proofs of your theorems. There are, however, some lemmas (for example, the Pumping Lemma and the Lifting Lemma) that are perhaps more famous and important than the theorems they were used to prove.

Finally, you should remember that the point of this lecture was not the specific statements we proved, but the different proof strategies, and their logical structure. Make sure you understand them clearly; you will be using them when you write your own proofs in homework and exams.

Proof Tips

Sometimes you can get some idea of how to structure your proof by looking at the form of the proposition you are trying to prove. Some examples:

- To prove something of the form $P \wedge Q$, first, prove P. Then, prove Q.
- To prove something of the form $P \vee Q$, one possibility is to guess which of P,Q is true and prove just that one. Another possibility is to try proving $(\neg P) \Longrightarrow Q$ or $(\neg Q) \Longrightarrow P$.
- To prove something of the form $P \Longrightarrow Q$, try a direct proof (start by assuming P, work out its consequences, and see if you can derive Q).
- To prove something of the form $(\forall x \in S)P(x)$, try a direct proof: consider a generic x, where you make absolutely no assumptions about the value of x other than that $x \in S$; then see if you can prove that $P(x)$ holds for that value of x. If you proved $P(x)$ without making any assumptions about x, your proof must apply to every possible value of x.

  In the next Note, we will see another technique to prove statements of the form $(\forall n \in \mathbb{N})P(n)$.
- To prove something of the form $(\exists x)P(x)$, try to find a value of x that makes $P(x)$ true, and just list it.
- To prove something of the form P, sometimes it is helpful to split by cases: look for some other proposition Q that allows you to prove both $Q \implies P$ and $(\neg Q) \implies P$. You can then use a proof by cases.
- If the proposition has no quantifiers (no $\forall$ or $\exists$ symbols), you could try a proof by enumeration: draw a truth table and show that the proposition is true in all cases. If you have a compound proposition built up out of $P_1, \ldots, P_k$ atomic propositions, then the truth table will have 2^k rows, so this technique is only feasible if the number of atomic propositions is not too large.

You can combine these techniques. For instance, if you are trying to prove something of the form $(\forall n \in \mathbb{N})(P(n) \Longrightarrow Q(n))$, you might try a direct proof: consider a generic value of $n \in \mathbb{N}$, where all you are allowed to assume about n is that $P(n)$ is true, and try to derive $Q(n)$.

Induction

Induction is an extremely powerful tool in mathematics. It is a way of proving propositions that hold for all natural numbers:

1) $\forall k \in \mathbb{N}, 0+1+2+3+\cdots+k = \frac{k(k+1)}{2}$

2) $\forall k \in \mathbb{N}$, the sum of the first k odd numbers is a perfect square.

3) Any graph with k vertices and k edges contains a cycle.

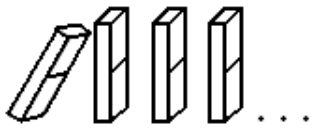
Each of these propositions is of the form $\forall k \in \mathbb{N}\ P(k)$. For example, in the first proposition, $P(k)$ is the statement $0+1+\cdots+k = \frac{k(k+1)}{2}$, $P(0)$ says $0 = \frac{0(0+1)}{2}$, $P(1)$ says $0+1 = \frac{1(1+1)}{2}$, etc. The *principle of induction* asserts that you can prove $P(k)$ is true $\forall k \in \mathbb{N}$, by following these three steps:

Base Case: Prove that $P(0)$ is true.

Inductive Hypothesis: Assume that $P(k)$ is true.

Inductive Step: Prove that $P(k+1)$ is true.

The principle of induction formally says that if $P(0)$ and $\forall n \in \mathbb{N}\ (P(n) \implies P(n+1))$, then $\forall n \in \mathbb{N}\ P(n)$. Intuitively, the base case says that $P(0)$ holds, while the inductive step says that $P(0) \implies P(1)$, and $P(1) \implies P(2)$, and so on. The fact that this "domino effect" eventually shows that $\forall n \in \mathbb{N}\ P(n)$ is what the principle of induction (or the induction axiom) states. In fact, dominoes are a wonderful analogy: we have a domino for each proposition $P(k)$. The dominoes are lined up so that if the k^{th} domino is knocked over, then it in turn knocks over the $k+1^{st}$. Knocking over the k^{th} domino corresponds to proving $P(k)$ is true. So the induction step corresponds to the fact that the k^{th} domino knocks over the $k+1^{st}$ domino. Now, if we knock over the first domino (the one numbered 0), then this sets off a chain reaction that knocks down all the dominoes.



Let's see some examples.

Theorem: $\forall k \in \mathbb{N}, \sum_{i=0}^{k} i = \frac{k(k+1)}{2}$.

Proof (by induction on k):

- Base Case: $P(0)$ asserts: $\sum_{i=0}^{0} i = \frac{0(0+1)}{2}$. This clearly holds, since the left and right hand sides both equal 0.

- Inductive Hypothesis: Assume $P(k)$ is true. That is, $\sum_{i=0}^{k} i = \frac{k(k+1)}{2}$.
- Inductive Step: We must show $P(k+1)$. That is, $\sum_{i=0}^{k+1} i = \frac{(k+1)(k+2)}{2}$:

$$\begin{aligned}
\sum_{i=0}^{k+1} i &= (\sum_{i=0}^{k} i) + (k+1) \\
&= \frac{k(k+1)}{2} + (k+1) && \text{(by the inductive hypothesis)} \\
&= (k+1)(\frac{k}{2}+1) \\
&= \frac{(k+1)(k+2)}{2}.
\end{aligned}$$

Hence, by the principle of induction, the theorem holds. ♠

Note the structure of the inductive step. You try to show $P(k+1)$ *under the assumption that* $P(k)$ is true. The idea is that $P(k+1)$ by itself is a difficult proposition to prove. Many difficult problems in EECS are solved by breaking the problem into smaller, easier ones. This is precisely what we did in the inductive step: $P(k+1)$ is difficult to prove, but we were able to recursively define it in terms of $P(k)$.

We will now look at another proof by induction, but first we will introduce some notation and a definition for divisibility. We say that integer a divides b (or b is divisible by a), written as $a|b$, if and only if for some integer q, $b = aq$.

Theorem: $\forall n \in \mathbb{N}$, $n^3 - n$ is divisible by 3.

Proof (by induction over n):

- Base Case: $P(0)$ asserts that $3|(0^3 - 0)$ or $3|0$, which is clearly true (since $0 = 3 \cdot 0$).
- Inductive Hypothesis: Assume $P(n)$ is true. That is, $3|(n^3 - n)$.
- Inductive Step: We must show that $P(n+1)$ is true, which asserts that $3|((n+1)^3 - (n+1))$. Let us expand this out:

$$\begin{aligned}
(n+1)^3 - (n+1) &= n^3 + 3n^2 + 3n + 1 - (n+1) \\
&= (n^3 - n) + 3n^2 + 3n \\
&= 3q + 3(n^2 + n),\ q \in \mathbb{Z} && \text{(by the inductive hypothesis)} \\
&= 3(q + n^2 + n)
\end{aligned}$$

Hence, by the principle of induction, $\forall n \in \mathbb{N}$, $3|(n^3 - n)$. ♠

The next example we will look at is an inequality between two functions of n. Such inequalities are useful in computer science when showing that one algorithm is more efficient than another. Notice that for this example, we have chosen as our base case $n = 2$, which is natural because the claim we are aiming to prove holds for all natural numbers greater than or equal to 2. If you think about the underlying induction principle,

it should be clear that this is perfectly valid, for the same reason that standard induction starting at $n = 0$ is valid (think back again to the domino analogy, where now the first domino is domino number 2).[1]

Theorem: $\forall n \in \mathbb{N}, n > 1 \implies n! < n^n$.

Proof (by induction over n):

- Base Case: $P(2)$ asserts that $2! < 2^2$, or $2 < 4$, which is clearly true.
- Inductive Hypothesis: Assume $P(n)$ is true (i.e., $n! < n^n$).
- Inductive Step: We must show $P(n+1)$, which states that $(n+1)! < (n+1)^{n+1}$. Let us begin with the left side of the inequality:

$$\begin{aligned}(n+1)! &= (n+1)\cdot n! \\ &< (n+1)\cdot n^n \qquad \text{(by the inductive hypothesis)} \\ &< (n+1)\cdot (n+1)^n \\ &= (n+1)^{n+1}\end{aligned}$$

Hence, by the induction principle, $\forall n \in \mathbb{N}$, if $n > 1$, then $n! < n^n$. ♠

In the middle of the last century, a colloquial expression in common use was "that is a horse of a different color", referring to something that is quite different from normal or common expectation. The famous mathematician George Polya (who was also a great expositor of mathematics for the lay public) gave the following proof to show that there is no horse of a different color!

Theorem: All horses are the same color.

Proof (by induction on the number of horses):

- Base Case: $P(1)$ is certainly true, since with just one horse, all horses have the same color.
- Inductive Hypothesis: Assume $P(n)$, which is the statement that n horses all have the same color.
- Inductive Step: Given a set of $n+1$ horses $\{h_1, h_2, \ldots, h_{n+1}\}$, we can exclude the last horse in the set and apply the inductive hypothesis just to the first n horses $\{h_1, \ldots, h_n\}$, deducing that they all have the same color. Similarly, we can conclude that the last n horses $\{h_2, \ldots, h_{n+1}\}$ all have the same color. But now the "middle" horses $\{h_2, \ldots, h_n\}$ (i.e., all but the first and the last) belong to both of these sets, so they have the same color as horse h_1 and horse h_{n+1}. It follows, therefore, that all $n+1$ horses have the same color. Thus, by the principle of induction, all horses have the same color. ♠

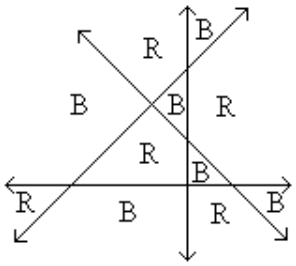
Clearly, it is not true that all horses are of the same color, so where did we go wrong in our induction proof? It is tempting to blame the induction hypothesis. But even though the induction hypothesis is false (for $n \geq 2$), that is not the flaw in the reasoning! Before reading on, think about this and see if you can understand why, and figure out the real flaw in the proof.

What makes the flaw in this proof a little tricky to pinpoint is that the induction step *is* valid for a "typical" value of n, say, $n = 3$. The flaw, however, is in the induction step when $n = 1$. In this case, for $n+1 = 2$ horses, there are *no* "middle" horses, and so the argument completely breaks down!

[1] Alternatively, we could insist on making the base case $n = 0$ (which holds vacuously here because $0 > 1$ is false). Then we would assert that $P(0) \implies P(1)$, since $P(1)$ holds (vacuously again), and that $P(1) \implies P(2)$ since $P(2)$ holds (as we show in the base case below). Then we would proceed as in the inductive step of the proof below. But this is all rather tedious.

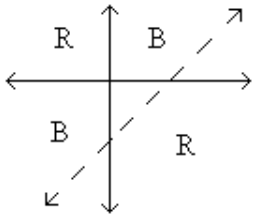
Some of you might still not feel completely convinced. Why is the above flaw more convincing than simply saying that the induction hypothesis is false? Saying that the induction hypothesis is false is like saying that the statement of the theorem is false, and so there is definitely a flaw in the proof. True, but our task was to pinpoint exactly where *in the proof* the flaw occurs. The point is that a valid induction proof involves only showing the base case, say $P(0)$, and that $\forall n\ P(n) \implies P(n+1)$. One way of saying that $P(n) \implies P(n+1)$ is to assume $P(n)$ is true and then show that $P(n+1)$ is true. If $P(n)$ is false, then $P(n) \implies P(n+1)$ vacuously. So just saying that the induction hypothesis $P(n)$ is false does not pinpoint the flaw in the proof.

Two Color Theorem: There is a famous theorem called the four color theorem. It states that any map can be colored with four colors such that any two adjacent countries (which share a border, but not just a point) must have different colors. The four color theorem is very difficult to prove, and several bogus proofs were claimed since the problem was first posed in 1852. It was not until 1976 that the theorem was finally proved (with the aid of a computer) by Appel and Haken. (For an interesting history of the problem, and a state-of-the-art proof, which is nonetheless still very challenging, see www.math.gatech.edu/∼thomas/FC/fourcolor.html). We consider a simpler scenario, where we divide the plane into regions by drawing straight lines. We want to know if we can color this map using no more than two colors (say, red and blue) such that no two regions that share a boundary have the same color. Here is an example of a two-colored map:



We will prove this "two color theorem" by induction on n, the number of lines:

- Base Case: Prove that $P(0)$ is true, which is the proposition that a map with $n = 0$ lines can be can be colored using no more than two colors. But this is easy, since we can just color the entire plane using one color.
- Inductive Hypothesis: Assume $P(n)$. That is, a map with n lines can be two-colored.
- Inductive Step: Prove $P(n+1)$. We are given a map with $n+1$ lines and wish to show that it can be two-colored. Let's see what happens if we remove a line. With only n lines on the plane, we know we can two-color the map (by the inductive hypothesis). Let us make the following observation: if we swap red $\leftrightarrow$ blue, we still have a two-coloring. With this in mind, let us place back the line we removed, and leave colors on one side of the line unchanged. On the other side of the line, swap red $\leftrightarrow$ blue. We claim that this is a valid two-coloring for the map with $n+1$ lines.

  Why does this work? We can say with certainty that regions which do not border the line are properly two-colored. But what about regions that do share a border with the line? We must be certain that any

two such regions have opposite coloring. But any two regions that border the line must have been the same region when the line was removed, so the reversal of color on one side of the line guarantees an opposite coloring. ♠

Strengthening the Inductive Hypothesis

Let us prove by induction the following proposition:

Theorem: For all $n \geq 1$, the sum of the first n odd numbers is a perfect square.

Proof: By induction on n.

- Base Case: $n = 1$. The first odd number is 1, which is a perfect square.
- Inductive Hypothesis: Assume that the sum of the first n odd numbers is a perfect square, say k^2.
- Inductive Step: The $n+1$-th odd number is $2n+1$, so the sum of the first $n+1$ odd numbers is k^2+2n+1. But now we are stuck. Why should k^2+2n+1 be a perfect square?

Here is an idea: let us show something stronger!

Theorem: For all $n \geq 1$, the sum of the first n odd numbers is n^2.

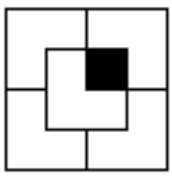
Proof: By induction on n.

- Base Case: $n = 1$. The first odd number is 1, which is 1^2.
- Inductive Hypothesis: Assume that the sum of the first n odd numbers is n^2.
- Inductive Step: The $(n+1)$-th odd number is $2n+1$, so the sum of the first $n+1$ odd numbers is $n^2+(2n+1) = (n+1)^2$. Thus by the principle of induction the theorem holds. ♠

See if you can understand what happened here. We could not prove a proposition, so we proved a harder proposition instead! Can you see why that can sometimes be easier when you are doing a proof by induction? When you are trying to prove a stronger statement by induction, you have to show something harder in the induction step, but you also get to assume something stronger in the induction hypothesis. Sometimes the stronger assumption helps you reach just that much further...

Here is another example:

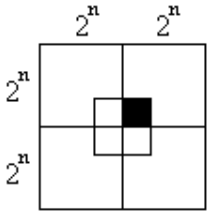
Imagine that we are given L-shaped tiles (i.e., a 2×2 square tile with a missing 1×1 square), and we want to know if we can tile a $2^n \times 2^n$ courtyard with a missing 1×1 square in the middle. Here is an example of a successful tiling in the case that $n = 2$:



Let us try to prove the proposition by induction on n.

- Base Case: Prove $P(1)$. This is the proposition that a 2×2 courtyard can be tiled with L-shaped tiles with a missing 1×1 square in the middle. But this is easy:

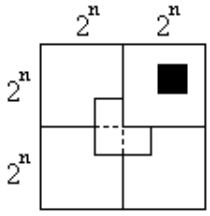
- Inductive Hypothesis: Assume $P(n)$ is true. That is, we can tile a $2^n \times 2^n$ courtyard with a missing 1×1 square in the middle.
- Inductive Step: We want to show that we can tile a $2^{n+1} \times 2^{n+1}$ courtyard with a missing 1×1 square in the middle. Let's try to reduce this problem so we can apply our inductive hypothesis. A $2^{n+1} \times 2^{n+1}$ courtyard can be broken up into four smaller courtyards of size $2^n \times 2^n$, each with a missing 1×1 square as follows:

  But the holes are not in the middle of each $2^n \times 2^n$ courtyard, so the inductive hypothesis does not help! How should we proceed? We should strengthen our inductive hypothesis!

What we are about to do is completely counter-intuitive. It's like attempting to lift 100 pounds, failing, and then saying "I couldn't lift 100 pounds. Let me try to lift 200," and then succeeding! Instead of proving that we can tile a $2^n \times 2^n$ courtyard with a hole in the middle, we will try to prove something stronger: that we can tile the courtyard with the hole being *anywhere we choose*. It is a trade-off: we have to prove more, but we also get to assume a stronger hypothesis. The base case is the same, so we will just work on the inductive hypothesis and step.

- Inductive Hypothesis (second attempt): Assume $P(n)$ is true, so that we can tile a $2^n \times 2^n$ courtyard with a missing 1×1 square anywhere.
- Inductive Step (second attempt): As before, we can break up the $2^{n+1} \times 2^{n+1}$ courtyard as follows.

  By placing the first tile as shown, we get four $2^n \times 2^n$ courtyards, each with a 1×1 hole; three of these courtyards have the hole in one corner, while the fourth has the hole in a position determined by the hole in the $2^{n+1} \times 2^{n+1}$ courtyard. The stronger inductive hypothesis now applies to each of these four courtyards, so that each of them can be successfully tiled. Thus, we have proven that we can tile a $2^{n+1} \times 2^{n+1}$ courtyard with a hole anywhere! Hence, by the induction principle, we have proved the (stronger) theorem. ♠

Strong Induction

Strong induction is very similar to simple induction, with the exception of the inductive hypothesis. With strong induction, instead of just assuming $P(k)$ is true, you assume the stronger statement that $P(0)$, $P(1)$,

..., and $P(k)$ are all true (i.e., $P(0) \wedge P(1) \wedge \cdots \wedge P(k)$ is true, or in more compact notation $\bigwedge_{i=0}^{k} P(i)$ is true). Strong induction sometimes makes the proof of the inductive step much easier since we get to assume a stronger statement, as illustrated in the next example.

Theorem: Every natural number $n > 1$ can be written as a product of primes.

Recall that a number $n \geq 2$ is prime if 1 and n are its only divisors. Let $P(n)$ be the proposition that n can be written as a product of primes. We will prove that $P(n)$ is true for all $n \geq 2$.

- Base Case: We start at $n = 2$. Clearly $P(2)$ holds, since 2 is a prime number.
- Inductive Hypothesis: Assume $P(k)$ is true for $2 \leq k \leq n$: i.e., every number $k : 2 \leq k \leq n$ can be written as a product of primes.
- Inductive Step: We must show that $n+1$ can be written as a product of primes. We have two cases: either $n+1$ is a prime number, or it is not. For the first case, if $n+1$ is a prime number, then we are done. For the second case, if $n+1$ is not a prime number, then by definition $n+1 = xy$, where $x,y \in \mathbb{Z}^+$ and $1 < x, y < n+1$. By the inductive hypothesis, x and y can each be written as a product of primes (since $x, y \leq n$). Therefore, $n+1$ can also be written as a product of primes. ♠

Why does this proof fail if we were to use simple induction? If we only assume $P(n)$ is true, then we cannot apply our inductive hypothesis to x and y. For example, if we were trying to prove $P(42)$, we might write $42 = 6 \times 7$, and then it is useful to know that $P(6)$ and $P(7)$ are true. However, with simple induction, we could only assume $P(41)$, i.e., that 41 can be written as a product of primes — a fact that is not useful in establishing $P(42)$.

Simple Induction vs. Strong Induction

We have seen that strong induction makes certain proofs easy when simple induction seems to fail. A natural question to ask then, is whether the strong induction axiom is logically stronger than the simple induction axiom. In fact, the two methods of induction are logically equivalent. Clearly anything that can be proven by simple induction can also be proven by strong induction (convince yourself of this!). For the other direction, suppose we can prove by strong induction that $\forall n\ P(n)$. Let $Q(k) = P(0) \wedge \cdots \wedge P(k)$. Let us prove $\forall k\ Q(k)$ by *simple* induction. The proof is modeled after the strong induction proof of $\forall n\ P(n)$. That is, we want to show $Q(k) \Rightarrow Q(k+1)$, or equivalently $P(0) \wedge \cdots \wedge P(k) \Rightarrow P(0) \wedge \cdots \wedge P(k) \wedge P(k+1)$. But this is true iff $P(0) \wedge \cdots \wedge P(k) \Rightarrow P(k+1)$. This is exactly what the strong induction proof of $\forall n\ P(n)$ establishes! Therefore, we can establish $\forall n\ Q(n)$ by simple induction. And clearly, proving $\forall n\ Q(n)$ also proves $\forall n\ P(n)$.

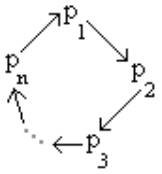
Well Ordering Principle

How can the induction axiom fail to be true? Recall that the axiom says the following: $[P(0) \wedge \forall n\ P(n) \Rightarrow P(n+1)] \Longrightarrow \forall n\ P(n)$. Assume for contradiction that $\neg(\forall n \in \mathbb{N}\ P(n))$. Then this means that $\exists n(\neg P(n))$, i.e., $P(n)$ is false for some n. Let m be the *smallest* n for which $P(n)$ is false. Since m is smallest, it must be the case that $P(m-1)$ is true. But this directly contradicts the fact that $P(m-1) \Longrightarrow P(m)$! It may seem as though we just proved the induction axiom. But what we have actually done is to show that the induction axiom follows from another axiom, which was used implicitly in defining m as "the smallest n for which $P(n)$ is false."

Well ordering principle: If $S \subseteq \mathbb{N}$ and $S \neq \emptyset$, then S has a smallest element.

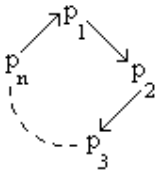
We assumed something when defining m that is usually taken for granted: that we can actually find a smallest number in any set of natural numbers. This property does *not* hold for, say, the real numbers; to see why, consider the set $\{x \in \mathbb{R} : 0 < x < 1\}$. Whatever number is claimed to be the smallest in this set, we can always find a smaller one. Again, the well ordering principle may seem obvious but it should not be taken for granted. It is only because the natural numbers (and any subset of the natural numbers) are well ordered that we can find a smallest element. Not only does the principle underlie the induction axioms, but it also has direct uses in its own right. Here is a simple example.

Round robin tournament: Suppose that, in a round robin tournament, we have a set of n players $\{p_1, p_2, \ldots, p_n\}$ such that p_1 beats p_2, p_2 beats p_3, ..., and p_n beats p_1. This is called a *cycle* in the tournament:

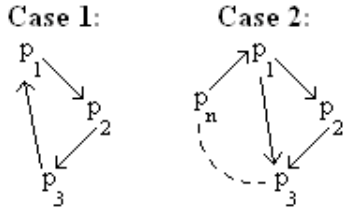


Claim: If there exists a cycle in a tournament, then there exists a cycle of length 3.

Proof: Assume for contradiction that the smallest cycle is:



with $n > 3$. Let us look at the game between p_1 and p_3. We have two cases: either p_3 beats p_1, or p_1 beats p_3. In the first case (where p_3 beats p_1), then we are done because we have a 3-cycle. In the second case (where p_1 beats p_3), we have a shorter cycle $\{p_3, p_4, \ldots, p_n\}$ and thus a contradiction. Therefore, if there exists a cycle, then there must exist a 3-cycle as well. ♠



Induction and Recursion

There is an intimate connection between induction and recursion in mathematics and computer science. A recursive definition of a function over the natural numbers specifies the value of the function at small values of n, and defines the value of $f(n)$ for a general n in terms of the value of $f(m)$ for $m < n$. Let us consider the example of the Fibonacci numbers, defined in a puzzle by Fibonacci (in the year 1202).

Fibonacci's puzzle: A certain man put a pair of rabbits in a place surrounded on all sides by a wall. How many pairs of rabbits can be produced from that pair in a year if it is supposed that every month each pair begets a new pair which from the second month on becomes productive?

Let $F(n)$ denote the number of pairs of rabbits in month n. According to the above specification, the initial conditions are $F(0) = 0$ and, when the pair of rabbits is introduced, $F(1) = 1$. Also $F(2) = 1$, since the pair is not yet productive. In month 3, according to the conditions, the pair of rabbits begets a new pair. So

$F(3) = 2$. In the fourth month, this new pair is not yet productive, but the original pair is, so $F(4) = 3$. What about $F(n)$ for a general value of n? This is a little tricky to figure out unless you look at it the right way. The number of pairs in month $n-1$ is $F(n-1)$. Of these how many were productive? Only those that were alive in the previous month - i.e. $F(n-2)$ of them. Thus there are $F(n-2)$ new pairs in the n-th month, and so $F(n) = F(n-1) + F(n-2)$. This completes the recursive definition of $F(n)$:

- $F(0) = 0$, and $F(1) = 1$
- For $n \geq 2$, $F(n) = F(n-1) + F(n-2)$

This admittedly simple model of population growth nevertheless illustrates a fundamental principle. Left unchecked, populations grow exponentially over time. [Exercise: can you show, for example, that $F(n) \geq 2^{(n-1)/2}$ for all $n \geq 3$?] Understanding the significance of this unchecked exponential population growth was a key step that led Darwin to formulate his theory of evolution. To quote Darwin: "There is no exception to the rule that every organic being increases at so high a rate, that if not destroyed, the earth would soon be covered by the progeny of a single pair."

Be sure you understand that a recursive definition is not circular — even though in the above example $F(n)$ is defined in terms of the function F, there is a clear ordering which makes everything well-defined. Here is a recursive program to evaluate $F(n)$:

```
function F(n)
  if n=0 then return 0
  if n=1 then return 1
  else return F(n-1) + F(n-2)
```

Can you figure out how long this program takes to compute $F(n)$? This is a very inefficient way to compute the n-th Fibonacci number. A much faster way is to turn this into an iterative algorithm (this should be a familiar example of turning a tail-recursion into an iterative algorithm):

```
function F2(n)
  if n=0 then return 0
  if n=1 then return 1
  a = 1
  b = 1
  for k = 2 to n do
    c = a
    a = a + b
    b = c
  od
  return a
```

Can you show by induction that this new function $F_2(n) = F(n)$? How long does this program take to compute $F(n)$?

The Stable Marriage Problem: An Application of Proof Techniques to Analysis of Algorithms

Consider a dating agency that must match up n men and n women. Each man has an ordered *preference list* of the n women, and each woman has a similar list of the n men (ties are not allowed). Is there a good algorithm that the agency can use to determine a good pairing?

Example

Consider for example $n = 3$ men (represented by numbers 1, 2, and 3) and 3 women (A, B, and C), and the following preference lists:

Men	Women		
1	A	B	C
2	B	A	C
3	A	B	C

Women	Men		
A	2	1	3
B	1	2	3
C	1	2	3

For instance, the preference lists above mean that woman A is man 1's top choice; woman B is his second choice; and so on.

What properties should a good pairing have? One possible criterion for a "good" pairing is one in which the number of first ranked choices is maximized. Another possibility is to minimize the number of last ranked choices. Or perhaps minimizing the sum of the ranks of the choices, which may be thought of as maximizing the average happiness. In this lecture we will focus on a very basic criterion: *stability*. A pairing is unstable if there is a man and a woman who prefer each other to their current partners. We will call such a pair a *rogue couple*. So a pairing of n men and n women is stable if it has no rogue couples.

An unstable pairing from the above example is: {(1,C), (2,B), (3,A)}. The reason is that 1 and B form a rogue couple, since 1 would rather be with B than C (his current partner), and since B would rather be with 1 than 2 (her current partner). This is trouble: Before long, 1 and B are going to spending many late nights doing CS70 problem sets together. Obviously, the existence of rogue couples is not a good thing if you are a matchmaker, since they will lead to instability or customer dissatisfaction. That is why we focus on stable pairings.

An example of a stable pairing is: {(2,A), (1,B), (3,C)}. Note that (1,A) is not a rogue couple. It is true that man 1 would rather be with woman A than his current partner. Unfortunately for him, she would rather be with her current partner than with him. Note also that both 3 and C are paired with their least favorite choice in this matching. Nonetheless, it is a stable pairing, since there are no rogue couples.

The problem facing us is to find a stable pairing, given the preference lists for all n men and all n women.

Does a Stable Pairing Always Exist?

Before we discuss how to find a stable pairing, let us ask a more basic question: do stable pairings always exist? Surely the answer is yes, since we could start with any pairing and make it more and more stable as follows: if there is a rogue couple, modify the current pairing so that they are together. Repeat. Surely this procedure must result in a stable pairing! Unfortunately this reasoning is not sound. To demonstrate this, let us consider a slightly different scenario, the roommates problem. Here we have $2n$ people who must be paired up to be roommates (the difference being that unlike the dating scenario, a person can be paired with any of the remaining $2n-1$). The point is that nothing about the above reasoning relied on the fact that men can only be paired with women in the dating scenario, so by the same reasoning we would expect that there would be a stable pairing for the roommates problem as well. The following counter-example illustrates the fallacy in the reasoning:

Roommates			
A	B	C	D
B	C	A	D
C	A	B	D
D	–	–	–

Visually, we have the following situation:

What is interesting about this problem is that there is no stable pairing (i.e., there is always a rogue couple). For example, the pairing {(A,B), (C,D)} contains the rogue couple B and C. Using the reasoning above, we might decide to pair B and C together, giving us the pairing: {(B,C), (A,D)}. But this pairing is also unstable because now A and C are a rogue couple. [Exercise: Verify that in this example there is *no* stable pairing!] Thus any proof that there must be a stable pairing in the dating problem must use the fact that there are two genders in an essential way.

In fact, we shall show that, when we have n men and n women, a stable pairing always exists. This fact is somewhat surprising, but true.

The Traditional Marriage Algorithm

We will study what is sometimes called the "Traditional Marriage Algorithm" (TMA), so-called because it is based on a 1950's model of dating where the men propose to the women, and the women accept or reject these proposals. We will prove that the Traditional Marriage Algorithm always finds a stable pairing and study its many interesting properties.

The Traditional Marriage Algorithm works like this:

Every Morning: Each man goes to the first woman on his list not yet crossed off and proposes to her.

Every Afternoon: Each woman says "Maybe, come back tomorrow" to the man she likes best among the men who have proposed to her (she now has him on a string) and "No, I will never marry you!" to all the rest.

Every Evening: Each rejected suitor crosses off the woman who rejected him from his list.

The above loop is repeated each successive day until there are no more rejected suitors. On this day, each woman marries the man she has on a string.

We wish to show that this algorithm always outputs a stable pairing. But how do we even know that it must terminate? Let us prove something stronger: the algorithm is guaranteed to terminate in at most n^2 days. Well, for each day (except the last), at least one woman is crossed off some man's list. Since there are n men, each starting with a list of size n, it follows that the algorithm must terminate after n^2 days.

To establish that the pairing output is stable, we need the following crucial lemma:

Improvement Lemma: If woman W has man M on a string on the kth day, then on every subsequent day she has someone on a string whom she likes at least as much as M.

Proof: Suppose that on day ℓ (for some $\ell \geq k$) W has some man M' on a string, where she likes M' at least as much as M. (Possibly $M = M'$.) Since she has M' on a string, that means that she told M' "maybe" on day ℓ. On day $\ell+1$, M' will come back and propose to W again, since he was told "maybe" the previous day. So W has the choice of at least one man on day $\ell+1$. Let M'' be her favorite choice among all of those who propose to her on day $\ell+1$. (Possibly $M'' = M'$.) She will tell M'' "maybe" on day $\ell+1$, so on day $\ell+1$ she will have M'' on a string. Note that since M' was one of the proposers on day $\ell+1$, this means that she must like M'' at least as much as she likes M', and as we stated before, she likes M' at least as much as she does M. Therefore on day $\ell+1$ she has a man (namely, M'') on a string who she likes at least as much as M. We have proven that if the property is true on day ℓ, then it is true on day $\ell+1$, so by induction on ℓ, it must be true for all $\ell \geq k$. □

Let us now proceed to prove that at the end of the algorithm all $2n$ people are paired up. Before reading the proof, see if you can convince yourself that this is true. The proof is remarkably short and elegant and is based crucially on the Improvement Lemma:

Lemma: The algorithm terminates with a pairing.

Proof: Suppose for contradiction that there is a man M who is left unpaired at the end of the algorithm. He must have proposed to every single woman on his list. By the Improvement Lemma, each of these women thereafter has someone on a string. Thus when the algorithm terminates, n women have n men on a string not including M. So there must be at least $n+1$ men. Contradiction. Our original assumption must have been wrong. It follows that each man is paired up with a female partner of his own, which means that all n women have a partner as well. □

Now, before we prove that the output of the algorithm is a stable pairing, let us first do a sample run-through of the stable marriage algorithm. We will use the preference lists given earlier, which are duplicated here for convenience:

Men	Women		
1	A	B	C
2	B	A	C
3	A	B	C

Women	Men		
A	2	1	3
B	1	2	3
C	1	2	3

The following table shows which men propose to which women on the given day (the circled men are the ones who were told "maybe"):

Days	Women	Proposals
1	A	①, 3
	B	②
	C	—
2	A	①
	B	②, 3
	C	—
3	A	①
	B	②
	C	③

Thus, the stable pairing which the algorithm outputs is: $\{$(1,A), (2,B), (3,C)$\}$.

Theorem: The pairing produced by the Traditional Marriage Algorithm is always stable.

Proof: We will show that no man M can be involved in a rogue couple. Consider any couple (M,W) in the pairing and suppose that M prefers some woman W^* to W. We will argue that W^* prefers her partner to M, so that (M,W^*) cannot be a rogue couple. Since W^* occurs before W in M's list, he must have proposed to her before he proposed to W. Therefore, according to the algorithm, W^* must have rejected him for somebody she prefers. By the Improvement Lemma, W^* likes her final partner at least as much, and therefore prefers him to M. Thus no man M can be involved in a rogue couple, and the pairing is stable. □

Optimality

Consider the situation in which there are 4 men and 4 women with the following preference lists:

Men	Women			
1	A	B	C	D
2	A	D	C	B
3	A	C	B	D
4	A	B	C	D

Women	Men			
A	1	3	2	4
B	4	3	2	1
C	2	3	1	4
D	3	4	2	1

For these preference lists, there are exactly two stable pairings: $S = \{$(1,A), (2,D), (3,C), (4,B)$\}$ and $T = \{$(1,A), (2,C), (3,D), (4,B)$\}$. The fact that there is more than one stable pairing brings up an interesting question. What is the best possible partner for each person? For instance, who is the best possible partner who man 2 could hope to end up with? The trivial answer is his first choice (i.e., woman A), but that is just not a realistic possibility for him. Pairing man 2 with woman A would simply not be stable, since he is so low on her preference list. And indeed there is no stable pairing in which 2 is paired with A. Examining the two stable pairings, we can see that the best possible realistic outcome for man 2 is to be matched to woman D.

Let us make some definitions to better express these ideas: we say the *optimal* woman for a man is the highest woman on his list whom he is paired with in any *stable* pairing. In other words, the optimal woman is the best that a man can do under the condition of stability. In the above example, woman D is the optimal woman for man 2. So the best that each man can hope for is to be paired with his optimal woman. But can all the men achieve this optimality *simultaneously*? In other words, is there a stable pairing such that each man is paired with his optimal woman? If such a pairing exists, we will call it a *male optimal* pairing. Turning to the example above, S is a male optimal pairing since A is 1's optimal woman, D is 2's optimal woman, C is 3's optimal woman, and B is 4's optimal woman. Similarly, we can define a female optimal pairing, which is the pairing in which each woman is paired with her optimal man. [Exercise: Check that T is a female optimal pairing.] We can also go in the opposite direction and define the *pessimal* woman for a man to be the lowest ranked woman whom he is ever paired with in some stable pairing. This leads naturally to the notion of a *male pessimal* pairing — can you define it, and also a female pessimal pairing?

Now, a natural question to ask is: Who is better off in the Traditional Marriage Algorithm: men or women? Think about this before you read on...

Theorem: The pairing output by the Traditional Marriage Algorithm is male optimal.

Proof: Suppose for the sake of contradiction that the pairing output by the TMA is *not* male optimal. This means there must exist an earliest day, let's say day k, in which some man was rejected by his optimal wife. Call that man M. (There might be multiple men who were rejected by their optimal wives on day k; in that case, we choose one of those men arbitrarily.) Let W be M's optimal wife. Since M was rejected by W on day k, on that day W must have accepted another man, let's call him M^*, who she likes better than M. Because no man was rejected by his optimal wife before day k, and because M^* was not rejected on day k, we conclude that M^* has not yet been rejected by his optimal wife on day k. Therefore, there are only two possibilities: (1) W is M^*'s optimal wife; or, (2) on day k, M^* has not yet proposed to his optimal wife and hence likes W better than his optimal wife.

Since W is M's optimal wife, this means there must exist some stable pairing, call it T, where they are paired together. In T, M^* must have been paired off with someone else, call her W^*. In other words, T looks like this: $T = \{\ldots,(M,W),\ldots,(M^*,W^*),\ldots\}$. Since T is stable, there are only two possibilities: (a) W^* is M^*'s optimal wife; or, (b) M^* likes his optimal wife better than W^*.

All in all, we have four cases. We will show that each of these cases is impossible or leads to a contradiction.

- Case (1)(a) is impossible: W and W^* are two different women, so they can't both be M^*'s optimal wife.
- In case (1)(b), M^* likes W, his optimal wife, better than W^*.
- In case (2)(a), M^* likes W better than W^*, his optimal wife.
- In case (2)(b), M^* likes W better than his optimal wife and his optimal wife better than W^*, so M^* likes W better than W^*.

In every possible case, M^* likes W better than W^*. And, as we mentioned earlier, W likes M^* better than M. But this means that (M^*,W) form a rogue couple in T: M^* likes W better than his wife in T, and W likes M^* better than her husband in T. The existence of a rogue couple in T contradicts the stability of T. Contradiction. Thus our original assumption—that the TMA-pairing is not male-optimal—must have been incorrect. □

What proof techniques did we use to prove this theorem? We used the well-ordering principle. How do we see it as a regular induction proof? This is a bit subtle to figure out. See if you can do so before reading on... the proof is really showing by induction on k the following statement: for every k, no man gets rejected by his optimal woman on the kth day. [Exercise: Can you complete the induction?]

So men appear to fare very well by following the Traditional Marriage Algorithm. How about the women? The following theorem reveals the sad truth:

Theorem: If a pairing is male optimal, then it is also female pessimal.

Proof: Let $T = \{\ldots,(M,W),\ldots\}$ be the male optimal pairing. Consider any stable pairing where W is paired with someone else, say, $S = \{\ldots,(M^*,W),\ldots,(M,W'),\ldots\}$. Since T is male optimal, we know that M prefers W (his mate in T) to his mate in S. Since S is stable, we know that (M,W) is not a rogue couple, so W must prefer her mate in S (namely, M^*) to M. This means that every other stable pairing must match W up to some other male who she likes at least as much as her mate in T. In other words, T pairs W up with her pessimal man. Since W was arbitrary, we see that T pairs every female with her pessimal man. Therefore, the male optimal pairing is female pessimal. □

All this seems a bit unfair to the women! Are there any lessons to be learned from this? Make the first move!

The Residency Match

Perhaps the most well-known application of the stable marriage algorithm is the residency match program, which pairs medical school graduates and residency slots (internships) at teaching hospitals. Graduates and hospitals submit their ordered preference lists, and the stable pairing produced by a computer matches students with residency programs.

The road to the residency match program was long and twisted. Medical residency programs were first introduced about a century ago. Since interns offered a source of cheap labor for hospitals, soon the number of residency slots exceeded the number of medical graduates, resulting in fierce competition. Hospitals tried to outdo each other by making their residency offers earlier and earlier. By the mid-40s, offers for residency were being made by the beginning of junior year of medical school, and some hospitals were contemplating even earlier offers — to sophomores! The American Medical Association finally stepped in and prohibited medical schools from releasing student transcripts and reference letters until their senior year. This sparked a new problem, with hospitals now making "short fuse" offers to make sure that if their offer was rejected they could still find alternate interns to fill the slot. Once again the competition between hospitals led to an unacceptable situation, with students being given only a few hours to decide whether they would accept an offer.

Finally, in the early 50s, this unsustainable situation led to the centralized system called the National Residency Matching Program (N.R.M.P.) in which the hospitals ranked the residents and the residents ranked the hospitals. The N.R.M.P. then produced a pairing between the applicants and the hospitals, though at first this pairing was not stable. It was not until 1952 that the N.R.M.P. switched to the Traditional Marriage Algorithm, resulting in a stable pairing. Until recently the algorithm was run with the hospitals doing the proposing, and so the pairings produced were hospital optimal. Only recently were the roles reversed such that the medical students were proposing to the hospitals. In the 1990's, there were other improvements made to the algorithm which the N.R.M.P. used. For example, the pairing takes into account preferences for married students for positions at the same or nearby hospitals.

Unsurprisingly, the Traditional Marriage Algorithm is also (reportedly) used by a large dating agency. But it is apparently also used by Akamai, a large web hosting company. Akamai receives a great many web requests, and needs to direct each web request to one of a pool of Akamai web servers. Web requests are best served by nearby servers, and servers that aren't currently busy are better suited to handle web requests than servers that are. Apparently, Akamai uses the Traditional Marriage Algorithm to match web requests to servers efficiently, where the web requests play the role of the men and the web servers are the girls.

Further reading (optional!)

Though it was in use 10 years earlier, the Traditional Marriage Algorithm was first properly analyzed by Gale and Shapley, in a famous paper dating back to 1962 that still stands as one of the great achievements in the analysis of algorithms. The full reference is:

D. Gale and L.S. Shapley, "College Admissions and the Stability of Marriage," *American Mathematical Monthly* **69** (1962), pp. 9–14.

Stable marriage and its numerous variants remains an active topic of research in computer science. Although it is by now twenty years old, the following very readable book covers many of the interesting developments since Gale and Shapley's algorithm:

D. Gusfield and R.W. Irving, *The Stable Marriage Problem: Structure and Algorithms*, MIT Press, 1989.

Modular Arithmetic

One way to think of modular arithmetic is that it limits numbers to a predefined range $\{0,1,\ldots,m-1\}$, and wraps around whenever you try to leave this range — like the hand of a clock (where $m=12$) or the days of the week (where $m=7$).

Example: Calculating the day of the week. Suppose that you have mapped the sequence of days of the week (Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday) to the sequence of numbers $(0,1,2,3,4,5,6)$ so that Sunday is 0, Monday is 1, etc. Suppose that today is Thursday ($=4$), and you want to calculate what day of the week will be 10 days from now. Intuitively, the answer is the remainder of $4+10=14$ when divided by 7, that is, 0 —Sunday. In fact, it makes little sense to add a number like 10 in this context, you should probably find *its* remainder modulo 7, namely 3, and then add this to 4, to find 7, which is 0.

What if we want to continue this in 10 day jumps? After 5 such jumps, we would have day $4+3\cdot 5=19$, which gives 5 modulo 7 (Friday).

This example shows that in certain circumstances it makes sense to do arithmetic within the confines of a particular number (7 in this example), that is, to do arithmetic by always finding the remainder of each number modulo 7, say, and repeating this for the results, and so on. As well as being efficient in the sense of keeping intermediate values as small as possible, this actually has several important applications, including error-correcting codes and cryptography, as we shall see later.

To define things more formally, for any integer m (such as 7) we say that x and y are *congruent modulo m* if they differ by a multiple of m, or in symbols,

$$x \equiv y \pmod{m} \qquad \Leftrightarrow \qquad m \text{ divides } (x-y).$$

For example, $29 \equiv 5 \pmod{12}$ because $29-5$ is a multiple of 12. We can also write $22 \equiv -2 \pmod{12}$. When we write

$$29 \equiv 5 \pmod{12}$$

we use "$\equiv$" instead of "$=$" to remind us that this is not the ordinary equality, and "$\pmod{12}$" to indicate that we are ignoring multiples of 12. An equivalent definition is to say that x and y are congruent modulo m (written $x \equiv y \pmod{m}$) iff they have the same remainder modulo m. Notice that "congruent modulo m" is an *equivalence relation*: it partitions the integers into m equivalence classes $0,1,2,\ldots,m-1$. We will also use $\bmod(x,m)$ to denote the function that, given integers x and $m \geq 1$, returns the remainder of x modulo m. So for example, $\bmod(23,7)=2$.

When computing modulo m, it is often convenient to reduce any intermediate results $\bmod m$ to simplify the calculation, as we did in the example above. This is justified by the following claim:

Theorem 5.1: If $a \equiv c \pmod{m}$ and $b \equiv d \pmod{m}$, then $a+b \equiv c+d \pmod{m}$ and $a \cdot b \equiv c \cdot d \pmod{m}$.

Proof: We know that $c = a+k\cdot m$ and $d = b+\ell \cdot m$, so $c+d = a+k\cdot m+b+\ell\cdot m = a+b+(k+\ell)\cdot m$, which means that $a+b \equiv c+d \pmod{m}$. The proof for multiplication is similar and left as an exercise. □

What this theorem tells us is that we can always reduce any arithmetic expression modulo m into a natural number smaller than m. As an example, consider the expression $(13+11)\cdot 18 \bmod 7$. Using the above Theorem several times we can write:

$$\begin{aligned}(13+11)\cdot 18 &\equiv (6+4)\cdot 4 \pmod 7\\ &\equiv 10\cdot 4 \pmod 7\\ &\equiv 3\cdot 4 \pmod 7\\ &\equiv 12 \pmod 7\\ &\equiv 5 \pmod 7.\end{aligned}$$

In summary, we can always do calculations modulo m by reducing intermediate results modulo m.

Inverses

Addition and multiplication $\bmod m$ is easy. To add two numbers a and b modulo m, we just add the numbers and then subtract m if necessary to reduce the result to a number between 0 and $m-1$. Multiplication can be similarly carried out by multiplying a and b and then calculating the remainder when the result is divided by m. Subtraction is equally easy. This is because subtracting b modulo m is the same as adding $-b \equiv m-b \pmod m$.

What about division? This is a bit harder. Over the reals, dividing by a number x is the same as multiplying by $y = 1/x$. Here y is that number such that $x\cdot y = 1$. Of course we have to be careful when $x = 0$, since such a y does not exist. Similarly, when we wish to divide by x mod m, we need to find y mod m such that $x\cdot y \equiv 1 \pmod m$; then dividing by x modulo m will be the same as multiplying by y modulo m. Such a y is called the *multiplicative inverse* of x modulo m. In our present setting of modular arithmetic, can we be sure that x has an inverse mod m, and if so, is it unique (modulo m) and can we compute it?

As a first example, take $x = 8$ and $m = 15$. Then $2x \equiv 16 \equiv 1 \pmod{15}$, so 2 is a multiplicative inverse of 8 mod 15. As a second example, take $x = 12$ and $m = 15$. Then the sequence $\{\bmod(ax,m) : a = 0,1,2,\ldots\}$ is periodic, and takes on the values $\{0,12,9,6,3\}$ (check this!). Thus 12 *has no multiplicative inverse mod* 15.

So when *does* x have a multiplicative inverse modulo m? The answer is: iff $\gcd(m,x) = 1$. This condition means that x and m share no common factors (except 1), and is often expressed by saying that x and m are *relatively prime*. Moreover, when the inverse exists it is unique.

Theorem 5.2: Let m,x be positive integers such that $\gcd(m,x) = 1$. Then x has a multiplicative inverse modulo m, and it is unique (modulo m).

Proof: Consider the sequence of m numbers $0,x,2x,\ldots(m-1)x$. We claim that these are all distinct modulo m. Since there are only m distinct values modulo m, it must then be the case that $ax \equiv 1 \pmod m$ for exactly one a (modulo m). This a is the unique multiplicative inverse.

To verify the above claim, suppose that $ax \equiv bx \pmod m$ for two distinct values a,b in the range $0 \le a,b \le m-1$. Then we would have $(a-b)x \equiv 0 \pmod m$, or equivalently, $(a-b)x = km$ for some integer k (possibly zero or negative). But since x and m are relatively prime, it follows that $a-b$ must be an integer multiple of m. This is not possible since a,b are distinct non-negative integers less than m. □

Actually it turns out that $\gcd(m,x) = 1$ is also a *necessary* condition for the existence of an inverse: i.e., if $\gcd(m,x) > 1$ then x has no multiplicative inverse modulo m. You might like to try to prove this using a similar idea to that in the above proof.

Since we know that multiplicative inverses are unique when $\gcd(m,x) = 1$, we shall write the inverse of x as $x^{-1} \bmod m$. But how do we compute x^{-1}, given x and m? For this we take a somewhat roundabout route.

First we shall consider the problem of computing the greatest common divisor $\gcd(a,b)$ of two integers a and b.

Computing the Greatest Common Divisor

The *greatest common divisor* of two natural numbers x and y, denoted $\gcd(x,y)$, is the largest natural number that divides them both. (Recall that 0 divides no number, and is divided by all.) How does one compute the gcd? By *Euclid's algorithm*, perhaps the first algorithm ever invented:

```
algorithm gcd(x,y)
  if y = 0 then return x
  else return gcd(y, mod(x,y))
```

We can express the very same algorithm a little more elegantly in Scheme:

```
(define (gcd x y)
  (if (= y 0)
    x
    (gcd y (remainder x y)) ) )
```

Note: This algorithm assumes that $x \geq y \geq 0$ and $x > 0$.

Before proving that this algorithm correctly computes the greatest common divisor, it will be helpful to prove a few lemmas.

Lemma 5.1: If $x \geq y \geq 0$ and $x > 0$, $\gcd(x,y) = \gcd(x-y,y)$.

Proof: If d divides both x and y, then it divides $x-y$ and y. If d divides $x-y$ and y, then it divides x and y. □

Lemma 5.2: If $x \geq y \geq 0$ and $x > 0$, $\gcd(x,y) = \gcd(\bmod(x,y),y)$.

Proof: Apply the prior lemma n times, where $n = \lfloor x/y \rfloor$. □

Theorem 5.3: Euclid's algorithm (above) correctly computes the gcd of x and y in time $O(n)$, where n is the total number of bits in the input (x,y).

Proof: Correctness is proved by (strong) induction on y, the smaller of the two input numbers. For each $y \geq 0$, let $P(y)$ denote the proposition that the algorithm correctly computes $\gcd(x,y)$ for all values of x such that $x \geq y$ (and $x > 0$). Certainly $P(0)$ holds, since $\gcd(x,0) = x$ and the algorithm correctly computes this in the `if`-clause. For the inductive step, we may assume that $P(z)$ holds for all $z < y$ (the inductive hypothesis); our task is to prove $P(y)$. By the inductive hypothesis, the recursive call `gcd(y, mod(x, y))` correctly returns $\gcd(y, \bmod(x,y))$ (just take $z = \bmod(x,y)$, and notice that $0 \leq z < y$). Now the lemma assures us that $\gcd(x,y) = \gcd(\bmod(x,y),y) = \gcd(y, \bmod(x,y))$, hence the `else`-clause of the algorithm will return the correct value $\gcd(x,y)$. This completes our verification of $P(y)$, and hence the induction proof.

Now for the $O(n)$ bound on the running time. It is obvious that the arguments of the recursive calls become smaller and smaller (because $y \leq x$ and $\bmod(x,y) < y$). The question is, how fast? We shall show that, in the computation of $\gcd(x,y)$, after two recursive calls the first (larger) argument is smaller than x by at least a factor of two (assuming $x > 0$). There are two cases:

1. $y \leq x/2$. Then the first argument in the next recursive call, y, is already smaller than x by a factor of 2, and thus in the next recursive call it will be even smaller.

2. $x \geq y > x/2$. Then in two recursive calls the first argument will be $\mod(x,y)$, which is smaller than $x/2$.

So, in both cases the first argument decreases by a factor of at least two every two recursive calls. Thus after at most $2n$ recursive calls, where n is the number of bits in x, the recursion will stop (note that the first argument is always a natural number). □

Note that the second part of the above proof only shows that the *number of recursive calls* in the computation is $O(n)$. We can make the same claim for the running time if we assume that each call only requires constant time. Since each call involves one integer comparison and one mod operation, it is reasonable to claim that its running time is constant. In a more realistic model of computation, however, we should really make the time for these operations depend on the size of the numbers involved: thus the comparison would require $O(n)$ elementary (bit) operations, and the mod (which is really a division) would require $O(n^2)$ operations, for a total of $O(n^2)$ operations in each recursive call. (Here n is the maximum number of bits in x or y, which is just the number of bits in x.) Thus in such a model the running time of Euclid's algorithm is really $O(n^3)$.

Back to Multiplicative Inverses

Let's now return to the question of computing the multiplicative inverse of x modulo m. For any pair of numbers x, y, suppose we could not only compute $d = \gcd(x,y)$, but also find integers a, b such that

$$d = ax + by. \tag{1}$$

(Note that this is not a modular equation; and the integers a, b could be zero or negative.) For example, we can write $1 = \gcd(35,12) = -1 \cdot 35 + 3 \cdot 12$, so here $a = -1$ and $b = 3$ are possible values for a, b.

If we could do this then we'd be able to compute inverses, as follows. If $\gcd(m,x) = 1$, apply the above procedure to the numbers m, x; this returns integers a, b such that

$$1 = am + bx.$$

But this means that $bx \equiv 1 \pmod{m}$, so b is a multiplicative inverse of x modulo m. Reducing b modulo m gives us the unique inverse we are looking for. In the above example, we see that 3 is the multiplicative inverse of 12 mod 35.

So, we have reduced the problem of computing inverses to that of finding integers a, b that satisfy equation (1). Now since this problem is a generalization of the basic gcd, it is perhaps not too surprising that we can solve it with a fairly simple extension of Euclid's algorithm. The following algorithm *extended-gcd* takes as input a pair of natural numbers $x \geq y$ as in Euclid's algorithm, and returns a triple of integers (d, a, b) such that $d = \gcd(x,y)$ and $d = ax + by$:

```
algorithm extended-gcd(x,y)
if y = 0 then return(x, 1, 0)
else
  (d, a, b) := extended-gcd(y, mod(x,y))
  return (d, b, a - ⌊x/y⌋ * b)
```

Note that this algorithm has the same form as the basic gcd algorithm we saw earlier; the only difference is that we now carry around in addition the required values a, b. You should hand-turn the algorithm on the input $(x,y) = (35,12)$ from our earlier example, and check that it delivers correct values for a, b.

Let's now look at why the algorithm works. In the base case ($y = 0$), we return the gcd value $d = x$ as before, together with values $a = 1$ and $b = 0$; and it's easy to see that in this case the returned values satisfy $ax + by = d$, since $1 \cdot x + 0 \cdot y = x = d$. If $y > 0$, we first recursively compute values (d, a, b) such that $d = \gcd(y, \bmod(x, y))$ and

$$d = ay + b \cdot (\bmod(x, y)). \tag{2}$$

Just as in our analysis of the vanilla algorithm, we know that this d will be equal to $\gcd(x, y)$. So the first component of the triple returned by the algorithm is correct.

What about the other two components? Let's call them A and B. What should their values be? Well, from the specification of the algorithm, they must be integers that satisfy

$$d = Ax + By. \tag{3}$$

To figure out what A and B should be, we need to rearrange equation (2), as follows:

$$\begin{aligned} d &= ay + b \cdot (\bmod(x, y)) \\ &= ay + b \cdot (x - \lfloor x/y \rfloor y) \\ &= bx + (a - \lfloor x/y \rfloor b)y. \end{aligned}$$

(In the second line here, we have used the fact that $\bmod(x, y) = x - \lfloor x/y \rfloor y$ — check this!) Comparing this last equation with equation (3), we see that we need to take $A = b$ and $B = a - \lfloor x/y \rfloor b$. This is exactly what the algorithm does, so we have concluded our proof of correctness.

Since the extended gcd algorithm has exactly the same recursive structure as the vanilla version, its running time will be the same up to constant factors (reflecting the increased time per recursive call). So once again the running time on n-bit numbers will be $O(n)$ arithmetic operations, and $O(n^3)$ bit operations. Combining this with our earlier discussion of inverses, we see that for any x, m with $\gcd(m, x) = 1$ we can compute $x^{-1} \bmod m$ in the same time bounds.

Public-Key Cryptography, RSA, and Modular Arithmetic

This lecture, we'll discuss a beautiful and important application of modular arithmetic: the *RSA public-key cryptosystem*, so named after its inventors Ronald Rivest, Adi Shamir, and Leonard Adleman.

The basic setting for cryptography is typically described via a cast of three characters: Alice and Bob, who wish to communicate confidentially over some (insecure) communication link, and Eve, an eavesdropper who is listening in and trying to discover what they are saying. Suppose Alice wants to transmit a message x (written in binary) to Bob. Alice will apply her encryption function E to x and send the encrypted message $E(x)$ over the communication link. Bob, upon receipt of $E(x)$, will then apply his decryption function D to it and thus recover the original message: i.e., $D(E(x)) = x$.

Since the link is insecure, Alice and Bob have to assume that Eve may get hold of $E(x)$. (Think of Eve as being a "sniffer" on the network.) Thus ideally we would like to know that the encryption function E is chosen so that just knowing $E(x)$ (without knowing the decryption function D) doesn't allow one to discover anything about the original message x.

For centuries cryptography was based on what are now called *private-key* protocols. In such a scheme, Alice and Bob meet beforehand and together choose a secret codebook, with which they encrypt all future correspondence between them. (This codebook plays the role of the functions E and D above.) Eve's only hope then is to collect some encrypted messages and use them to at least partially figure out the codebook.

Public-key schemes, such as RSA, are significantly more subtle and tricky, but simultaneously also more useful: they allow Alice to send Bob a message without ever having met him before! This almost sounds impossible, because in this scenario there is a symmetry between Bob and Eve: why should Bob have any advantage over Eve in terms of being able to understand Alice's message? The central idea behind the RSA cryptosystem is that Bob is able to implement a digital lock, to which only he has the key. Now by making this digital lock public, he gives Alice (or, indeed, anybody else) a way to send him a secure message which only he can open. Intuitively, Alice (or anyone) can "apply the lock" to her message, but thereafter only Bob can "open the lock" and recover the original message.

Here is how the digital lock is implemented in the RSA scheme. Each person has a public key known to the whole world (the "lock"), and a private key known only to him- or herself (the "key to the lock"). When Alice wants to send a message x to Bob, she encodes it using Bob's public key. Bob then decrypts it using his private key, thus retrieving x. Eve is welcome to see as many encrypted messages for Bob as she likes, but she will not be able to decode them (under certain simple assumptions explained below).

The RSA scheme is based heavily on modular arithmetic. Let p and q be two large primes (typically having, say, 1024 bits each), and let $N = pq$. We will think of messages to Bob as numbers modulo N. (Larger messages can always be broken into smaller pieces and sent separately.)

Also, let e be any number that is relatively prime to $(p-1)(q-1)$. (Typically e is a small value such as 3.) Then Bob's public key is the pair of numbers (N, e). This pair is published to the whole world. (Note, however, that the numbers p and q are not public; this point is crucial and we return to it below.)

What is Bob's private key? Bob's private key is the number d, which the RSA scheme sets to the inverse of

e modulo $(p-1)(q-1)$. (This inverse is guaranteed to exist because e and $(p-1)(q-1)$ have no common factor.)

We are now in a position to describe the RSA scheme, by defining its encryption and decryption functions:

- **[Encryption]:** When Alice wants to send a message x (assumed to be an integer modulo N) to Bob, she computes the value $E(x) = \mod(x^e, N)$ and sends this to Bob. Thus $E(x) \equiv x^e \pmod{N}$.
- **[Decryption]:** Upon receiving the value $y = E(x)$, Bob computes $D(y) = \mod(y^d, N)$; as we shall prove later, this will be equal to the original message x.

Example: Let $p = 5$, $q = 11$, and $N = pq = 55$. (In practice, p and q would be much larger.) Then we can choose $e = 3$, which is relatively prime to $(p-1)(q-1) = 40$. Therefore Bob's public key is $(N, e) = (55, 3)$. His private key is $d = 27$, since $e^{-1} \equiv 3^{-1} \equiv 27 \pmod{40}$. For any message x that Alice (or anybody else) wishes to send to Bob, the encryption of x is $y = E(x) = \mod(x^3, 55)$, and the decryption of y is $D(y) = \mod(y^{27}, 55)$. So, for example, if the message is $x = 13$, then the encryption is $y \equiv 13^3 \equiv 52 \pmod{55}$, and this is decrypted as $52^{27} \equiv 13 \pmod{55}$.

How do we know that the RSA scheme works? We need to check that Bob really does recover the original message x. The following theorem establishes this fact.

Theorem 6.1: Under the above definitions of the encryption and decryption functions E and D, we have $D(E(x)) = x$ for every possible message $x \in \{0, 1, \ldots, N-1\}$.

The proof of this theorem makes use of a standard fact from number theory known as Fermat's Little Theorem and a simple corollary to it, both shown below.

Theorem 6.2: **[Fermat's Little Theorem]** For any prime p and any a satisfying $\gcd(a, p) = 1$, we have $a^{p-1} \equiv 1 \pmod{p}$.

Proof: Consider the function $f : \{1, 2, \ldots, p-1\} \to \{1, 2, \ldots, p-1\}$ given by $f(x) = \mod(ax, p)$. We proved in Theorem 5.2 that the function f is bijective. In other words, for every $x_1, x_2 \in \{1, 2, \ldots, p-1\}$ with $x_1 \neq x_2$, we have $f(x_1) \neq f(x_2)$.

It follows that the list of numbers

$$\mod(a, p), \mod(2a, p), \mod(3a, p), \ldots, \mod((p-1), p)$$

runs through exactly the numbers $1, 2, 3, \ldots, p-1$ (hitting each such number exactly once), albeit in some other order.

Recall that multiplication is commutative, so re-ordering a set of factors does not affect their product. Thus, the product of the first list is the same as the product of the second list. This remains true modulo p. So

$$\mod(a, p) \times \mod(2a, p) \times \mod(3a, p) \times \cdots \times \mod((p-1), p) \equiv 1 \times 2 \times 3 \times \cdots \times (p-1) \pmod{p}.$$

In other words,

$$(a) \times (2a) \times (3a) \times \cdots \times ((p-1)a) \equiv 1 \times 2 \times 3 \times \cdots \times (p-1) \pmod{p}.$$

This equation merits careful examination. Since p is prime, it has no common factor with 2, so 2 has an inverse modulo p. This means that we can multiply both sides by $2^{-1} \pmod{m}$, effectively cancelling the factor of 2 on both sides. The same goes for each of $3, 4, \ldots, p-1$: since p is prime, none of these has any common factor with p, so they all have inverses and can be cancelled out.

Cancelling out these common factors, what's left? On the left-hand side we have $p-1$ factors of a, and everything on the right-hand side gets cancelled out. In other words, we find

$$a^{p-1} \equiv 1 \pmod{p},$$

which is what we set out to prove. □

Lemma 6.1: For any prime p and any a, b, we have $a^{1+b(p-1)} \equiv a \pmod{p}$.

Proof: There are two cases. If $a \equiv 0 \pmod{p}$, then the lemma is certainly true, for 0 to any power is still 0. If $a \not\equiv 0 \pmod{p}$, then $\gcd(a,p) = 1$, so by Fermat's Little Theorem, $a^{1+b(p-1)} \equiv a^1 \times (a^{p-1})^b \equiv a \times 1^b \equiv a \times 1 \equiv a \pmod{p}$. In either case, the lemma holds. □

Lemma 6.2: For any two different primes p, q and any x, k, we have $x^{1+k(p-1)(q-1)} \equiv x \pmod{pq}$.

Proof: Applying Lemma 6.1 with $a = x$ and $b = k(q-1)$, we see that $x^{1+k(p-1)(q-1)} \equiv x \pmod{p}$. Therefore, $x^{1+k(p-1)(q-1)} - x$ is a multiple of p.

A second application of Lemma 6.1 shows that $x^{1+k(p-1)(q-1)} \equiv x \pmod{q}$. Therefore, $x^{1+k(p-1)(q-1)} - x$ is also a multiple of q.

However, any value that is divisible by both p and q must be divisible by their product pq, so this difference must be a multiple of pq. It follows that $x^{1+k(p-1)(q-1)} \equiv x \pmod{pq}$, as was to be shown. □

Finally, we are ready to to prove Theorem 6.1.

Proof: Let $x \in \{0, 1, \ldots, N-1\}$ be arbitrary. Notice that

$$D(E(x)) \equiv (x^e)^d \equiv x^{ed} \pmod{pq}.$$

Let's look at the exponent, namely ed. By the definition of d, we know that $ed \equiv 1 \pmod{(p-1)(q-1)}$. Therefore we can write ed in the form $ed = 1 + k(p-1)(q-1)$ for some integer k. Now by Lemma 6.2, $x^{1+k(p-1)(q-1)} \equiv x \pmod{pq}$. This means that

$$D(E(x)) \equiv x^{ed} \equiv x^{1+k(p-1)(q-1)} \equiv x \pmod{pq}.$$

Since $D(E(x)) \equiv x \pmod{N}$, and since both x and $D(E(x))$ are numbers in the set $\{0, 1, \ldots, N-1\}$, it follows that $D(E(x)) = x$, which proves the theorem. □

So we have seen that the RSA protocol is correct, in the sense that Alice can encrypt messages in such a way that Bob can reliably decrypt them again. But how do we know that it is secure, i.e., that Eve cannot get any useful information by observing the encrypted messages? The security of RSA hinges upon the following simple assumption:

> Given N, e, and $y = \mod(x^e, N)$, there is no efficient algorithm for determining x.

This assumption is quite plausible. How might Eve try to guess x? She could experiment with all possible values of x, each time checking whether $x^e \equiv y \pmod{N}$; but she would have to try on the order of N values of x, which is completely unrealistic if N is a number with (say) 2048 bits. Alternatively, she could try to factor N to retrieve p and q, and then figure out d by computing the inverse of $e \pmod{(p-1)(q-1)}$; but this approach requires Eve to be able to factor N into its prime factors, a problem which is believed to be impossible to solve efficiently for large values of N. We should point out that the security of RSA has not been formally proved: it rests on the assumptions that breaking RSA is essentially tantamount to factoring N, and that factoring is hard.

We close this note with a brief discussion of implementation issues for RSA. Since we have argued that breaking RSA is impossible because factoring would take a very long time, we should check that the computations that Alice and Bob themselves have to perform are much simpler, and can be done efficiently. There are really only two non-trivial things that Alice and Bob have to do:

1. Bob has to find prime numbers p and q, each having many (say, 1024) bits.

2. Both Alice and Bob have to compute exponentials modulo N. (Alice has to compute $\bmod(x^e, N)$ and Bob has to compute $\bmod(y^d, N)$.)

We briefly discuss the implementation of each of these points in turn.

To find large prime numbers, we use the fact that, given a positive integer n, there is an efficient algorithm that determines whether or not n is prime. (Here "efficient" means a running time of $O((\log n)^k)$ for some small k, i.e., a low-degree power of the number of bits in n. Notice the dramatic contrast here with factoring: we can tell efficiently whether or not n is prime, but in the case that it is not prime we cannot efficiently find its factors. The security of RSA hinges crucially on this distinction.) Given that we can test for primes, Bob just needs to generate some random integers n with the right number of bits, and test them until he finds two primes p and q. This works because of the following basic fact from number theory (which we will not prove), which says that a reasonably large fraction of positive integers are prime:

Theorem 6.3: **Prime Number Theorem** Let $\pi(n)$ denote the number of primes that are less than or equal to n. Then for all $n \geq 17$, we have $\pi(n) \geq \frac{n}{\ln n}$. (And in fact, $\lim_{n\to\infty} \frac{\pi(n)}{n/\ln n} = 1$.)

Setting $n = 2^{1024}$, for example, the Prime Number Theorem says that roughly one in every 710 of all 1024-bit numbers are prime. Therefore, if we keep picking random 1024-bit numbers and testing them, we would expect to have to try only about 710 numbers until we find a prime.

We turn now to the second operation: exponentiation. How does Alice compute the quantity $x^e \pmod N$? Note that here e could be a very large number (with a similar number of bits to N), so repeatedly multiplying by x $e-1$ times is out of the question. Fortunately, Alice can do much better using the following trick known as "repeated squaring." Let $e_k e_{k-1} \ldots e_1 e_0$ be the binary representation of e (so that e_0 is the least significant bit, and e has $k+1$ bits in all). Then we can compute x_e as follows:

- compute the powers $x, x^2, x^4, \ldots, x^{2k} \pmod N$ by repeated squaring
- compute x^e by multiplying (modulo N) those powers x^{2i} for which $e_i = 1$.

Example: Let's compute $x^{27} \pmod N$ using the above scheme. Here $e = 27$, which has binary representation 11011. The above scheme first computes the powers x, x^2, x^4, x^8, x^{16} (all reduced modulo N) by repeated squaring. It then uses these powers to compute x^e as the product

$$x^{27} = x^{16} \times x^8 \times x^2 \times x.$$

Note that the number of multiplications required here is only seven (four for the powers, and three for the final product). This is much less than the $e - 1 = 26$ multiplications that would be required in the naive computation.

How long does this procedure take in general? The number of powers to be computed is equal to the number of bits in e, and each is computed using a single multiplication (squaring). Then x^e itself is computed by multiplying together some subset of these powers. Thus the total number of multiplications required is at most twice the number of bits in e. Since $e < N$, the entire computation can be performed using $O(\log N)$ multiplications of $O(\log N)$-bit numbers, i.e., in time $O((\log N)^3)$. Recall that $\log N$ is a relatively small quantity (say, about 1024). Hence the running time here is very reasonable. The same procedure can be used by Bob when computing the decryption function y^d.

To summarize, then, in the RSA protocol Bob need only perform simple calculations such as multiplication, exponentiation, and primality testing to implement his digital lock. Similarly, Alice and Bob need only perform simple calculations to lock and unlock the message respectively—operations that any pocket computing device could handle. By contrast, to unlock the message without the key, Eve would have to perform

operations like factoring large numbers, which (according at least to a widely accepted belief) requires more computational power than the world's most powerful computers combined! This compelling guarantee of security explains why the RSA cryptosystem is such a revolutionary development in cryptography, and why it is so widely used in practice.

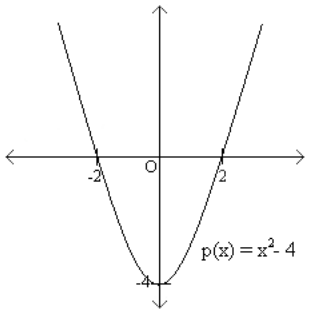
One parting caution: there are many non-trivial details in using the mathematics shown here to build a secure system, so don't assume that just because you have seen this math, you now know everything you need to know to build a secure communication system. (You can learn more about these issues in CS161.)

Credits: This note is partly based on Section 1.4 of "Algorithms," by S. Dasgupta, C. Papadimitriou and U. Vazirani, McGraw-Hill, 2007.

Polynomials

Recall from your high school math that a *polynomial* in a single variable is of the form $p(x) = a_d x^d + a_{d-1}x^{d-1} + \ldots + a_0$. Here the *variable* x and the *coefficients* a_i are usually real numbers. For example, $p(x) = 5x^3 + 2x + 1$, is a polynomial of *degree* $d = 3$. Its coefficients are $a_3 = 5$, $a_2 = 0$, $a_1 = 2$, and $a_0 = 1$. Polynomials have some remarkably simple, elegant and powerful properties, which we will explore in this note.

First, a definition: we say that a is a *root* of the polynomial $p(x)$ if $p(a) = 0$. For example, the degree 2 polynomial $p(x) = x^2 - 4$ has two roots, namely 2 and -2, since $p(2) = p(-2) = 0$. If we plot the polynomial $p(x)$ in the x-y plane, then the roots of the polynomial are just the places where the curve crosses the x axis:

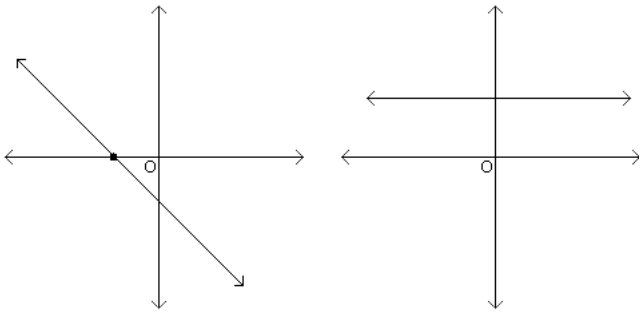


We now state two fundamental properties of polynomials that we will prove in due course.

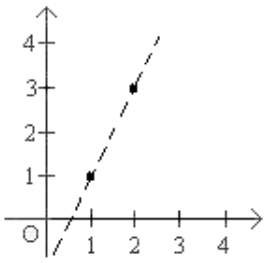
Property 1: A non-zero polynomial of degree d has at most d roots.

Property 2: Given $d+1$ pairs $(x_1, y_1), \ldots, (x_{d+1}, y_{d+1})$, with all the x_i distinct, there is a unique polynomial $p(x)$ of degree (at most) d such that $p(x_i) = y_i$ for $1 \leq i \leq d+1$.

Let us consider what these two properties say in the case that $d = 1$. A graph of a linear (degree 1) polynomial $y = a_1 x + a_0$ is a line. Property 1 says that if a line is not the x-axis (i.e. if the polynomial is not $y = 0$), then it can intersect the x-axis in at most one point.



Property 2 says that two points uniquely determine a line.



Polynomial Interpolation

Property 2 says that two points uniquely determine a degree 1 polynomial (a line), three points uniquely determine a degree 2 polynomial, four points uniquely determine a degree 3 polynomial, and so on. Given $d+1$ pairs $(x_1,y_1),\ldots,(x_{d+1},y_{d+1})$, how do we determine the polynomial $p(x)=a_dx^d+\ldots+a_1x+a_0$ such that $p(x_i)=y_i$ for $i=1$ to $d+1$? We will give two different efficient algorithms for reconstructing the coefficients $a_0,\ldots,a_d$, and therefore the polynomial $p(x)$.

In the first method, we write a system of $d+1$ linear equations in $d+1$ variables: the coefficients of the polynomial $a_0,\ldots,a_d$. The ith equation is: $a_dx_i^d+a_{d-1}x_i^{d-1}+\ldots+a_0=y_i$.

Since x_i and y_i are constants, this is a linear equation in the $d+1$ unknowns $a_0,\ldots,a_d$. Now solving these equations gives the coefficients of the polynomial $p(x)$. For example, given the 3 pairs $(-1,2)$, $(0,1)$, and $(2,5)$, we will construct the degree 2 polynomial $p(x)$ which goes through these points. The first equation says $a_2(-1)^2+a_1(-1)+a_0=2$. Simplifying, we get $a_2-a_1+a_0=2$. Applying the same technique to the second and third equations, we get the following system of equations:

$$\begin{aligned} a_2-a_1+a_0&=2\\ a_0&=1\\ 4a_2+2a_1+a_0&=5 \end{aligned}$$

Substituting for a_0 and multiplying the first equation by 2 we get:

$$\begin{aligned} 2a_2-2a_1&=2\\ 4a_2+2a_1&=4 \end{aligned}$$

Then, adding down we find that $6a_2=6$, so $a_2=1$, and plugging back in we find that $a_1=0$. Thus, we have determined the polynomial $p(x)=x^2+1$. To justify this method more carefully, we must show that the equations always have a solution and that it is unique. This involves showing that a certain determinant is non-zero. We will leave that as an exercise, and turn to the second method.

The second method is called *Lagrange interpolation*: Let us start by solving an easier problem. Suppose that we are told that $y_1=1$ and $y_j=0$ for $2\le j\le d+1$. Now can we reconstruct $p(x)$? Yes, this is easy! Consider $q(x)=(x-x_2)(x-x_3)\cdots(x-x_{d+1})$. This is a polynomial of degree d (the x_i's are constants, and x appears d times). Also, we clearly have $q(x_j)=0$ for $2\le j\le d+1$. But what is $q(x_1)$? Well, $q(x_1)=(x_1-x_2)(x_1-x_3)\cdots(x_1-x_{d+1})$, which is some constant not equal to 0 (since the x_i are all distinct). Thus if we let $p(x)=q(x)/q(x_1)$ (dividing is ok since $q(x_1)\neq 0$), we have the polynomial we are looking for. For example, suppose you were given the pairs $(1,1)$, $(2,0)$, and $(3,0)$. Then we can construct the degree $d=2$ polynomial $p(x)$ by letting $q(x)=(x-2)(x-3)=x^2-5x+6$, and $q(x_1)=q(1)=2$. Thus, we can now construct $p(x)=q(x)/q(x_1)=(x^2-5x+6)/2$.

Of course the problem is no harder if we single out some arbitrary index i instead of 1. In other words, if we want to find a polynomial such that $y_i=1$ and $y_j=0$ for $j\neq i$, we can do that. Let us introduce some

notation: let us denote by $\Delta_i(x)$ the degree d polynomial that goes through these $d+1$ points, i.e., $\Delta_i(x_i) = 1$ and $\Delta_i(x_j) = 0$ when $j \neq i$. Then

$$\Delta_i(x) = \frac{\prod_{j \neq i}(x - x_j)}{\prod_{j \neq i}(x_i - x_j)}.$$

Let us now return to the original problem. Given $d+1$ pairs $(x_1, y_1), \ldots, (x_{d+1}, y_{d+1})$, we first construct the $d+1$ polynomials $\Delta_1(x), \ldots, \Delta_{d+1}(x)$. Now the polynomial we are looking for is

$$p(x) = \sum_{i=1}^{d+1} y_i \Delta_i(x).$$

Why does this work? First notice that $p(x)$ is a polynomial of degree d as required, since it is the sum of polynomials of degree d. And when it is evaluated at x_i, d of the $d+1$ terms in the sum evaluate to 0 and the ith term evaluates to y_i times 1, as required.

As an example, suppose we want to find the degree-2 polynomial $p(x)$ that passes through the three points $(1,1)$, $(2,2)$ and $(3,4)$. We have $d = 2$ and $x_i = i$. The three polynomials Δ_i are thus as follows:

$$\begin{aligned}
\Delta_1(x) &= \frac{(x-2)(x-3)}{(1-2)(1-3)} = \frac{(x-2)(x-3)}{2} = \frac{1}{2}x^2 - \frac{5}{2}x + 3; \\
\Delta_2(x) &= \frac{(x-1)(x-3)}{(2-1)(2-3)} = \frac{(x-1)(x-3)}{-1} = -x^2 + 4x - 3; \\
\Delta_3(x) &= \frac{(x-1)(x-2)}{(3-1)(3-2)} = \frac{(x-1)(x-2)}{2} = \frac{1}{2}x^2 - \frac{3}{2}x + 1.
\end{aligned}$$

The polynomial $p(x)$ is therefore given by

$$p(x) = 1 \cdot \Delta_1(x) + 2 \cdot \Delta_2(x) + 4 \cdot \Delta_3(x) = \frac{1}{2}x^2 - \frac{1}{2}x + 1.$$

You should verify that this polynomial does indeed pass through the above three points.

Property 2 and Uniqueness

We have shown how to find a polynomial $p(x)$ that passes through any given $(d+1)$ points. This proves part of Property 2 (the existence of the polynomial). How do we prove the second part, that the polynomial is unique? Suppose for contradiction that there is another polynomial $q(x)$ that also passes through the $d+1$ points. Now consider the polynomial $r(x) = p(x) - q(x)$. This is a non-zero polynomial of degree at most d. So by Property 1 it can have at most d roots. But on the other hand $r(x_i) = p(x_i) - q(x_i) = 0$ for $i = 1, \ldots, d+1$, so $r(x)$ has $d+1$ distinct roots. Contradiction. Therefore $p(x)$ is the unique polynomial that satisfies the $d+1$ conditions.

Property 1

Now let us turn to Property 1. We will prove this property in two steps.

Theorem 7.1: a is a root of $p(x)$ if and only if the polynomial $x - a$ divides $p(x)$.

Proof: Dividing $p(x)$ by the polynomial $x - a$ yields

$$p(x) = (x-a)q(x) + r(x)$$

for some polynomials $q(x)$ and $r(x)$, where $q(x)$ is the quotient and $r(x)$ is the remainder. The degree of $r(x)$ is necessarily smaller than the degree of the divisor $(x-a)$. Therefore $r(x)$ must have degree 0 and therefore

is some constant c, so $r(x) = c$. Now substituting $x = a$, we get $p(a) = 0 \cdot q(a) + r(a) = c$. If a is a root of $p(x)$, then $p(a) = 0$, so $c = 0$ and therefore $p(x) = (x-a)q(x)$, thus showing that $(x-a)$ divides $p(x)$.

On the other hand, if $x-a$ divides $p(x)$, then we know that $r(x) = 0$ and $p(x) = (x-a)q(x)$, hence $p(a) = 0 \cdot q(a) = 0$ and in particular a is a root of $p(x)$. □

Theorem 7.2: If $a_1, \ldots, a_d$ are d distinct roots of a polynomial $p(x)$ of degree d, then $p(x)$ has no other roots.

Proof: We will show that $p(x) = c(x-a_1)(x-a_2)\cdots(x-a_d)$ for some constant c. First, observe that $p(x) = (x-a_1)q_1(x)$ for some polynomial $q_1(x)$ of degree $d-1$, since a_1 is a root. Also $0 = p(a_2) = (a_2-a_1)q_1(a_2)$ since a_2 is a root. But since $a_2 - a_1 \neq 0$, it follows that $q_1(a_2) = 0$, i.e., a_2 is a root of q_1. So $q_1(x) = (x-a_2)q_2(x)$, for some polynomial $q_2(x)$ of degree $d-2$. Proceeding in this manner by induction, we find that $p(x) = (x-a_1)(x-a_2)\cdots(x-a_d)q_d(x)$ for some polynomial $q_d(x)$ of degree 0. A polynomial of degree 0 must be of the form $q_d(x) = c$ for some constant c, so we've shown that $p(x) = c(x-a_1)(x-a_2)\cdots(x-a_d)$ for some constant c, as claimed.

The theorem follows immediately. If a is any other value, different from $a_1, \ldots, a_d$, then $a - a_i \neq 0$ for all i and hence $p(a) = c(a-a_1)(a-a_2)\cdots(a-a_d) \neq 0$. In other words, no other value a can be a root of the polynomial $p(x)$. □

This completes the proof that a polynomial of degree d has at most d roots.

Finite Fields

Property 1 and Property 2 were stated under the assumption that the coefficients of the polynomials and the variable x range over the real numbers. These properties also hold if we use the set of rational numbers, or even the set of complex numbers, instead of the real numbers.

However, the properties do not hold if the values are restricted to the set of natural numbers or integers. Let us try to understand this a little more closely. The only properties of numbers that we used in polynomial interpolation and in the proof of Property 1 is that we can add, subtract, multiply and divide any pair of numbers as long as we are not dividing by 0. We cannot subtract two natural numbers and guarantee that the result is a natural number. And dividing two integers does not usually result in an integer. As a result, our proof of Property 1 does not generalize to the case where the values are restricted to $\mathbb{N}$ or $\mathbb{Z}$.

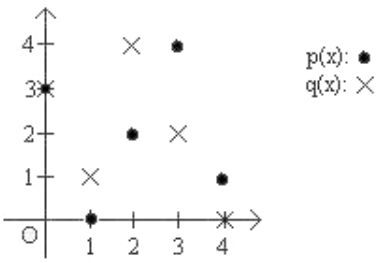
But if we work with numbers modulo a prime m, then we can add, subtract, multiply and divide (by any non-zero number modulo m). To check this, recall from our discussion of modular arithmetic in the previous notes that x has an inverse mod m if $\gcd(m,x) = 1$. Thus if m is prime *all* the numbers $\{1, \ldots, m-1\}$ have an inverse mod m. So both Property 1 and Property 2 hold if the coefficients and the variable x are restricted to take on values modulo m. This remarkable fact that these properties hold even when we restrict ourselves to a *finite* set of values is the key to several applications that we will presently see.

First, let's see an example of these properties holding in the case of polynomials of degree $d = 1$ modulo 5. Consider the polynomial $p(x) \equiv 4x + 3 \pmod 5$. The roots of this polynomial are all values x such that $4x + 3 \equiv 0 \pmod 5$. Solving for x, we get that $4x \equiv 2 \pmod 5$, or $x \equiv 3 \pmod 5$. (In this last step we multiplied through by the inverse of 4 mod 5, which is 4.) Thus, we found only 1 root for a degree 1 polynomial. Now, given the points $(0,3)$ and $(1,2)$, we will reconstruct the degree 1 polynomial $p(x)$ modulo 5. Using Lagrange interpolation, we get that $\Delta_1(x) \equiv \frac{x-x_2}{x_1-x_2} \equiv \frac{x-1}{0-1} \equiv -(x-1) \pmod 5$, and $\Delta_2(x) \equiv \frac{x-x_1}{x_2-x_1} \equiv \frac{x-0}{1-0} \equiv x \pmod 5$. Thus, $p(x) \equiv 3 \cdot \Delta_1(x) + 2 \cdot \Delta_2(x) \equiv -3(x-1) + 2x \equiv -x + 3 \equiv 4x + 3 \pmod 5$.

When we work with numbers modulo a prime m, we are working over finite fields, denoted by F_m or

$GF(m)$ (for Galois Field). In order for a set to be called a field, it must satisfy certain axioms which are the building blocks that allow for these amazing properties and others to hold. Intuitively, a field is a set where we can add, subtract, multiply, and divide any pair of elements from the set, and we will get another element in the set (as long as we don't try to divide by 0). If you would like to learn more about fields and the axioms they satisfy, you can visit Wikipedia's site and read the article on fields: `http://en.wikipedia.org/wiki/Field_%28mathematics%29`. While you are there, you can also read the article on Galois Fields and learn more about some of their applications and elegant properties which will not be covered in this lecture: `http://en.wikipedia.org/wiki/Galois_field`.

We said above that it is remarkable that Properties 1 and 2 continue to hold when we restrict all values to a finite set modulo a prime number m. To see why this is remarkable let us see what the graph of a linear polynomial (degree 1) looks like modulo 5. There are now only 5 possible choices for x, and only 5 possible choices for y. Consider the polynomials $p(x) \equiv 2x+3 \pmod 5$ and $q(x) \equiv 3x-2 \pmod 5$ over $GF(5)$. We can represent these polynomials in the x-y plane as follows:



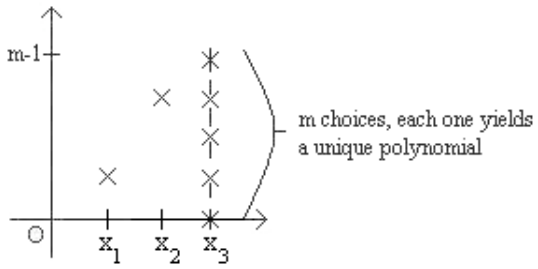
Notice that these two "lines" intersect in exactly one point, even though the picture looks nothing at all like lines in the Euclidean plane! Modulo 5, two lines can still intersect in at most one point, and that is thanks to the properties of addition, subtraction, multiplication, and division modulo 5.

Finally, you might wonder why we chose m to be a prime. Let us briefly consider what would go wrong if we chose m not to be prime, for example $m = 6$. Now we can no longer divide by 2 or 3. In the proof of Property 1, we asserted that $p(a) = c(a-a_1)(a-a_2)\cdots(a-a_d) \neq 0$ if $a \neq a_i$ for all i. But when we are working modulo 6, if $a - a_1 \equiv 2 \pmod 6$ and $a - a_2 \equiv 3 \pmod 6$, these factors are non-zero, but $(a-a_1)(a-a_2) \equiv 2 \cdot 3 \equiv 0 \pmod 6$. Working modulo a prime ensures that this disaster cannot happen.

Counting

How many polynomials of degree (at most) 2 are there modulo m? This is easy: there are 3 coefficients, each of which can take on m distinct values, so there are a total of $m \times m \times m = m^3$ such polynomials. Equivalently, each of the polynomials is uniquely specified by its values at three points, say at $x = 1$, $x = 2$ and $x = 3$; and there are m^3 choices for these three values, each of which yields a distinct polynomial.

Now suppose we are given three pairs $(x_1, y_1), (x_2, y_2), (x_3, y_3)$. Then by Property 2, there is a unique polynomial of degree 2 such that $p(x_i) = y_i$ for $1 \leq i \leq 3$. Suppose we were only given two pairs $(x_1, y_1), (x_2, y_2)$; how many distinct degree 2 polynomials are there that go through these two points? Here is a slick way of working this out. Fix any x_3, and notice that there are exactly m choices for y_3. Once three points are specified, by Property 2 there is a unique polynomial of degree 2 that goes through these three points. Since this is true for each of the m ways of choosing y_3, it follows that there are m polynomials of degree at most 2 that go through the 2 points $(x_1, y_1), (x_2, y_2)$. This is illustrated below:



What if you were only given one point? Well, there are m choices for the second point, and for each of these there are m choices for the third point, yielding a total of m^2 polynomials of degree at most 2 that go through the point given. A pattern begins to emerge, as is summarized in the following table:

Polynomials of degree $\leq d$ over F_m	
# of points given	# of polynomials
$d+1$	1
d	m
$d-1$	m^2
$\vdots$	$\vdots$
$d-k$	m^{k+1}

The reason that we can count the number of polynomials is because we are working over a finite field. If we were working over an infinite field such as the rationals, there would be infinitely many polynomials of degree d that can go through d points. Think of a line, which has degree one. If you were just given one point, there would be infinitely many possibilities for the second point, each of which uniquely defines a line.

Secret Sharing

In the late 1950's and into the 1960's, during the Cold War, President Dwight D. Eisenhower approved instructions and authorized top commanding officers for the use of nuclear weapons under very urgent emergency conditions. Such measures were set up in order to defend the United States in case of an attack in which there was not enough time to confer with the President and decide on an appropriate response. This would allow for a rapid response in case of a Soviet attack on U.S. soil. This is a perfect situation in which a secret sharing scheme could be used to ensure that a certain number of officials must come together in order to successfully launch a nuclear strike, so that for example no single person has the power and control over such a devastating and destructive weapon.

Suppose the U.S. government decides that a nuclear strike can be initiated only if at least k major officials agree to it, for some $k > 1$. Suppose that missiles are protected by a secret launch code; the missile will only launch if it is supplied with the proper launch code. Let's devise a scheme such that (1) any group of k of these officials can pool their information to figure out the launch code and initiate the strike, but (2) no group of $k-1$ or fewer have any information about the launch code (not even partial information), even if they pool their knowledge. For example, a group of $k-1$ conspiring officials should not be able to tell whether the secret launch code is odd or even; whether it is a prime number or not; whether it is divisible by some number a; or whether its least significant bit is 0 or 1. How can we accomplish this?

We'll presume that there are n officials indexed from 1 to n and that the secret launch code is some natural number s. Let q be a prime number larger than n and s. We will work over $GF(q)$ from now on, i.e., we will be working modulo q.

The scheme is simple. We pick a random polynomial $P(x)$ of degree $k-1$ such that $P(0) = s$. Then, we give the share $P(1)$ to the first official, $P(2)$ to the second official, ..., $P(n)$ to the nth official.

This satisfies our two desiderata:

1. Any k officials, having the values of the polynomial at k points, can use Lagrange interpolation to find $P(x)$. Once they know the polynomial $P(x)$, they can recover the secret s by computing $s = P(0)$.

2. What about some group of $k-1$ conspiring officials? They don't have enough information to recover the polynomial $P(x)$. All they know is that there is some polynomial of degree $k-1$ passing through their $k-1$ points. However, for each possible value $P(0) = b$, there is a unique polynomial that is consistent with the information of the $k-1$ officials, and that also satisfies the constraint that $P(0) = b$. This means that any conjectured value of the secret is consistent with the information available to the $k-1$ conspirators, so the conspirators cannot rule out any hypothesized value for $P(0)$ as impossible. In short, the conspirators learn nothing about $P(0) = s$.

This scheme is known as Shamir secret sharing, in honor of its inventor, Adi Shamir.

Example. Suppose you are in charge of setting up a secret sharing scheme, with secret $s = 1$, where you want to distribute $n = 5$ shares to 5 people such that any $k = 3$ or more people can figure out the secret, but two or fewer cannot. We will need a polynomial of degree $k-1 = 2$. Let's say we are working over $GF(7)$ and you randomly choose the polynomial $P(x) = 3x^2 + 5x + 1$. (Notice: $P(0) = 1 = s$, the secret.) So you know everything there is to know about the secret and the polynomial, but what about the people that receive the shares? Well, the shares handed out are $P(1) \equiv 3 \cdot 1^2 + 5 \cdot 1 + 1 \equiv 9 \equiv 2 \pmod 7$ to the first official, $P(2) \equiv 3 \cdot 2^2 + 5 \cdot 2 + 1 \equiv 23 \equiv 2 \pmod 7$ to the second, $P(3) \equiv 3 \cdot 3^2 + 5 \cdot 3 + 1 \equiv 43 \equiv 1 \pmod 7$ to the third, $P(4) \equiv 6 \pmod 7$ to the fourth, and $P(5) \equiv 3$ to the fifth official. Let's say that officials 3, 4, and 5 get together. We expect them to be able to recover the secret. Using Lagrange interpolation, they compute the following functions:

$$\Delta_3(x) \equiv \frac{(x-4)(x-5)}{(3-4)(3-5)} \equiv \frac{(x-4)(x-5)}{2} \equiv 4(x-4)(x-5) \pmod 7$$
$$\Delta_4(x) \equiv \frac{(x-3)(x-5)}{(4-3)(4-5)} \equiv \frac{(x-3)(x-5)}{-1} \equiv -(x-3)(x-5) \pmod 7$$
$$\Delta_5(x) \equiv \frac{(x-3)(x-4)}{(5-3)(5-4)} \equiv \frac{(x-3)(x-4)}{2} \equiv 4(x-3)(x-4) \pmod 7.$$

They then compute the polynomial $P(x) = 1 \cdot \Delta_3(x) + 6 \cdot \Delta_4(x) + 3 \cdot \Delta_5(x) \equiv 3x^2 + 5x + 1$ (you should verify this computation!). Now they simply compute $P(0)$ and discover that the secret is 1.

Let's see what happens if two officials try to get together, say persons 1 and 5. They both know that the polynomial looks like $P(x) = a_2x^2 + a_1x + s$. They also know the following equations:

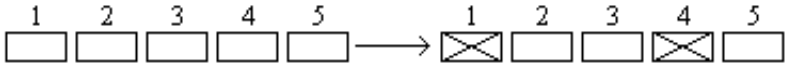
$$P(1) \equiv a_2 + a_1 + s \equiv 2 \pmod 7$$
$$P(5) \equiv 4a_2 + 5a_1 + s \equiv 3 \pmod 7$$

But that is all they have, 2 equations with 3 unknowns, and thus they cannot find out the secret. This is the case no matter which two officials get together. Notice that since we are working over $GF(7)$, the two people could have guessed the secret ($0 \leq s \leq 6$) and identified a unique degree 2 polynomial that's consistent with their guess (by Property 2). But the two people combined have the same chance of guessing what the secret is as they do individually. This is important, as it implies that two people have no more information about the secret than one person does—in particular, these two people have no information about the secret, not even partial information.

Error Correcting Codes

Erasure Errors

We will consider the situation where we wish to transmit information over an unreliable communication channel. This is exemplified by the internet, where the information (say a file) is broken up into fixed-length packets, and the unreliability is manifest in the fact that some of the packets may be lost during transmission, as shown below:



Suppose that, in the absence of packet loss, it would take n packets to send the entire message—but in reality up to k packets may be lost during transmission. We will show how to encode the initial message consisting of n packets into a redundant encoding consisting of $n+k$ packets such that the recipient can reconstruct the message from *any* n received packets. We will assume that the packets are labeled and thus the recipient knows exactly which packets were dropped during transmission.

In our scheme, the contents of each packet is a number modulo q, where q is a prime. For example, a 32-bit string can be regarded as a number between 0 and $2^{32}-1$; then we could choose q to be any prime larger than 2^{32} and view it as a number modulo q. The properties of polynomials over $GF(q)$ (i.e., with coefficients and values reduced modulo q) are perfectly suited to solve this problem and are the backbone of this error-correcting scheme. To see this, let us denote the message to be sent by $m_1,\ldots,m_n$. Then we can form a polynomial $P(x)$ of degree $n-1$ by setting $P(x) = m_n x^{n-1} + \cdots + m_3 x^2 + m_2 x + m_1$. This polynomial encodes all of the information in the message: if we can somehow communicate the polynomial to the recipient over the unreliable communication channel, then the recipient can recover the message $m_1,\ldots,m_n$ from $P(x)$. Therefore, we can consider the message to be given by the polynomial $P(x)$.

We generate the encoded packets to be sent over the channel by evaluating $P(x)$ at $n+k$ points. In particular, we will transmit the following $n+k$ packets: $c_1 = P(1)$, $c_2 = P(2)$, $\ldots$, $c_{n+k} = P(n+k)$. In particular, we transmit $n+k$ numbers, where each number is in the range $0 \ldots q-1$. Notice that the transmission is redundant: it contains more packets than the original message, to deal with lost packets. (A technical restriction: since we are working modulo q, we must make sure that $n+k \leq q$, to ensure that the numbers $1,2,\ldots,n+k$ are all distinct modulo q. However, this condition does not impose a serious constraint, since in practice q will typically be very large.)

Here is the key point: the recipient can uniquely reconstruct $P(x)$ from its values at any n distinct points, since it has degree $n-1$. This means that the recipient can reconstruct $P(x)$ from any n of the transmitted packets (e.g., by using Lagrange interpolation). The coefficients of this reconstructed polynomial $P(x)$ reveal the original message $m_1,\ldots,m_n$. Therefore, as long as no more than k packets are lost—as long as the recipient receives at least n of the $n+k$ encoded packets—the recipient will be able to recover the message that the sender wanted to send.

Example. Suppose Alice wants to send Bob a message of $n = 4$ packets and she wants to guard against up to $k = 2$ lost packets. Then, assuming the packets can be coded as integers between 0 and 6, Alice can work over $GF(7)$ (since $7 \geq n + k = 6$). Suppose the message that Alice wants to send to Bob is $m_1 = 5$, $m_2 = 0$, $m_3 = 4$, and $m_4 = 1$. The polynomial with these coefficients is $P(x) = x^3 + 4x^2 + 5$.

Alice must evaluate $P(x)$ at $n + k = 6$ points, namely, at $x = 1, 2, \ldots, 6$. The encoded packets are

$$\begin{aligned} c_1 &= P(1) = 3 \\ c_2 &= P(2) = 1 \\ c_3 &= P(3) = 5 \\ c_4 &= P(4) = 0 \\ c_5 &= P(5) = 6 \\ c_6 &= P(6) = 1. \end{aligned}$$

(Remember that all arithmetic is done in $GF(7)$, i.e., modulo 7, so everything is reduced modulo 7.) Since $k = 2$, Alice must evaluate $P(x)$ at 2 extra points: Suppose packets 2 and 6 are dropped, in which case we have the following situation:

From the values that Bob received (3, 5, 0, and 6), he uses Lagrange interpolation and computes the following delta functions:

$$\begin{aligned} \Delta_1(x) &= \frac{(x-3)(x-4)(x-5)}{-24} \\ \Delta_3(x) &= \frac{(x-1)(x-4)(x-5)}{4} \\ \Delta_4(x) &= \frac{(x-1)(x-3)(x-5)}{-3} \\ \Delta_5(x) &= \frac{(x-1)(x-3)(x-4)}{8}. \end{aligned}$$

He then reconstructs the polynomial $P(x) = 3\Delta_1(x) + 5\Delta_3(x) + 0\Delta_4(x) + 6\Delta_5(x) = x^3 + 4x^2 + 5$. Bob then reads off the coefficients to recover the original message Alice wanted to send: $m_1 = 5$, $m_2 = 0$, $m_3 = 4$, $m_4 = 1$. More generally, no matter which two packets were dropped, following the same method Bob could still have reconstructed $P(x)$ and thus the original message.

Let us consider what would happen if Alice sent one fewer packet. If Alice only sent $c_1, \ldots, c_5$, then after 2 packets are lost, Bob would only receive c_j for 3 distinct values j. Thus, Bob would not be able to reconstruct $P(x)$ (since there are exactly q polynomials of degree at most 3 that agree with the 3 encoded packets which Bob received). This error-correcting scheme is therefore optimal: it can recover the n characters of the transmitted message from any n received characters, but recovery from any fewer characters is impossible.

General Errors

The main shortcoming of the scheme described above is that it only deals with lost packets. What if up to k packets might be corrupted or modified in transmission, so the affected packet is received incorrectly by Alice? The scheme above cannot handle that. In fact, it turns out that, with some additional work, the

scheme can be made to handle corrupted packets. We will need to send $n+2k$ encoded packets, instead of just $n+k$, and we will need a more sophisticated decoding procedure. We won't cover the details in this class, but many of the main ideas still apply, and again polynomials and modular arithmetic hold the key.

Counting

In the next major topic of the course, we will be looking at probability. Suppose you toss a fair coin a thousand times. How likely is it that you get exactly 500 heads? And what about 1000 heads? It turns out that the chances of 500 heads are roughly 5%, whereas the chances of 1000 heads are so infinitesimally small that we may as well say that it is impossible. But before you can learn to compute or estimate odds or probabilities you must learn to count! That is the subject of this note.

We will learn how to count the number of outcomes while tossing coins, rolling dice and dealing cards. Many of the questions we will be interested in can be cast in the following simple framework, called the *occupancy model*:

Balls & Bins: We have a set of k balls. We wish to place them into n bins. How many different possible outcomes are there?

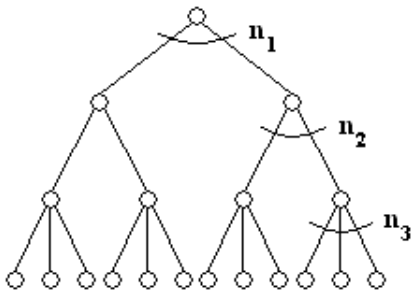
How do we represent coin tossing and card dealing in this framework? Consider the case of $n = 2$ bins labelled H and T, corresponding to the two outcomes of a coin toss. The placement of the k balls correspond to the outcomes of k successive coin tosses. To model card dealing, consider the situation with 52 bins corresponding to a deck of cards. Here the balls correspond to successive cards in a deal.

The two examples illustrate two different constraints on ball placements. In the coin tossing case, different balls can be placed in the same bin. This is called *sampling with replacement*. In the cards case, no bin can contain more than one ball (i.e., the same card cannot be dealt twice). This is called *sampling without replacement*. As an exercise, what are n and k for rolling dice? Is it sampling with or without replacement?

We are interested in counting the number of ways of placing k balls in n bins in each of these scenarios. This is easy to do by applying the first rule of counting:

First Rule of Counting: If an object can be made by a succession of k choices, where there are n_1 ways of making the first choice, and *for every* way of making the first choice there are n_2 ways of making the second choice, and *for every* way of making the first and second choice there are n_3 ways of making the third choice, and so on up to the n_k-th choice, then the total number of distinct objects that can be made in this way is the product $n_1 \times n_2 \times n_3 \times \cdots \times n_k$.

Here is another way of picturing this rule: consider a tree with branching factor n_1 at the root, n_2 at every node at the second level, ..., n_k at every node at the k-th level. Then the number of leaves in the tree is the product $n_1 \times n_2 \times n_3 \times \cdots \times n_k$. For example, if $n_1 = 2$, $n_2 = 2$, and $n_3 = 3$, then there are 12 leaves (i.e., outcomes):



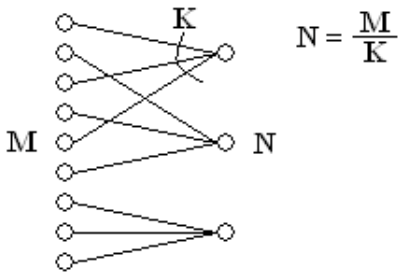
Let us apply this counting rule to figuring out the number of ways of placing k balls in n bins with replacement, if the balls are numbered from 1 to k. This is easy; it is just n^k: n choices for the first ball, n for the second, and so on.

The rule is more interesting in the case of sampling without replacement. Now there are n ways of placing the first ball, and *no matter* where it is placed there are exactly $n-1$ bins in which the second ball may be placed (exactly which $n-1$ depends upon which bin the first ball was placed in, but the number of choices is always $n-1$), and so on. So as long as $k \leq n$, the number of placements is $n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$. (By convention we define $0! = 1$.)

Counting Unordered Sets

When dealing a hand of cards, say a poker hand, it is more natural to count the number of distinct hands (i.e., the set of 5 cards dealt in the hand), without regard to the order in which the cards in our hand were dealt to us. To count this number we use the second rule of counting:

Second Rule of Counting: If an object is made by a succession of choices, and the order in which the choices is made does not matter, count the number of ordered objects (pretending that the order matters), and divide by the number of ordered objects per unordered object. Note that this rule can only be applied if the number of ordered objects is the same for every unordered object.

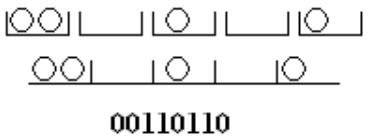


Let us continue with our example of a poker hand. We wish to calculate the number of ways of choosing 5 cards out of a deck of 52 cards. So we first count the number of ways of dealing a 5-card hand pretending that we care which order the cards are dealt in. This is exactly $\frac{52!}{47!}$ as we computed above. Now we ask: for a given poker hand, in how many ways could it have been dealt? The 5 cards in the given hand could have been dealt in any one of 5! ways. Therefore, by the second rule of counting, the number of poker hands is $\frac{52!}{47!\times 5!}$.

This quantity $\frac{n!}{(n-k)!k!}$ is used so often that there is special notation for it: $\binom{n}{k}$, pronounced *n choose k*. This is the number of ways of placing k balls in n bins (without replacement), where the order of placement does not matter. Equivalently, it's the number of ways of choosing k objects out of a total of n objects, where the order of the choices does not matter.

What about the case of sampling with replacement? How many ways are there of placing k balls in n bins with replacement when the order does not matter? Let us try to use the second rule of counting. There are n^k ordered placements. How many ordered placements are there per unordered placement? Unfortunately this depends on which unordered placement we are considering. For example, when $k = 2$ (two balls), if the two balls are in distinct bins then there are two corresponding ordered placements, while if they are in the same bin then there is just one corresponding ordered placement. Thus we have to consider these two cases separately. In the first case, there are n ways to place the first ball, and $n-1$ ways to place the second ball, giving us $n(n-1)$ corresponding ordered placements; by the second rule of counting, we divide by 2 and get $\frac{n(n-1)}{2}$ unordered placements of the balls in distinct bins. In the second case, there are n ways to place both

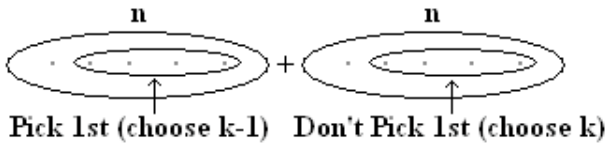
balls in the same bin; by the second rule of counting, we divide by 1 and get n unordered placements for the balls in the same bin. Putting both cases together, there are $\frac{n(n-1)}{2} + n = \frac{n(n+1)}{2}$ ways to place two balls into n bins where order does not matter. For larger values of k, it seems hopelessly complicated. Yet there is a remarkably slick way of calculating this number. Represent each of the balls by a 0 and the separations between boxes by 1's. So we have k 0's and $(n-1)$ 1's. Each placement of the k balls in the n boxes corresponds uniquely to a binary string with k 0's and $(n-1)$ 1's. Here is a sample placement of $k = 4$ balls into $n = 5$ bins and how it can be represented as a binary string:



But the number of such binary strings is easy to count: we have $n+k-1$ positions, and we must choose which k of them contain 0's. So the answer is $\binom{n+k-1}{k}$.

Combinatorial Proofs

Combinatorial arguments are interesting because they rely on intuitive counting arguments rather than algebraic manipulation. For example, it is true that $\binom{n}{k} = \binom{n}{n-k}$. Though you may be able to prove this fact rigorously by definition of $\binom{n}{k}$ and algebraic manipulation, some proofs are actually much more tedious and difficult. Instead, we will try to discuss what each term means, and then see why the two sides are equal. When we write $\binom{n}{k}$, we are really counting how many ways we can choose k objects from n objects. But each time we choose any k objects, we must also leave behind $n-k$ objects, which is the same as choosing $n-k$ (to leave behind). Thus, $\binom{n}{k} = \binom{n}{n-k}$. Some facts are less trivial. For example, it is true that $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$. The two terms on the right hand side are splitting up choosing k from n objects into two cases: we either choose the first element, or we do not. To count the number of ways where we choose the first element, we have $k-1$ objects left to choose, and only $n-1$ objects to choose from, and hence $\binom{n-1}{k-1}$ ways. For the number of ways where we don't choose the first element, we have to pick k objects from $n-1$ this time, giving $\binom{n-1}{k}$ ways. [Exercise: Check algebraically that the above formula holds.]



We can also prove even more complex facts, such as $\binom{n}{k+1} = \binom{n-1}{k} + \binom{n-2}{k} + \cdots + \binom{k}{k}$. What does the right hand side really say? It is splitting up the process into cases according to the first (i.e., lowest-numbered) object we select. These cases can be summarized as follows:

If the first element selected is…	then the number of combinations is…
element 1	$\binom{n-1}{k}$
element 2	$\binom{n-2}{k}$
element 3	$\binom{n-3}{k}$
⋮	⋮
element $n-k$	$\binom{k}{k}$

(Note that the lowest-numbered object we select cannot be higher than $n-k$ as we have to select k distinct objects.)

The last combinatorial proof we will do is the following: $\binom{n}{0}+\binom{n}{1}+\cdots+\binom{n}{n}=2^n$. To see this, imagine that we have a set S with n elements. On the left hand side, the i^{th} term counts the number of ways of choosing a subset of S of size exactly i; so the sum on the left hand side counts the total number of subsets (of any size) of S. But we claim that the right hand side (2^n) does indeed also count the total number of subsets. To see this, just identify a subset with an n-bit vector, where in each position j we put a 1 if the jth element is in the subset, and a 0 otherwise. So the number of subsets is equal to the number of n-bit vectors, which is 2^n. Let us look at an example, where $S=\{1,2,3\}$ (so $n=3$). Enumerate all $2^3=8$ possible subsets of S: $\{\{\},\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}\}$. The term $\binom{3}{0}$ counts the number of ways to choose a subset of S with 0 elements; there is only one such subset, namely the empty set. There are $\binom{3}{1}=3$ ways of choosing a subset with 1 element, $\binom{3}{2}=3$ ways of choosing a subset with 2 elements, and $\binom{3}{3}=1$ way of choosing a subset with 3 elements (namely, the subset consisting of the whole of S). Summing, we get $1+3+3+1=8$, as expected.

Introduction to Discrete Probability

Probability theory has its origins in gambling — analyzing card games, dice, roulette wheels. Today it is an essential tool in engineering and the sciences. No less so in electrical engineering and computer science, where its use is widespread in algorithms, systems, learning theory, artificial intelligence, communications, control theory and signal processing.

Here are some typical statements that you might see concerning probability:

1. The chance of getting a flush in a 5-card poker hand is about 2 in 1000.

2. The chance that this randomized primality testing algorithm outputs "prime" when the input is not prime is at most one in a trillion.

3. In this load-balancing scheme, the probability that any processor has to deal with more than 12 requests is negligible.

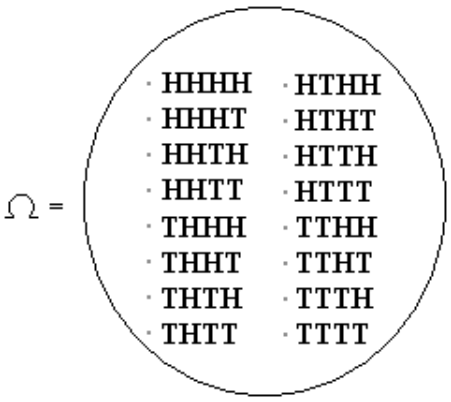
4. The average time between system failures is about 3 days.

5. There is a 30% chance of a magnitude 8.0 earthquake in Northern California before 2030.

Implicit in all such statements is the notion of an underlying **probability space**. This may be the result of a random experiment that we have ourselves constructed (as in 1, 2 and 3 above), or some model we build of the real world (as in 4 and 5 above). None of these statements makes sense unless we specify the probability space we are talking about: for this reason, statements like 5 (which are typically made without this context) are almost content-free.

Let us try to understand all this more clearly. The first important notion here is that of a **random experiment**. An example of such an experiment is tossing a coin 4 times, or dealing a poker hand. In the first case an outcome of the experiment might be $HTHT$ or it might be $HHHT$. The question we are interested in might be "what is the chance that there are exactly two H's?" Well, the number of outcomes that meet this condition is $\binom{4}{2} = \frac{4!}{2!2!} = 6$ (corresponding to choosing the positions of the two H's in the sequence of four tosses); these outcomes are $HHTT$, $HTHT$, $HTTH$, $THHT$, $THTH$, $TTHH$. On the other hand, the total number of distinct outcomes for this experiment is $2^4 = 16$. If the coin is fair then all these 16 outcomes are equally likely, so the chance that there are exactly two H's is $6/16 = 3/8$.

Now some terminology. The outcome of a random experiment is called a **sample point**. Thus $HTHT$ is an example of a sample point. The **sample space**, often denoted by Ω, is the set of all possible outcomes. In our example the sample space has 16 elements:



A **probability space** is a sample space Ω, together with a **probability** $\Pr[\omega]$ for each sample point ω, such that

- $0 \leq \Pr[\omega] \leq 1$ for all $\omega \in \Omega$.
- $\sum_{\omega \in \Omega} \Pr[\omega] = 1$, i.e., the sum of the probabilities of all outcomes is 1.

The easiest way to assign probabilities to sample points is to give all of them the same probability (as we saw earlier in the coin tossing example): if $|\Omega| = N$, then $\Pr[x] = \frac{1}{N}$ $\forall x \in \Omega$. This is known as a uniform distribution. We will see examples of non-uniform probability assignments soon.

Here's another example: dealing a poker hand. In this case, our sample space $\Omega = \{$all possible poker hands$\}$, which corresponds to choosing $k = 5$ objects without replacement from a set of size $n = 52$ where order does not matter. Hence, as we saw in the previous Note, $|\Omega| = \binom{52}{5} = \frac{52\times51\times50\times49\times48}{5\times4\times3\times2\times1} = 2,598,960$. Since the probability of each outcome is equally likely, the probability of any particular hand is the reciprocal of the size of the sample space. For instance, $\Pr[\{5\heartsuit, 3\clubsuit, 7\spadesuit, 8\clubsuit, K\heartsuit\}] = \frac{1}{2,598,960}$.

As we saw in the coin tossing example above, after performing an experiment we are often interested only in knowing whether a certain event occurred. Thus we considered the event that there were exactly two H's in the four tosses of the coin. Here are some more examples of events we might be interested in:

- The sum of the rolls of 2 dice is ≥ 10.
- The poker hand idealt to you is a flush (i.e., all 5 cards have the same suit).
- In n coin tosses, at least $\frac{n}{3}$ of the tosses come up tails.

Let us now formalize the notion of an event. Formally, an **event** A is just a subset of the sample space, $A \subseteq \Omega$. As we saw above, the event "exactly 2 H's in four tosses of the coin" is the subset $\{HHTT, HTHT, HTTH, THHT, THTH, TTHH\} \subseteq \Omega$.

How should we define the probability of an event A? Naturally, we should just *add up* the probabilities of the sample points in A.

For any event $A \subseteq \Omega$, we define the **probability of** A to be

$$\Pr[A] = \sum_{\omega \in A} \Pr[\omega].$$

Thus the probability of getting exactly two H's in four coin tosses can be calculated using this definition as follows. The event A consists of all sequences that have exactly two H's, and so $|A| = \binom{4}{2} = 6$. There are

$|\Omega| = 2^4 = 16$ possible outcomes for flipping four coins. Thus, each sample point $\omega \in A$ has probability $\frac{1}{16}$; and, as we saw above, there are six sample points in A, giving us $\Pr[A] = 6 \cdot \frac{1}{16} = \frac{3}{8}$.

We will now look at examples of probability spaces and typical events that may occur in such experiments.

1. Flip a fair coin. Here $\Omega = \{H, T\}$, and $\Pr[H] = \Pr[T] = \frac{1}{2}$.

2. Flip a fair coin three times. Here $\Omega = \{(t_1, t_2, t_3) : t_i \in \{H, T\}\}$, where t_i gives the outcome of the ith toss. Thus Ω consists of $2^3 = 8$ points, each with equal probability $\frac{1}{8}$. More generally, if we flip the coin n times, we get a sample space of size 2^n (corresponding to all sequences of length n over the alphabet $\{H, T\}$), each sample point having probability $\frac{1}{2^n}$. We can look, for example, at the event A that all three coin tosses are the same. Then $A = \{HHH, TTT\}$, with each sample point having probability $\frac{1}{8}$. Thus, $\Pr[A] = \Pr[HHH] + \Pr[TTT] = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$.

3. Flip a biased coin once. Suppose the bias is two-to-one in favor of Heads, i.e., it comes up Heads with probability $\frac{2}{3}$ and Tails with probability $\frac{1}{3}$. The sample space here is exactly the same as in the first example, but the probabilities are different: $\Pr[H] = \frac{2}{3}, \Pr[T] = \frac{1}{3}$. This is the first example in which the sample points have non-uniform probabilities.

4. Flip the biased coin in the previous example three times. The sample space is exactly the same as in the second example, but it is not immediately obvious how to assign probabilities to the sample points. This is because the bias of the coin only tells us how to assign probabilities to the outcome of *one* flip, not the outcome of *multiple* flips. It is, however, clear that the probabilities of the outcomes should not be uniform. We will return to this example after learning the important notion of *independence* in the next Note.

5. Roll two fair dice. Then $\Omega = \{(i, j) : 1 \leq i, j \leq 6\}$. Each of the 36 outcomes has equal probability, $\frac{1}{36}$. We can look at the event A that the sum of the dice is at least 10, and the event B that there is at least one 6. In this example (and in 1 and 2 above), our probability space is **uniform**, i.e., all the sample points have the *same* probability (which must be $\frac{1}{|\Omega|}$, where $|\Omega|$ denotes the size of Ω). In such circumstances, the probability of any event A is clearly just

$$\Pr[A] = \frac{\text{\# of sample points in } A}{\text{\# of sample points in } \Omega} = \frac{|A|}{|\Omega|}.$$

   So for uniform spaces, computing probabilities reduces to *counting* sample points! Using this observation, it is now easy to compute the probabilities of the two events A and B above: $\Pr[A] = \frac{6}{36} = \frac{1}{6}$, and $\Pr[B] = \frac{11}{36}$.

6. **Card Shuffling.** Shuffle a deck of cards. Here Ω consists of the 52! permutations of the deck, each with equal probability $\frac{1}{52!}$. [Note that we're really talking about an idealized mathematical model of shuffling here; in real life, there will always be a bit of bias in our shuffling. However, the mathematical model is close enough to be useful.]

7. **Poker Hands.** Shuffle a deck of cards, and then deal a poker hand. Here Ω consists of all possible five-card hands, each with equal probability (because the deck is assumed to be randomly shuffled). As we saw above, the number of such hands is $\binom{52}{5}$. What is the probability that our poker hand is a flush? [For those who are not addicted to gambling, a *flush* is a hand in which all cards have the same suit, say Hearts.] To compute this probability, we just need to figure out how many poker hands are flushes. Well, there are 13 cards in each suit, so the number of flushes in each suit is $\binom{13}{5}$. The total number of flushes is therefore $4 \cdot \binom{13}{5}$. So we have

$$\Pr[\text{hand is a flush}] = \frac{4 \cdot \binom{13}{5}}{\binom{52}{5}} = \frac{4 \cdot 13! \cdot 5! \cdot 47!}{5! \cdot 8! \cdot 52!} = \frac{4 \cdot 13 \cdot 12 \cdot 11 \cdot 10 \cdot 9}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48} \approx 0.002.$$

8. **Balls and Bins.** Throw 20 (labeled) balls into 10 (labeled) bins, so that each ball is equally likely to land in any bin, regardless of what happens to the other balls. (Thus, in the terminology of the previous Note, we are in the situation of "sampling with replacement" and order does matter.) Here $\Omega = \{(b_1, b_2, \ldots, b_{20}) : 1 \leq b_i \leq 10\}$; the component b_i denotes the bin in which ball i lands. There are 10^{20} possible outcomes (why?), each with probability $\frac{1}{10^{20}}$. More generally, if we throw m balls into n bins, we have a sample space of size n^m. [Note that example 2 above is the special case with $m = 3$ and $n = 2$, and example 4 is the special case $m = 2$, $n = 6$.] Let A be the event that bin 1 is empty. Again, we just need to count how many outcomes have this property. And this is exactly the number of ways all 20 balls can fall into the remaining nine bins, which is 9^{20}. Hence $\Pr[A] = \frac{9^{20}}{10^{20}} = (\frac{9}{10})^{20} \approx 0.12$. What is the probability that bin 1 contains at least one ball? This is easy: this event, call it $\bar{A}$, is the *complement* of A, i.e., it consists of precisely those sample points that are not in A. So $\Pr[\bar{A}] = 1 - \Pr[A] \approx 0.88$. More generally, if we throw m balls into n bins, we have

$$\Pr[\text{bin 1 is empty}] = \left(\frac{n-1}{n}\right)^m = \left(1 - \frac{1}{n}\right)^m.$$

As we shall see, balls and bins is another probability space that shows up very often in Computer Science: for example, we can think of it as modeling a load balancing scheme, in which each job is sent to a random processor.

Birthday Paradox

The "birthday paradox" is a remarkable phenomenon that examines the chances that two people in a group have the same birthday. It is a "paradox" not because of a logical contradiction, but because it goes against intuition. For ease of calculation, we take the number of days in a year to be 365. If we consider the case where there are n people in a room, then $|\Omega| = 365^n$. Let $A =$ "At least two people have the same birthday," and let $\bar{A} =$ "No two people have the same birthday." It is clear that $\Pr[A] = 1 - \Pr[\bar{A}]$. We will calculate $\Pr[\bar{A}]$, since it is easier, and then find out $\Pr[A]$. How many ways are there for no two people to have the same birthday? Well, there are 365 choices for the first person, 364 for the second, $\ldots$, $365 - n + 1$ choices for the n^{th} person, for a total of $365 \times 364 \times \cdots \times (365 - n + 1)$. (Note that this is just sampling without replacement with 365 bins and n balls; as we saw in the previous Note, the number of outcomes is $\frac{365!}{(365-n)!}$, which is what we just got.) Thus we have $\Pr[\bar{A}] = \frac{|\bar{A}|}{|\Omega|} = \frac{365 \times 364 \times \cdots \times (365-n+1)}{365^n}$. Then $\Pr[A] = 1 - \frac{365 \times 364 \times \cdots \times (365-n+1)}{365^n}$. This allows us to compute $\Pr[A]$ as a function of the number of people, n. Of course, as n increases $\Pr[A]$ increases. In fact, with $n = 23$ people you should be willing to bet that at least two people do have the same birthday, since then $\Pr[A]$ is larger than 50%! For $n = 60$ people, $\Pr[A]$ is over 99%.

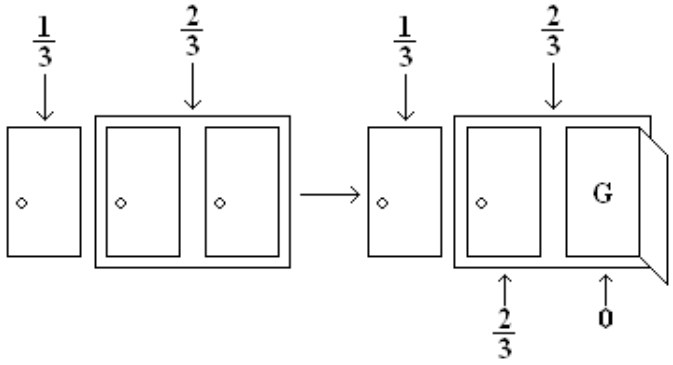
The Monty Hall Problem

In an (in)famous 1970s game show hosted by one Monty Hall, a contestant was shown three doors; behind one of the doors was a prize, and behind the other two were goats. The contestant picks a door (but doesn't open it). Then Hall's assistant (Carol), opens one of the other two doors, revealing a goat (since Carol knows where the prize is, she can always do this). The contestant is then given the option of sticking with his current door, or switching to the other unopened one. He wins the prize if and only if his chosen door is the correct one. The question, of course, is: Does the contestant have a better chance of winning if he switches doors?

Intuitively, it seems obvious that since there are only two remaining doors after the host opens one, they must have equal probability. So you may be tempted to jump to the conclusion that it should not matter

whether or not the contestant stays or switches. We will see that actually, the contestant has a better chance of picking the car if he or she uses the switching strategy. We will first give an intuitive pictorial argument, and then take a more rigorous probability approach to the problem.

To see why it is in the contestant's best interests to switch, consider the following. Initially when the contestant chooses the door, he or she has a $\frac{1}{3}$ chance of picking the car. This must mean that the other doors combined have a $\frac{2}{3}$ chance of winning. But after Carol opens a door with a goat behind it, how do the probabilities change? Well, the door the contestant originally chose still has a $\frac{1}{3}$ chance of winning, and the door that Carol opened has no chance of winning. What about the last door? It must have a $\frac{2}{3}$ chance of containing the car, and so the contestant has a higher chance of winning if he or she switches doors. This argument can be summed up nicely in the following picture:



What is the sample space here? Up to the point where the contestant makes his final decision, there are three random choices made: the game show host's choice of where to put the car, the contestant's initial choice of door, and Carol's choice of which door to open. We can therefore describe the outcome of the game using a triple of the form (i, j, k), where $i, j \in \{1,2,3\}$ and $k \in \{\text{Heads, Tails}\}$. The values i, j respectively specify the location of the prize and the initial door chosen by the contestant. The value k represents the result of a coin flip by Carol, with the interpretation that if there are two doors Carol can choose from, she picks the lower-numbered door if the coin flip is Heads and the higher-numbered door if it is Tails. (If Carol has no choice as to which door to open, then she will just ignore the coin flip and open that door.) Note that the sample space has exactly $3 \times 3 \times 2 = 18$ sample points.

Now the assignment of probabilities. A reasonable probability model is to think of all the outcomes as equally likely, so that the probability of each of the outcomes is $1/18$.

Let's return to the Monty Hall problem. Recall that we want to investigate the relative merits of the "sticking" strategy and the "switching" strategy. Let's suppose the contestant decides to switch doors. The event A we are interested in is the event that the contestant wins. Which sample points (i, j, k) are in A? Well, since the contestant is switching doors, the event A does not include any sample points where the contestant's initial choice j is equal to the prize door i. And all outcomes where $i \neq j$ correspond to a win for the contestant, because Carol must open the second non-prize door, leaving the contestant to switch to the prize door. So A consists of all outcomes (i, j, k) in which $i \neq j$. How many of these outcomes are there? Well, there are 6 pairs of (i, j) in which $i \neq j$, and for each such pair, k can be Heads or Tails. So there are a total of $6 \times 2 = 12$ outcomes in A. So $\Pr[A] = \frac{12}{18} = \frac{2}{3}$. That is, using the switching strategy, the contestant wins with probability $\frac{2}{3}$! It should be intuitively clear (and easy to check formally — try it!) that under the sticking strategy his probability of winning is $\frac{1}{3}$. (In this case, he is really just picking a single random door.) So by switching, the contestant actually improves his odds by a huge amount!

This is one of many examples that illustrate the importance of doing probability calculations systematically, rather than "intuitively." Recall the key steps in all our calculations:

- What is the **sample space** (i.e., the experiment and its set of possible outcomes)?

- What is the **probability** of each outcome (sample point)?
- What is the **event** we are interested in (i.e., which subset of the sample space)?
- Finally, compute the probability of the event by adding up the probabilities of the sample points inside it.

Whenever you meet a probability problem, you should always go back to these basics to avoid potential pitfalls. Even experienced researchers make mistakes when they forget to do this — witness many erroneous "proofs", submitted by mathematicians to newspapers at the time, of their (erroneous) claim that the switching strategy in the Monty Hall problem does not improve the odds.

Conditional Probability

A pharmaceutical company is marketing a new test for a certain medical disorder. According to clinical trials, the test has the following properties:

1. When applied to an affected person, the test comes up positive in 90% of cases, and negative in 10% (these are called "false negatives").
2. When applied to a healthy person, the test comes up negative in 80% of cases, and positive in 20% (these are called "false positives").

Suppose that the incidence of the disorder in the US population is 5%. When a random person is tested and the test comes up positive, what is the probability that the person actually has the disorder? (Note that this is presumably *not* the same as the simple probability that a random person has the disorder, which is just $\frac{1}{20}$.) The implicit probability space here is the entire US population with uniform probabilities.

This is an example of a *conditional probability*: we are interested in the probability that a person has the disorder (event A) *given that* he/she tests positive (event B). Let's write this as $\Pr[A|B]$.

How should we define $\Pr[A|B]$? Well, since event B is guaranteed to happen, we should look not at the whole sample space Ω, but at the smaller sample space consisting only of the sample points in B. What should the conditional probabilities of these sample points be? If they all simply inherit their probabilities from Ω, then the sum of these probabilities will be $\sum_{\omega \in B} \Pr[\omega] = \Pr[B]$, which in general is less than 1. So we should *normalize* the probability of each sample point by $\frac{1}{\Pr[B]}$. In other words, for each sample point $\omega \in B$, the new probability becomes

$$\Pr[\omega|B] = \frac{\Pr[\omega]}{\Pr[B]}.$$

Now it is clear how to define $\Pr[A|B]$: namely, we just sum up these normalized probabilities over all sample points that lie in both A and B:

$$\Pr[A|B] := \sum_{\omega \in A \cap B} \Pr[\omega|B] = \sum_{\omega \in A \cap B} \frac{\Pr[\omega]}{\Pr[B]} = \frac{\Pr[A \cap B]}{\Pr[B]}.$$

Definition 11.1 (conditional probability): For events A, B in the same probability space, such that $\Pr[B] > 0$, the *conditional probability of A given B* is

$$\Pr[A|B] := \frac{\Pr[A \cap B]}{\Pr[B]}.$$

Let's go back to our medical testing example. The sample space here consists of all people in the US. Let N denote the number of people in the US (so $N \approx 250$ million). The population consists of four disjoint subsets:

TP: the true positives (90% of $\frac{N}{20}$, i.e., $\frac{9N}{200}$ of them);

FP: the false positives (20% of $\frac{19N}{20}$, i.e., $\frac{19N}{100}$ of them);

TN: the true negatives (80% of $\frac{19N}{20}$, i.e., $\frac{76N}{100}$ of them);

FN: the false negatives (10% of $\frac{N}{20}$, i.e., $\frac{N}{200}$ of them).

We choose a person at random. Recall that A is the event that the person so chosen is affected, and B the event that he/she tests positive. Note that B is the union of the disjoint sets TP and FP, so

$$|B| = |TP| + |FP| = \tfrac{9N}{200} + \tfrac{19N}{100} = \tfrac{47N}{200}.$$

Thus we have

$$\Pr[A] = \tfrac{1}{20} \qquad \text{and} \qquad \Pr[B] = \tfrac{47}{200}.$$

Now when we condition on the event B, we focus in on the smaller sample space consisting only of those $\frac{47N}{200}$ individuals who test positive. To compute $\Pr[A|B]$, we need to figure out $\Pr[A \cap B]$ (the part of A that lies in B). But $A \cap B$ is just the set of people who are both affected and test positive, i.e., $A \cap B = TP$. So we have

$$\Pr[A \cap B] = \frac{|TP|}{N} = \frac{9}{200}.$$

Finally, we conclude from the definition of conditional probability that

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{9/200}{47/200} = \frac{9}{47} \approx 0.19.$$

This seems bad: if a person tests positive, there's only about a 19% chance that he/she actually has the disorder! This sounds worse than the original claims made by the pharmaceutical company, but in fact it's just another view of the same data.

[Incidentally, note that $\Pr[B|A] = \frac{9/200}{1/20} = \frac{9}{10}$; so $\Pr[A|B]$ and $\Pr[B|A]$ can be very different. Of course, $\Pr[B|A]$ is just the probability that a person tests positive given that he/she has the disorder, which we knew from the start was 90%.]

To complete the picture, what's the (unconditional) probability that the test gives a correct result (positive or negative) when applied to a random person? Call this event C. Then

$$\Pr[C] = \tfrac{|TP|+|TN|}{N} = \tfrac{9}{200} + \tfrac{76}{100} = \tfrac{161}{200} \approx 0.8.$$

So the test is about 80% effective overall, a more impressive statistic.

But how impressive is it? Suppose we ignore the test and just pronounce everybody to be healthy. Then we would be correct on 95% of the population (the healthy ones), and wrong on the affected 5%. In other words, this trivial test is 95% effective! So we might ask if it is worth running the test at all. What do you think?

Here are a couple more examples of conditional probabilities, based on some of our sample spaces from the previous lecture note.

1. **Balls and bins.** Suppose we toss $m = 3$ (labelled) balls into $n = 3$ bins; this is a uniform sample space with $3^3 = 27$ points. We already know that the probability the first bin is empty is $(1 - \frac{1}{3})^3 = (\frac{2}{3})^3 = \frac{8}{27}$. What is the probability of this event *given that* the second bin is empty? Call these events A, B

respectively. To compute $\Pr[A|B]$ we need to figure out $\Pr[A \cap B]$. But $A \cap B$ is the event that both the first two bins are empty, i.e., all three balls fall in the third bin. So $\Pr[A \cap B] = \frac{1}{27}$ (why?). Therefore,

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{1/27}{8/27} = \frac{1}{8}.$$

Not surprisingly, $\frac{1}{8}$ is quite a bit less than $\frac{8}{27}$: knowing that bin 2 is empty makes it significantly less likely that bin 1 will be empty.

2. **Dice.** Roll two fair dice. Let A be the event that their sum is even, and B the event that the first die is even. By symmetry it's easy to see that $\Pr[A] = \frac{1}{2}$ and $\Pr[B] = \frac{1}{2}$. Moreover, a little counting gives us that $\Pr[A \cap B] = \frac{1}{4}$. What is $\Pr[A|B]$? Well,

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{1/4}{1/2} = \frac{1}{2}.$$

In this case, $\Pr[A|B] = \Pr[A]$, i.e., conditioning on B does not change the probability of A.

Bayesian Inference

The medical test problem is a canonical example of an *inference* problem: given a noisy observation (the result of the test), we want to figure out the likelihood of something not directly observable (whether a person is healthy). To bring out the common structure of such inference problems, let us redo the calculations in the medical test example but only in terms of events without explicitly mentioning the sample points of the underlying sample space.

Recall: A is the event the person is affected, B is the event that the test is positive. What are we given?

- $\Pr[A] = 0.05$, (5% of the U.S. population is affected)
- $\Pr[B|A] = 0.9$ (90% of the affected people test positive)
- $\Pr[B|\overline{A}] = 0.2$ (20% of healthy people test positive)

We want to calculate $\Pr[A|B]$. We can proceed as follows:

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{\Pr[B|A]\Pr[A]}{\Pr[B]} \tag{1}$$

and

$$\Pr[B] = \Pr[A \cap B] + \Pr[\overline{A} \cap B] = \Pr[B|A]\Pr[A] + \Pr[B|\overline{A}](1 - \Pr[A]) \tag{2}$$

Combining equations (1) and (2), we have expressed $\Pr[A|B]$ in terms of $\Pr[A], \Pr[B|A]$ and $\Pr[B|\overline{A}]$:

$$\Pr[A|B] = \frac{\Pr[B|A]\Pr[A]}{\Pr[B|A]\Pr[A] + \Pr[B|\overline{A}](1 - \Pr[A])} \tag{3}$$

This equation is useful for many inference problems. We are given $\Pr[A]$, which is the (unconditional) probability that the event of interest A happens. We are given $\Pr[B|A]$ and $\Pr[B|\overline{A}]$, which quantify how noisy the observation is. (If $\Pr[B|A] = 1$ and $\Pr[B|\overline{A}] = 0$, for example, the observation is completely noiseless.)

Now we want to calculate $\Pr[A|B]$, the probability that the event of interest happens given we made the observation. Equation (3) allows us to do just that.

Equation (3) is at the heart of a subject called *Bayesian inference*, used extensively in fields such as machine learning, communications and signal processing. The equation can be interpreted as a way to *update knowledge* after making an observation. In this interpretation, $\Pr[A]$ can be thought of as a *prior* probability: our assessment of the likelihood of an event of interest A *before* making an observation. It reflects our prior knowledge. $\Pr[A|B]$ can be interpreted as the *posterior* probability of A after the observation. It reflects our new knowledge.

Of course, equations (1), (2) and (3) are derived from the basic axioms of probability and the definition of conditional probability, and are therefore true with or without the above Bayesian inference interpretation. However, this interpretation is very useful when we apply probability theory to study inference problems.

Bayes' Rule and Total Probability Rule

Equations (1) and (2) are very useful in their own right. The first is called **Bayes' Rule** and the second is called the **Total Probability Rule**. Bayes' Rule is useful when one wants to calculate $\Pr[A|B]$ but one is given $\Pr[B|A]$ instead, i.e. it allows us to "flip" things around. The Total Probability Rule is an application of the strategy of "dividing into cases" we learned in Note 2 to calculating probabilities. We want to calculate the probability of an event B. There are two possibilities: either an event A happens or A does not happen. If A happens the probability that B happens is $\Pr[B|A]$. If A does not happen, the probability that B happens is $\Pr[B|\overline{A}]$. If we know or can easily calculate these two probabilities and also $\Pr[A]$, then the total probability rule yields the probability of event B.

Independent events

Definition 11.2 (independence): Two events A, B in the same probability space are *independent* if $\Pr[A \cap B] = \Pr[A] \times \Pr[B]$.

The intuition behind this definition is the following. Suppose that $\Pr[B] > 0$ and A, B are independent. Then we have

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{\Pr[A] \times \Pr[B]}{\Pr[B]} = \Pr[A].$$

Thus independence has the natural meaning that "the probability of A is not affected by whether or not B occurs." (By a symmetrical argument, we also have $\Pr[B|A] = \Pr[B]$ provided $\Pr[A] > 0$.) For events A, B such that $\Pr[B] > 0$, the condition $\Pr[A|B] = \Pr[A]$ is actually *equivalent* to the definition of independence.

Examples: In the balls and bins example above, events A, B are *not* independent. In the dice example, events A, B *are* independent.

The above definition generalizes to any finite set of events:

Definition 11.3 (mutual independence): Events $A_1, \ldots, A_n$ are *mutually independent* if for every subset $I \subseteq \{1, \ldots, n\}$,

$$\Pr[\textstyle\bigcap_{i \in I} A_i] = \prod_{i \in I} \Pr[A_i].$$

Note that we need this property to hold for *every* subset I.

For mutually independent events $A_1, \ldots, A_n$, it is not hard to check from the definition of conditional probability that, for any $1 \leq i \leq n$ and any subset $I \subseteq \{1, \ldots, n\} \setminus \{i\}$, we have

$$\Pr[A_i | \textstyle\bigcap_{j \in I} A_j] = \Pr[A_i].$$

Note that the independence of every pair of events (so-called *pairwise independence*) does *not* necessarily imply mutual independence. For example, it is possible to construct three events A, B, C such that each *pair* is independent but the triple A, B, C is *not* mutually independent.

Combinations of events

In most applications of probability in Computer Science, we are interested in things like $\Pr[\bigcup_{i=1}^n A_i]$ and $\Pr[\bigcap_{i=1}^n A_i]$, where the A_i are simple events (i.e., we know, or can easily compute, the $\Pr[A_i]$). The intersection $\bigcap_i A_i$ corresponds to the logical AND of the events A_i, while the union $\bigcup_i A_i$ corresponds to their logical OR. As an example, if A_i denotes the event that a failure of type i happens in a certain system, then $\bigcup_i A_i$ is the event that the system fails.

In general, computing the probabilities of such combinations can be very difficult. In this section, we discuss some situations where it can be done.

Intersections of events

From the definition of conditional probability, we immediately have the following *product rule* (sometimes also called the *chain rule*) for computing the probability of an intersection of events.

Theorem 11.1: **[Product Rule]** For any events A, B, we have

$$\Pr[A \cap B] = \Pr[A] \Pr[B|A].$$

More generally, for any events $A_1, \ldots, A_n$,

$$\Pr[\textstyle\bigcap_{i=1}^n A_i] = \Pr[A_1] \times \Pr[A_2|A_1] \times \Pr[A_3|A_1 \cap A_2] \times \cdots \times \Pr[A_n|\textstyle\bigcap_{i=1}^{n-1} A_i].$$

Proof: The first assertion follows directly from the definition of $\Pr[B|A]$ (and is in fact a special case of the second assertion with $n = 2$).

To prove the second assertion, we will use induction on n (the number of events). The base case is $n = 1$, and corresponds to the statement that $\Pr[A] = \Pr[A]$, which is trivially true. For the inductive step, let $n > 1$ and assume (the inductive hypothesis) that

$$\Pr[\textstyle\bigcap_{i=1}^{n-1} A_i] = \Pr[A_1] \times \Pr[A_2|A_1] \times \cdots \times \Pr[A_{n-1}|\textstyle\bigcap_{i=1}^{n-2} A_i].$$

Now we can apply the definition of conditional probability to the two events A_n and $\bigcap_{i=1}^{n-1} A_i$ to deduce that

$$\begin{aligned}
\Pr[\textstyle\bigcap_{i=1}^n A_i] &= \Pr[A_n \cap (\textstyle\bigcap_{i=1}^{n-1} A_i)] \\
&= \Pr[A_n|\textstyle\bigcap_{i=1}^{n-1} A_i] \times \Pr[\textstyle\bigcap_{i=1}^{n-1} A_i] \\
&= \Pr[A_n|\textstyle\bigcap_{i=1}^{n-1} A_i] \times \Pr[A_1] \times \Pr[A_2|A_1] \times \cdots \times \Pr[A_{n-1}|\textstyle\bigcap_{i=1}^{n-2} A_i],
\end{aligned}$$

where in the last line we have used the inductive hypothesis. This completes the proof by induction. □

The product rule is particularly useful when we can view our sample space as a sequence of choices. The next few examples illustrate this point.

1. **Coin tosses.** Toss a fair coin three times. Let A be the event that all three tosses are heads. Then $A = A_1 \cap A_2 \cap A_3$, where A_i is the event that the ith toss comes up heads. We have

$$\begin{aligned}
\Pr[A] &= \Pr[A_1] \times \Pr[A_2|A_1] \times \Pr[A_3|A_1 \cap A_2] \\
&= \Pr[A_1] \times \Pr[A_2] \times \Pr[A_3] \\
&= \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} = \tfrac{1}{8}.
\end{aligned}$$

The second line here follows from the fact that the tosses are mutually independent. Of course, we already know that $\Pr[A] = \frac{1}{8}$ from our definition of the probability space in the previous lecture note. The above is really a check that the space behaves as we expect.[1]

It seems reasonable that the tosses should remain mutually independent, even if the coin is biased, since no coin toss is affected by any of the other tosses. If the coin is biased with heads probability p, this independence assumption implies

$$\Pr[A] = \Pr[A_1] \times \Pr[A_2] \times \Pr[A_3] = p^3.$$

As another example, let B denote the event that the first coin toss comes up tails and the next two coin tosses come up heads. Then $B = \overline{A_1} \cap A_2 \cap A_3$, and these events are independent, so

$$\Pr[B] = \Pr[\overline{A_1}] \times \Pr[A_2] \times \Pr[A_3] = p^2(1-p),$$

since $\Pr[\overline{A_1}] = 1 - \Pr[A_1] = 1 - p$. More generally, the probability of any sequence of n tosses containing r heads and $n-r$ tails is $p^r(1-p)^{n-r}$. The notion of independence is the key concept that enables us to assign probabilities to these outcomes.

2. **Balls and bins.** Let A be the event that bin 1 is empty. We saw in the previous lecture note (by counting) that $\Pr[A] = (1-\frac{1}{n})^m$, where m is the number of balls and n is the number of bins. The product rule gives us a different way to compute the same probability. We can write $A = \bigcap_{i=1}^m A_i$, where A_i is the event that ball i misses bin 1. Clearly $\Pr[A_i] = 1 - \frac{1}{n}$ for each i. Also, the A_i are mutually independent since ball i chooses its bin regardless of the choices made by any of the other balls. So
$$\Pr[A] = \Pr[A_1] \times \cdots \times \Pr[A_m] = \left(1 - \frac{1}{n}\right)^m.$$

3. **Card shuffling.** We can look at the sample space as a sequence of choices as follows. First the top card is chosen uniformly from all 52 cards, i.e., each card with probability $\frac{1}{52}$. Then (conditional on the first card), the second card is chosen uniformly from the remaining 51 cards, each with probability $\frac{1}{51}$. Then (conditional on the first two cards), the third card is chosen uniformly from the remaining 50, and so on. The probability of any given permutation, by the product rule, is therefore
$$\frac{1}{52} \times \frac{1}{51} \times \frac{1}{50} \times \cdots \times \frac{1}{2} \times \frac{1}{1} = \frac{1}{52!}.$$
Reassuringly, this is in agreement with our definition of the probability space in the previous lecture note, based on counting permutations.

4. **Poker hands.** Again we can view the sample space as a sequence of choices. First we choose one of the cards (note that it is not the "first" card, since the cards in our hand have no ordering) uniformly from all 52 cards. Then we choose another card from the remaining 51, and so on. For any given poker hand, the probability of choosing it is (by the product rule):
$$\frac{5}{52} \times \frac{4}{51} \times \frac{3}{50} \times \frac{2}{49} \times \frac{1}{48} = \frac{1}{\binom{52}{5}},$$
just as before. Where do the numerators 5, 4, 3, 2, 1 come from? Well, for the given hand the first card we choose can be any of the five in the hand: i.e., five choices out of 52. The second can be any

[1] Strictly speaking, we should really also have checked from our original definition of the probability space that $\Pr[A_1]$, $\Pr[A_2|A_1]$ and $\Pr[A_3|A_1 \cap A_2]$ are all equal to $\frac{1}{2}$.

of the remaining four in the hand: four choices out of 51. And so on. This arises because the order of the cards in the hand is irrelevant.

Let's use this view to compute the probability of a flush in a different way. Clearly this is $4 \times \Pr[A]$, where A is the probability of a Hearts flush. And we can write $A = \bigcap_{i=1}^{5} A_i$, where A_i is the event that the ith card we pick is a Heart. So we have

$$\Pr[A] = \Pr[A_1] \times \Pr[A_2|A_1] \times \cdots \times \Pr[A_5|\bigcap_{i=1}^{4} A_i].$$

Clearly $\Pr[A_1] = \frac{13}{52} = \frac{1}{4}$. What about $\Pr[A_2|A_1]$? Well, since we are conditioning on A_1 (the first card is a Heart), there are only 51 remaining possibilities for the second card, 12 of which are Hearts. So $\Pr[A_2|A_1] = \frac{12}{51}$. Similarly, $\Pr[A_3|A_1 \cap A_2] = \frac{11}{50}$, and so on. So we get

$$4 \times \Pr[A] = 4 \times \frac{13}{52} \times \frac{12}{51} \times \frac{11}{50} \times \frac{10}{49} \times \frac{9}{48},$$

which is exactly the same fraction we computed in the previous lecture note.

So now we have two methods of computing probabilities in many of our sample spaces. It is useful to keep these different methods around, both as a check on your answers and because in some cases one of the methods is easier to use than the other.

Unions of events

You are in Las Vegas, and you spy a new game with the following rules. You pick a number between 1 and 6. Then three dice are thrown. You win if and only if your number comes up on at least one of the dice.

The casino claims that your odds of winning are 50%, using the following argument. Let A be the event that you win. We can write $A = A_1 \cup A_2 \cup A_3$, where A_i is the event that your number comes up on die i. Clearly $\Pr[A_i] = \frac{1}{6}$ for each i. Therefore,

$$\Pr[A] = \Pr[A_1 \cup A_2 \cup A_3] = \Pr[A_1] + \Pr[A_2] + \Pr[A_3] = 3 \times \frac{1}{6} = \frac{1}{2}.$$

Is this calculation correct? Well, suppose instead that the casino rolled six dice, and again you win if and only if your number comes up at least once. Then the analogous calculation would say that you win with probability $6 \times \frac{1}{6} = 1$, i.e., certainly! The situation becomes even more ridiculous when the number of dice gets bigger than 6.

The problem is that the events A_i are *not disjoint*: there are some sample points that lie in more than one of the A_i. (We could get really lucky and our number could come up on two of the dice, or all three.) So if we add up the $\Pr[A_i]$ we are counting some sample points more than once.

Fortunately, there is a formula for this, known as the *Principle of Inclusion/Exclusion*:

Theorem 11.2: **[Inclusion/Exclusion]** For events $A_1, \ldots, A_n$ in some probability space, we have

$$\Pr[\bigcup_{i=1}^{n} A_i] = \sum_{i=1}^{n} \Pr[A_i] - \sum_{\{i,j\}} \Pr[A_i \cap A_j] + \sum_{\{i,j,k\}} \Pr[A_i \cap A_j \cap A_k] - \cdots \pm \Pr[\bigcap_{i=1}^{n} A_i].$$

[In the above summations, $\{i, j\}$ denotes all unordered pairs with $i \neq j$, $\{i, j, k\}$ denotes all unordered triples of distinct elements, and so on.]

In other words, to compute $\Pr[\bigcup_i A_i]$, we start by summing the event probabilities $\Pr[A_i]$, then we *subtract* the probabilities of all pairwise intersections, then we *add* back in the probabilities of all three-way intersections, and so on.

We won't prove this formula here; but you might like to verify it for the special case $n = 3$ by drawing a Venn diagram and checking that every sample point in $A_1 \cup A_2 \cup A_3$ is counted exactly once by the formula. You might also like to prove the formula for general n by induction (in similar fashion to the proof of the Product Rule above).

Taking the formula on faith, what is the probability we get lucky in the new game in Vegas?

$$\Pr[A_1 \cup A_2 \cup A_3] = \Pr[A_1] + \Pr[A_2] + \Pr[A_3] - \Pr[A_1 \cap A_2] - \Pr[A_1 \cap A_3] - \Pr[A_2 \cap A_3] + \Pr[A_1 \cap A_2 \cap A_3].$$

Now the nice thing here is that the events A_i are mutually independent (the outcome of any die does not depend on that of the others), so $\Pr[A_i \cap A_j] = \Pr[A_i]\Pr[A_j] = (\frac{1}{6})^2 = \frac{1}{36}$, and similarly $\Pr[A_1 \cap A_2 \cap A_3] = (\frac{1}{6})^3 = \frac{1}{216}$. So we get

$$\Pr[A_1 \cup A_2 \cup A_3] = \left(3 \times \tfrac{1}{6}\right) - \left(3 \times \tfrac{1}{36}\right) + \tfrac{1}{216} = \tfrac{91}{216} \approx 0.42.$$

So your odds are quite a bit worse than the casino is claiming!

[Actually, for this example there is an easier way to compute this probability, by looking at the complement event. Let W denote the event that you win, and $\overline{W}$ the complement event, i.e., $\overline{W}$ is the event that you lose. Then

$$\Pr[W] = 1 - \Pr[\overline{W}] = 1 - \Pr[\overline{A_1} \cap \overline{A_2} \cap \overline{A_3}] = 1 - \Pr[\overline{A_1}] \times \Pr[\overline{A_2}] \times \Pr[\overline{A_3}] = 1 - \tfrac{5}{6} \times \tfrac{5}{6} \times \tfrac{5}{6} = \tfrac{91}{216}.$$

But this complement trick does not always work, and then it is useful to know about the inclusion/exclusion formula.]

When n is large (i.e., we are interested in the union of many events), the Inclusion/Exclusion formula is essentially useless because it involves computing the probability of the intersection of every non-empty subset of the events, and there are $2^n - 1$ of these! Sometimes we can just look at the first few terms of it and forget the rest: note that successive terms actually give us an overestimate and then an underestimate of the answer, and these estimates both get better as we go along.

However, in many situations we can get a long way by just looking at the *first* term:

1. **Disjoint events.** If the events A_i are all *disjoint* (i.e., no pair of them contain a common sample point — such events are also called *mutually exclusive*), then

   $$\Pr[\textstyle\bigcup_{i=1}^n A_i] = \sum_{i=1}^n \Pr[A_i].$$

   [Note that we have already used this fact several times in our examples, e.g., in claiming that the probability of a flush is four times the probability of a Hearts flush — clearly flushes in different suits are disjoint events.]

2. **Union bound.** Always, it is the case that

   $$\Pr[\textstyle\bigcup_{i=1}^n A_i] \leq \sum_{i=1}^n \Pr[A_i].$$

   This merely says that adding up the $\Pr[A_i]$ can only *over*estimate the probability of the union. Crude as it may seem, in the next lecture note we'll see how to use the union bound effectively in a Computer Science example.

A Killer Application

In this lecture, we will see a "killer app" of elementary probability in Computer Science. Suppose a hash function distributes keys evenly over a table of size n. How many (randomly chosen) keys can we hash before the probability of a collision exceeds (say) $\frac{1}{2}$? As we shall see, this question can be tackled by an analysis of the balls-and-bins probability space which we have already encountered.

Application: Hash functions

As you may already know, a hash table is a data structure that supports the storage of sets of keys from a (large) universe U (say, the names of all 250 million people in the US). The operations supported are ADDing a key to the set, DELETEing a key from the set, and testing MEMBERship of a key in the set. The hash function h maps U to a table T of modest size. To ADD a key x to our set, we evaluate $h(x)$ (i.e., apply the hash function to the key) and store x at the location $h(x)$ in the table T. All keys in our set that are mapped to the same table location are stored in a simple linked list. The operations DELETE and MEMBER are implemented in similar fashion, by evaluating $h(x)$ and searching the linked list at $h(x)$. Note that the efficiency of a hash function depends on having only few *collisions*—i.e., keys that map to the same location. This is because the search time for DELETE and MEMBER operations is proportional to the length of the corresponding linked list.

The question we are interested in here is the following: suppose our hash table T has size n, and that our hash function h distributes U evenly over T.[1] Assume that the keys we want to store are chosen uniformly at random and independently from the universe U. What is the largest number, m, of keys we can store before the probability of a collision reaches $\frac{1}{2}$?

Let's begin by seeing how this problem can be put into the balls and bins framework. The balls will be the m keys to be stored, and the bins will be the n locations in the hash table T. Since the keys are chosen uniformly and independently from U, and since the hash function distributes keys evenly over the table, we can see each key (ball) as choosing a hash table location (bin) uniformly and independently from T. Thus the probability space corresponding to this hashing experiment is exactly the same as the balls and bins space.

We are interested in the event A that there is no collision, or equivalently, that all m balls land in different bins. Clearly $\Pr[A]$ will decrease as m increases (with n fixed). Our goal is to find the largest value of m such that $\Pr[A]$ remains above $\frac{1}{2}$.

Let's fix the value of m and try to compute $\Pr[A]$. Since our probability space is uniform (each outcome has probability $\frac{1}{n^m}$), it's enough just to count the number of outcomes in A. In how many ways can we arrange m balls in n bins so that no bin contains more than one ball? Well, there are n places to put the first ball, then $n-1$ remaining places for the second ball (since it cannot go in the same bin as the first), $n-2$ places for

[1] In other words, $|U| = \alpha n$ (the size of U is an integer multiple α of the size of T), and for each $y \in T$, the number of keys $x \in U$ for which $h(x) = y$ is exactly α.

the third ball, and so on. Thus the total number of such arrangements is

$$n \times (n-1) \times (n-2) \times \cdots \times (n-m+2) \times (n-m+1).$$

This formula is valid as long as $m \leq n$: if $m > n$ then clearly the answer is zero. From now on, we'll assume that $m \leq n$.

Now we can calculate the probability of no collision:

$$\begin{aligned}
\Pr[A] &= \frac{n(n-1)(n-2)\dots(n-m+1)}{n^m} \\
&= \frac{n}{n} \times \frac{n-1}{n} \times \frac{n-2}{n} \times \cdots \times \frac{n-m+1}{n} \\
&= \left(1-\frac{1}{n}\right) \times \left(1-\frac{2}{n}\right) \times \cdots \times \left(1-\frac{m-1}{n}\right). \qquad (1)
\end{aligned}$$

Before going on, let's pause to observe that we could compute $\Pr[A]$ in a different way, as follows. View the probability space as a sequence of choices, one for each ball. For $1 \leq i \leq m$, let A_i be the event that the ith ball lands in a different bin from balls $1, 2, \dots, i-1$. Then

$$\begin{aligned}
\Pr[A] = \Pr[\textstyle\bigcap_{i=1}^{n} A_i] &= \Pr[A_1] \times \Pr[A_2|A_1] \times \Pr[A_3|A_1 \cap A_2] \times \cdots \times \Pr[A_m|\textstyle\bigcap_{i=1}^{m-1} A_i] \\
&= 1 \times \frac{n-1}{n} \times \frac{n-2}{n} \times \cdots \times \frac{n-m+1}{n} \\
&= \left(1-\frac{1}{n}\right) \times \left(1-\frac{2}{n}\right) \times \cdots \times \left(1-\frac{m-1}{n}\right).
\end{aligned}$$

Fortunately, we get the same answer as before! [You should make sure you see how we obtained the conditional probabilities in the second line above. For example, $\Pr[A_3|A_1 \cap A_2]$ is the probability that the third ball lands in a different bin from the first two balls, *given that* those two balls also landed in different bins. This means that the third ball has $n-2$ possible bin choices out of a total of n.]

Essentially, we are now done with our problem: equation (1) gives an exact formula for the probability of no collision when m keys are hashed. All we need to do now is plug values $m = 1, 2, 3, \dots$ into (1) until we find that $\Pr[A]$ drops below $\frac{1}{2}$. The corresponding value of m (minus 1) is what we want.

But this is not really satisfactory: it would be much more useful to have a formula that gives the "critical" value of m directly, rather than having to compute $\Pr[A]$ for $m = 1, 2, 3, \dots$. Note that we would have to do this computation separately for each different value of n we are interested in: i.e., whenever we change the size of our hash table.

So what remains is to "turn equation (1) around", so that it tells us the value of m at which $\Pr[A]$ drops below $\frac{1}{2}$. To do this, let's take logs: this is a good thing to do because it turns the product into a sum, which is easier to handle. We get

$$\ln(\Pr[A]) = \ln\left(1-\frac{1}{n}\right) + \ln\left(1-\frac{2}{n}\right) + \cdots + \ln\left(1-\frac{m-1}{n}\right), \qquad (2)$$

where "ln" denotes natural (base e) logarithm. Now we can make use of a standard approximation for logarithms: namely, if x is small then $\ln(1-x) \approx -x$. This comes from the Taylor series expansion

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots.$$

So by replacing $\ln(1-x)$ by $-x$ we are making an error of at most $\frac{x^2}{2} + \frac{x^3}{3} + \cdots$, which is at most $2x^2$ when $x \leq \frac{1}{2}$. In other words, we have

$$-x - 2x^2 \leq \ln(1-x) \leq -x.$$

And if x is small then the error term $2x^2$ will be much smaller than the main term $-x$. Rather than carry around the error term $2x^2$ everywhere, in what follows we'll just write $\ln(1-x) \approx -x$, secure in the knowledge that we could make this approximation precise if necessary.

Now let's plug this approximation into equation (2):

$$\begin{aligned}\ln(\Pr[A]) &\approx -\frac{1}{n} - \frac{2}{n} - \frac{3}{n} - \cdots - \frac{m-1}{n} \\ &= -\frac{1}{n}\sum_{i=1}^{m-1} i \\ &= -\frac{m(m-1)}{2n} \\ &\approx -\frac{m^2}{2n}. \end{aligned} \tag{3}$$

Note that we've used the approximation for $\ln(1-x)$ with $x = \frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \ldots, \frac{m-1}{n}$. So our approximation should be good provided all these are small, i.e., provided n is fairly big and m is quite a bit smaller than n. Once we're done, we'll see that the approximation is actually pretty good even for modest sizes of n.

Now we can undo the logs in (3) to get our expression for $\Pr[A]$:

$$\Pr[A] \approx \mathrm{e}^{-\frac{m^2}{2n}}.$$

The final step is to figure out for what value of m this probability becomes $\frac{1}{2}$. So we want the largest m such that $\mathrm{e}^{-\frac{m^2}{2n}} \geq \frac{1}{2}$. This means we must have

$$-\frac{m^2}{2n} \geq \ln(\tfrac{1}{2}) = -\ln 2, \tag{4}$$

or equivalently

$$m \leq \sqrt{(2\ln 2)n} \approx 1.177\sqrt{n}.$$

So the bottom line is that we can hash approximately $\boxed{m = \lfloor 1.177\sqrt{n} \rfloor \text{ keys}}$ before the probability of a collision reaches $\frac{1}{2}$.

Recall that our calculation was only approximate; so we should go back and get a feel for how much error we made. We can do this by using equation (1) to compute the exact value $m = m_0$ at which $\Pr[A]$ drops below $\frac{1}{2}$, for a few sample values of n. Then we can compare these values with our estimate $m = 1.177\sqrt{n}$.

n	10	20	50	100	200	365	500	1000	10^4	10^5	10^6
$1.177\sqrt{n}$	3.7	5.3	8.3	11.8	16.6	22.5	26.3	37.2	118	372	1177
exact m_0	4	5	8	12	16	22	26	37	118	372	1177

From the table, we see that our approximation is very good even for small values of n. When n is large, the error in the approximation becomes negligible.

Why $\frac{1}{2}$?

Our hashing question asked when the probability of a collision rises to $\frac{1}{2}$. Is there anything special about $\frac{1}{2}$? Not at all. What we did was to (approximately) compute $\Pr[A]$ (the probability of no collision) as a function of m, and then find the largest value of m for which our estimate is smaller than $\frac{1}{2}$. If instead we were

interested in keeping the collision probability below (say) 0.05 (= 5%), we would just replace $\frac{1}{2}$ by 0.95 in equation (4). If you work through the last piece of algebra again, you'll see that this gives us the critical value $m = \sqrt{(2\ln(20/19))n} \approx 0.32\sqrt{n}$, which of course is a bit smaller than before because our desired collision probability is now smaller. But no matter what desired probability we specify, the critical value of m will always be $c\sqrt{n}$ for some constant c (which depends on the desired probability).

The birthday paradox revisited

Recall from a previous lecture the birthday "paradox": what is the probability that, in a group of m people, no two people have the same birthday? The problem we have solved above is essentially just a generalization of the birthday problem: the bins are the birthdays and the balls are the people, and we want the probability that there is no collision. The above table at $n = 365$ tells us for what value of m this probability drops below $\frac{1}{2}$: namely, 23.

The coupon collector's problem

Suppose that when we buy a box of cereal, as a marketing ploy, the cereal manufacturer has included a random baseball card. Suppose that there are n baseball players who appear on some card, and each cereal box contains a card chosen uniformly and independently at random from these n possibilities.

Assume that Joe DiMaggio is one of the n baseball players, and I am a huge fan of him: I really want his baseball card. How many boxes of cereal will I have to buy, to have at least a 50% chance of obtaining a Joe DiMaggio card? This problem can be analyzed as follows. Suppose we buy m boxes of cereal. Let E_{JD} be the event that I do not receive a Joe DiMaggio card in any of the m boxes. The card in each of the m boxes is independent, and for each box the chances that I don't receive Joe DiMaggio's card from that box is $1 - \frac{1}{n}$. Therefore,

$$\Pr[E_{\text{JD}}] = \left(1 - \frac{1}{n}\right)^m.$$

Note that the Taylor's series for e^{-x} is

$$\mathrm{e}^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots,$$

so $1 - x \leq \mathrm{e}^{-x}$ (if $x \leq 1$), and $1 - x \approx \mathrm{e}^{-x}$ is a good approximation if x is small. Plugging in this approximation above, we find

$$\Pr[E_{\text{JD}}] \approx \left(\mathrm{e}^{-1/n}\right)^m = \mathrm{e}^{-m/n}.$$

(In fact, this is an upper bound: $\Pr[E_{\text{JD}}] \leq \mathrm{e}^{-m/n}$.) So if we want this probability to be about $1/2$, we set

$$\mathrm{e}^{-m/n} = \tfrac{1}{2}$$

and solve for m, obtaining

$$m = n\ln 2 \approx 0.69n.$$

In other words, if we buy $m = 0.69n$ cereal boxes, we expect to have about a $\frac{1}{2}$ probability of finding a Joe DiMaggio card. (In fact, we are guaranteed at least a $\frac{1}{2}$ probability, possibly higher.)

Next suppose that what I really want is a complete collection: I want at least one of each of the n cards. How many boxes of cereal will I have to buy, to have at least a 50% chance of obtaining a complete collection? We will analyze this as follows. Let E denote the event that I do not receive a complete collection, after

buying m boxes of cereal. Let E_i denote the event that I don't receive any card containing the ith baseball player, after buying those m boxes of cereal. Note that

$$E = E_1 \cup E_2 \cup \cdots \cup E_n.$$

Moreover, our calculation above shows that

$$\Pr[E_i] \leq \mathrm{e}^{-m/n}.$$

Let's make this $= \frac{1}{2n}$, for reasons to be explained in a moment. Setting $\mathrm{e}^{-m/n} = \frac{1}{2n}$ and solving for m, we obtain $m = n\ln(2n)$. In other words, if we buy $m = n\ln(2n)$ cereal boxes, then the probability of missing player i is at most $\frac{1}{2n}$, i.e., $\Pr[E_i] \leq \frac{1}{2n}$ for all $i = 1, 2, \ldots, n$. Now we can apply a "Union Bound", to calculate an upper bound on $\Pr[E]$:

$$\begin{aligned}
\Pr[E] &= \Pr[E_1 \cup E_2 \cup \cdots \cup E_n] \\
&\leq \Pr[E_1] + \Pr[E_2] + \cdots + \Pr[E_n] && \text{(by the Union Bound)} \\
&\leq \frac{1}{2n} + \frac{1}{2n} + \cdots + \frac{1}{2n} && \text{(by above)} \\
&\leq \frac{1}{2}.
\end{aligned}$$

In other words, if we buy $m = n\ln(2n)$ cereal boxes, then we have at least a $\frac{1}{2}$ probability of obtaining a complete collection of all n baseball players. Interestingly, we need considerably more than n cereal boxes to assemble a complete collection: a good incentive to buy lots of cereal, and a great marketing ploy for cereal manufacturers.

This has some applications in analysis of network protocols. For instance, here is a simplified description of the BitTorrent peer-to-peer file-sharing protocol. When someone uploads a file to BitTorrent, it breaks up the file into many chunks, randomly selects many peers in the network, and sends each peer one chunk of the file. When another BitTorrent client wants to download the file, it queries its peers to find all peers that have a copy of some chunk from the file, and repeatedly asks a random such peer to return its chunk, until the client has all of the chunks of the file. Thus, in each iteration the client obtains another random chunk of the file. If the file consists of n chunks, and in each iteration the client obtains a random chunk, then we can expect that downloaders will have to wait a long time to obtain the whole file: they'll have to perform about $n\ln(2n)$ iterations before they have a decent chance of collecting all the chunks. This is inefficient, so in practice BitTorrent is forced to go out of its way to include special mechanisms to speed up downloading: BitTorrent clients download chunks randomly until they have a large fraction of the file, then they switch to searching for the specific chunks they are missing. [There are other solutions. You might enjoy pondering how to use error-correcting codes to avoid this inefficiency.]

Perhaps the most interesting aspect of the coupon collector's problem above is that it illustrates the use of the union bound to upper-bound the probability that something bad happens. In this case, the bad event is that we fail to obtain a complete collection of baseball cards, but in general, this methodology is a powerful way to prove that some bad event is not too likely to happen.

Random Variables: Distribution and Expectation

Random Variables

Question: The homeworks of 20 students are collected in, randomly shuffled and returned to the students. How many students receive their own homework?

To answer this question, we first need to specify the probability space: plainly, it should consist of all 20! permutations of the homeworks, each with probability $\frac{1}{20!}$. [Note that this is the same as the probability space for card shuffling, except that the number of items being shuffled is now 20 rather than 52.] It helps to have a picture of a permutation. Think of 20 books lined up on a shelf, labeled from left to right with $1, 2, \ldots, 20$. A permutation π is just a reordering of the books, which we can describe just by listing their labels from left to right. Let's denote by π_i the label of the book that is in position i. We are interested in the number of books that are still in their original position, i.e., in the number of i's such that $\pi_i = i$. These are often known as *fixed points* of the permutation.

Of course, our question does not have a simple numerical answer (such as 6), because the number depends on the particular permutation we choose (i.e., on the sample point). Let's call the number of fixed points X. To make life simpler, let's also shrink the class size down to 3 for a while. The following table gives a complete listing of the sample space (of size $3! = 6$), together with the corresponding value of X for each sample point. [We use our bookshelf convention for writing a permutation: thus, for example, the permutation 312 means that book 3 is on the left, book 1 in the center, and book 2 on the right. You should check you agree with this table.]

permutation π	value of X
123	3
132	1
213	1
231	0
312	0
321	1

Thus we see that X takes on values 0, 1 or 3, depending on the sample point. A quantity like this, which takes on some numerical value at each sample point, is called a *random variable* (or *r.v.*) on the sample space.

Definition 13.1 (random variable): A *random variable* X on a sample space Ω is a function that assigns to each sample point $\omega \in \Omega$ a real number $X(\omega)$.

Until further notice, we'll restrict out attention to random variables that are *discrete*, i.e., they take values in a range that is finite or countably infinite.

The r.v. X in our permutation example above is completely specified by its values at all sample points, as given in the above table. (Thus, for example, $X(123) = 3$ etc.)

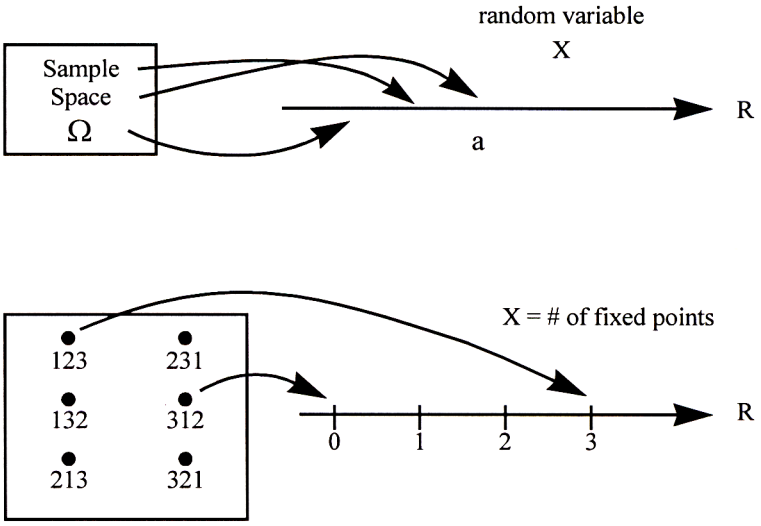


Figure 1: Visualization of how a random variable is defined on the sample space.

A random variable can be visualized in general by the picture in Figure 1[1]. Note that the term "random variable" is really something of a misnomer: it is a function so there is nothing random about it and it is definitely not a variable! What is random is which sample point of the experiment is realized and hence the value that the random variable maps the sample point to.

Distribution

When we introduced the basic probability space in Note 10, we defined two things: 1) the sample space Ω consisting of all the possible outcomes (sample points) of the experiment; 2) the probability of each of the sample points. Analogously, there are two things important about any random variable: 1) the set of values that it can take ; 2) the probabilities with which it takes on the values. Since a random variable is defined on a probability space, we can calculate these probabilities given the probabilities of the sample points. Let a be any number in the range of a random variable X. Then the set

$$\{\omega \in \Omega : X(\omega) = a\}$$

is an *event* in the sample space (do you see why?). We usually abbreviate this event to simply "$X = a$". Since $X = a$ is an event, we can talk about its probability, $\Pr[X = a]$. The collection of these probabilities, for all possible values of a, is known as the *distribution* of the r.v. X.

Definition 13.2 (distribution): The *distribution* of a discrete random variable X is the collection of values $\{(a, \Pr[X = a]) : a \in \mathscr{A}\}$, where $\mathscr{A}$ is the set of all possible values taken by X.

Thus the distribution of the random variable X in our permutation example above is

$$\Pr[X = 0] = \frac{1}{3}; \qquad \Pr[X = 1] = \frac{1}{2}; \qquad \Pr[X = 3] = \frac{1}{6};$$

and $\Pr[X = a] = 0$ for all other values of a.

[1]This and other figures in this note are inspired by figures in Chapter 2 of "Introduction to Probability" by D. Bertsekas and J. Tsitsiklis.

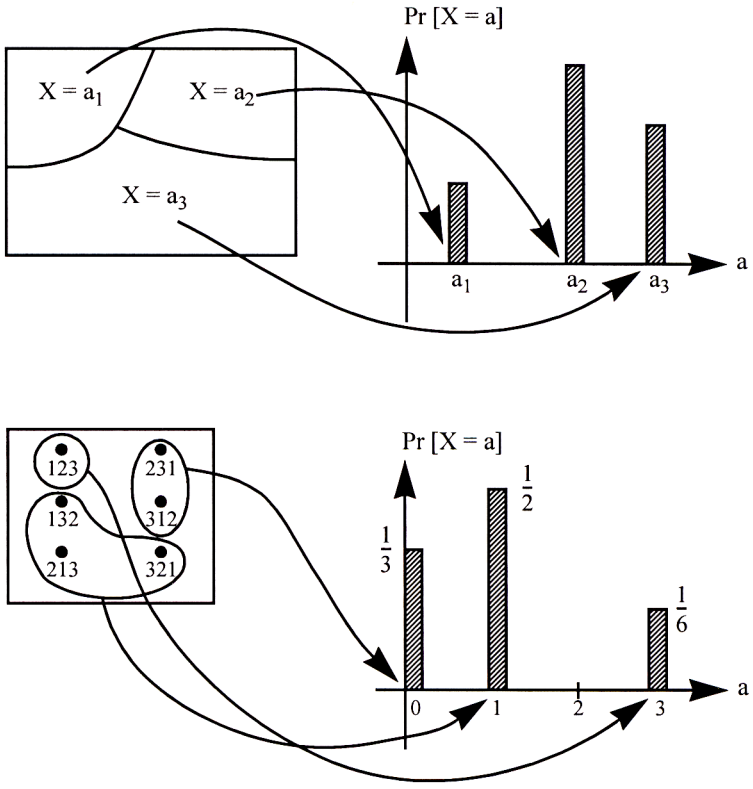


Figure 2: Visualization of how the distribution of a random variable is defined.

The distribution of a random variable can be visualized as a bar diagram, shown in Figure 2. The x-axis represents the values that the random variable can take on. The height of the bar at a value a is the probability $\Pr[X = a]$. Each of these probabilities can be computed by looking at the probability of the corresponding event in the sample space.

Note that the collection of events $X = a$, $a \in \mathcal{A}$, satisfy two important properties:

- any two events $X = a_1$ and $X = a_2$ with $a_1 \neq a_2$ are disjoint.
- the union of all these events is equal to the entire sample space Ω.

The collection of events thus form a *partition* of the sample space (see Figure 2). Both properties follow directly from the fact that X is a function defined on Ω, i.e. X assigns a unique value to each and every possible sample point in Ω. As a consequence, the sum of the probabilities $\Pr[X = a]$ over all possible values of a is exactly 1. So when we sum up the probabilities of the events $X = a$, we are really summing up the probabilities of all the sample points.

Example: The binomial distribution: This is one of the most important distributions in probability. It can be defined in terms of a coin-tossing experiment. Consider n independent tosses of a biased coin with Heads probability p. Each sample point is a sequence of tosses. For example, when $n = 3$, the sample space Ω is $\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.

Let X be the number of Heads. Note that this is a function on the sample space: for each sample point ω, $X(\omega)$ is the number of Heads in ω. For example, $X(THH) = 2$ To compute the distribution of X, we first enumerate the possible values X can take on. They are simply $0, 1, \ldots, n$. Then we compute the probability of each event $X = i$ for $i = 0, \ldots, n$. The probability of the event $X = i$ is the sum of the probabilities of all

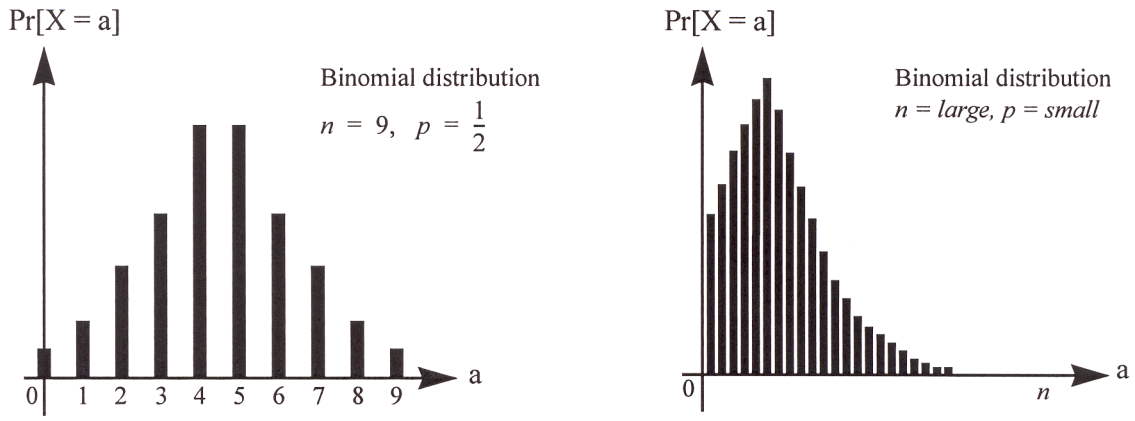


Figure 3: The binomial distributions for two choices of (n, p).

the sample points with i Heads. Any such sample point has a probability $p^i(1-p)^{n-i}$. There are exactly $\binom{n}{i}$ of these sample points. So

$$\Pr[X = i] = \binom{n}{i} p^i (1-p)^{n-i} \qquad i = 0, 1, \ldots n \tag{1}$$

This is the *binomial* distribution with parameters n and p. A random variable with this distribution is called a *binomial* random variable (for brevity, we will say $X \sim \text{Bin}(n, p)$). An example of a binomial distribution is shown in Figure 3.

Although we define the binomial distribution in terms of an experiment involving tossing coins, this distribution is useful for modeling many real-world problems. Consider for example the error correction problem studied in Note 8. Recall that we wanted to encode n packets into $n+k$ packets such that the recipient can reconstruct the original n packets from any n packets received. But in practice, the number of packet losses is random, so how do we choose k, the amount of redundancy? If we model each packet getting lost with probability p and the losses are independent, then if we transmit $n+k$ packets, the number of packets received is a random variable X and $X \sim \text{Bin}(n+k, 1-p)$. (We are tossing a coin $n+k$ times, and each coin turns out to be a Head (packet received) with probability $1-p$). So the probability of successfully decoding the original data is:

$$\Pr[X \geq n] = \sum_{i=n}^{n+k} \binom{n+k}{i} (1-p)^i p^{n+k-i}.$$

We can choose k such that this probability is no less than, say, 0.95.

Expectation

The distribution of a r.v. contains *all* the probabilistic information about the r.v. In most applications, however, the complete distribution of a r.v. is very hard to calculate: for example, suppose we go back to our original question with 20 students. In principle, we'd have to enumerate $20! \approx 2.4 \times 10^{18}$ sample points, compute the value of X at each one, and count the number of points at which X takes on each of its possible values! (In practice we could streamline this calculation quite a bit, but it is still tedious.) Moreover, even when we can compute the complete distribution of a r.v., it's not always very informative.

For these reasons, we seek to *compress* the distribution into a more compact, convenient form that is also easier to compute. The most widely used such form is the *expectation* (or *mean*) of the r.v.

Definition 13.3 (expectation): The *expectation* of a discrete random variable X is defined as

$$\mathbb{E}(X) = \sum_{a \in \mathscr{A}} a \times \Pr[X = a],$$

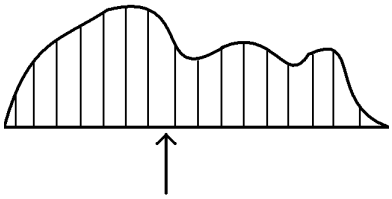
where the sum is over all possible values taken by the r.v.

For our running permutation example, the expectation is

$$\mathbb{E}(X) = \left(0 \times \frac{1}{3}\right) + \left(1 \times \frac{1}{2}\right) + \left(3 \times \frac{1}{6}\right) = 0 + \frac{1}{2} + \frac{1}{2} = 1.$$

In other words, the expected number of fixed points in a permutation of three items is exactly 1.

The expectation can be seen in some sense as a "typical" value or the center of mass of the r.v. (though note that it may not actually be a value that the r.v. ever takes on). It serves as a "balance" for the distribution of a random variable.



The question of how typical the expectation is for a given r.v. is a very important one that we shall return to in a later lecture.

Here are some simple examples of expectations.

1. **Single die.** Throw one fair die. Let X be the number that comes up. Then X takes on values $1, 2, \ldots, 6$ each with probability $\frac{1}{6}$, so

   $$\mathbb{E}(X) = \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = \frac{7}{2}.$$

   Note that X never actually takes on its expected value $\frac{7}{2}$.

2. **Two dice.** Throw two fair dice. Let X be the sum of their scores. Then the distribution of X is

a	2	3	4	5	6	7	8	9	10	11	12
$\Pr[X = a]$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

   The expectation is therefore

   $$\mathbb{E}(X) = \left(2 \times \frac{1}{36}\right) + \left(3 \times \frac{1}{18}\right) + \left(4 \times \frac{1}{12}\right) + \cdots + \left(12 \times \frac{1}{36}\right) = 7.$$

3. **Roulette.** A roulette wheel is spun. You bet \$1 on Black. If a black number comes up, you receive your stake plus \$1; otherwise you lose your stake. Let X be your net winnings in one game. Then X can take on the values $+1$ and -1, and $\Pr[X = 1] = \frac{18}{38}$, $\Pr[X = -1] = \frac{20}{38}$. [Recall that a roulette wheel has 38 slots: the numbers $1, 2, \ldots, 36$, half of which are red and half black, plus 0 and 00, which are green.] Thus

   $$\mathbb{E}(X) = \left(1 \times \frac{18}{38}\right) + \left(-1 \times \frac{20}{38}\right) = -\frac{1}{19};$$

   i.e., you expect to lose about a nickel per game. Notice how the zeros tip the balance in favor of the casino!

There is an alternative but equivalent way of defining expectation that is useful in some problems. Recall from Definition 13.3 that

$$\mathbb{E}(X) = \sum_{a \in \mathscr{A}} a \times \Pr[X = a].$$

Let's look at one term $a \times \Pr[X = a]$ in the above sum. Notice that $\Pr[X = a]$, by definition, is the sum of $\Pr[\omega]$ over those sample points ω for which $X(\omega) = a$. And we know that every sample point $\omega \in \Omega$ is in exactly one of these events $X = a$. This means we can write out the above definition in an equivalent form as

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega) \times \Pr[\omega]. \tag{2}$$

Linearity of expectation

So far, we've computed expectations by brute force: i.e., we have written down the whole distribution and then added up the contributions for all possible values of the r.v. The real power of expectations is that in many real-life examples they can be computed much more easily using a simple shortcut. The shortcut is the following:

Theorem 13.1: For any two random variables X and Y on the same probability space, we have

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Also, for any constant c, we have

$$\mathbb{E}(cX) = c\mathbb{E}(X).$$

A note: here $X+Y$ denotes the random variable that is the sum of X and Y, i.e., it takes on the value $X(\omega)+Y(\omega)$ at the outcome ω.

Proof: Let's write out $\mathbb{E}(X+Y)$ using the alternative definition of expectation as given by equation (2):

$$\begin{aligned}
\mathbb{E}(X+Y) &= \sum_{\omega \in \Omega} (X+Y)(\omega) \times \Pr[\omega] \\
&= \sum_{\omega \in \Omega} (X(\omega)+Y(\omega)) \times \Pr[\omega] \\
&= \sum_{\omega \in \Omega} \Big(X(\omega) \times \Pr[\omega]\Big) + \sum_{\omega \in \Omega} \Big(Y(\omega) \times \Pr[\omega]\Big) \\
&= \mathbb{E}(X) + \mathbb{E}(Y).
\end{aligned}$$

In the last step, we used equation (2) twice.

This completes the proof of the first equality. The proof of the second equality is much simpler and is left as an exercise. □

Theorem 13.1 is very powerful: it says that the expectation of a sum of r.v.'s is the sum of their expectations, no matter what those r.v.'s may be. We can use Theorem 13.1 to conclude things like $\mathbb{E}(3X-5Y) = 3\mathbb{E}(X) - 5\mathbb{E}(Y)$. This property is known as *linearity of expectation*. One important caveat: Theorem 13.1 does *not* say that $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, or that $\mathbb{E}(\frac{1}{X}) = \frac{1}{\mathbb{E}(X)}$ etc. These claims are not true in general. It is only sums and differences and constant multiples of random variables that behave so nicely.

Now let's see some examples of Theorem 13.1 in action.

4. **Two dice again.** Here's a much less painful way of computing $\mathbb{E}(X)$, where X is the sum of the scores of the two dice. Note that $X = X_1 + X_2$, where X_i is the score on die i. We know from example 1 above that $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \frac{7}{2}$. So by Theorem 13.1 we have $\mathbb{E}(X) = \mathbb{E}(X_1) + \mathbb{E}(X_2) = 7$.

5. **More roulette.** Suppose we play the above roulette game not once, but 1000 times. Let X be our expected net winnings. Then $X = X_1 + X_2 + \cdots + X_{1000}$, where X_i is our net winnings in the ith play. We know from earlier that $\mathbb{E}(X_i) = -\frac{1}{19}$ for each i. Therefore, by Theorem 13.1, $\mathbb{E}(X) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_{1000}) = 1000 \times (-\frac{1}{19}) = -\frac{1000}{19} \approx 53$. So if you play 1000 games, you expect to lose about \$53.

6. **Homeworks.** Let's go back and answer our original question about the class of 20 students. Recall that the r.v. X is the number of students who receive their own homework after shuffling (or equivalently, the number of fixed points). To take advantage of Theorem 13.1, we need to write X as the *sum* of simpler r.v.'s. But since X *counts* the number of times something happens, we can write it as a sum using the following trick:

$$X = X_1 + X_2 + \cdots + X_{20}, \qquad \text{where } X_i = \begin{cases} 1 & \text{if student } i \text{ gets her own homework;} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

[You should think about this equation for a moment. Remember that all the X's are random variables. What does an equation involving random variables mean? What we mean is that, *at every sample point* ω, we have $X(\omega) = X_1(\omega) + X_2(\omega) + \cdots + X_{20}(\omega)$. Do you see why this is true?]

A 0/1-valued random variable such as X_i is called an *indicator* random variable of the corresponding event (in this case, the event that student i gets her own homework). For indicator r.v.'s, the expectation is particularly easy to calculate. Namely,

$$\mathbb{E}(X_i) = (0 \times \Pr[X_i = 0]) + (1 \times \Pr[X_i = 1]) = \Pr[X_i = 1].$$

But in our case, we have

$$\Pr[X_i = 1] = \Pr[\text{student } i \text{ gets her own homework}] = \frac{1}{20}.$$

Now we can apply Theorem 13.1 to (3), to get

$$\mathbb{E}(X) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_{20}) = 20 \times \frac{1}{20} = 1.$$

So we see that the expected number of students who get their own homeworks in a class of size 20 is 1. But this is exactly the same answer as we got for a class of size 3! And indeed, we can easily see from the above calculation that we would get $\mathbb{E}(X) = 1$ for *any* class size n: this is because we can write $X = X_1 + X_2 + \cdots + X_n$, and $\mathbb{E}(X_i) = \frac{1}{n}$ for each i.

So *the expected number of fixed points in a random permutation of n items is always 1*, regardless of n. Amazing, but true.

7. **Coin tosses.** Toss a fair coin 100 times. Let the r.v. X be the number of Heads. As in the previous example, to take advantage of Theorem 13.1 we write

$$X = X_1 + X_2 + \cdots + X_{100},$$

where X_i is the indicator r.v. of the event that the ith toss is Heads. In other words, $X_i = 1$ if the ith toss comes up Heads, and $X_i = 0$ otherwise. Since the coin is fair, we have

$$\mathbb{E}(X_i) = \Pr[X_i = 1] = \Pr[i\text{th toss is Heads}] = \tfrac{1}{2}.$$

Using Theorem 13.1, we therefore get

$$\mathbb{E}(X) = \sum_{i=1}^{100} \mathbb{E}(X_i) = 100 \times \tfrac{1}{2} = 50.$$

More generally, the expected number of Heads in n tosses of a fair coin is $\frac{n}{2}$. And in n tosses of a biased coin with Heads probability p, the expected number of Heads is np (do you see why?). So the expectation of a r.v. $X \sim \text{Bin}(n,p)$ is np. Note that it would have been harder to reach the same conclusion by computing this directly from definition of expectation.

8. **Balls and bins.** Throw m balls into n bins. Let the r.v. X be the number of balls that land in the first bin. Then X behaves exactly like the number of Heads in n tosses of a biased coin, with Heads probability $\frac{1}{n}$ (do you see why?). So from example 7 we get $\mathbb{E}(X) = \frac{m}{n}$.

   In the special case $m = n$, the expected number of balls in any bin is 1. If we wanted to compute this directly from the distribution of X, we'd get into a messy calculation involving binomial coefficients.

   Here's another example on the same sample space. Let the r.v. Y be the number of empty bins. The distribution of Y is horrible to contemplate: to get a feel for this, you might like to write it down for $m = n = 3$ (3 balls, 3 bins). However, computing the expectation $\mathbb{E}(Y)$ is a piece of cake using Theorem 13.1. As usual, let's write

$$Y = Y_1 + Y_2 + \cdots + Y_n, \tag{4}$$

   where Y_i is the indicator r.v. of the event "bin i is empty". Again as usual, the expectation of Y_i is easy:

$$\mathbb{E}(Y_i) = \Pr[Y_i = 1] = \Pr[\text{bin } i \text{ is empty}] = \left(1 - \frac{1}{n}\right)^m;$$

   recall that we computed this probability (quite easily) in an earlier lecture. Applying Theorem 13.1 to (4) we therefore have

$$\mathbb{E}(Y) = \sum_{i=1}^{n} \mathbb{E}(Y_i) = n\left(1 - \frac{1}{n}\right)^m,$$

   a very simple formula, very easily derived.

   Let's see how it behaves in the special case $m = n$ (same number of balls as bins). In this case we get $\mathbb{E}(Y) = n(1 - \frac{1}{n})^n$. Now the quantity $(1 - \frac{1}{n})^n$ can be approximated (for large enough values of n) by the number $\frac{1}{\text{e}}$.[2] So we see that

$$\mathbb{E}(Y) \to \frac{n}{\text{e}} \approx 0.368n \quad \text{as } n \to \infty.$$

   The bottom line is that, if we throw (say) 1000 balls into 1000 bins, the expected number of empty bins is about 368.

[2] More generally, it is a standard fact that for any constant c,

$$(1 + \tfrac{c}{n})^n \to \text{e}^c \quad \text{as } n \to \infty.$$

We just used this fact in the special case $c = -1$. The approximation is actually very good even for quite small values of n — try it yourself! E.g., for $n = 20$ we already get $(1 - \frac{1}{n})^n$=0.358, which is very close to $\frac{1}{\text{e}} = 0.367\ldots$. The approximation gets better and better for larger n.

Some Important Distributions

The first important distribution we learned about in the last Lecture Note is the *binomial distribution* $\text{Bin}(n,p)$. This is the distribution of the number of Heads in n tosses of a biased coin with probability p to be Heads. Its expectation is np. In this Note, we explore two other important distributions: the Geometric and the Poisson distributions. The first one is also associated with a coin tossing experiment.

Geometric Distribution

Question: A biased coin with Heads probability p is tossed repeatedly until the first Head appears. What is the distribution and the expected number of tosses?

As always, our first step in answering the question must be to define the sample space Ω. A moment's thought tells us that

$$\Omega = \{H, TH, TTH, TTTH, \ldots\},$$

i.e., Ω consists of all sequences over the alphabet $\{H,T\}$ that end with H and contain no other H's. This is our first example of an *infinite* sample space (though it is still discrete).

What is the probability of a sample point, say $\omega = TTH$? Since successive coin tosses are independent (this is implicit in the statement of the problem), we have

$$\Pr[TTH] = (1-p) \times (1-p) \times p = (1-p)^2 p.$$

And generally, for any sequence $\omega \in \Omega$ of length i, we have $\Pr[\omega] = (1-p)^{i-1}p$. To be sure everything is consistent, we should check that the probabilities of all the sample points add up to 1. Since there is exactly one sequence of each length $i \geq 1$ in Ω, we have

$$\sum_{\omega \in \Omega} \Pr[\omega] = \sum_{i=1}^{\infty} (1-p)^{i-1} p = p \sum_{i=0}^{\infty} (1-p)^i = p \times \frac{1}{1-(1-p)} = 1,$$

as expected. [In the second-last step here, we used the formula for summing a geometric series.]

Now let the random variable X denote the number of tosses in our sequence (i.e., $X(\omega)$ is the length of ω). Its distribution has a special name: it is called the *geometric distribution with parameter p* (where p is the probability that the coin comes up Heads on each toss).

Definition 14.1 (geometric distribution): A random variable X for which

$$\Pr[X = i] = (1-p)^{i-1} p \qquad \text{for } i = 1, 2, 3, \ldots$$

is said to have the *geometric distribution with parameter p*. This is abbreviated as $X \sim Geom(p)$.

If we plot the distribution of X (i.e., the values $\Pr[X = i]$ against i) we get a curve that decreases monotonically by a factor of $1-p$ at each step. See Figure 1.

Our next goal is to compute $\mathbb{E}(X)$. Despite the fact that X counts something, there's no obvious way to write it as a sum of simple r.v.'s as we did in many examples in an earlier lecture note. (Try it!) In a later lecture,

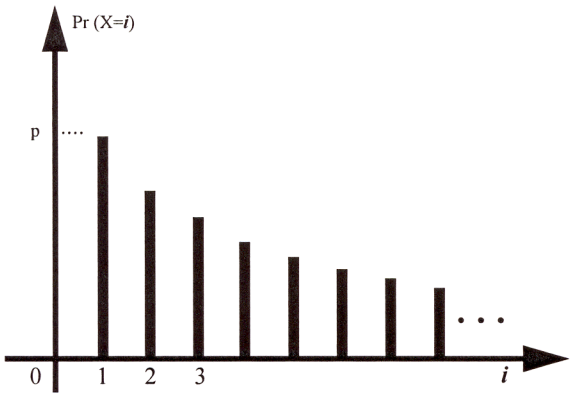


Figure 1: The Geometric distribution.

we will give a slick way to do this calculation. For now, let' s just dive in and try a direct computation of $\mathbb{E}(X)$. Note that the distribution of X is quite simple:

$$\Pr[X = i] = (1-p)^{i-1}p \qquad \text{for } i = 1,2,3,\ldots$$

So from the definition of expectation we have

$$\mathbb{E}(X) = (1\times p) + (2\times(1-p)p) + (3\times(1-p)^2p) + \cdots = p\sum_{i=1}^{\infty} i(1-p)^{i-1}.$$

This series is a blend of an arithmetic series (the i part) and a geometric series (the $(1-p)^{i-1}$ part). There are several ways to sum it. Here is one way, using an auxiliary trick (given in the following Theorem) that is often very useful. [Ask your TA about other ways.]

Theorem 14.1: Let X be a random variable that takes on only non-negative integer values. Then

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} \Pr[X \geq i].$$

Proof: For notational convenience, let's write $p_i = \Pr[X = i]$, for $i = 0,1,2,\ldots$. From the definition of expectation, we have

$$\begin{aligned}
\mathbb{E}(X) &= (0\times p_0) + (1\times p_1) + (2\times p_2) + (3\times p_3) + (4\times p_4) + \cdots \\
&= p_1 + (p_2+p_2) + (p_3+p_3+p_3) + (p_4+p_4+p_4+p_4) + \cdots \\
&= (p_1+p_2+p_3+p_4+\cdots) + (p_2+p_3+p_4+\cdots) + (p_3+p_4+\cdots) + (p_4+\cdots) + \cdots \\
&= \Pr[X\geq 1] + \Pr[X\geq 2] + \Pr[X\geq 3] + \Pr[X\geq 4] + \cdots.
\end{aligned}$$

In the third line, we have regrouped the terms into convenient infinite sums. You should check that you understand how the fourth line follows from the third. Our "..." notation here is a little informal, but the meaning should be clear. Here is a more rigorous, but perhaps less clear proof:

$$\sum_{i=1}^{\infty}\Pr[X\geq i] = \sum_{i=1}^{\infty}\sum_{j=i}^{\infty}\Pr[X=j] = \sum_{1\leq i\leq j<\infty}\Pr[X=j] = \sum_{j=1}^{\infty}\sum_{i=1}^{j}\Pr[X=j] = \sum_{j=1}^{\infty} j\times\Pr[X=j] = \mathbb{E}(X).$$

□

An alternative statement of Theorem 14.1 is

$$\mathbb{E}(X) = \sum_{i=0}^{\infty} \Pr[X > i].$$

Using Theorem 14.1, it is easy to compute $\mathbb{E}(X)$. The key observation is that, for our coin-tossing r.v. X,

$$\Pr[X \geq i] = (1-p)^{i-1}. \tag{1}$$

Why is this? Well, the event "$X \geq i$" means that at least i tosses are required. This is exactly equivalent to saying that the first $i-1$ tosses are all Tails. And the probability of this event is precisely $(1-p)^{i-1}$. Now, plugging equation (1) into Theorem 14.1, we get

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} \Pr[X \geq i] = \sum_{i=1}^{\infty} (1-p)^{i-1} = 1 + (1-p) + (1-p)^2 + (1-p)^3 + \cdots = \frac{1}{1-(1-p)} = \frac{1}{p}.$$

Here we have used the sum of the geometric series, namely, $1 + x + x^2 + x^3 + \cdots = 1/(1-x)$ if $-1 < x < 1$. So, *the expected number of tosses of a biased coin until the first Head appears is* $\frac{1}{p}$. For a fair coin, the expected number of tosses is 2.

For posterity, let's record two important facts we've learned about the geometric distribution:

Theorem 14.2: For a random variable X having the geometric distribution with parameter p,

1. $\mathbb{E}(X) = \frac{1}{p}$; and
2. $\Pr[X \geq i] = (1-p)^{i-1}$ for $i = 1, 2, \ldots$.

The geometric distribution occurs very often in applications because frequently we are interested in how long we have to wait before a certain event happens: how many runs before the system fails, how many shots before one is on target, how many poll samples before we find a Democrat, how many retransmissions of a packet before successfully reaching the destination, etc. The next section discusses a rather more involved application, which is important in its own right.

The Coupon Collector's Problem Revisited

Let's go back to the coupon collector's problem we discussed in Note 12. We have calculated the number of cereal boxes we have to buy to guarantee that with probability $1/2$ we have collected at least one copy of every card. This number is fixed, not random. Suppose now we run the experiment until we have collected one copy of every card. The number of cereal boxes we have bought is now a random variable. We can ask about its expectation.

Question: We are trying to collect a set of n different baseball cards. We get the cards by buying boxes of cereal: each box contains exactly one card, and it is equally likely to be any of the n cards. What is the expected number of boxes we need to buy until we have collected at least one copy of every card?

The sample space here is similar in flavor to that for our previous coin-tossing example, though rather more complicated. It consists of all sequences ω over the alphabet $\{1, 2, \ldots, n\}$, such that

1. ω contains each symbol $1, 2, \ldots, n$ at least once; and
2. the final symbol in ω occurs only once.

[Check that you understand this!] For any such ω, the probability is just $\Pr[\omega] = \frac{1}{n^i}$, where i is the length of ω (why?). However, it is very hard to figure out how many sample points ω are of length i (try it for the case $n = 3$). So we will have a hard time figuring out the distribution of the random variable X, which is the length of the sequence (i.e., the number of boxes bought).

Fortunately, we can compute the expectation $\mathbb{E}(X)$ very easily, using (guess what?) linearity of expectation, plus the fact we have just learned about the expectation of the geometric distribution. As usual, we would like to write

$$X = X_1 + X_2 + \ldots + X_n \tag{2}$$

for suitable simple random variables X_i. But what should the X_i be? A natural thing to try is to make X_i equal to the number of boxes we buy while trying to get the ith new card (starting immediately after we've got the $(i-1)$st new card). With this definition, make sure you believe equation (2) before proceeding.

What does the distribution of X_i look like? Well, X_1 is trivial: no matter what happens, we always get a new card in the first box (since we have none to start with). So $\Pr[X_1 = 1] = 1$, and thus $\mathbb{E}(X_1) = 1$.

How about X_2? Each time we buy a box, we'll get the same old card with probability $\frac{1}{n}$, and a new card with probability $\frac{n-1}{n}$. So we can think of buying boxes as flipping a biased coin with Heads probability $p = \frac{n-1}{n}$; then X_1 is just the number of tosses until the first Head appears. So X_1 has the geometric distribution with parameter $p = \frac{n-1}{n}$, and

$$\mathbb{E}(X_2) = \frac{n}{n-1}.$$

How about X_3? This is very similar to X_2 except that now we only get a new card with probability $\frac{n-2}{n}$ (since there are now two old ones). So X_3 has the geometric distribution with parameter $p = \frac{n-2}{n}$, and

$$\mathbb{E}(X_3) = \frac{n}{n-2}.$$

Arguing in the same way, we see that, for $i = 1, 2, \ldots, n$, X_i has the geometric distribution with parameter $p = \frac{n-i+1}{n}$, and hence that

$$\mathbb{E}(X_i) = \frac{n}{n-i+1}.$$

Finally, applying linearity of expectation to equation (2), we get

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{2} + \frac{n}{1} = n \sum_{i=1}^{n} \frac{1}{i}. \tag{3}$$

This is an exact expression for $\mathbb{E}(X)$. We can obtain a tidier form by noting that the sum in it actually has a very good approximation[1], namely:

$$\sum_{i=1}^{n} \frac{1}{i} \approx \ln n + \gamma,$$

where $\gamma = 0.5772\ldots$ is *Euler's constant*.

Thus *the expected number of cereal boxes needed to collect n cards is about* $n(\ln n + \gamma)$. This is an excellent approximation to the exact formula (3) even for quite small values of n. So for example, for $n = 100$, we expect to buy about 518 boxes.

[1] This is another of the little tricks you might like to carry around in your toolbox.

The Poisson distribution

Throw n balls into $\frac{n}{\lambda}$ bins (where λ is a constant). Let X be the number of balls that land in bin 1. Then X has the binomial distribution with parameters n and $p = \frac{\lambda}{n}$, and its expectation is $\mathbb{E}(X) = np = \lambda$. (Why?)

Let's look in more detail at the distribution of X (which is a special case of the binomial distribution, in which the parameter p is of the form $\frac{\lambda}{n}$). For convenience, we'll write $p_i = \Pr[X = i]$ for $i = 0, 1, 2, \ldots$. Beginning with p_0, we have

$$p_0 = \Pr[\text{all balls miss bin 1}] = \left(1 - \frac{\lambda}{n}\right)^n \to \mathrm{e}^{-\lambda} \quad \text{as } n \to \infty.$$

So the probability of no balls landing in bin 1 will be very close to the constant value $\mathrm{e}^{-\lambda}$ when n is large.

What about the other p_i? Well, we know from the binomial distribution that

$$p_i = \binom{n}{i} (\frac{\lambda}{n})^i (1 - \frac{\lambda}{n})^{n-i}.$$

Since we know how p_0 behaves, let's look at the ratio $\frac{p_1}{p_0}$:

$$\frac{p_1}{p_0} = \frac{n \times \frac{\lambda}{n} \times (1 - \frac{\lambda}{n})^{n-1}}{(1 - \frac{\lambda}{n})^n} = \frac{\lambda}{1 - \frac{\lambda}{n}} = \frac{n\lambda}{n - \lambda} \to \lambda \quad \text{as } n \to \infty.$$

[Recall that we are assuming λ is a constant.] So, since $p_0 \to \mathrm{e}^{-\lambda}$, we see that $p_1 \to \lambda \mathrm{e}^{-\lambda}$ as $n \to \infty$.

Now let's look at the ratio $\frac{p_2}{p_1}$:

$$\frac{p_2}{p_1} = \frac{\binom{n}{2} \times (\frac{\lambda}{n})^2 \times (1 - \frac{\lambda}{n})^{n-2}}{n \times (\frac{\lambda}{n}) \times (1 - \frac{\lambda}{n})^{n-1}} = \frac{n-1}{2} \times \frac{\lambda}{n} \times \frac{1}{(1 - \frac{\lambda}{n})} = \frac{n-1}{n - \lambda} \times \frac{\lambda}{2} \to \frac{\lambda}{2} \quad \text{as } n \to \infty.$$

So $p_2 \to \frac{\lambda^2}{2} \mathrm{e}^{-\lambda}$ as $n \to \infty$.

For each value of i, something very similar happens to the ratio $\frac{p_i}{p_{i-1}}$:

$$\frac{p_i}{p_{i-1}} = \frac{\binom{n}{i} \times (\frac{\lambda}{n})^i \times (1 - \frac{\lambda}{n})^{n-i}}{\binom{n}{i-1} (\frac{\lambda}{n})^{i-1} (1 - \frac{\lambda}{n})^{n-i+1}} = \frac{n-i+1}{i} \times \frac{\lambda}{n} \times \frac{n}{n - \lambda} = \frac{n-i+1}{n - \lambda} \times \frac{\lambda}{i} \to \frac{\lambda}{i} \quad \text{as } n \to \infty.$$

Putting this together, we see that, for each fixed value i,

$$p_i \to \frac{\lambda^i}{i!} \mathrm{e}^{-\lambda} \qquad \text{as } n \to \infty.$$

[You should check this!] In other words, when n is large compared to i, the probability that exactly i balls fall into bin 1 is very close to $\frac{\lambda^i}{i!} \mathrm{e}^{-\lambda}$. This motivates the following definition:

Definition 14.2 (Poisson distribution): A random variable X for which

$$\Pr[X = i] = \frac{\lambda^i}{i!} \mathrm{e}^{-\lambda} \qquad \text{for } i = 0, 1, 2, \ldots \tag{4}$$

is said to have the *Poisson distribution with parameter* λ. This is abbreviated as $X \sim \text{Poiss}(\lambda)$.

To make sure this definition is valid, we had better check that (4) is in fact a distribution, i.e., that the probabilities sum to 1. We have

$$\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} e^{-\lambda} = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda} \times e^{\lambda} = 1.$$

[In the second-last step here, we used the Taylor series expansion $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$.] So this does indeed meet the requirements to be a distribution.

What is the expectation of a Poisson random variable X? This is a simple hands-on calculation, starting from the definition of expectation:

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{i=0}^{\infty} i \times \Pr[X = i] \\
&= \sum_{i=0}^{\infty} i \frac{\lambda^i}{i!} e^{-\lambda} \\
&= e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^i}{(i-1)!} \\
&= \lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} \\
&= \lambda e^{-\lambda} e^{\lambda} \\
&= \lambda.
\end{aligned}$$

So *the expectation of a Poisson r.v. X with parameter λ is $\mathbb{E}(X) = \lambda$.*

A plot of the Poisson distribution reveals a curve that rises monotonically to a single peak and then decreases monotonically. The peak is as close as possible to the expected value, i.e., at $i = \lfloor \lambda \rfloor$.

We have seen that the Poisson distribution arises as the limit of the number of balls in bin 1 when n balls are thrown into $\frac{n}{\lambda}$ bins. In other words, it is the limit of the binomial distribution with parameters n and $p = \frac{\lambda}{n}$ as $n \to \infty$, with λ being a fixed constant.

The Poisson distribution is also a very widely accepted model for so-called "rare events", such as misconnected phone calls, radioactive emissions, crossovers in chromosomes, etc. This model is appropriate whenever the occurrences can be assumed to happen randomly with some constant density λ in a continuous region (of time or space), such that occurrences in disjoint subregions are independent. One can then show that the number of occurrences in a region of unit size should obey the Poisson distribution with parameter λ.

To see this in a concrete setting, suppose we want to model the number of cell phone users initiating calls in a network during a time period, of duration say 1 minute. There are many paying customers in the network, and all of them can potentially make a call during this time period. However, only a very small fraction of them actually will. Under this scenario, it seems reasonable to make two assumptions:

- The probability of having more than 1 customer initiating a call in any small time interval is negligible.
- The initiation of calls in disjoint time intervals are independent events.

Then if we divide the one-minute time period into n disjoint intervals, than the number of calls X in that time period can be modeled as binomially distributed with parameter n and probability of success p, the probability of having a call initiated in a time interval of length $1/n$. But what should p be in terms of the relevant parameters of the problem? If calls are initiated at an average rate of λ calls per minute, then

$\mathbb{E}(X) = \lambda$ and so $np = \lambda$, i.e. $p = \frac{\lambda}{n}$. So $X \sim \text{Bin}(n, \frac{\lambda}{n})$. By choosing n large, it is thus reasonable to model the number of call initiations during a one-minute period to be Poisson with parameter λ.

The Poisson distribution arises naturally in several important contexts. Along with the binomial and geometric distributions (which you have already seen) and the normal distribution (which we shall meet soon), the Poisson distribution is one of the four distributions you are most likely to find yourself working with.

Variance

Question: At each time step, I flip a fair coin. If it comes up Heads, I walk one step to the right; if it comes up Tails, I walk one step to the left. How far do I expect to have traveled from my starting point after n steps?

Denoting a right-move by $+1$ and a left-move by -1, we can describe the probability space here as the set of all words of length n over the alphabet $\{\pm 1\}$, each having equal probability $\frac{1}{2^n}$. For instance, one possible outcome is $(+1,+1,-1,\ldots,-1)$. Let the r.v. X denote our position (relative to our starting point 0) after n moves. Thus

$$X = X_1 + X_2 + \cdots + X_n,$$

where $X_i = \begin{cases} +1 & \text{if } i\text{th toss is Heads;} \\ -1 & \text{otherwise.} \end{cases}$

Now obviously we have $\mathbb{E}(X) = 0$. The easiest way to see this is to note that $\mathbb{E}(X_i) = (\frac{1}{2} \times 1) + (\frac{1}{2} \times (-1)) = 0$, so by linearity of expectation $\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(X_i) = 0$. Thus after n steps, my expected position is 0. But of course this is not very informative, and is due to the fact that positive and negative deviations from 0 cancel out.

What the above question is really asking is: What is the expected value of $|X|$, our *distance* from 0? Unfortunately, computing the expected value of $|X|$ turns out to be a little awkward, due to the absolute value operator. Therefore, rather than consider the r.v. $|X|$, we will instead look at the r.v. X^2. Notice that this also has the effect of making all deviations from 0 positive, so it should also give a good measure of the distance traveled. However, because it is the *squared* distance, we will need to take a square root at the end.

Let's calculate $\mathbb{E}(X^2)$:

$$\begin{aligned} \mathbb{E}(X^2) &= \mathbb{E}((X_1 + X_2 + \cdots + X_n)^2) \\ &= \mathbb{E}(\sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} X_i X_j) \\ &= \sum_{i=1}^{n} \mathbb{E}(X_i^2) + \sum_{i \neq j} \mathbb{E}(X_i X_j) \end{aligned}$$

In the last line here, we used linearity of expectation. To proceed, we need to compute $\mathbb{E}(X_i^2)$ and $\mathbb{E}(X_i X_j)$ (for $i \neq j$). Let's consider first X_i^2. Since X_i can take on only values ± 1, clearly $X_i^2 = 1$ always, so $\mathbb{E}(X_i^2) = 1$. What about $\mathbb{E}(X_i X_j)$? Well, $X_i X_j = +1$ when $X_i = X_j = +1$ or $X_i = X_j = -1$, and otherwise $X_i X_j = -1$. Also,

$$\Pr[(X_i = X_j = +1) \cup (X_i = X_j = -1)] = \Pr[X_i = X_j = +1] + \Pr[X_i = X_j = -1] = \frac{1}{4} + \frac{1}{4} = \frac{1}{2},$$

so $X_i X_j = 1$ with probability $\frac{1}{2}$. Therefore $X_i X_j = -1$ with probability $\frac{1}{2}$ also. Hence $\mathbb{E}(X_i X_j) = 0$.

Plugging these values into the above equation gives

$$\mathbb{E}(X^2) = (n \times 1) + 0 = n.$$

So we see that *our expected squared distance from 0 is n*. One interpretation of this is that we might expect to be a distance of about $\sqrt{n}$ away from 0 after n steps. However, we have to be careful here: we **cannot** simply argue that $\mathbb{E}(|X|) = \sqrt{\mathbb{E}(X^2)} = \sqrt{n}$. (Do you see why not?) We will see later in the lecture how to make precise deductions about $|X|$ from knowledge of $\mathbb{E}(X^2)$.

For the moment, however, let's agree to view $\mathbb{E}(X^2)$ as an intuitive measure of "spread" of the r.v. X. In fact, for a more general r.v. with expectation $\mathbb{E}(X) = \mu$, what we are really interested in is $\mathbb{E}((X-\mu)^2)$, the expected squared distance *from the mean*. In our random walk example, we had $\mu = 0$, so $\mathbb{E}((X-\mu)^2)$ just reduces to $\mathbb{E}(X^2)$.

Definition 15.1 (variance): For a r.v. X with expectation $\mathbb{E}(X) = \mu$, the *variance* of X is defined to be

$$\mathrm{Var}(X) = \mathbb{E}((X-\mu)^2).$$

The square root $\sigma(X) := \sqrt{\mathrm{Var}(X)}$ is called the *standard deviation* of X.

The point of the standard deviation is merely to "undo" the squaring in the variance. Thus the standard deviation is "on the same scale as" the r.v. itself. Since the variance and standard deviation differ just by a square, it really doesn't matter which one we choose to work with as we can always compute one from the other immediately. We shall usually use the variance. For the random walk example above, we have that $\mathrm{Var}(X) = n$, and the standard deviation of X, $\sigma(X)$, is $\sqrt{n}$.

The following easy observation gives us a slightly different way to compute the variance that is simpler in many cases.

Theorem 15.1: For a r.v. X with expectation $\mathbb{E}(X) = \mu$, we have $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mu^2$.

Proof: From the definition of variance, we have

$$\mathrm{Var}(X) = \mathbb{E}((X-\mu)^2) = \mathbb{E}(X^2 - 2\mu X + \mu^2) = \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + \mu^2 = \mathbb{E}(X^2) - 2\mu^2 + \mu^2 = \mathbb{E}(X^2) - \mu^2.$$

In the third step here, we used linearity of expectation. □

Let's see some examples of variance calculations.

1. **Fair die.** Let X be the score on the roll of a single fair die. Recall from an earlier lecture that $\mathbb{E}(X) = \frac{7}{2}$. So we just need to compute $\mathbb{E}(X^2)$, which is a routine calculation:

   $$\mathbb{E}(X^2) = \frac{1}{6}\left(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2\right) = \frac{91}{6}.$$

   Thus from Theorem 15.1

   $$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}.$$

   More generally, if X is a random variable that takes on values $1, \ldots, n$ with equal probability $1/n$ (i.e. X has a uniform distribution) , the mean, variance and standard deviation of X are:

   $$\mathbb{E}(X) = \frac{n+1}{2}, \qquad \mathrm{Var}(X) = \frac{n^2-1}{12}, \qquad \sigma(X) = \sqrt{\frac{n^2-1}{12}}.$$

   (You should verify these.)

2. **Biased coin.** Let X the the number of Heads in n tosses of a biased coin with Heads probability p (i.e., X has the binomial distribution with parameters n, p). We already know that $\mathbb{E}(X) = np$. Let

$$X_i = \begin{cases} 1 & \text{if } i\text{th toss is Head,} \\ 0 & \text{otherwise.} \end{cases}$$

We can then write $X = X_1 + X_2 + \cdots + X_n$, and then

$$\begin{aligned} \mathbb{E}(X^2) &= \mathbb{E}((X_1 + X_2 + \cdots + X_n)^2) \\ &= \sum_{i=1}^{n} \mathbb{E}(X_i^2) + \sum_{i \neq j} \mathbb{E}(X_i X_j) \\ &= (n \times p) + (n(n-1) \times p^2) \\ &= n^2 p^2 + np(1-p). \end{aligned}$$

In the third line here, we have used the facts that $\mathbb{E}(X_i^2) = p$, and that

$$\mathbb{E}(X_i X_j) = \Pr[X_i = X_j = 1] = \Pr[X_i = 1] \cdot \Pr[X_j = 1] = p^2,$$

(since $X_i = 1$ and $X_j = 1$ are independent events). Note that there are $n(n-1)$ pairs i, j with $i \neq j$.

Finally, we get that $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = np(1-p)$. As an example, for a fair coin the expected number of Heads in n tosses is $\frac{n}{2}$, and the standard deviation is $\frac{\sqrt{n}}{2}$. Since the maximum number of Heads is n, the standard deviation is much less than this maximum number for large n. This is in contrast to the previous example of the uniformly distributed random variable, where the standard deviation $\sigma(X) = \sqrt{(n^2-1)/12} \approx n/\sqrt{12}$ is of the same order as the largest value n. In this sense, the spread of a binomially distributed r.v. is much smaller than that of a uniformly distributed r.v.

Also, notice that in fact $\mathrm{Var}(X) = \sum_i \mathrm{Var}(X_i)$, and the same was true in the random walk example. This is no coincidence. We will explore for what kinds of random variables this is true later in the next lecture.

3. **Geometric distribution.** What is the variance of a geometric r.v. $X \sim \mathrm{Geom}(p)$? We can calculate $\mathbb{E}(X^2)$:

$$\mathbb{E}(X^2) = p + 4p(1-p) + 9p(1-p)^2 + 16p(1-p)^3 + 25p(1-p)^4 + \cdots$$

If we multiply this be $(1-p)$ and subtract the result from $\mathbb{E}(X^2)$, we get:

$$\begin{aligned} \mathbb{E}(X^2) &= p \;+\; 4p(1-p) \;+\; 9p(1-p)^2 \;+\; 16p(1-p)^3 \;+\; 25p(1-p)^4 \;+\; \cdots \\ (1-p)\mathbb{E}(X^2) &= \qquad p(1-p) \;+\; 4p(1-p)^2 \;+\; 9p(1-p)^3 \;+\; 16p(1-p)^4 \;+\; \cdots \\ p\mathbb{E}(X^2) &= p \;+\; 3p(1-p) \;+\; 5p(1-p)^2 \;+\; 7p(1-p)^3 \;+\; 9p(1-p)^4 \;+\; \cdots. \end{aligned}$$

Rearranging terms, we get:

$$\begin{aligned} &= 2[p + 2p(1-p) + 3p(1-p)^2 + 4p(1-p)^3 + 5p(1-p)^4 + \cdots] \\ &\quad - [p + p(1-p) + p(1-p)^2 + p(1-p)^3 + p(1-p)^4 + \cdots] \\ &= 2\mathbb{E}(X) - 1. \end{aligned}$$

In the last line, we used the fact that

$$\mathbb{E}(X) = p + 2p(1-p) + 3p(1-p)^2 + 4p(1-p)^3 + 5p(1-p)^4 + \cdots$$

and that

$$p + p(1-p) + p(1-p)^2 + p(1-p)^3 + p(1-p)^4 + \cdots = 1$$

is the sum of the probabilities of the outcomes in a geometric distribution. Thus, we get

$$p\mathbb{E}(X^2) = 2\mathbb{E}(X) - 1 = \frac{2}{p} - 1 = \frac{2-p}{p},$$

recalling that $\mathbb{E}(X) = \frac{1}{p}$ for a geometric r.v. Dividing both sides by p, we get

$$\mathbb{E}(X^2) = \frac{2-p}{p^2}.$$

Finally, we can compute the variance:

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

4. **Poisson distribution.** What is the variance of a Poisson r.v. X? We can calculate $\mathbb{E}(X^2)$:

$$\mathbb{E}(X^2) = \sum_{i=0}^{\infty} i^2 \mathrm{e}^{-\lambda} \frac{\lambda^i}{i!} = \lambda \sum_{i=1}^{\infty} i \mathrm{e}^{-\lambda} \frac{\lambda^{i-1}}{(i-1)!} = \lambda \left(\sum_{i=1}^{\infty} (i-1) \mathrm{e}^{-\lambda} \frac{\lambda^{i-1}}{(i-1)!} + \sum_{i=1}^{\infty} \mathrm{e}^{-\lambda} \frac{\lambda^{i-1}}{(i-1)!} \right) = \lambda(\lambda+1).$$

[Check you follow each of these steps. In the last step, the two sums are respectively $\mathbb{E}(X)$ and $\sum_i \Pr[X=i] = 1$.]

Finally, we get $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \lambda$. So, for a Poisson random variable, the expectation and variance are equal.

5. **Number of fixed points.** Let X be the number of fixed points in a random permutation of n items (i.e., the number of students in a class of size n who receive their own homework after shuffling). We saw in an earlier lecture that $\mathbb{E}(X) = 1$ (regardless of n). To compute $\mathbb{E}(X^2)$, write $X = X_1 + X_2 + \cdots + X_n$, where

$$X_i = \begin{cases} 1 & \text{if } i \text{ is a fixed point;} \\ 0 & \text{otherwise.} \end{cases}$$

Then as usual we have

$$\mathbb{E}(X^2) = \sum_{i=1}^{n} \mathbb{E}(X_i^2) + \sum_{i \neq j} \mathbb{E}(X_i X_j). \tag{1}$$

Since X_i is an indicator r.v., we have that $\mathbb{E}(X_i^2) = \Pr[X_i = 1] = \frac{1}{n}$. Since both X_i and X_j are indicators, we can compute $\mathbb{E}(X_i X_j)$ as follows:

$$\mathbb{E}(X_i X_j) = \Pr[X_i = 1 \cap X_j = 1] = \Pr[\text{both } i \text{ and } j \text{ are fixed points}] = \frac{1}{n(n-1)}.$$

[Check that you understand the last step here.] Plugging this into equation (1) we get

$$\mathbb{E}(X^2) = (n \times \tfrac{1}{n}) + (n(n-1) \times \tfrac{1}{n(n-1)}) = 1 + 1 = 2.$$

Thus $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = 2 - 1 = 1$. In other words, the variance and the mean are both equal to 1. Like the mean, the variance is also independent of n. Intuitively at least, this means that it is unlikely that there will be more than a small number of fixed points even when the number of items, n, is very large.

Chebyshev's Inequality

We have seen that, intuitively, the variance (or, more correctly the standard deviation) is a measure of "spread", or deviation from the mean. Our next goal is to make this intuition quantitatively precise. What we can show is the following:

Theorem 15.2: **[Chebyshev's Inequality]** For a random variable X with expectation $\mathbb{E}(X) = \mu$, and for any $\alpha > 0$,

$$\Pr[|X - \mu| \geq \alpha] \leq \frac{\mathrm{Var}(X)}{\alpha^2}.$$

Before proving Chebyshev's inequality, let's pause to consider what it says. It tells us that the probability of any given deviation, α, from the mean, either above it or below it (note the absolute value sign), is at most $\frac{\mathrm{Var}(X)}{\alpha^2}$. As expected, this deviation probability will be small if the variance is small. An immediate corollary of Chebyshev's inequality is the following:

Corollary 15.3: For a random variable X with expectation $\mathbb{E}(X) = \mu$, and standard deviation $\sigma = \sqrt{\mathrm{Var}(X)}$,

$$\Pr[|X - \mu| \geq \beta\sigma] \leq \frac{1}{\beta^2}.$$

Proof: Plug $\alpha = \beta\sigma$ into Chebyshev's inequality. □

So, for example, we see that the probability of deviating from the mean by more than (say) two standard deviations on either side is at most $\frac{1}{4}$. In this sense, the standard deviation is a good working definition of the "width" or "spread" of a distribution.

We should now go back and prove Chebyshev's inequality. The proof will make use of the following simpler bound, which applies only to *non-negative* random variables (i.e., r.v.'s which take only values ≥ 0).

Theorem 15.4: **[Markov's Inequality]** For a *non-negative* random variable X with expectation $\mathbb{E}(X) = \mu$, and any $\alpha > 0$,

$$\Pr[X \geq \alpha] \leq \frac{\mathbb{E}(X)}{\alpha}.$$

Proof: From the definition of expectation, we have

$$\begin{aligned}
\mathbb{E}(X) &= \sum_a a \times \Pr[X = a] \\
&= \sum_{a<\alpha} a \times \Pr[X = a] + \sum_{a \geq \alpha} a \times \Pr[X = a] \\
&\geq \sum_{a \geq \alpha} a \times \Pr[X = a] \\
&\geq \alpha \sum_{a \geq \alpha} \Pr[X = a] \\
&= \alpha \Pr[X \geq \alpha].
\end{aligned}$$

The crucial step here is the third line, where we have used the fact that X takes on only non-negative values and consequently $a \geq 0$ and $\Pr[X = a] \geq 0$. (This step would not be valid if X could be negative, since if X can take on negative values, then we might have $a \times \Pr[X = a] < 0$.) □

Now we can prove Chebyshev's inequality quite easily.

Proof of Theorem 15.2: Define the r.v. $Y = (X - \mu)^2$. Note that $\mathbb{E}(Y) = \mathbb{E}((X - \mu)^2) = \mathrm{Var}(X)$. Also, notice that the probability we are interested in, $\Pr[|X - \mu| \geq \alpha]$, is exactly the same as $\Pr[Y \geq \alpha^2]$. (This is

because the inequality $|X-\mu| \geq \alpha$ is true if and only if $(X-\mu)^2 \geq \alpha^2$ is true, i.e., if and only if $Y \geq \alpha^2$ is true.) Moreover, Y is obviously non-negative, so we can apply Markov's inequality to it to get

$$\Pr[Y \geq \alpha^2] \leq \frac{\mathbb{E}(Y)}{\alpha^2} = \frac{\mathrm{Var}(X)}{\alpha^2}.$$

This completes the proof. □

Let's apply Chebyshev's inequality to answer our question about the random walk at the beginning of this lecture note. Recall that X is our position after n steps, and that $\mathbb{E}(X) = 0$, $\mathrm{Var}(X) = n$. Corollary 15.3 says that, for any $\beta > 0$, $\Pr[|X| \geq \beta\sqrt{n}] \leq \frac{1}{\beta^2}$. Thus for example, if we take $n = 10^6$ steps, the probability that we end up more than 10000 steps away from our starting point is at most $\frac{1}{100}$.

Here are a few more examples of applications of Chebyshev's inequality (you should check the algebra in them):

1. **Coin tosses.** Let X be the number of Heads in n tosses of a fair coin. The probability that X deviates from $\mu = \frac{n}{2}$ by more than $\sqrt{n}$ is at most $\frac{1}{4}$. The probability that it deviates by more than $5\sqrt{n}$ is at most $\frac{1}{100}$.

2. **Poisson distribution.** Let X be a Poisson r.v. with parameter λ. The probability that X deviates from λ by more than $2\sqrt{\lambda}$ is at most $\frac{1}{4}$.

3. **Fixed points.** Let X be the number of fixed points in a random permutation of n items; recall that $\mathbb{E}(X) = \mathrm{Var}(X) = 1$. Thus the probability that more than (say) 10 students get their own homeworks after shuffling is at most $\frac{1}{100}$, however large n is[1].

[1] In more detail: $\Pr[X > 10] = \Pr[X \geq 11] \leq \Pr[|X-1] \geq 10] \leq \frac{\mathrm{Var}(X)}{100} = \frac{1}{100}$.

Polling and the Law of Large Numbers

Polling

Question: We want to estimate the proportion p of Democrats in the US population, by taking a small random sample. How large does our sample have to be to guarantee that our estimate will be within (say) 0.1 of the true value with probability at least 0.95?

This is perhaps the most basic statistical estimation problem, and it shows up everywhere. We will develop a simple solution that uses only Chebyshev's inequality. More refined methods can be used to get sharper results.

Let's denote the size of our sample by n (to be determined), and the number of Democrats in it by the random variable S_n. (The subscript n just reminds us that the r.v. depends on the size of the sample.) Then our estimate will be the value $A_n = \frac{1}{n}S_n$.

Now as has often been the case, we will find it helpful to write $S_n = X_1 + X_2 + \cdots + X_n$, where

$$X_i = \begin{cases} 1 & \text{if person } i \text{ in sample is a Democrat;} \\ 0 & \text{otherwise.} \end{cases}$$

Note that each X_i can be viewed as a coin toss, with Heads probability p (though of course we do not know the value of p). And the coin tosses are independent.[1] Hence, S_n is a binomial random variable with parameters n and p.

What is the expectation of our estimate?

$$\mathbb{E}(A_n) = \mathbb{E}(\tfrac{1}{n}S_n) = \tfrac{1}{n}\mathbb{E}(S_n) = \tfrac{1}{n} \times (np) = p.$$

So for any value of n, our estimate will always have the correct expectation p. [Such a r.v. is often called an *unbiased estimator* of p.] Now presumably, as we increase our sample size n, our estimate should get more and more accurate. This will show up in the fact that the *variance* decreases with n: i.e., as n increases, the probability that we are far from the mean p will get smaller.

To see this, we need to compute $\mathrm{Var}(A_n)$. But $A_n = \frac{1}{n}S_n$, which is just a constant times a binomial random variable.

Theorem 16.1: For any random variable X and constant c, we have

$$\mathrm{Var}(cX) = c^2\mathrm{Var}(X).$$

The proof of this theorem follows directly from the definition of the variance. (Try it yourself.) Now to compute $\mathrm{Var}(A_n)$:

$$\mathrm{Var}(A_n) = \mathrm{Var}(\tfrac{1}{n}S_n) = (\tfrac{1}{n})^2\mathrm{Var}(S_n) = (\tfrac{1}{n})^2\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n},$$

[1] We are assuming here that the sampling is done "with replacement"; i.e., we select each person in the sample from the entire population, including those we have already picked. So there is a small chance that we will pick the same person twice.

where we have written σ^2 for the variance of each of the X_i. The third equality follows from the calculation we did for the binomial random variable in the last lecture note. So we see that *the variance of A_n decreases linearly with n*. This fact ensures that, as we take larger and larger sample sizes n, the probability that we deviate much from the expectation p gets smaller and smaller.

Let's now use Chebyshev's inequality to figure out how large n has to be to ensure a specified accuracy in our estimate of the proportion of Democrats p. A natural way to measure this is for us to specify two parameters, ε and δ, both in the range $(0,1)$. The parameter ε controls the *error* we are prepared to tolerate in our estimate, and δ controls the *confidence* we want to have in our estimate. A more precise version of our original question is then the following:

Question: For the Democrat-estimation problem above, how large does the sample size n have to be in order to ensure that

$$\Pr[|A_n - p| \geq \varepsilon] \leq \delta \ ?$$

In our original question, we had $\varepsilon = 0.1$ and $\delta = 0.05$.

Let's apply Chebyshev's inequality to answer our more precise question above. Since we know $\mathrm{Var}(A_n)$, this will be quite simple. From Chebyshev's inequality, we have

$$\Pr[|A_n - p| \geq \varepsilon] \leq \frac{\mathrm{Var}(A_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

To make this less than the desired value δ, we need to set

$$n \geq \sigma^2 \times \frac{1}{\varepsilon^2 \delta}. \tag{1}$$

Now recall that $\sigma^2 = \mathrm{Var}(X_i)$ is the variance of a single sample X_i. So, since X_i is a 0/1-valued r.v., we have $\sigma^2 = p(1-p)$, and inequality (1) becomes

$$n \geq p(1-p) \times \frac{1}{\varepsilon^2 \delta}. \tag{2}$$

Plugging in $\varepsilon = 0.1$ and $\delta = 0.05$, we see that a sample size of $n = 2000p(1-p)$ is sufficient.

At this point you should be worried. Why? Because our formula for the sample size contains p, and this is precisely the quantity we are trying to estimate! But we can get around this. The largest possible value of $p(1-p)$ is $1/4$ (achieved when $p = 1/2$.) Hence, if we pick $n = 2000 \times (1/4) = 500$, then no matter what the value of p is, the sample size is sufficient.

Estimating a general expectation

What if we wanted to estimate something a little more complex than the proportion of Democrats in the population, such as the average wealth of people in the US? Then we could use exactly the same scheme as above, except that now the r.v. X_i is the wealth of the ith person in our sample. Clearly $\mathbb{E}(X_i) = \mu$, the average wealth (which is what we are trying to estimate). And our estimate will again be $A_n = \frac{1}{n}\sum_{i=1}^n X_i$, for a suitably chosen sample size n. We again have $\mathbb{E}(A_n) = \mu$. And as long as $\mathrm{Var}(\sum_{=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i)$, we have as before $\mathrm{Var}(A_n) = \frac{\sigma^2}{n}$, where $\sigma^2 = \mathrm{Var}(X_i)$ is the common variance of the X_i's. From equation (1), it is enough for the sample size n to satisfy

$$n \geq \sigma^2 \times \frac{1}{\varepsilon^2 \delta}. \tag{3}$$

Here ε and δ are the desired error and confidence respectively, as before. Now of course we don't know σ^2, appearing in equation (3). In practice, we would use an upper bound on σ^2 (just as we used the fact that

$\sigma^2 = p(1-p) \leq 1/4$ in the Democrats problem). Plugging these bounds into equation (3) will ensure that our sample size is large enough.

Let us recapitulate the three properties we used about the random variables X_i's in the above derivation:

(1) $\mathbb{E}(X_i) = \mu, i = 1, \ldots, n.$

(2) $\text{Var}(X_i) = \sigma^2, i = 1, \ldots, n.$

(3) $\text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i).$

The first two properties hold if the X_i's have the same distribution. The third property holds if the random variables are *mutually independent*. We have already defined the notion of independence for *events*, and we will define independence for *random variables* in the next lecture note. Intuitively, two random variables are independent if the events "associated" with them are independent. In the polling example when the X_i's are indicator random variables for flipping Heads, the random variables are independent if the flips are independent. Random variables which have the same distribution and are independent are called *independent identically distributed* (abbreviated as i.i.d.).

As a further example, suppose we are trying to estimate the average rate of emission from a radioactive source, and we are willing to assume that the emissions follow a Poisson distribution with some unknown parameter λ — of course, this λ is precisely the expectation we are trying to estimate. Now in this case we have $\sigma^2 = \lambda$ (see the previous lecture note), so a sample size of $n = \frac{\lambda}{\varepsilon^2 \delta}$ suffices. (Again, in practice we would use an upper bound on λ.)

The Law of Large Numbers

The estimation method we used in the previous two sections is based on a principle that we accept as part of everyday life: namely, the Law of Large Numbers (LLN). This asserts that, if we observe some random variable many times, and take the average of the observations, then this average will converge to a *single value*, which is of course the expectation of the random variable. In other words, averaging tends to smooth out any large fluctuations, and the more averaging we do the better the smoothing.

Theorem 16.2: **[Law of Large Numbers]** Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with common expectation $\mu = \mathbb{E}(X_i)$. Define $A_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then for any $\alpha > 0$, we have

$$\Pr[|A_n - \mu| \geq \alpha] \to 0 \qquad \text{as } n \to \infty.$$

Proof: Let $\text{Var}(X_i) = \sigma^2$ be the common variance of the r.v.'s; we assume that σ^2 is finite[2]. With this (relatively mild) assumption, the LLN is an immediate consequence of Chebyshev's Inequality. For, as we have seen above, $\mathbb{E}(A_n) = \mu$ and $\text{Var}(A_n) = \frac{\sigma^2}{n}$, so by Chebyshev we have

$$\Pr[|A_n - \mu| \geq \alpha] \leq \frac{\text{Var}(A_n)}{\alpha^2} = \frac{\sigma^2}{n\alpha^2} \to 0 \qquad \text{as } n \to \infty.$$

This completes the proof. □

Notice that the LLN says that the probability of *any* deviation α from the mean, however small, tends to zero as the number of observations n in our average tends to infinity. Thus by taking n large enough, we can make the probability of any given deviation as small as we like.

[2]If σ^2 is not finite, the LLN still holds but the proof is much trickier.

Multiple Random Variables and Applications to Inference

In many probability problems, we have to deal with *multiple* r.v.'s defined on the same probability space. We have already seen examples of that: for example, we saw that computing the expectation and variance of a binomial r.v. X is easier if we express it as a sum $X = \sum_{i=1}^{n} X_i$, where X_i represents the result of the ith trial. Multiple r.v.'s arise naturally in the case of inference problems, where we observe certain quantities and use our observations to draw inferences about other hidden quantities. This Note starts by developing some of the basics of handling multiple r.v.'s, then applies those concepts to several examples of inference problems.

Joint Distributions

Consider two random variables X and Y defined on the same probability space. By linearity of expectation, we know that $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$. Since $\mathbb{E}(X)$ can be calculated if we know the distribution of X and $\mathbb{E}(Y)$ can be calculated if we know the distribution of Y, this means that $\mathbb{E}(X+Y)$ can be computed knowing only the individual distributions of X and Y. In particular, to compute $\mathbb{E}(X+Y)$, no information is needed about the *relationship* between X and Y. However, this happy situation is unusual. For instance, consider the situation where we need to compute, say, $\mathbb{E}((X+Y)^2)$, as arose when we computed the variance of a binomial r.v. Now we need information about the association or relationship between X and Y, if we want to compute $\mathbb{E}((X+Y)^2)$. This is because $\mathbb{E}((X+Y)^2) = \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2)$, and $\mathbb{E}(XY)$ depends on the relationship between X and Y. How can we capture such a relationship, mathematically?

Recall that the distribution of a single random variable X is the collection of the probabilities of all events $X = a$, for all possible values of a that X can take on. When we have two random variables X and Y, we can think of (X,Y) as a "two-dimensional" random variable, in which case the events of interest are $X = a \cap Y = b$ for all possible values of (a,b) that (X,Y) can take on. Thus, a natural generalization of the notion of distribution to multiple random variables is the following.

Definition 17.1 (joint distribution): The *joint distribution* of two discrete random variables X and Y is the collection of values $\{(a,b,\Pr[X = a \cap Y = b]) : (a,b) \in \mathscr{A} \times \mathscr{B}\}$, where $\mathscr{A}$ and $\mathscr{B}$ are the sets of all possible values taken by X and Y respectively.

This notion obviously generalizes to three or more random variables. Since we will write $\Pr[X = a \cap Y = b]$ quite often, we will abbreviate it to $\Pr[X = a, Y = b]$.

Just like the distribution of a single random variable, the joint distribution is *normalized*, i.e.

$$\sum_{a \in \mathscr{A}, b \in \mathscr{B}} \Pr[X = a, Y = b] = 1.$$

This follows from noticing that the events $X = a \cap Y = b$ (where a ranges over $\mathscr{A}$ and b ranges over $\mathscr{B}$) partition the sample space.

The joint distribution between two random variables fully describes their statistical relationships, and provides enough information for computing any probabilities and expectations involving the two random vari-

Y \ X	0	1	2
0	0.1	0.2	0.15
1	0.05	0.05	0.2
2	0.1	0.1	0.05

Figure 1: A tabular representation of a joint distribution.

ables. For example,

$$\mathbb{E}(XY) = \sum_c c \times \Pr[XY = c] = \sum_a \sum_b ab \times \Pr[X = a, Y = b].$$

More generally, if f is any function on $\mathbb{R} \times \mathbb{R}$,

$$\mathbb{E}(f(X,Y)) = \sum_c c \times \Pr[f(X,Y) = c] = \sum_a \sum_b f(a,b) \times \Pr[X = a, Y = b].$$

Moreover, the individual distributions of X and Y can be recovered from the joint distribution as follows:

$$\Pr[X = a] = \sum_{b \in \mathscr{B}} \Pr[X = a, Y = b] \qquad \forall a \in \mathscr{A}, \tag{1}$$

$$\Pr[Y = b] = \sum_{a \in \mathscr{A}} \Pr[X = a, Y = b] \qquad \forall b \in \mathscr{B}. \tag{2}$$

The first follows from the fact that the events $Y = b$ (where b ranges over $\mathscr{B}$) form a partition of the sample space Ω, and so the events $X = a \cap Y = b$ (where b ranges over $\mathscr{B}$) are disjoint and their union yields the event $X = a$. The second fact follows for similar reasons.

Pictorially, one can think of the joint distribution values as entries filling a table, with the columns indexed by the values that X can take on and the rows indexed by the values Y can take on (see Figure 1). To get the distribution of X, all one needs to do is to sum the entries in each of the columns. To get the distribution of Y, just sum the entries in each of the rows. This process is sometimes called *marginalization* and the individual distributions are sometimes called *marginal* distributions to differentiate them from the joint distribution.

Independent Random Variables

Independence for random variables is defined in analogous fashion to independence for events:

Definition 17.2 (independent r.v.'s): Random variables X and Y on the same probability space are said to be *independent* if the events $X = a$ and $Y = b$ are independent for all values a, b. Equivalently, the joint distribution of independent r.v.'s decomposes as

$$\Pr[X = a, Y = b] = \Pr[X = a] \Pr[Y = b] \quad \forall a, b.$$

Note that for independent r.v.'s, the joint distribution is fully specified by the marginal distributions.

Mutual independence of more than two r.v.'s is defined similarly. A very important example of independent r.v.'s is a sequence of indicator r.v.'s for independent events. Thus, for example, if X_i is an indicator r.v. for the ith toss of a coin being Heads, then $X_1, X_2, \ldots, X_n$ are mutually independent r.v.'s.

We saw that the expectation of a sum of r.v.'s is the sum of the expectations of the individual r.v.'s. This is not true in general for variance. However, it turns out to be true if the random variables are independent. To see this, first we look at the expectation of a product of independent r.v.'s (which is a quantity that frequently shows up in variance calculations, as we have seen).

Theorem 17.1: For *independent* random variables X, Y, we have $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Proof: We have

$$\begin{aligned}
\mathbb{E}(XY) &= \sum_a \sum_b ab \times \Pr[X = a, Y = b] \\
&= \sum_a \sum_b ab \times \Pr[X = a] \times \Pr[Y = b] \\
&= \sum_a a \times \Pr[X = a] \times \sum_b b \times \Pr[Y = b] \\
&= \left(\sum_a a \times \Pr[X = a] \right) \times \left(\sum_b b \times \Pr[Y = b] \right) \\
&= \mathbb{E}(X) \times \mathbb{E}(Y),
\end{aligned}$$

as claimed. In the second line we made crucial use of independence. □

For example, this theorem would have allowed us to conclude immediately in our random walk example at the beginning of Lecture Note 15 that $\mathbb{E}(X_i X_j) = \mathbb{E}(X_i)\mathbb{E}(X_j) = 0$, without the need for a calculation.

We now use the above theorem to conclude a nice property of the variance of independent random variables.

Theorem 17.2: For *independent* random variables X, Y, we have $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Proof: From the alternative formula for variance in Theorem 15.1, we have, using linearity of expectation extensively,

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbb{E}((X+Y)^2) - (\mathbb{E}(X+Y))^2 \\
&= \mathbb{E}(X^2) + \mathbb{E}(Y^2) + 2\mathbb{E}(XY) - (\mathbb{E}(X) + \mathbb{E}(Y))^2 \\
&= (\mathbb{E}(X^2) - \mathbb{E}(X)^2) + (\mathbb{E}(Y^2) - \mathbb{E}(Y)^2) + 2(\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2(\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)).
\end{aligned}$$

Now *because X, Y are independent*, by Theorem 17.1 the final term in this expression is zero. Hence we get our result.

Here is an alternative proof. We can assume, without loss of generality, that $\mathbb{E}(X) = \mathbb{E}(Y) = \mathbb{E}(X+Y) = 0$. In this case, $\mathrm{Var}(X) = \mathbb{E}(X^2)$, $\mathrm{Var}(Y) = \mathbb{E}(Y^2)$, and

$$\mathrm{Var}(X+Y) = \mathbb{E}((X+Y)^2) = \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2) = \mathrm{Var}(X) + 2\mathbb{E}(X)\mathbb{E}(Y) + \mathrm{Var}(Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

(If $\mathbb{E}(X) \neq 0$ or $\mathbb{E}(Y) \neq 0$, replace X with $X' = X - \mathbb{E}(X)$ and Y with $Y' = Y - \mathbb{E}(Y)$, noting that substitution leaves all the variances unchanged. Then the above proof goes through.) □

Note: The expression $\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$ appearing in the above proof is called the *covariance* of X and Y, and is a measure of the dependence between X, Y. It is zero when X, Y are independent.

Theorem 17.2 can be used to simplify several of our variance calculations in Lecture Note 15. For example, in the random walk example of that Lecture Note, since the X_i are independent r.v.'s with $\mathrm{Var}(X_i) = 1$ for each i, we have $\mathrm{Var}(X) = \mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = 1 + \cdots + 1 = n \times 1 = n$. And in the biased coin example (example 2) the X_i are independent with $\mathrm{Var}(X_i) = p(1-p)$, so we have

$\text{Var}(X) = \text{Var}(X_1 + \cdots + X_n) = \text{Var}(X_1) + \cdots + \text{Var}(X_n) = np(1-p)$. Note, however, that we don't get any simplification in example 4 of Lecture Note 15 (fixed points) because the X_i are *not* independent.

It is very important to remember that **neither** Theorem 17.1 **nor** Theorem 17.2 is true in general, without the assumption that X, Y are independent. Here is a simple example. Let X be an indicator r.v. with $\Pr[X = 1] = p$. Now $\mathbb{E}(X \times X) = \mathbb{E}(X^2) = p$, but $\mathbb{E}(X) \times \mathbb{E}(X) = \mathbb{E}(X)^2 = p^2$, and these two values are not equal (because of course X and X are not independent!).

Another note of caution: it is *not* true that $\text{Var}(cX) = c\text{Var}(X)$ for a constant c. In fact, the following is true:

Theorem 17.3: For any random variable X and constant c, $\text{Var}(cX) = c^2\text{Var}(X)$.

Proof: From the definition of variance, we have

$$\text{Var}(cX) = \mathbb{E}((cX - \mathbb{E}(cX))^2) = \mathbb{E}((cX - c\mathbb{E}(X))^2) = \mathbb{E}(c^2(X - \mathbb{E}(X))^2) = c^2\text{Var}(X).$$

□

As an aside, we can contrast the following two cases.

- If $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables, and we let $X = X_1 + \cdots + X_n$ be their sum, then $\text{Var}(X) = \text{Var}(X_1) + \cdots + \text{Var}(X_n) = n \times \text{Var}(X_1)$.
- On the other hand, if we let $Y = X_1 + X_1 + \cdots + X_1 = nX_1$ be the sum of n copies of the same r.v., then $\text{Var}(Y) = n^2\,\text{Var}(X_1)$—which is much larger than $\text{Var}(X)$.

So if we sum n independent and identically distributed random variables, the standard deviation is proportional to $\sqrt{n}$, whereas if we sum n dependent random variables, the standard deviation can be much larger: it might be proportional to n.

Conditional Distribution and Expectation

In an earlier lecture, we discussed the concept of conditional probability of an event A given an event B. This concept allows us to define a notion of *conditional distribution* of a random variable given another random variable.

Definition 17.3 (conditional distribution): The *conditional distribution* of X given $Y = b$ is the collection of values $\{(a, \Pr[X = a|Y = b]) : a \in \mathscr{A}\}$, where $\mathscr{A}$ is the set of all possible values taken by X.

The conditional distribution can be calculated from the joint and marginal distributions:

$$\Pr[X = a|Y = b] = \frac{\Pr[X = a, Y = b]}{\Pr[Y = b]}.$$

It follows from eqn. (2) that

$$\sum_{a \in \mathscr{A}} \Pr[X = a|Y = b] = 1,$$

so the conditional distribution is normalized, just like a (unconditional) distribution. Note that if X and Y are independent r.v.'s, $\Pr[X = a|Y = b] = \Pr[X = a]$ for every a, b, i.e., the conditional and unconditional distributions of X are the same.

One can also naturally talk about the conditional distribution of multiple random variables. For example, the conditional distribution of X and Y given $Z = c$ is simply given by

$$\{(a, b, \Pr[X = a, Y = b|Z = c]) : a \in \mathscr{A}, b \in \mathscr{B}\}.$$

Conditional distributions have exactly the same properties as (unconditional) distributions. Therefore whatever we do with distributions we can do with conditional distributions. For example, one can compute expectations of conditional distributions. This leads to the concept of *conditional expectation*.

Definition 17.4 (conditional expectation): Let X and Y be two r.v.'s defined on the same probability space. The conditional expectation of X given $Y = b$ is defined to be:

$$\mathbb{E}(X|Y=b) := \sum_{a \in \mathscr{A}} a \times \Pr[X=a|Y=b].$$

Recall that conditional probabilities often help us to calculate probabilities of events by means of the total probability law. Similarly, conditional expectations are often useful to compute expectations via the *total expectation law*:

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{a \in \mathscr{A}} a \Pr[X=a] \\
&= \sum_{a \in \mathscr{A}} a \sum_{b \in \mathscr{B}} \Pr[Y=b] \Pr[X=a|Y=b] \\
&= \sum_{b \in \mathscr{B}} \Pr[Y=b] \sum_{a \in \mathscr{A}} a \Pr[X=a|Y=b] \\
&= \sum_{b \in \mathscr{B}} \Pr[Y=b] \times \mathbb{E}(X|Y=b).
\end{aligned}$$

This formula is quite intuitive: to calculate the expectation of the r.v. X, first calculate the conditional expectation of X given each of the various values of Y. Then sum them, weighted by the probabilities Y takes on the various values. This formula is another instance of the *divide into cases* strategy.

For example, let us use the total expectation law to give an alternative way of computing the expectation of a geometrically distributed r.v. X. (In Lecture Note 14 we did this in a different way.) Recall that X is the number of independent trials until we get our first success. Let Y be the indicator r.v. of the event that the first trial is successful. Using the total expectation law,

$$\begin{aligned}
\mathbb{E}(X) &= \Pr[Y=1] \times \mathbb{E}(X|Y=1) + \Pr[Y=0] \times \mathbb{E}(X|Y=0) \\
&= p\mathbb{E}(X|Y=1) + (1-p)\mathbb{E}(X|Y=0). \qquad (3)
\end{aligned}$$

Now, if $Y = 1$, the first trial is already successful, and $X = 1$ with probability 1. Hence, $\mathbb{E}(X|Y=1) = 1$. What about if $Y = 0$? If the first trial is unsuccessful, we are back to square one and have to continue trying. Hence the number of additional trials after the first trial is another geometric r.v. with the same parameter p, and $\mathbb{E}(X|Y=0) = 1 + \mathbb{E}(X)$. Substituting into (3), we get:

$$\mathbb{E}(X) = p + (1-p)(1+\mathbb{E}(X)).$$

Upon solving this equation, we get $\mathbb{E}(X) = \frac{1}{p}$.

It is interesting to see that while linearity of expectation was a useful tool to compute the expectation of a binomially distributed r.v., the total expectation rule is a more natural tool to compute the expectation of a geometrically distributed r.v. Both are tools that allow us to compute expectations without directly computing distributions.

Inference

One of the major uses of probability is to provide a systematic framework to perform *inference under uncertainty*. A few specific applications are:

- **communications**: Information bits are sent over a noisy physical channel (wireless, DSL phone line, etc.). From the received symbols, the recipient wants to make a decision about what bits were transmitted.
- **control**: A spacecraft needs to be landed on the moon. From noisy measurements by motion sensors, the spacecraft control software needs to estimate the current position of the spacecraft relative to the moon surface so that the craft can be maneuvered appropriately.
- **object recognition**: From an image containing an object, we would like automated tools to recognize what type of object it is.
- **speech recognition**: From hearing noisy utterances, one wants to recognize what is being said.
- **investing**: By observing past performance of a stock, one wants to estimate its intrinsic quality and hence make a decision on whether and how much to invest in it.

All of the above problems can be modeled with the following ingredients:

- There is a random variable X representing some hidden quantity, which cannot be directly observed but in which we are interested. X might be the value of an information bit in a communication scenario, position of the spacecraft in the control application, or the object class in the recognition problem.
- There are random variables $Y_1, Y_2, \ldots, Y_n$ representing the observations. They might be the outputs of a noisy channel at different times, the pixel values of an image, the values of the stocks on successive days, etc.
- The distribution of X, called the *prior* distribution. This can be interpreted as the knowledge about X *before* seeing the observations.
- The conditional distribution of $Y_1, \ldots, Y_n$ given X. This models the noise or randomness in the observations.

Since the observations are noisy, there is in general no hope of knowing what the *exact* value of X is given the observations. Instead, we would like to know which values of X are most likely, and how likely they are. All available knowledge about X can be summarized by the *conditional distribution* of X given the observations. We don't know what the exact value of X is, but the conditional distribution tells us which values of X are more likely and which are less likely. Based on this information, intelligent decisions can be made.

Inference Example 1: Multi-armed Bandits

Question: You walk into a casino. There are several slot machines (bandits). You know some slot machines have odds very favorable to you, some have less favorable odds, and some have very poor odds. However, you don't know which are which. You start playing on them, and by observing the outcomes, you want to learn which is which so that you can intelligently figure out which machine to play on (or not play at all, which may be the most intelligent decision).

Stripped-down version: Suppose there are n biased coins. Coin i has probability p_i of coming up Heads; however, you don't know which is which. You randomly pick one coin and flip it. If the coin comes up Heads you win \$1, and if it comes up Tails you lose \$1. What is the probability of winning? What is the probability of winning on the next flip given you have observed a Heads with this coin? Given you have observed two Heads in a row? Would you bet on the next flip?

Modeling using Random Variables

Let X be the coin randomly chosen, and Y_j be the indicator r.v. for the event that the jth flip of this randomly chosen coin comes up Heads. Since we don't know which coin we have chosen, X is the hidden quantity. The Y_j's are the observations.

Predicting the First Flip

The first question asks for $\Pr[Y_1 = 1]$. First we calculate the joint distribution of X and Y_1:

$$\Pr[X = i, Y_1 = H] = \Pr[X = i] \Pr[Y_1 = H | X = i] = \frac{p_i}{n}. \tag{4}$$

Applying (2), we get:

$$\Pr[Y_1 = H] = \sum_{i=1}^{n} \Pr[X = i, Y_1 = H] = \frac{1}{n} \sum_{i=1}^{n} p_i. \tag{5}$$

Note that combining the above two equations, we are in effect using the fact that:

$$\Pr[Y_1 = H] = \sum_{i=1}^{n} \Pr[X = i] \Pr[Y_1 = H | X = i]. \tag{6}$$

This is just the *Total Probability Rule* for events applied to random variables. Once you get familiar with this type of calculation, you can bypass the intermediate calculation of the joint distribution and directly write this down.

Predicting the Second Flip after Observing the First

Now, given that we observed $Y_1 = H$, we have learned something about the randomly chosen coin X. This knowledge is captured by the conditional distribution

$$\Pr[X = i | Y_1 = H] = \frac{\Pr[X = i, Y_1 = H]}{\Pr[Y_1 = H]} = \frac{p_i}{\sum_{j=1}^{n} p_j},$$

using eqns. (4) and (5).

Note that when we substitute eqn. (4) into the above equation, we are in effect using:

$$\Pr[X = i | Y_1 = H] = \frac{\Pr[X = i] \Pr[Y_1 = H | X = i]}{\Pr[Y_1 = H]}.$$

This is just Bayes' rule for events applied to random variables. Just as for events, this rule tells us how to update our knowledge, based on the observation. In particular, $\{(i, \Pr[X = i]) : i = 1, \ldots, n\}$ is the *prior distribution* of the hidden X; $\{(i, \Pr[X = i | Y_1 = H]) : i = 1, \ldots, n\}$ is the *posterior* distribution of X given the observation. Bayes' rule updates the prior distribution to yield the posterior distribution.

Now we can calculate the probability of Heads on the second flip, if we use the same coin we used for the first flip:

$$\Pr[Y_2 = H | Y_1 = H] = \sum_{i=1}^{n} \Pr[X = i | Y_1 = H] \Pr[Y_2 = H | X = i, Y_1 = H]. \tag{7}$$

This can be interpreted as the total probability rule (6) but in a new probability space *where we condition on the event* $Y_1 = H$. In other words, the original probability space uses $\Pr[\cdot]$ for its probability assignment;

the new probability space uses $\Pr[\cdot|Y_1 = H]$ for its probability assignment. Since the new probability space is, well, a probability space, all our formulas (such as the total probability rule) can be applied to the new probability space $\Pr[\cdot|Y_1 = H]$ and they will remain valid in this new context. (It is a good exercise to verify the formula (7) from first principles.)

Now let us calculate the various probabilities on the right hand side of (7). The probability $\Pr[X = i|Y_1 = H]$ is just the posterior distribution of X given the observation, which we have already calculated above. What about the probability $\Pr[Y_2 = H|X = i, Y_1 = H]$? There are two conditioning events: $X = i$ and $Y_1 = H$. But here is the thing: once we know that the unknown coin is coin i, then knowing the first flip is a Head is redundant and provides no further statistical information about the outcome of the second flip: the probability of getting a Heads on the second flip is just p_i. In other words,

$$\Pr[Y_2 = H|X = i, Y_1 = H] = \Pr[Y_2 = H|X = i] = p_i. \tag{8}$$

The events $Y_1 = H$ and $Y_2 = H$ are said to be independent *conditional on* the event $X = i$. Since in fact $Y_1 = a$ and $Y_2 = b$ are independent given $X = i$ for all a, b, i, we will say that the *random variables* Y_1 and Y_2 are independent given the *random variable* X.

Definition 17.5 (Conditional Independence): Two events A and B are said to be *conditionally independent* given a third event C if

$$\Pr[A \cap B|C] = \Pr[A|C] \times \Pr[B|C].$$

Two random variables X and Y are said to be *conditionally independent* given a third random variable Z if for every a, b, c,

$$\Pr[X = a, Y = b|Z = c] = \Pr[X = a|Z = c] \times \Pr[Y = b|Z = c].$$

Going back to our coin example, note that the r.v.'s Y_1 and Y_2 are definitely *not* independent. Knowing the outcome of Y_1 tells us some information about the identity of the coin (X) and hence allows us to infer something about Y_2. However, *if we already know* X, then the outcomes of the different flips Y_1 and Y_2 *are* independent. Therefore, Y_1 and Y_2 are conditionally independent given X.

Now substituting (8) into (7), we get the probability of winning using this coin in the second flip:

$$\Pr[Y_2 = H|Y_1 = H] = \sum_{i=1}^{n} \Pr[X = i|Y_1 = H] \Pr[Y_2 = H|X = i] = \frac{\sum_{i=1}^n p_i^2}{\sum_{i=1}^n p_i}.$$

It can be shown (using the Cauchy-Schwarz inequality) that $n\sum_i p_i^2 \geq (\sum_i p_i)^2$, which implies that

$$\Pr[Y_2 = H|Y_1 = H] = \frac{\sum_{i=1}^n p_i^2}{\sum_{i=1}^n p_i} \geq \frac{\sum_{i=1}^n p_i}{n} = \Pr[Y_1 = H].$$

Thus our observation of a Heads on the first flip increases the probability that the second toss is Heads. This, of course, is intuitively reasonable, because the posterior distribution puts larger weight on the coins with larger values of p_i.

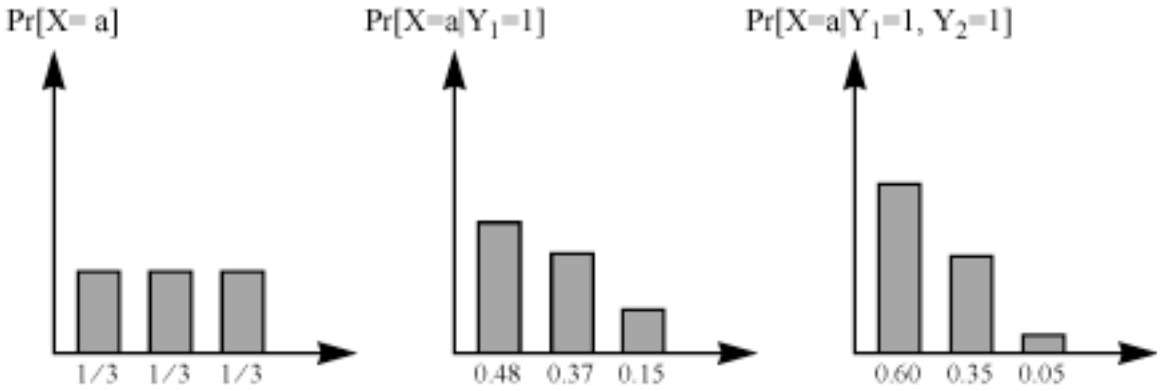


Figure 2: The conditional distributions of X given no observations, 1 Heads, and 2 Heads.

Predicting the Third Flip After Observing the First Two

Using Bayes' rule and the total probability rule, we can compute the posterior distribution of X given that we observed two Heads in a row:

$$\begin{aligned}
\Pr[X=i|Y_1=H,Y_2=H] &= \frac{\Pr[X=i]\Pr[Y_1=H,Y_2=H|X=i]}{\Pr[Y_1=H,Y_2=H]} \\
&= \frac{\Pr[X=i]\Pr[Y_1=H,Y_2=H|X=i]}{\sum_{j=1}^n \Pr[X=j]\Pr[Y_1=H,Y_2=H|X=j]} \\
&= \frac{\Pr[X=i]\Pr[Y_1=H|X=i]\Pr[Y_2=H|X=i]}{\sum_{j=1}^n \Pr[X=j]\Pr[Y_1=H|X=j]\Pr[Y_2=H|X=j]} \\
&= \frac{p_i^2}{\sum_{j=1}^n p_j^2}
\end{aligned}$$

The probability of getting a win on the third flip using the same coin is then:

$$\begin{aligned}
\Pr[Y_3=H|Y_1=H,Y_2=H] &= \sum_{i=1}^n \Pr[X=i|Y_1=H,Y_2=H]\Pr[Y_3=H|X=i,Y_1=H,Y_2=H] \\
&= \sum_{i=1}^n \Pr[X=i|Y_1=H,Y_2=H]\Pr[Y_3=H|X=i] \\
&= \frac{\sum_{i=1}^n p_i^3}{\sum_{i=1}^n p_i^2}.
\end{aligned}$$

Again, it can be shown that $\frac{\sum_{i=1}^n p_i^3}{\sum_{i=1}^n p_i^2} \geq \frac{\sum_{i=1}^n p_i^2}{\sum_{i=1}^n p_i}$, so the probability of seeing another Heads on the next flip has again increased. If we continue this process further (conditioning on having seen more and more Heads), the probability of Heads on the next flip will keep increasing towards the limit $p_{\max} = \max(p_1,\ldots,p_n)$.

As a numerical illustration, suppose $n=3$ and the three coins have Heads probabilities $p_1 = 2/3$, $p_2 = 1/2$, $p_3 = 1/5$. The conditional distributions of X after observing no flip, one Heads and two Heads in a row are shown in Figure 2. Note that as more Heads are observed, the conditional distribution is increasingly concentrated on coin 1 with $p_1 = 2/3$: we are increasingly certain that the coin chosen is the best coin. The corresponding probabilities of winning on the next flip after observing no flip, one Heads and two Heads in a row are 0.46, 0.54, and 0.58 respectively. The conditional probability of winning gets better and better (approaching 2/3 in the limit).

Iterative Update

A potential criticism of the method outline above is that it does not allow us to re-use past calculations, when we make a new observation. Suppose that we've observed that the first two coin tosses came up Heads, and calculated the conditional distribution on X given those observations. Now we observe the third coin toss is Heads, and want to update the distribution of X. It would be nice if we could re-use our past work. The method described in the past few pages has us start over from scratch; here is an alternative approach which allows us to update the distribution more efficiently.

Let's work with a simple, concrete example: say we have there are two possibilities for the Heads probability of the coin we select, $p_1 = 3/4$ or $p_2 = 1/2$. Previously, we assumed that the prior distribution on X was uniform, but let's remove that assumption and allow an arbitrary prior:

$$\Pr[X = 1] = q, \qquad \Pr[X = 2] = 1 - q.$$

Now suppose that we toss this coin and observe it to come up Heads. We can calculate the posterior distribution $\Pr[X = i|Y_1 = H]$:

$$\Pr[Y_1 = H] = \sum_j \Pr[Y_1 = H|X = j]\Pr[X = j] = \frac{3}{4}q + \frac{1}{2}(1-q) = \frac{1}{4}q + \frac{1}{2},$$

$$\Pr[X = 1|Y_1 = H] = \frac{\Pr[Y_1 = H|X = 1]\Pr[X = 1]}{\Pr[Y_1 = H]} = \frac{\frac{3}{4}q}{\frac{1}{4}q + \frac{1}{2}} = \frac{3q}{q+2},$$

$$\Pr[X = 2|Y_1 = H] = \frac{\Pr[Y_1 = H|X = 2]\Pr[X = 2]}{\Pr[Y_1 = H]} = \frac{\frac{1}{2}(1-q)}{\frac{1}{4}q + \frac{1}{2}} = \frac{2-2q}{q+2}.$$

So the distribution gets updated like this:

$$\begin{array}{c} (q, 1-q) \\ \Big\downarrow \quad \text{Update}_H \\ \left(\frac{3q}{q+2}, \frac{2-2q}{q+2}\right) \end{array}$$

For instance, if we had a uniform prior (both coins equally likely), after observing a single Heads the update rule would compute the posterior distribution as

$$\begin{array}{c} \left(\frac{1}{2}, \frac{1}{2}\right) \\ \Big\downarrow \quad \text{Update}_H \\ \left(\frac{3}{5}, \frac{2}{5}\right) \end{array}$$

since $\text{Update}_H(\frac{1}{2}, \frac{1}{2}) = (\frac{3/2}{5/2}, \frac{1}{5/2}) = (\frac{3}{5}, \frac{2}{5})$.

What if we toss the coin again, and see Heads again? Fortunately, the following procedure works. We treat the posterior distribution as our new prior, and apply the update rule we calculated above to this prior. For instance, if we saw a second Heads, we would compute the posterior like this:

$$\begin{array}{c} \left(\frac{1}{2}, \frac{1}{2}\right) \\ \Big\downarrow \quad \text{Update}_H \\ \left(\frac{3}{5}, \frac{2}{5}\right) \\ \Big\downarrow \quad \text{Update}_H \\ \left(\frac{9}{13}, \frac{4}{13}\right) \end{array}$$

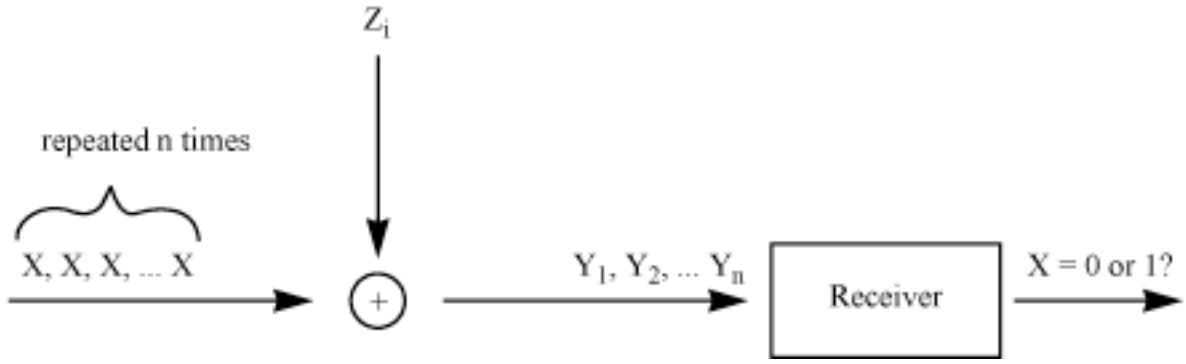


Figure 3: The system diagram for the communication problem.

since $\text{Update}_H(\frac{3}{5},\frac{2}{5}) = (\frac{9/5}{13/5},\frac{4/5}{13/5}) = (\frac{9}{13},\frac{4}{13})$. In other words, after tossing the coin twice and seeing two Heads, there is a $\frac{9}{13}$ probability that we've been flipping coin 1 (with $3/4$ Heads probability), and a $\frac{4}{13}$ probability that we've been flipping coin 2 (the fair coin).

It may be surprising that we can separate the computation into two individual update steps. However, it can be proven that this always works, whenever the observations are conditionally independent given the hidden variable[1]. Formally, $\text{Update}_{HH}(q,1-q) = \text{Update}_H(\text{Update}_H(q,1-q))$. The ability to break the calculation down into individual steps that update the distribution as each observation arrives is known as *recursive Bayesian estimation*[2].

In many practical applications, this is an important optimization, because it means we don't need to retain a history of all past observations: we can simply keep track of the current distribution (conditioned upon all the observations we've seen so far) and update the distribution whenever we make a new observation. Also, the incremental approach can be computationally more efficient, since we do only a constant amount of computation per observation.

Inference Example 2: Communication over a Noisy Channel

Question: I have one bit of information that I want to communicate over a noisy channel. The noisy channel flips each one of my transmitted symbols independently with probability $p < 0.5$. How much improvement in reliability do I get by repeating my transmission n times?

Comment: In an earlier lecture note, we also considered a communication problem and gave some examples of error-correcting codes. However, the models for the communication channel are different. There, we put a bound on the maximum number of flips (errors) the channel can make. Here, we do not put such bounds *a priori* but instead impose a bound on the *probability* that each bit is flipped (so that the *expected* number of bits flipped is np). Since there is no bound on the maximum number of flips the channel can make, there is no guarantee that the receiver will always decode correctly. Instead, one has to be satisfied with being able to decode correctly *with high probability*, e.g., probability of error < 0.01.

[1] In this specific example, we can do a simple sanity check on the answer we obtained this way. Note that $\Pr[Y_1 = H] = (3/4+1/2)/2 = 5/8$ and $\Pr[Y_1 = H, Y_2 = H] = ((3/4)^2+(1/2)^2)/2 = 13/32$. Therefore, $\Pr[X = 1|Y_1 = H, Y_2 = H] = \frac{(3/4)^2/2}{13/32} = \frac{9}{13}$ and $\Pr[X = 2|Y_1 = H, Y_2 = H] = \frac{(1/2)^2/2}{13/32} = \frac{4}{13}$, which matches what we calculated using two update steps.

[2] "Iterative Bayesian estimation" might be just as good a name, but for whatever reason, "recursive Bayesian estimation" is the standard term, despite the lack of any fundamental reliance upon recursion.

Modeling

The situation is shown in Figure 3.

Let X (= 0 or 1) be the value of the information bit I want to transmit. Assume that X is equally likely to be 0 or 1 (this is the prior distribution). The received symbol on the ith repetition of X is

$$Y_i = X + Z_i \mod 2, \qquad i = 1, 2, \ldots, n$$

with $Z_i = 1$ with probability p and $Z_i = 0$ with probability $1 - p$. Note that Y_i is different from X if and only if $Z_i = 1$. Thus, the transmitted symbol is flipped with probability p. The Z_i's are assumed to be mutually independent across different repetitions of X and also independent of X. The Z_i's can be interpreted as *noise*.

Note that the received symbols Y_i's are *not* independent; they all contain information about the transmitted bit X. However, *given* X, they are (conditionally) independent since they then only depend on the noise Z_i.

Decision Rule

First, we have to figure out what *decision rule* to use at the receiver, i.e., given each of the 2^n possible received sequences, $Y_1 = b_1, Y_2 = b_2, \ldots, Y_n = b_n$, how should the receiver guess what value of X was transmitted?

A natural rule is the *maximum a posteriori* (MAP) rule: guess the value a^* for which the conditional probability of $X = a^*$ given the observations is the largest among all a. More explicitly:

$$a^* = \begin{cases} 0 & \text{if } \Pr[X = 0 | Y_1 = b_1, \ldots, Y_n = b_n] \geq \Pr[X = 1 | Y_1 = b_1, \ldots Y_n = b_n] \\ 1 & \text{otherwise.} \end{cases}$$

Now, let's reformulate this rule so that it looks cleaner. By Bayes' rule, we have

$$\Pr[X = 0 | Y_1 = b_1, \ldots, Y_n = b_n] = \frac{\Pr[X = 0] \Pr[Y_1 = b_1, \ldots, Y_n = b_n | X = 0]}{\Pr[Y_1 = b_1, \ldots, Y_n = b_n]} \tag{9}$$

$$= \frac{\Pr[X = 0] \Pr[Y_1 = b_1 | X = 0] \Pr[Y_2 = b_2 | X = 0] \ldots, \Pr[Y_n = b_n | X = 0]}{\Pr[Y_1 = b_1, \ldots, Y_n = b_n]} \tag{10}$$

In the second step, we are using the fact that the observations Y_i's are conditionally independent given X. (Do you see why this is true?) Similarly,

$$\Pr[X = 1 | Y_1 = b_1, \ldots Y_n = b_n] = \frac{\Pr[X = 1] \Pr[Y_1 = b_1, \ldots, Y_n = b_n | X = 1]}{\Pr[Y_1 = b_1, \ldots, Y_n = b_n]} \tag{11}$$

$$= \frac{\Pr[X = 1] \Pr[Y_1 = b_1 | X = 1] \Pr[Y_2 = b_2 | X = 1] \ldots \Pr[Y_n = b_n | X = 1]}{\Pr[Y_1 = b_1, \ldots, Y_n = b_n]}. \tag{12}$$

An equivalent way of describing the MAP rule is that it computes the ratio of these conditional probabilities and checks if it is greater than or less than 1. If it is greater than (or equal to) 1, then guess that a 0 was transmitted; otherwise guess that a 1 was transmitted. (This ratio indicates how likely a 0 is compared to a 1, and is called the *likelihood ratio*.) Dividing (10) and (12), and recalling that we are assuming $\Pr[X = 1] = \Pr[X = 0]$, the likelihood ratio L is:

$$L = \prod_{i=1}^{n} \frac{\Pr[Y_i = b_i | X = 0]}{\Pr[Y_i = b_i | X = 1]}. \tag{13}$$

Note that we didn't have to compute $\Pr[Y_1 = b_1, \ldots, Y_n = b_n]$, since it appears in both of the conditional probabilities and got canceled out when computing the ratio.

Now,

$$\frac{\Pr[Y_i = b_i | X = 0]}{\Pr[Y_i = b_i | X = 1]} = \begin{cases} \frac{p}{1-p} & \text{if } b_i = 1 \\ \frac{1-p}{p} & \text{if } b_i = 0. \end{cases}$$

Therefore, if n_0 counts the number of 0's received and n_1 the number of 1's received, we have

$$L = \left(\frac{1-p}{p}\right)^{n_0} \left(\frac{p}{1-p}\right)^{n_1} = \left(\frac{1-p}{p}\right)^{n_0 - n_1}.$$

In other words, L has a factor of $p/(1-p)$ for every 1 received and a factor of $(1-p)/p$ for every 0 received. The former factor is < 1; the latter factor is > 1. So the likelihood ratio L is greater than 1 if and only if the number of 0's is greater than the number of 1's. Thus, the decision rule is simply a *majority* rule: guess that a 0 was transmitted if the number of 0's in the received sequence is at least as large as the number of 1's, otherwise guess that a 1 was transmitted.

Note that in deriving this rule, we assumed that the prior distribution on X is uniform, i.e., $\Pr[X = 0] = \Pr[X = 1] = 0.5$. When the prior distribution on X is not uniform, the MAP rule is no longer a simple majority rule. It is a good exercise to derive the MAP rule in the general case.

Error Probability Analysis

What is the probability that the guess is incorrect? This is just the event E that the number of flips by the noisy channel is greater than $n/2$. So the error probability of our majority rule is:

$$\Pr[E] = \Pr\left[\sum_{i=1}^{n} Z_i > \frac{n}{2}\right] = \sum_{k=\lceil n/2 \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k},$$

recognizing that the random variable $S := \sum_{i=1}^{n} Z_i$ has a binomial distribution with parameters n and p.

This gives an expression for the probability that the bit is received incorrectly. This probability can be numerically evaluated for given values of n. Given a target error probability of, say, 0.01, one can then compute the smallest number of repetitions needed to achieve the target error probability.[3]

However, the formula above does not give us much intuition about how the error probability varies as a function of n. As in the hashing application we looked at earlier in the course, we are interested in a more explicit relationship between n and the error probability to get a better intuition for the problem. The above expression is too cumbersome for this purpose. Instead, notice that $n/2$ is greater than the mean np of S and hence the error event is related to the tail of the distribution of S. One can therefore apply Chebyshev's inequality to bound the error probability:

$$\Pr[S > \frac{n}{2}] \leq \Pr[|S - np| > n(\frac{1}{2} - p)] \leq \frac{\mathrm{Var}(S)}{n^2(\frac{1}{2} - p)^2} = \frac{p(1-p)}{(\frac{1}{2} - p)^2} \cdot \frac{1}{n}.$$

The important thing to note is that the error probability decreases with n, so indeed by repeating more times the performance improves (as one would expect!). For a given target error probability of, say, 0.01, one needs to repeat no more than

$$100 \cdot \frac{(\frac{1}{2} - p)^2}{p(1-p)}$$

[3] Needless to say, one does not want to repeat more times than is necessary, as the more bits you send, the longer it takes, so repeating more than necessary will cause unnecessary slowdowns in communication.

times. For $p = 0.25$, this evaluates to 300 repetitions.

It is possible to compare the bound above with the actual error probability. If you do so, you will see that the bound above is rather pessimistic, and actually one can achieve an error probability of 0.01 with many fewer repetitions. In an upper-division course such as CS 174 or EE 126, you can learn about much better bounds on error probabilities like this.

A Brief Introduction to Continuous Probability

Up to now we have focused exclusively on *discrete* probability spaces Ω, where the number of sample points $\omega \in \Omega$ is either finite or countably infinite (such as the integers). As a consequence we have only been able to talk about *discrete* random variables, which take on only a finite (or countably infinite) number of values.

But in real life many quantities that we wish to model probabilistically are *continuous-valued*; examples include the position of a particle in a box, the time at which a certain incident happens, or the direction of travel of a meteorite. In this lecture, we discuss how to extend the concepts we've seen in the discrete setting to this continuous setting. As we shall see, everything translates in a natural way once we have set up the right framework. The framework involves some elementary calculus.

Continuous uniform probability spaces

Suppose we spin a "wheel of fortune" and record the position of the pointer on the outer circumference of the wheel. Assuming that the circumference is of length ℓ and that the wheel is unbiased, the position is presumably equally likely to take on any value in the real interval $[0,\ell]$. How do we model this experiment using a probability space?

Consider for a moment the (almost) analogous discrete setting, where the pointer can stop only at a finite number m of positions distributed evenly around the wheel. (If m is very large, then presumably this is in some sense similar to the continuous setting.) Then we would model this situation using the discrete sample space $\Omega = \{0, \frac{\ell}{m}, \frac{2\ell}{m}, \ldots, \frac{(m-1)\ell}{m}\}$, with uniform probabilities $\Pr[\omega] = \frac{1}{m}$ for each $\omega \in \Omega$. In the continuous world, however, we get into trouble if we try the same approach. If we let ω range over all real numbers in $[0,\ell]$, what value should we assign to each $\Pr[\omega]$? By uniformity this probability should be the same for all ω, but then if we assign to it any positive value, the sum of all probabilities $\Pr[\omega]$ for $\omega \in \Omega$ will be ∞! Thus $\Pr[\omega]$ must be zero for all $\omega \in \Omega$. But if all of our sample points have probability zero, then we are unable to assign meaningful probabilities to any events!

To rescue this situation, consider instead any non-empty *interval* $[a,b] \subseteq [0,\ell]$. Can we assign a non-zero probability value to this interval? Since the total probability assigned to $[0,\ell]$ must be 1, and since we want our probability to be uniform, the natural assignment of probability to the interval $[a,b]$ is

$$\Pr[[a,b]] = \frac{\text{length of } [a,b]}{\text{length of } [0,\ell]} = \frac{b-a}{\ell}. \tag{1}$$

In other words, the probability of an interval is proportional to its length.

Note that intervals are subsets of the sample space Ω and are therefore *events*. So in continuous probability, we are assigning probabilities to certain basic events, in contrast to discrete probability, where we assigned probability to *points* in the sample space. But what about probabilities of other events? Actually, by specifying the probability of intervals we have also specified the probability of any event E which can be written as the disjoint union of (a finite or countably infinite number of) intervals, $E = \cup_i E_i$. For then we can write $\Pr[E] = \sum_i \Pr[E_i]$, in analogous fashion to the discrete case. Thus for example the probability that the pointer

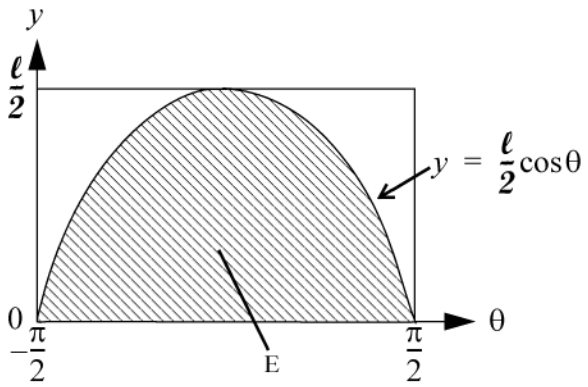


Figure 1: The event E as a subset of the sample space Ω.

ends up in the first or third quadrants of the wheel is $\frac{\ell/4}{\ell} + \frac{\ell/4}{\ell} = \frac{1}{2}$. For all practical purposes, such events are all we really need.[1]

An example: Buffon's needle

Here is a simple application of continuous probability to the analysis of a classical procedure for estimating the value of π known as *Buffon's needle*, after its 18th century inventor Georges-Louis Leclerc, Comte de Buffon.

Here we are given a needle of length ℓ, and a board ruled with horizontal lines at distance ℓ apart. The experiment consists of throwing the needle randomly onto the board and observing whether or not it crosses one of the lines. We shall see below that (assuming a perfectly random throw) the probability of this event is exactly $2/\pi$. This means that, if we perform the experiment many times and record the *proportion* of throws on which the needle crosses a line, then the Law of Large Numbers (Lecture Note 16) tells us that we will get a good estimate of the quantity $2/\pi$, and therefore also of π; and we can use Chebyshev's inequality as in the other estimation problems we considered in that same Lecture Note to determine how many throws we need in order to achieve specified accuracy and confidence.

To analyze the experiment, we first need to specify the probability space. Note that the position where the needle lands is completely specified by two numbers: the vertical distance y between the midpoint of the needle and the closest horizontal line, and the angle θ between the needle and the vertical. The vertical distance y ranges between 0 and $\ell/2$, while θ ranges between $-\pi/2$ and $\pi/2$. Thus, the sample space is the rectangle $\Omega = [-\pi/2, \pi/2] \times [0, \ell/2]$. Note that, compared to the wheel-of-fortune example, the sample space is two-dimensional rather than one-dimensional. But like the wheel-of-fortune example, the sample space is also continuous.

Now let E denote the event that the needle crosses a line. It is a subset of the sample space Ω. We need to identify this subset explicitly. By elementary geometry the vertical distance of the endpoint of the needle from its midpoint is $\frac{\ell}{2}\cos\theta$, so the needle will cross the line if and only if $y \leq \frac{\ell}{2}\cos\theta$. The event E is sketched in Figure 1.

Since we are assuming a completely random throw, probability of the event E is:

$$\Pr[E] = \frac{\text{area of } E}{\text{area of } \Omega}.$$

[1] A formal treatment of which events can be assigned a well-defined probability requires a discussion of *measure theory*, which is beyond the scope of this course.

This is the two-dimensional generalization of equation (1) in the wheel-of-fortune example, where the probability of landing in an interval is proportional to the length of the interval.

The area of the whole sample space is $\pi\ell/2$. The area of E is:

$$\int_{-\pi/2}^{\pi/2} \frac{\ell}{2}\cos\theta\, d\theta = \left[\frac{\ell}{2}\sin\theta\right]_{-\pi/2}^{\pi/2} = \ell.$$

Hence,

$$\Pr[E] = \frac{\ell}{\pi\ell/2} = \frac{2}{\pi}.$$

This is exactly what we claimed at the beginning of the section!

Continuous random variables

Recall that in the discrete setting we typically work with *random variables* and their distributions, rather than directly with probability spaces and events. This is even more so in continuous probability, since numerical quantities are almost always involved. In the wheel-of-fortune example, the position X of the pointer is a random variable. In the Buffon needle example, the vertical distance Y and the angle Θ are random variables. In fact, they are all *continuous* random variables. These random variables are all relatively simple, in the sense that they are all *uniformly* distributed on the range of values they can take on. (Because of this simplicity, we didn't even need to worry about the random variables explicitly when calculating probabilities in these examples, and instead reason directly with the sample space.) But more complicated random variables do not have a uniform distribution. How, precisely, should we define the distribution of a general continuous random variable? In the discrete case the distribution of a r.v. X is described by specifying, for each possible value a, the probability $\Pr[X = a]$. But for the r.v. X corresponding to the position of the pointer, we have $\Pr[X = a] = 0$ for every a, so we run into the same problem as we encountered above in defining the probability space.

The resolution is essentially the same: instead of specifying $\Pr[X = a]$, we instead specify $\Pr[a < X \leq b]$ for all intervals $[a,b]$.[2] To do this formally, we need to introduce the concept of a *probability density function* (sometimes referred to just as a "density", or a "pdf").

Definition 18.1 (Density): A *probability density function* for a random variable X is a function $f : \mathbb{R} \to \mathbb{R}$ satisfying

$$\Pr[a < X \leq b] = \int_a^b f(x)dx \qquad \text{for all } a \leq b. \tag{2}$$

Let's examine this definition. Note that the definite integral is just the area under the curve f between the values a and b (Figure 2(a)). Thus f plays a similar role to the "histogram" we sometimes draw to picture the distribution of a discrete random variable.

In order for the definition to make sense, f must obey certain properties. Some of these are technical in nature, which basically just ensure that the integral is always well defined; we shall not dwell on this issue here since all the densities that we will meet will be well behaved. What about some more basic properties of f? First, it must be the case that f is a non-negative function; for if f took on negative values we could find an interval in which the integral is negative, so we would have a negative probability for some event! Second, since the r.v. X must take on some value everywhere in the space, we must have

$$\int_{-\infty}^{\infty} f(x)dx = \Pr[-\infty < X < \infty] = 1. \tag{3}$$

[2] Note that it does not matter whether or not we include the endpoints a,b; since $\Pr[X = a] = \Pr[X = b] = 0$, we have $\Pr[a < X < b] = \Pr[a < X \leq b] = \Pr[a < X < b]$.

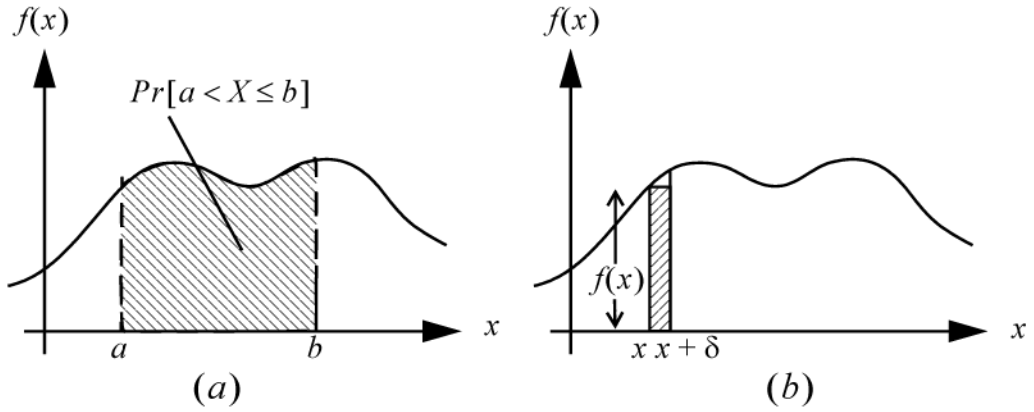


Figure 2: (a) The area under the density curve between a and b is the probability that the random variable lies in that range. (b) For small δ, the area under the curve between x and $x+\delta$ can be well approximated by the area of the rectangle of height $f(x)$ and width δ.

In other words, the total area under the curve f must be 1.

A caveat is in order here. Following the "histogram" analogy above, it is tempting to think of $f(x)$ as a "probability." However, $f(x)$ doesn't itself correspond to the probability of anything! For one thing, there is no requirement that $f(x)$ be bounded by 1 (and indeed, we shall see examples of densities in which $f(x)$ is greater than 1 for some x). To connect $f(x)$ with probabilities, we need to look at a very small interval $[x, x+\delta]$ close to x. Assuming that the interval $[x, x+\delta]$ is so small that the function f doesn't change much over that interval. we have

$$\Pr[x < X \leq x+\delta] = \int_x^{x+\delta} f(z)dz \approx \delta f(x). \tag{4}$$

This approximation is illustrated in Figure 2(b). Equivalently,

$$f(x) \approx \frac{\Pr[x < X \leq x+\delta]}{\delta}. \tag{5}$$

The approximation in (5) becomes more accurate as δ becomes small. Hence, more formally, we can relate density and probability by taking limits:

$$f(x) = \lim_{\delta \to 0} \frac{\Pr[x < X \leq x+\delta]}{\delta}. \tag{6}$$

Thus we can interpret $f(x)$ as the "probability per unit length" in the vicinity of x. Note that while the equation (2) allows us to compute probabilities given the probability density function, the equation (6) allows us to compute the probability density function given probabilities. Both relationships are useful in problems.

Now let's go back and put our wheel-of-fortune r.v. X into this framework. What should be the density of X? Well, we want X to have non-zero probability only on the interval $[0, \ell]$, so we should certainly have $f(x) = 0$ for $x < 0$ and for $x > \ell$. Within the interval $[0, \ell]$ we want the distribution of X to be uniform, which means we should take $f(x) = c$ for $0 \leq x \leq \ell$. What should be the value of c? This is determined by the requirement (3) that the total area under f is 1. The area under the above curve is $\int_{-\infty}^{\infty} f(x)dx = \int_0^{\ell} c dx = c\ell$, so we must take $c = \frac{1}{\ell}$. Summarizing, then, the density of the uniform distribution on $[0, \ell]$ is given by

$$f(x) = \begin{cases} 0 & \text{for } x < 0; \\ 1/\ell & \text{for } 0 \leq x \leq \ell; \\ 0 & \text{for } x > \ell. \end{cases}$$

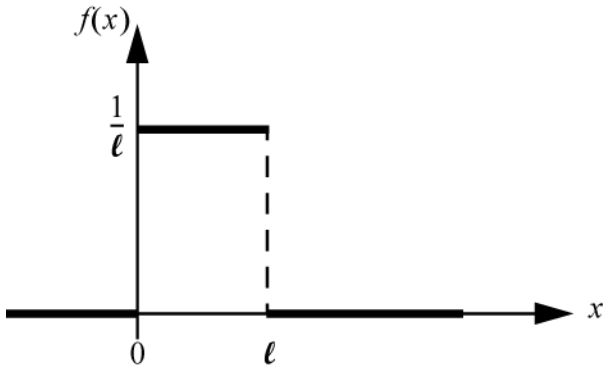


Figure 3: The density function of the wheel-of-fortune r.v. X.

This is plotted in Figure 3. Note that $f(x)$ can certainly be greater than 1, depending on the value of ℓ.

Another Example: Suppose you throw a dart and it lands uniformly at random on a target which is a disk of unit radius. What is the probability density function of the distance of the dart from the center of the disk?

Let X be the distance of the dart from the center of the disk. We first calculate the probability that X is between x and $x+\delta$. If x is negative or greater than or equal to 1, this probability is zero, so we focus on the case that x is between 0 and 1. The event in question is that the dart lands in the ring (annulus) shown in Figure 4. Since the dart lands uniformly at random on the disk, the probability of the event is just the ratio of the area of the ring and the area of the disk. Hence,

$$\begin{aligned} \Pr[x < X \le x+\delta] &= \frac{\pi[(x+\delta)^2 - x^2]}{\pi(1)^2} \\ &= x^2 + 2\delta x + \delta^2 - x^2 = 2\delta x - \delta^2. \end{aligned} \tag{7}$$

Using equation (6), we can now compute the probability density function of X:

$$f(x) = \lim_{\delta \to 0} \frac{\Pr[x < X \le x+\delta]}{\delta} = f(x) = \lim_{\delta \to 0} \frac{2\delta x - \delta^2}{\delta} = 2x.$$

Summarizing, we have

$$f(x) = \begin{cases} 0 & \text{for } x \le 0; \\ 2x & \text{for } 0 \le x < 1; \\ 0 & \text{for } x \ge 1. \end{cases}$$

It is plotted in Figure 5(a). Note that although the dart lands uniformly inside the target, the distance X from the center is *not* uniformly distributed in the range from 0 to 1. This is because an ring farther away from the center has a larger area than an ring closer to the center with the same width δ. Hence the probability of landing in the ring farther away from the center is larger.

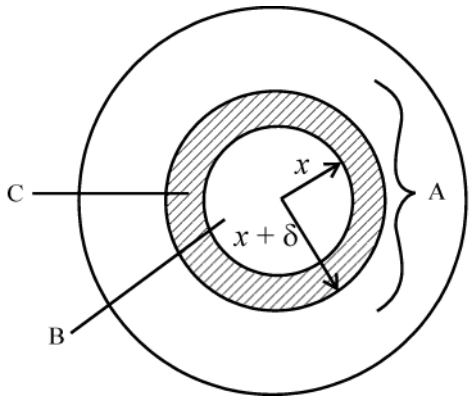


Figure 4: The sample space is the disk of unit radius. The event of interest is the ring.

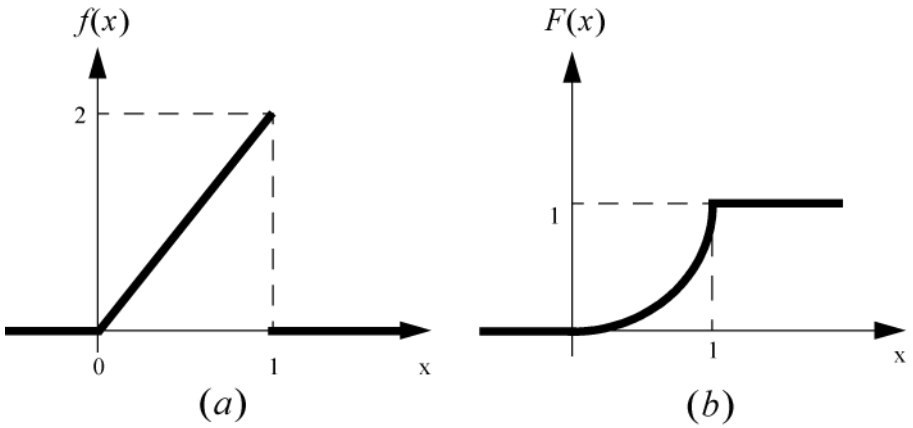


Figure 5: (a) The probability density function and (b) the cumulative distribution function of the distance X from the target center.

Cumulative Distribution Function

Let us re-interpret equation (7) in the dart throwing example above. In words, we are saying:

$$\begin{aligned}
\Pr[x < X \leq x+\delta] &= \frac{\text{area of ring}}{\text{area of target}} \\
&= \frac{(\text{area of disk of radius } x+\delta) \ - \ (\text{area of disk of radius } x)}{\text{area of target}} \\
&= \frac{\text{area of disk of radius } x+\delta}{\text{area of target}} - \frac{\text{area of disk of radius } x}{\text{area of target}} \\
&= \Pr[X \leq x+\delta] - \Pr[X \leq x].
\end{aligned}$$

This last equality can be understood directly as follows. The event A that $X \leq x+\delta$ (dart lands inside disk of radius $x+\delta$) can be decomposed as a union of two events: 1) the event B that $X \leq x$ (dart lands inside disk of radius x), and 2) the event C that $x < X \leq x+\delta$ (dart lands inside ring). The two events are *disjoint*. (See Figure 4.) Hence,

$$\Pr[A] = \Pr[B] + \Pr[C]$$

or

$$\Pr[x < X \leq x+\delta] = \Pr[X \leq x+\delta] - \Pr[X \leq x], \tag{8}$$

which is exactly the same as above.

Clearly, the reasoning leading to (8) has nothing much to do the particulars of this example but in fact (8) holds true for *any* random variable X. All we needed are the facts that $A = B \cup C$ and B and C are disjoint events, and the facts are true in general.

Substituting (8) into (6), we obtain:

$$f(x) = \lim_{\delta \to 0} \frac{\Pr[X \leq x+\delta] - \Pr[X \leq x]}{\delta}.$$

What does this equation remind you of? To make things even more explicit, let us define the function

$$F(x) = \Pr[X \leq x]. \tag{9}$$

Then we have:

$$f(x) = \lim_{\delta \to 0} \frac{F(x+\delta) - F(x)}{\delta} = \frac{d}{dx} F(x). \tag{10}$$

This equation provides an alternative way of solving the dart throwing example by first computing the function F and then differentiating F to get the probability density function. This way, we can avoid taking limits. In that example, F is given by (please check!):

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0; \\ x^2 & \text{for } 0 \leq x < 1; \\ 1 & \text{for } x \geq 1. \end{cases}$$

This function is plotted in Figure 5(b).

The function F has a name: it is called the *cumulative distribution function* of the random variable X (sometimes abbreviated as cdf). It is called "cumulative" because at each value x, $F(x)$ is the cumulative probability up to x. Note that the cumulative distribution function and the probability density function of a random variable contains *exactly* the same information. Given the cumulative distribution function F, one

can differentiate to get the probability density function f. Given the probability density function f, one can integrate to get the cumulative distribution function:

$$F(x) = \int_{-\infty}^{x} f(a)da.$$

So strictly speaking, one does not need to introduce the concept of cumulative distribution function at all. However, for many problems, the cumulative distribution function is easier to compute first and from that one can then compute the probability density function.

Summarizing, we have:

Definition 18.2 (Cumulative Distribution Function): The *cumulative distribution function* for a random variable X is a function $F : \mathbb{R} \to \mathbb{R}$ defined to be:

$$F(x) = \Pr[X \leq x]. \tag{11}$$

Its relationship with the probability density function f of X is given by

$$f(x) = \frac{d}{dx}F(x), \qquad F(x) = \int_{-\infty}^{x} f(a)da.$$

Expectation and variance of a continuous random variable

By analogy with the discrete case, we define the expectation of a continuous r.v. as follows:

Definition 18.3 (Expectation): The expectation of a continuous random variable X with probability density function f is

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

Note that the integral plays the role of the summation in the discrete formula $\mathbb{E}(X) = \sum_a a \Pr[X = a]$.

Example: Let X be a uniform r.v. on the interval $[0, \ell]$. Then

$$\mathbb{E}(X) = \int_0^{\ell} x\frac{1}{\ell}dx = \left[\frac{x^2}{2\ell}\right]_0^{\ell} = \frac{\ell}{2}.$$

This is certainly what we would expect!

Example: Consider the dart-throwing example again. The expectation of the distance X of the dart from the origin is

$$\mathbb{E}(X) = \int_0^1 x \cdot 2x dx = \left[\frac{2}{3}x^3\right]_0^1 = \frac{2}{3}.$$

Note that the expected distance is greater than $\frac{1}{2}$.

We will see more examples of expectations of continuous r.v.'s in the next section.

Since variance is really just another expectation, we can immediately port its definition to the continuous setting as well:

Definition 18.4 (Variance): The variance of a continuous random variable X with probability density function f is

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \left(\int_{-\infty}^{\infty} xf(x)dx\right)^2.$$

Example: Let's calculate the variance of the uniform r.v. X on the interval $[0,\ell]$. From the above definition, and plugging in our previous value for $\mathbb{E}(X)$, we get

$$\mathrm{Var}(X) = \int_0^\ell x^2 \frac{1}{\ell} dx - \mathbb{E}(X)^2 = \left[\frac{x^3}{3\ell}\right]_0^\ell - \left(\frac{\ell}{2}\right)^2 = \frac{\ell^2}{3} - \frac{\ell^2}{4} = \frac{\ell^2}{12}.$$

The factor of $\frac{1}{12}$ here is not particularly intuitive, but the fact that the variance is proportional to ℓ^2 should come as no surprise. Like its discrete counterpart, this distribution has large variance.

Two more important continuous distributions

We have already seen one important continuous distribution, namely the uniform distribution. In this section we will see two more: the *exponential* distribution and the *normal* (or *Gaussian*) distribution. These three distributions cover the vast majority of continuous random variables arising in applications.

Exponential distribution: The exponential distribution is a continuous version of the geometric distribution, which we have already met. Recall that the geometric distribution describes the number of tosses of a coin until the first Head appears; the distribution has a single parameter p, which is the bias (Heads probability) of the coin. Of course, in real life applications we are usually not waiting for a coin to come up Heads but rather waiting for a system to fail, a clock to ring, an experiment to succeed etc.

In such applications we are frequently not dealing with discrete events or discrete time, but rather with *continuous* time: for example, if we are waiting for an apple to fall off a tree, it can do so at any time at all, not necessarily on the tick of a discrete clock. This situation is naturally modeled by the exponential distribution, defined as follows:

Definition 18.5 (Exponential distribution): For any $\lambda > 0$, a continuous random variable X with pdf f given by

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

is called an *exponential* random variable with parameter λ.

Like the geometric, the exponential distribution has a single parameter λ, which characterizes the *rate* at which events happen. We shall illuminate the connection between the geometric and exponential distributions in a moment.

First, let's do some basic computations with the exponential distribution. We should check first that it is a valid distribution, i.e., that it satisfies (3):

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^\infty \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_0^\infty = 1,$$

as required. Next, what is its expectation? We have

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^\infty \lambda x e^{-\lambda x} dx = \left[-xe^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x} dx = 0 + \left[-\frac{e^{-\lambda x}}{\lambda}\right]_0^\infty = \frac{1}{\lambda},$$

where for the first integral we used integration by parts.

To compute the variance, we need to evaluate

$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^\infty \lambda x^2 e^{-\lambda x} dx = \left[-x^2 e^{-\lambda x}\right]_0^\infty + \int_0^\infty 2xe^{-\lambda x} dx = 0 + \frac{2}{\lambda}\mathbb{E}(X) = \frac{2}{\lambda^2},$$

where again we used integration by parts. The variance is therefore

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Let us now explore the connection with the geometric distribution. Note first that the exponential distribution satisfies, for any $t \geq 0$,

$$\Pr[X > t] = \int_t^\infty \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_t^\infty = e^{-\lambda t}. \tag{12}$$

In other words, the probability that we have to wait more than time t for our event to happen is $e^{-\lambda t}$, which is an exponential decay with rate λ.

Now consider a discrete-time setting in which we perform one trial every δ seconds (where δ is very small—in fact, we will take $\delta \to 0$ to make time "continuous"), and where our success probability is $p = \lambda\delta$. Making the success probability proportional to δ makes sense, as it corresponds to the natural assumption that there is a fixed *rate* of success *per unit time*, which we denote by $\lambda = p/\delta$. The number of trials until we get a success has the geometric distribution with parameter p, so if we let the r.v. Y denote the time (in seconds) until we get a success we have

$$\Pr[Y > k\delta] = (1-p)^k = (1-\lambda\delta)^k \qquad \text{for any } k \geq 0.$$

Hence, for any $t > 0$, we have

$$\Pr[Y > t] = \Pr[Y > (\tfrac{t}{\delta})\delta] = (1-\lambda\delta)^{t/\delta} \approx e^{-\lambda t},$$

where this final approximation holds in the limit as $\delta \to 0$ with $\lambda = p/\delta$ fixed. (We are ignoring the detail of rounding $\frac{t}{\delta}$ to an integer since we are taking an approximation anyway.)

Comparing this expression with (12) we see that this distribution has the same form as the exponential distribution with parameter λ, where λ (the success rate per unit time) plays an analogous role to p (the probability of success on each trial)—though note that λ is not constrained to be ≤ 1. Thus we may view the exponential distribution as a continuous time analog of the geometric distribution.

Normal Distribution: The last continuous distribution we will look at, and by far the most prevalent in applications, is called the *normal* or *Gaussian* distribution. It has two parameters, μ and σ.

Definition 18.6 (Normal distribution): For any μ and $\sigma > 0$, a continuous random variable X with pdf f given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

is called a *normal* random variable with parameters μ and σ. In the special case $\mu = 0$ and $\sigma = 1$, X is said to have the *standard normal* distribution.

A plot of the pdf f reveals a classical "bell-shaped" curve, centered at (and symmetric around) $x = \mu$, and with "width" determined by σ. (The precise meaning of this latter statement will become clear when we discuss the variance below.)

Let's run through the usual calculations for this distribution. First, let's check equation (3):

$$\int_{-\infty}^{\infty} f(x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2} dx = 1. \tag{13}$$

The fact that this integral evaluates to 1 is a routine exercise in integral calculus, and is left as an exercise (or feel free to look it up in any standard book on probability or on the internet).

What are the expectation and variance of a normal r.v. X? Let's consider first the standard normal. By definition, its expectation is

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} xe^{-x^2/2}dx = \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{0} xe^{-x^2/2}dx + \int_{0}^{\infty} xe^{-x^2/2}dx \right) = 0.$$

The last step follows from the fact that the function $e^{-x^2/2}$ is symmetrical about $x = 0$, so the two integrals are the same except for the sign. For the variance, we have

$$\begin{aligned} \mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}dx \\ &= \frac{1}{\sqrt{2\pi}} \left[-xe^{-x^2/2} \right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}dx \\ &= 0 + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}dx = 1. \end{aligned}$$

In the first line here we used the fact that $\mathbb{E}(X) = 0$; in the second line we used integration by parts; and in the last line we used (13) in the special case $\mu = 0$, $\sigma = 1$. So the standard normal distribution has expectation $\mathbb{E}(X) = 0 = \mu$ and variance $\mathrm{Var}(X) = 1 = \sigma^2$.

Now suppose X has normal distribution with general parameters μ, σ. We claim that the r.v. $Y = \frac{X-\mu}{\sigma}$ has the standard normal distribution. To see this, note that

$$\Pr[a \le Y \le b] = \Pr[\sigma a + \mu \le X \le \sigma b + \mu] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\sigma a+\mu}^{\sigma b+\mu} e^{-(x-\mu)^2/2\sigma^2}dx = \frac{1}{\sqrt{2\pi}} \int_{a}^{b} e^{-y^2/2}dy,$$

by a simple change of variable in the integral. Hence Y is indeed standard normal. Note that Y is obtained from X just by shifting the origin to μ and scaling by σ. (And we shall see in a moment that μ is the mean and σ the standard deviation, so this operation is very natural.)

Now we can read off the expectation and variance of X from those of Y. For the expectation, using linearity, we have

$$0 = \mathbb{E}(Y) = \mathrm{E}\left(\frac{X-\mu}{\sigma} \right) = \frac{\mathbb{E}(X) - \mu}{\sigma},$$

and hence $\mathbb{E}(X) = \mu$. For the variance we have

$$1 = \mathrm{Var}(Y) = \mathrm{Var}\left(\frac{X-\mu}{\sigma} \right) = \frac{\mathrm{Var}(X)}{\sigma^2},$$

and hence $\mathrm{Var}(X) = \sigma^2$.

The bottom line, then, is that the normal distribution has expectation μ and variance σ^2. (This explains the notation for the parameters μ, σ.)

The fact that the variance is σ^2 (so that the standard deviation is σ) explains our earlier comment that σ determines the "width" of the normal distribution. Namely, by Chebyshev's inequality, a constant fraction of the distribution lies within distance (say) 2σ of the expectation μ.

Note: The above analysis shows that, by means of a simple origin shift and scaling, we can relate any normal distribution to the standard normal. This means that, when doing computations with normal distributions, it's enough to do them for the standard normal. For this reason, books and online sources of mathematical formulas usually contain tables describing the density of the standard normal. From this, one can read off the corresponding information for any normal r.v. X with parameters μ, σ^2, from the formula

$$\Pr[X \le a] = \Pr[Y \le \tfrac{a-\mu}{\sigma}],$$

where Y is standard normal.

Some basic facts that are useful to know, when dealing with a normal distribution:

- A normally distributed r.v. falls within $\pm\sigma$ of the mean about 68% of the time.
  (i.e., $\Pr[\mu-\sigma \le X \le \mu+\sigma] \approx 0.68$)
- A normally distributed r.v. falls within $\pm 2\sigma$ of the mean about 95% of the time.
  (i.e., $\Pr[\mu-2\sigma \le X \le \mu+2\sigma] \approx 0.95$)
- A normally distributed r.v. falls within $\pm 3\sigma$ of the mean about 99.7% of the time.
  (i.e., $\Pr[\mu-3\sigma \le X \le \mu+3\sigma] \approx 0.997$)

The normal distribution is ubiquitous throughout the sciences and the social sciences, because it is the standard model for any aggregate data that results from a large number of independent observations of the same random variable (such as the heights of females in the US population, or the observational error in a physical experiment). Such data, as is well known, tends to cluster around its mean in a "bell-shaped" curve, with the correspondence becoming more accurate as the number of observations increases. A theoretical explanation of this phenomenon is the Central Limit Theorem, which we next discuss.

The Central Limit Theorem

Recall from Lecture Note 16 the Law of Large Numbers for i.i.d. r.v.'s X_i's: it says that the probability of *any* deviation α of the sample average $A_n := \frac{1}{n}\sum_{i=1}^n X_i$ from the mean, however small, tends to zero as the number of observations n in our average tends to infinity. Thus by taking n large enough, we can make the probability of any given deviation as small as we like.

Actually we can say something much stronger than the Law of Large Numbers: namely, the distribution of the sample average A_n, for large enough n, looks like a *normal distribution* with mean μ and variance $\frac{\sigma^2}{n}$. (Of course, we already know that these are the mean and variance of A_n; the point is that the distribution becomes normal!) The fact that the standard deviation decreases with n (specifically, as $\frac{\sigma}{\sqrt{n}}$) means that the distribution approaches a sharp spike at μ.

Recall from the last section that the density of the normal distribution is a symmetrical bell-shaped curve centered around the mean μ. Its height and width are determined by the standard deviation σ as follows: the height at the mean is about $0.4/\sigma$; 50% of the mass is contained in the interval of width 0.67σ either side of the mean, and 99.7% in the interval of width 3σ either side of the mean. (Note that, to get the correct scale, deviations are on the order of σ rather than σ^2.)

To state the Central Limit Theorem precisely (so that the limiting distribution is a constant rather than something that depends on n), we shift the mean of A_n to 0 and scale it so that its variance is 1, i.e., we replace A_n by

$$A_n' = \frac{(A_n-\mu)\sqrt{n}}{\sigma} = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}.$$

The Central Limit Theorem then says that the distribution of A_n' converges to the *standard normal* distribution.

Theorem 18.1: **[Central Limit Theorem]** Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with common expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = \mathrm{Var}(X_i)$ (both assumed to be $< \infty$). Define $A_n' = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$. Then as $n \to \infty$, the distribution of A_n' approaches the standard normal distribution in the sense that, for any real α,

$$\Pr[A_n' \le \alpha] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-x^2/2} dx \quad \text{as } n \to \infty.$$

The Central Limit Theorem is a very striking fact. What it says is the following. If we take an average of n observations of absolutely any r.v. X, then the distribution of that average will be approximately a bell-shaped curve centered at $\mu = \mathbb{E}(X)$. Thus all trace of the distribution of X disappears as n gets large: all distributions, no matter how complex,[3] look like the normal distribution when they are averaged. The only effect of the original distribution is through the variance σ^2, which determines the width of the curve for a given value of n, and hence the rate at which the curve shrinks to a spike.

[3] We do need to assume that the mean and variance of X are finite.

Infinity and Countability

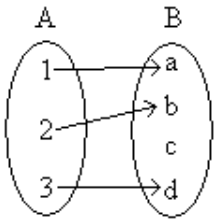
Bijections

Consider a function (or mapping) f that maps elements of a set A (called the *domain* of f) to elements of set B (called the *range* of f). For each element $x \in A$ ("input"), f must specify one element $f(x) \in B$ ("output"). Recall that we write this as $f : A \to B$. We say that f is a *bijection* if every element $a \in A$ has a unique *image* $b = f(a) \in B$, and every element $b \in B$ has a unique *pre-image* $a \in A$ such that $f(a) = b$.

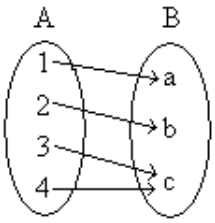
f is a *one-to-one function* (or an *injection*) if f maps distinct inputs to distinct outputs. More rigorously, f is one-to-one if the following holds: $\forall x, y \;.\; x \neq y \Rightarrow f(x) \neq f(y)$.

The next property we are interested in is functions that are *onto* (or *surjective*). A function that is onto essentially "hits" every element in the range (i.e., each element in the range has at least one pre-image). More precisely, a function f is onto if the following holds: $\forall y \; \exists x \;.\; f(x) = y$. Here are some examples to help visualize one-to-one and onto functions:

One-to-one



Onto



Note that according to our definition a function is a bijection iff it is both one-to-one and onto.

Cardinality

How can we determine whether two sets have the same *cardinality* (or "size")? The answer to this question, reassuringly, lies in early grade school memories: by demonstrating a *pairing* between elements of the two sets. More formally, we need to demonstrate a *bijection* f between the two sets. The bijection sets up a one-to-one correspondence, or pairing, between elements of the two sets. We know how this works for finite sets. In this lecture, we will see what it tells us about *infinite* sets.

Are there more natural numbers $\mathbb{N}$ than there are positive integers $\mathbb{Z}^+$? It is tempting to answer yes, since every positive integer is also a natural number, but the natural numbers have one extra element 0, which is not an element of $\mathbb{Z}^+$. Upon more careful observation, though, we see that we can generate a mapping between the natural numbers and the positive integers as follows:

$$\begin{array}{lcccccccc} \mathbb{N} & & 0 & 1 & 2 & 3 & 4 & 5 & \ldots \\ \downarrow & & \searrow & \searrow & \searrow & \searrow & \searrow & \searrow & \\ \mathbb{Z}^+ & & & 1 & 2 & 3 & 4 & 5 & 6 \quad \ldots \end{array}$$

Why is this mapping a bijection? Clearly, the function $f : \mathbb{N} \to \mathbb{Z}^+$ is onto because every positive integer is hit. And it is also one-to-one because no two natural numbers have the same image. (The image of n is $f(n) = n+1$, so if $f(n) = f(m)$ then we must have $n = m$.) Since we have shown a bijection between $\mathbb{N}$ and $\mathbb{Z}^+$, this tells us that there are as many natural numbers as there are positive integers! Informally, we have proved that "$\infty + 1 = \infty$."

What about the set of *even* natural numbers $2\mathbb{N} = \{0, 2, 4, 6, ...\}$? In the previous example the difference was just one element. But in this example, there seem to be twice as many natural numbers as there are even natural numbers. Surely, the cardinality of $\mathbb{N}$ must be larger than that of $2\mathbb{N}$ since $\mathbb{N}$ contains all of the odd natural numbers as well. Though it might seem to be a more difficult task, let us attempt to find a bijection between the two sets using the following mapping:

$\mathbb{N}$	0	1	2	3	4	5	...
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	
$2\mathbb{N}$	0	2	4	6	8	10	...

The mapping in this example is also a bijection. f is clearly one-to-one, since distinct natural numbers get mapped to distinct even natural numbers (because $f(n) = 2n$). f is also onto, since every n in the range is hit: its pre-image is $\frac{n}{2}$. Since we have found a bijection between these two sets, this tells us that in fact $\mathbb{N}$ and $2\mathbb{N}$ have the same cardinality!

What about the set of all integers, $\mathbb{Z}$? At first glance, it may seem obvious that the set of integers is larger than the set of natural numbers, since it includes negative numbers. However, as it turns out, it is possible to find a bijection between the two sets, meaning that the two sets have the same size! Consider the following mapping:

$$0 \to 0, \quad 1 \to -1, \quad 2 \to 1, \quad 3 \to -2, \quad 4 \to 2, \quad \ldots, \quad 124 \to 62, \quad \ldots$$

In other words, our function is defined as follows:

$$f(x) = \begin{cases} \frac{x}{2} & \text{if } x \text{ is even,} \\ \frac{-(x+1)}{2} & \text{if } x \text{ is odd.} \end{cases}$$

We will prove that this function $f : \mathbb{N} \to \mathbb{Z}$ is a bijection, by first showing that it is one-to-one and then showing that it is onto.

Proof (one-to-one): Suppose $f(x) = f(y)$. Then they both must have the same sign. Therefore either $f(x) = \frac{x}{2}$ and $f(y) = \frac{y}{2}$, or $f(x) = \frac{-(x+1)}{2}$ and $f(y) = \frac{-(y+1)}{2}$. In the first case, $f(x) = f(y) \Rightarrow \frac{x}{2} = \frac{y}{2} \Rightarrow x = y$. Hence $x = y$. In the second case, $f(x) = f(y) \Rightarrow \frac{-(x+1)}{2} = \frac{-(y+1)}{2} \Rightarrow x = y$. In both cases $f(x) = f(y) \Rightarrow x = y$, so taking the contrapositive, f must be one-to-one.

Proof (onto): If $y \in \mathbb{Z}$ is non-negative, then $f(2y) = y$, and $2y \geq 0$. Therefore, y has a pre-image in $\mathbb{N}$. If y is negative, then $f(-(2y+1)) = y$, and $-(2y+1) \geq 0$. Therefore, y has a pre-image in $\mathbb{N}$. Thus every $y \in \mathbb{Z}$ has a preimage, so f is onto.

Since f is a bijection, this tells us that $\mathbb{N}$ and $\mathbb{Z}$ have the same cardinality.

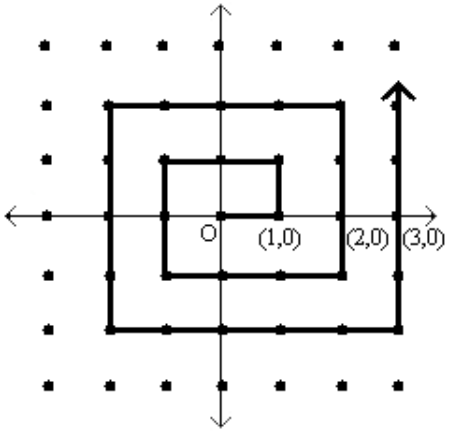
Now for an important definition. We say that a set S is **countable** if there is a bijection between S and $\mathbb{N}$ or some subset of $\mathbb{N}$. Thus any finite set S is countable (since there is a bijection between S and the subset $\{0, 1, 2, \ldots, m-1\}$, where $m = |S|$ is the size of S). And we have already seen three examples of countable infinite sets: $\mathbb{Z}^+$ and $2\mathbb{N}$ are obviously countable since they are themselves subsets of $\mathbb{N}$; and $\mathbb{Z}$ is countable because we have just seen a bijection between it and $\mathbb{N}$.

What about the set of all rational numbers? Recall that $\mathbb{Q} = \{\frac{x}{y} \mid x, y \in \mathbb{Z},\ y \neq 0\}$. Surely there are more

rational numbers than natural numbers? After all, there are infinitely many rational numbers between any two natural numbers. Surprisingly, the two sets have the same cardinality! To see this, let us introduce a slightly different way of comparing the cardinality of two sets.

If there is a one-to-one function $f : A \to B$, then the cardinality of A is less than or equal to that of B. Now to show that the cardinality of A and B are the same we can show that $|A| \leq |B|$ and $|B| \leq |A|$. This corresponds to showing that there is a one-to-one function $f : A \to B$ and a one-to-one function $g : B \to A$. The existence of these two one-to-one functions implies that there is a bijection $h : A \to B$, thus showing that A and B have the same cardinality. The proof of this fact, which is called the Cantor-Bernstein theorem, is actually quite hard, and we will skip it here.

Back to comparing the natural numbers and the rational numbers. First it is obvious that $|\mathbb{N}| \leq |\mathbb{Q}|$ because $\mathbb{N} \subseteq \mathbb{Q}$. So our goal now is to prove that also $|\mathbb{Q}| \leq |\mathbb{N}|$. To do this, we must exhibit a one-to-one function $f : \mathbb{Q} \to \mathbb{N}$. The following picture of a spiral conveys the idea of this function:



Each rational number $\frac{a}{b}$ (written in its lowest terms, so that $\gcd(a,b) = 1$) is represented by the point (a,b) in the infinite two-dimensional grid shown (which corresponds to $\mathbb{Z} \times \mathbb{Z}$, the set of all pairs of integers). Note that not all points on the grid are valid representations of rationals: e.g., all points on the x-axis have $b = 0$ so none are valid (except for $(0,0)$, which we take to represent the rational number 0); and points such as $(2,8)$ and $(-1,-4)$ are not valid either as the rational number $\frac{1}{4}$ is represented by $(1,4)$. But $\mathbb{Z} \times \mathbb{Z}$ certainly contains all rationals under this representation, so if we come up with an injection from $\mathbb{Z} \times \mathbb{Z}$ to $\mathbb{N}$ then this will also be an injection from $\mathbb{Q}$ to $\mathbb{N}$ (why?).

The idea is to map each pair (a,b) to its position along the spiral, starting at the origin. (Thus, e.g., $(0,0) \to 0$, $(1,0) \to 1$, $(1,1) \to 2$, $(0,1) \to 3$, and so on.) This mapping certainly maps every rational number to a natural number, because every rational appears somewhere (exactly once) in the grid, and the spiral hits every point in the grid. Why is this mapping an injection? Well, we just have to check that no two rational numbers map to the same natural number. But that is true because no two pairs lie at the same position on the spiral. (Note that the mapping is *not* onto because some positions along the spiral do not correspond to valid representations of rationals; but that is fine.)

This tells us that $|\mathbb{Q}| \leq |\mathbb{N}|$. Since also $|\mathbb{N}| \leq |\mathbb{Q}|$, as we observed earlier, by the Cantor-Bernstein Theorem $\mathbb{N}$ and $\mathbb{Q}$ have the same cardinality.

Our next example concerns the set of all binary strings (of any finite length), denoted $\{0,1\}^*$. Despite the fact that this set contains strings of unbounded length, it turns out to have the same cardinality as $\mathbb{N}$. To see this, we set up a direct bijection $f : \{0,1\}^* \to \mathbb{N}$ as follows. Note that it suffices to *enumerate* the elements of $\{0,1\}^*$ in such a way that each string appears exactly once in the list. We then get our bijection by setting $f(n)$ to be the nth string in the list. How do we enumerate the strings in $\{0,1\}^*$? Well, it's natural to list them in increasing order of length, and then (say) in *lexicographic* order (or, equivalently, numerically increasing

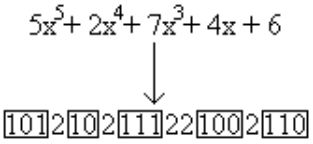
order when viewed as binary numbers) within the strings of each length. This means that the list would look like

$$\varepsilon, 0, 1, 00, 01, 10, 11, 000, 001, 010, 011, 100, 101, 110, 111, 1000, \ldots,$$

where ε denotes the empty string (the only string of length 0). It should be clear that this list contains each binary string once and only once, so we get a bijection with $\mathbb{N}$ as desired.

Our final countable example is the set of all polynomials with natural number coefficients, which we denote $\mathbb{N}(x)$. To see that this set is countable, we will make use of (a variant of) the previous example. Note first that, by essentially the same argument as for $\{0,1\}^*$, we can see that the set of all *ternary* strings $\{0,1,2\}^*$ (that is, strings over the alphabet $\{0,1,2\}$) is countable. To see that $\mathbb{N}(x)$ is countable, it therefore suffices to exhibit an injection $f : \mathbb{N}(x) \to \{0,1,2\}^*$, which in turn will give an injection from $\mathbb{N}(x)$ to $\mathbb{N}$. (It is obvious that there exists an injection from $\mathbb{N}$ to $\mathbb{N}(x)$, since each natural number n is itself trivially a polynomial, namely the constant polynomial n itself.)

How do we define f? Let's first consider an example, namely the polynomial $p(x) = 5x^5 + 2x^4 + 7x^3 + 4x + 6$. We can list the coefficients of $p(x)$ as follows: $(5,2,7,0,4,6)$. We can then write these coefficients as binary strings: $(101_2, 10_2, 111_2, 0_2, 100_2, 110_2)$. Now, we can construct a ternary string where a "2" is inserted as a separator between each binary coefficient (ignoring coefficients that are 0). Thus we map $p(x)$ to a ternary string as illustrated below:

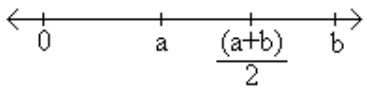


It is easy to check that this is an injection, since the original polynomial can be uniquely recovered from this ternary string by simply reading off the coefficients between each successive pair of 2's. (Notice that this mapping $f : \mathbb{N}(x) \to \{0,1,2\}^*$ is not onto (and hence not a bijection) since many ternary strings will not be the image of any polynomials; this will be the case, for example, for any ternary strings that contain binary subsequences with leading zeros.)

Hence we have an injection from $\mathbb{N}(x)$ to $\mathbb{N}$, so $\mathbb{N}(x)$ is countable.

Cantor's Diagonalization

We have established that $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ all have the same cardinality. What about the real numbers, the set of all points on the real line? Surely they are countable too? After all, the rational numbers, like the real numbers, are dense (i.e., between any two rational numbers there is a rational number):



In fact, between any two *real* numbers there is always a rational number. It is really surprising, then, that there are more real numbers than rationals. That is, there is no bijection between the rationals (or the natural numbers) and the reals. In fact, we will show something even stronger: even the real numbers in the interval $[0,1]$ are uncountable!

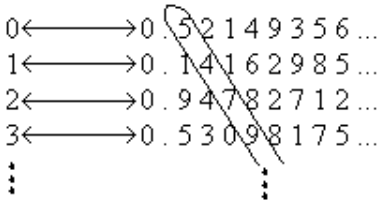
Recall that a real number can be written out in an infinite decimal expansion. A real number in the interval $[0,1]$ can be written as $0.d_1d_2d_3...$ Note that this representation is not unique; for example, $1 = 0.999...$, so the same real number can sometimes be expressed in two different ways.[1] For definiteness we shall assume

[1] To see this, write $x = .999\ldots$. Then $10x = 9.999\ldots$, so $9x = 9$, and thus $x = 1$.

that every real number is represented as a recurring decimal and we prefer to end with all 9's where possible (i.e., we choose the representation $0.999\ldots$ rather than 1).

Theorem: The set $[0,1]$ of real numbers is not countable.

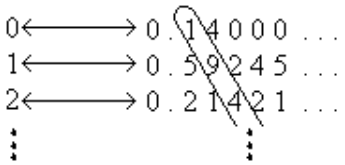
Cantor's Diagonalization Proof: Suppose towards a contradiction that there is a bijection $f : \mathbb{N} \to [0,1]$. This gives an infinite list of real numbers, namely, $f(0)$, $f(1)$, ... We can enumerate this infinite list as follows:



The number circled in the diagonal is some real number $r = 0.5479\ldots$, since it is an infinite decimal expansion. Now consider the real number s obtained by modifying every digit of r, say by replacing each digit d with $d+2$ modulo 10; thus in our example above, $s = 0.7691\ldots$. We claim that s does not occur in our infinite list of real numbers. Suppose for contradiction that it did, and that it was the nth number in the list. Then r and s differ in the nth digit: the nth digit of s is the nth digit of r plus 2 mod 10. So we have a real number s that is not in the range of f. But this contradicts the assertion that f is a bijection. Thus the real numbers are not countable.

Let us remark that the reason that we modified each digit by adding 2 modulo 10 as opposed to adding 1 is that the same real number can have two decimal expansions; for example 0.999... = 1.000.... But if two real numbers differ by more than 1 in any digit they cannot be equal. Thus we are completely safe in our assertion.

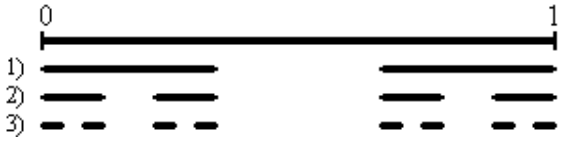
With Cantor's diagonalization method, we proved that $\mathbb{R}$ is uncountable. What happens if we apply the same method to $\mathbb{Q}$, in a (futile) attempt to show the rationals are uncountable? Well, suppose for a contradiction that our bijective function $f : \mathbb{N} \to \mathbb{Q} \cap [0,1]$ produces the following mapping:



This time, let us consider the number q obtained by modifying every digit of the diagonal, say by replacing each digit d with $d+2$ modulo 10. Then in the above example q = 0.316..., and we want to try to show that it does not occur in our infinite list of rational numbers. However, we do not know if q is rational (in fact, it is extremely unlikely for the decimal expansion of q to be periodic). This is why the method fails when applied to the rationals. When dealing with the reals, the modified diagonal number was guaranteed to be a real number.

The Cantor Set

The Cantor set is a remarkable set construction involving the real numbers in the interval $[0,1]$. The set is defined by repeatedly removing the middle thirds of line segments infinitely many times, starting with the original interval. For example, the first iteration would involve the removal of the interval $(\frac{1}{3}, \frac{2}{3})$, leaving $[0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$. The first three iterations are illustrated below:



The Cantor set contains all points that have *not* been removed: $C = \{x : x \text{ is never thrown out}\}$. How much of the original unit interval is left after this process is repeated infinitely? Well, we start with 1, and after the first iteration we remove $\frac{1}{3}$ of the interval, leaving us with $\frac{2}{3}$. For the second iteration, we keep $\frac{2}{3} \times \frac{2}{3}$ of the original interval. As we repeat the iterations infinitely, we are left with:

$$1 \longrightarrow \tfrac{2}{3} \longrightarrow \tfrac{2}{3} \times \tfrac{2}{3} \longrightarrow \tfrac{2}{3} \times \tfrac{2}{3} \times \tfrac{2}{3} \longrightarrow \cdots \longrightarrow \lim_{n\to\infty} (\tfrac{2}{3})^n = 0$$

According to the calculations, it looks like we have removed everything from the original interval! Does this mean that the Cantor set is empty? No, it doesn't. What it means is that the *measure* of the Cantor set is zero; the Cantor set consists of isolated points and does not contain any non-trivial intervals. In fact, not only is the Cantor set non-empty, it is uncountable![2]

To see why, let us first make a few observations about ternary strings. In ternary notation, all strings consist of digits (called "trits") from the set $\{0,1,2\}$. All real numbers in the interval $[0,1]$ can be written in ternary notation. (For example, $\frac{1}{3}$ can be written as 0.1, or equivalently as $0.0222\ldots$, and $\frac{2}{3}$ can be written as 0.2 or as $0.1222\ldots$.) Thus, in the first iteration, the middle third removed contains all ternary numbers of the form 0.1xxxxx. The ternary numbers left after the first removal can all be expressed either in the form 0.0xxxxx... or 0.2xxxxx... (We have to be a little careful here with the endpoints of the intervals; but we can handle them by writing $\frac{1}{3}$ as $0.02222\ldots$ and $\frac{2}{3}$ as 0.2.) The second iteration removes ternary numbers of the form 0.01xxxxx and 0.21xxxxx (i.e., any number with 1 in the second position). The third iteration removes 1's in the third position, and so on. Therefore, what remains is all ternary numbers with only 0's and 2's. Thus we have shown that

$$C = \{x \in [0,1] : x \text{ has a ternary representation consisting only of 0's and 2's}\}.$$

Finally, using this characterization, we can set up an *onto* map f from C to $[0,1]$. Since we already know that $[0,1]$ is uncountable, this implies that C is uncountable also. The map f is defined as follows: for $x \in C$, $f(x)$ is defined as the binary decimal obtained by dividing each digit of the ternary representation of x by 2. Thus, for example, if $x = 0.0220$ (in ternary), then $f(x)$ is the binary decimal 0.0110. But the set of all binary decimals $0.xxxxx\ldots$ is in one-to-one correspondence with the real interval $[0,1]$, and the map f is onto because every binary decimal is the image of some ternary string under f (obtained by doubling every binary digit).[3] This completes the proof that C is uncountable.

Power Sets and Higher Orders of Infinity

Let S be any set. Then the *power set* of S, denoted by $\mathscr{P}(S)$, is the set of all subsets of S. More formally, it is defined as: $\mathscr{P}(S) = \{T : T \subseteq S\}$. For example, if $S = \{1,2,3\}$, then $\mathscr{P}(S) = \{\{\}, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$.

What is the cardinality of $\mathscr{P}(S)$? If $|S| = k$ is finite, then $|\mathscr{P}(S)| = 2^k$. To see this, let us think of each subset of S corresponding to a k-bit string. In the example above, the subset $\{1,3\}$ corresponds to the string

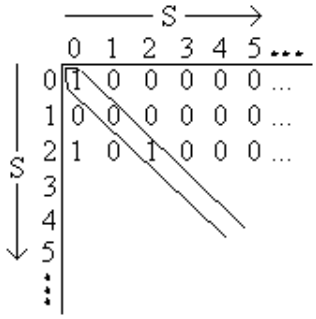
[2] It's actually easy to see that C contains at least countably many points, namely the endpoints of the intervals in the construction—i.e., numbers such as $\frac{1}{3}$, $\frac{2}{3}$, $\frac{1}{9}$, $\frac{1}{27}$ etc. It's less obvious that C also contains various other points, such as $\frac{1}{4}$ and $\frac{3}{10}$. (It's a fascinating exercise to figure out why.)

[3] Note that f is *not* injective; for example, the ternary strings $0.20222\ldots$ and 0.22 map to binary strings $0.10111\ldots$ and 0.11 respectively, which denote the same real number. Thus f is not a bijection. However, the current proof shows that the cardinality of C is at least that of $[0,1]$, while it is obvious that the cardinality of C is at most that of $[0,1]$ since $C \subset [0,1]$. Hence C has the same cardinality as $[0,1]$ (and as $\mathbb{R}$).

101. A 1 in the ith position indicates that the ith element of S is in the subset and a 0 indicates that it is not. Now the number of binary strings of length k is 2^k, since there are two choices for each bit position. Thus $|\mathscr{P}(S)| = 2^k$. So for finite sets S, the cardinality of the power set of S is exponentially larger than the cardinality of S. What about infinite (countable) sets? We claim that there is no bijection from S to $\mathscr{P}(S)$, so $\mathscr{P}(S)$ is not countable. Thus for example the set of all subsets of natural numbers is not countable, even though the set of natural numbers itself is countable.

Theorem: Let S be countably infinite. Then $|\mathscr{P}(S)| > |S|$.

Proof: Suppose towards a contradiction that there is a bijection $f : S \to \mathscr{P}(S)$. Recall that we can represent a subset by a binary string, with one bit for each element of S. (So, since S is infinite, the string will be infinitely long. Contrast the case of $\{0,1\}^*$ discussed earlier, which consists of all binary strings of *finite* length.) Consider the following diagonalization picture in which the function f maps natural numbers x to binary strings which correspond to subsets of S (e.g., $2 \to 10100... = \{0,2\}$):



In this case, we have assigned the following mapping: $0 \to \{0\}$, $1 \to \{\}$, $2 \to \{0,2\}$, ... (i.e., the nth row describes the nth subset as follows: if there is a 1 in the kth column, then k is in this subset, else it is not.) Using a similar diagonalization argument to the earlier one, flip each bit along the diagonal: $1 \to 0$, $0 \to 1$, and let b denote the resulting binary string. First, we must show that the new element is a subset of S. Clearly it is, since b is an infinite binary string which corresponds to a subset of S. Now suppose b were the nth binary string. This cannot be the case though, since the nth bit of b differs from the nth bit of the diagonal (the bits are flipped). So it's not on our list, but it should be, since we assumed that the list enumerated all possible subsets of S. Thus we have a contradiction, implying that $\mathscr{P}(S)$ is uncountable.

Thus we have seen that the cardinality of $\mathscr{P}(\mathbb{N})$ (the power set of the natural numbers) is strictly larger than the cardinality of $\mathbb{N}$ itself. The cardinality of $\mathbb{N}$ is denoted $\aleph_0$ (pronounced "aleph null"), while that of $\mathscr{P}(\mathbb{N})$ is denoted $2^{\aleph_0}$. It turns out that in fact $\mathscr{P}(\mathbb{N})$ has the same cardinality as $\mathbb{R}$ (the real numbers), and indeed as the real numbers in $[0,1]$. This cardinality is known as $\mathbf{c}$, the "cardinality of the continuum." So we know that $2^{\aleph_0} = \mathbf{c} > \aleph_0$. Even larger infinite cardinalities (or "orders of infinity"), denoted $\aleph_1, \aleph_2, \ldots$, can be defined using the machinery of set theory; these obey (to the uninitiated somewhat bizarre) rules of arithmetic. Several fundamental questions in modern mathematics concern these objects. For example, the famous "continuum hypothesis" asserts that $\mathbf{c} = \aleph_1$ (which is equivalent to saying that there are no sets with cardinality between that of the natural numbers and that of the real numbers).

Self-Reference and Computability

The Liar's Paradox

Propositions are statements that are either true or false. We saw before that some statements are not well defined or too imprecise to be called propositions. But here is a statement that is problematic for more subtle reasons: "All Cretans are liars." So said a Cretan in antiquity, thus giving rise to the so-called liar's paradox which has amused and confounded people over the centuries. Actually the above statement isn't really a paradox; it simply yields a contradiction if we assume it is true, but if it is false then there is no problem.

A stronger formulation of this paradox is the following statement: "This statement is false." Is the statement true? If the statement is true, then what it asserts must be true; namely that it is false. But if it is false, then it must be true. So it really is a paradox. Around a century ago, this paradox found itself at the center of foundational questions about mathematics and computation.

We will now study how this paradox relates to computation. Before doing so, let us consider another manifestation of the paradox, created by the great logician Bertrand Russell. In a village with just one barber, every man keeps himself clean-shaven. Some of the men shave themselves, while others go to the barber. The barber proclaims: "I shave all and only those men who do not shave themselves." It seems reasonable then to ask the question: Does the barber shave himself? Thinking more carefully about the question though, we see that we are presented with a logically impossible scenario. If the barber does not shave himself, then according to what he announced, he shaves himself. If the barber does shave himself, then according to his statement he does not shave himself!

The Halting Problem

Are there tasks that a computer cannot perform? For example, we would like to ask the following basic question when compiling a program: does it go into an infinite loop? In 1936, Alan Turing showed that there is no program that can perform this test. The proof of this remarkable fact is very elegant and combines two ingredients: self-reference (as in the liar's paradox), and the fact that programs and data are two sides of the same coin. In computers, a program is represented by a string of bits just as integers, characters, and other data are. The only difference is in how the string of bits is interpreted.

We will now examine the Halting Problem. Given the description of a program and its input, we would like to know if the program ever halts when it is executed on the given input. In other words, we would like to write a program TestHalt that behaves as follows:

$$\text{TestHalt}(P,I) = \begin{cases} \text{"halts"}, & \text{if program } P \text{ eventually halts, when run on input } I; \\ \text{"loops"}, & \text{if program } P \text{ never terminates, when run on input } I. \end{cases}$$

Why can't such a program exist? We're going to use the fact that a program is just a bit string, so it can be input as data. This means that it is perfectly valid to consider the behavior of TestHalt(P,P), which will output "halts" if P halts on input P, and "loops" if P loops forever on input P. With this idea, we will now prove that there cannot exist any program TestHalt that satisfies the specification above.

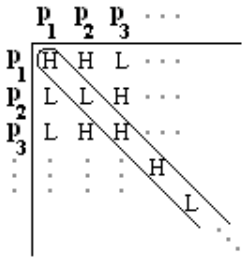
Proof: Define the program Turing, as follows:

Turing(P):
1. If TestHalt(P,P) = "halts", then loop forever.
2. Otherwise, halt.

So if the program P halts when given P as input, then Turing(P) loops forever; otherwise, Turing(P) halts. Assuming we have the program TestHalt, we can easily use it as a subroutine in line 1 of the above program.

Now let us look at the behavior of Turing(Turing), i.e., look at the result of running the program Turing on the input Turing. There are two cases: either it halts, or it does not. If Turing(Turing) halts, then it must be the case that TestHalt(Turing, Turing) returned "loops." But that would mean that Turing(Turing) would go into an infinite loop on line 1, and in particular should not have halted. In the second case, if Turing(Turing) does not halt, then it must be the case that TestHalt(Turing, Turing) returned "halts", which would mean that Turing(Turing) should have halted on line 2. In both cases, we arrive at a contradiction—which must mean that our initial assumption, namely that the program TestHalt exists, was wrong. Thus, TestHalt cannot exist, so it is impossible for a program to check if any general program halts. □

What proof technique did we use? This was actually a proof by diagonalization. Why? Since the set of all computer programs is countable (they are, after all, just finite-length strings over some alphabet, and the set of all finite-length strings is countable), we can enumerate all programs as follows (where P_i represents the ith program):



The (i,j)th entry is H if program P_i halts on input P_j, and L if it does not halt. Now if the program Turing exists it must occur somewhere on our list of programs, say as P_n. But this cannot be, since if the nth entry in the diagonal is H, meaning that P_n halts on P_n, then by its definition Turing loops on P_n; and if the entry is L, then by definition Turing halts on P_n. Thus the behavior of Turing is different from that of P_n, and hence Turing does not appear on our list. Since the list contains all possible programs, we must conclude that the program Turing does not exist. And since Turing is constructed by a simple modification of TestHalt, we can conclude that TestHalt does not exist either. Hence the Halting Problem cannot be solved.

In fact, there are many more cases of questions we would like to answer about a program, but cannot. For example, we cannot know if a program ever outputs anything. We cannot know if it ever executes a specific line of code. We cannot even check to see if the program is a virus. These issues are explored in greater detail in the advanced course CS172.

Introduction to Sets

A *set* is a well defined collection of objects considered as a whole. These objects are called *elements* or *members* of a set, and they can be anything, including numbers, letters, people, cities, and even other sets. By convention, sets are usually denoted by capital letters and can be described or defined by listing their elements and surrounding the list by curly braces. For example, we can describe the set A to be the set whose members are the first five prime numbers, or we can explicitly write: $A = \{2,3,5,7,11\}$. If x is an element of A, we write $x \in A$. Similarly, if y is not an element of A, then we write $y \notin A$. Two sets A and B are said to be equal, written as $A = B$, if they have the same elements. The order and repetition of elements do not matter, so $\{\text{red, white, blue}\} = \{\text{blue, white, red}\} = \{\text{red, blue, white}\}$. Sometimes, more complicated sets can be defined by using a different notation. For example, the set of all rational numbers denoted by $\mathbb{Q}$ can be written as: $\{\frac{a}{b} \mid a,b \text{ are integers, } b \neq 0\}$. In English, this is read as "the set of all fractions such that the numerator is an integer and the denominator is a non-zero integer."

Cardinality

We can also talk about the size of a set, or its *cardinality*. If $A = \{1,2,3,4\}$, then the cardinality of A, denoted by $|A|$, is 4. It is possible for the cardinality of a set to be 0. This set is called the *empty set*, denoted by the symbol $\emptyset$. A set can also have an infinite number of elements, such as the sets of all integers, prime numbers, or odd numbers.

Subsets and Proper Subsets

If every element of a set A is also in a set B, then we say that A is a *subset* of B, written $A \subseteq B$, or A is contained in B. We can also write $B \supseteq A$, meaning that B is a *superset* of A, or B contains A. A *proper subset* is a set A that is strictly contained in B, written as $A \subset B$, meaning that A excludes at least one element of B. For example, consider the set $B = \{1,2,3,4,5\}$. Then $\{1,2,3\}$ is both a subset and a proper subset of B, while $\{1,2,3,4,5\}$ is a subset but not a proper subset of B. Here are a few basic properties regarding subsets:

- The empty set is a proper subset of any nonempty set A: $\emptyset \subset A$.
- The empty set is a subset of every set B: $\emptyset \subseteq B$.
- Every set A is a subset of itself: $A \subseteq A$.

Intersections and Unions

The *intersection* of a set A with a set B, written as $A \cap B$, is the set of all elements which are members of both A and B. Two sets are said to be *disjoint* if $A \cap B = \emptyset$. The *union* of a set A with a set B, written as $A \cup B$, is the set of all elements which are either members of A or B (or both). For example, if A is the set of all positive even numbers and B is the set of all positive odd numbers, then $A \cap B = \emptyset$, and $A \cup B = \mathbb{Z}^+$, or the set of all positive integers. Here are a few properties of intersections and unions:

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

Complements

If A and B are two sets, then the *relative complement* of A in B, written as $B - A$ or $B \backslash A$, is the set of elements in B but not in A: $B \backslash A = \{x \in B \mid x \notin A\}$. For example, if $B = \{1,2,3\}$ and $A = \{3,4,5\}$, then $B \backslash A = \{1,2\}$. For another example, if $\mathbb{R}$ is the set of real numbers and $\mathbb{Q}$ is the set of rational numbers, then $\mathbb{R} \backslash \mathbb{Q}$ is the set of irrational numbers. Here are some important properties of complements:

- $A \backslash A = \emptyset$
- $A \backslash \emptyset = A$
- $\emptyset \backslash A = \emptyset$

Significant Sets

In mathematics, some sets are referred to so commonly that they are denoted by special symbols. Some of these numerical sets include:

- $\mathbb{P}$ denotes the set of all prime numbers: $\{2,3,5,7,11,\ldots\}$.
- $\mathbb{N}$ denotes the set of all natural numbers: $\{0,1,2,3,\ldots\}$.
- $\mathbb{Z}$ denotes the set of all integer numbers: $\{\ldots,-2,-1,0,1,2,\ldots\}$.
- $\mathbb{Z}^+$ denotes the set of all positive integer numbers: $\{1,2,3,\ldots\}$. Similarly, $\mathbb{Z}^-$ denotes the set of all negative integer numbers.
- $\mathbb{Q}$ denotes the set of all rational numbers: $\{\frac{a}{b} \mid a,b \in \mathbb{Z},\ b \neq 0\}$.
- $\mathbb{R}$ denotes the set of all real numbers.
- $\mathbb{C}$ denotes the set of all complex numbers.

In addition, the *Cartesian product* (also called the *cross product*) of two sets A and B, written as $A \times B$, is the set of all pairs whose first component is an element of A and whose second component is an element of B. In set notation, $A \times B = \{(a,b) \mid a \in A,\ b \in B\}$. For example, if $A = \{1,2,3\}$ and $B = \{u,v\}$, then $A \times B = \{(1,u),(1,v),(2,u),(2,v),(3,u),(3,v)\}$. Given a set S, another significant set is the *power set* of S, denoted by $\mathscr{P}(S)$, which is the set of all subsets of S: $\mathscr{P}(S) = \{T \mid T \subseteq S\}$. For example, if $S = \{1,2,3\}$, then the power set of S is: $\mathscr{P}(S) = \{\{\},\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}\}$. It is interesting to note that, if $|S| = k$, then $|\mathscr{P}(S)| = 2^k$.