Appendix A

Mathematical Tools

Can one learn mathematics by reading it? I am inclined to say no. Reading has an edge over listening because reading is more active—but not much. Reading with pencil and paper on the side is very much better—it is a big step in the right direction. The very best way to read a book, however, with, to be sure, pencil and paper on the side, is to keep the pencil busy on the paper and throw the book away.

Paul Halmos

In this appendix, we provide a mathematical toolbox that we use throughout the book—asymptotic notation, approximation methods and inequalities, the theory of discrete probability, the binomial, Gaussian, and Poisson distributions, methods for approximating asymptotic integrals, and the theory of finite groups. Where possible, we look “under the hood” so you can see how the relevant theorems are proved.

A.1 The Story of O

Like physics, the theory of computational complexity is concerned with scaling—how the resources we need to solve a problem increases with its size. In order to make qualitative distinctions between how different functions grow, we use $O$ and its companion symbols, the trademarks of asymptotic analysis.

Suppose that a careful analysis of an algorithm reveals that its running time $T(n)$ on instances of size $n$ is

$$T(n) = an^2 + bn + c$$

for some constants $a, b, c$ where $a > 0$. These constants depend on details of the implementation, the hardware we’re running the algorithm on, and the definition of elementary operations. What we really care about is that when $n$ is large, $T(n)$ is dominated by its quadratic term. In particular, there is a constant $d$ such that, for all $n > 0$,

$$T(n) \leq dn^2.$$  

We write this as $T(n) = O(n^2)$, and read “$T$ is big-oh of $n^2$. “
The general definition of $O$ is the following. Let $f$ and $g$ be two functions defined on the natural numbers. We say that $f(n) = O(g(n))$, or $f = O(g)$ for short, if there are constants $C$ and $n_0$ such that

$$f(n) \leq C g(n) \quad \text{for all } n > n_0. \quad (A.1)$$

We can also express this in terms of the limiting ratio between $f$ and $g$: there is a constant $C$ such that

$$\limsup_{n \to \infty} \frac{f(n)}{g(n)} \leq C.$$

Using the same definition, we can say that $f(n) = O(1)$ if $f(n)$ is bounded by a constant. We can also use $O$ to state that a real-valued function $f(n)$ decays to zero at a certain rate as $n \to \infty$. For instance, $f(n) = O(1/n)$ means that $f(n) \leq C/n$ for some constant $C$.

Here are some examples:

$$an^2 + bn + c = O(n^k) \text{ for any } k \geq 2$$
$$\sqrt{3n + 10} = O(\sqrt{n})$$
$$\log(n^7) = O(\log n)$$
$$n^k = O(2^n) \text{ for any constant } k$$
$$n! = O(n^n)$$
$$e^{\sin n} = O(1)$$
$$e^{-n} = O(n^{-c}) \text{ for any } c > 0 \quad (A.2)$$

We can also use $O$ inside arithmetic expressions. For instance, Stirling's approximation for the factorial is

$$n! = (1 + O(n^{-1})) \sqrt{2\pi n} n^n e^{-n}. \quad (A.3)$$

This means that, in the limit $n \to \infty$, the multiplicative error in this approximation is at most proportional to $n^{-1}$. In other words, there is a constant $C$ such that

$$n! = (1 + \epsilon) \sqrt{2\pi n} n^n e^{-n} \quad \text{where } |\epsilon| < Cn^{-1}. \quad (A.3)$$

Note that in cases like this, we use $O(n^{-1})$ to denote an error that could be positive or negative.

**Exercise A.1** Show that if $f_1 = O(g)$ and $f_2 = O(g)$ then $f_1 + f_2 = O(g)$.

**Exercise A.2** Show that the relation $O$ is transitive. That is, if $f = O(g)$ and $g = O(h)$ then $f = O(h)$.

**Exercise A.3** When we say $f(n) = O(\log n)$, why don't we need to state the base of the logarithm?

**Exercise A.4** What is wrong with the following argument? For any $k$, we have $k n = O(n)$. Hence

$$\sum_{k=1}^{n} kn = \sum_{k=1}^{n} O(n) = n O(n) = O(n^2).$$
Exercise A.5 Is $2^O(n)$ the same as $O(2^n)$? Why or why not?

Other sciences such as physics often use $f = O(g)$ to indicate that $f$ is proportional to $g$ when $n$ is large. However, in computer science it denotes an upper bound, and most of the bounds shown in (A.2) above are rather generous. You can think of $f = O(g)$ as the statement “$f$ grows at most as fast as $g$ does,” or “asymptotically, ignoring multiplicative constants, $f \leq g.$”

The analogous symbols for $\geq$ and $=$ are $\Omega$ and $\Theta$ respectively. We say that $f = \Omega(g)$ if and only if $g = O(f).$ In other words, there exist constants $C > 0$ and $n_0$ such that

$$f(n) \geq Cg(n) \quad \text{for all } n > n_0. \quad \text{(A.4)}$$

Alternately, $f = \Omega(g)$ if there is a constant $C > 0$ such that

$$\lim_{n \to \infty} \frac{f(n)}{g(n)} \geq C.$$

Thus $f(n)$ grows at least as fast, ignoring constants, as $g(n)$ does. We write $f = \Theta(g)$ if $f = O(g)$ and $g = O(f).$ Typically, this means that there is a constant $C > 0$ such that

$$\lim_{n \to \infty} \frac{f(n)}{g(n)} = C,$$

so that $f(n)$ is proportional to $g(n)$ when $n$ is large. It is possible, though rare in practice, for this ratio to oscillate between two constants without ever settling down.

We also have asymptotic versions of $<$ and $>.$ We say $f = o(g)$ if $f = O(g)$ but $f \neq \Theta(g),$ i.e., $f$ is dwarfed by $g$ in the limit $n \to \infty$:

$$\limsup_{n \to \infty} \frac{f(n)}{g(n)} = 0,$$

On the other side, we write $f = \omega(g)$ if $g = o(f),$ or equivalently

$$\liminf_{n \to \infty} \frac{f(n)}{g(n)} = \infty,$$

Exercise A.6 Which of the examples in (A.2) remain correct if we replace $O$ by $\Theta$? In which cases can we replace $O$ with $o$?

Exercise A.7 Give functions $f$ and $g$ such that $f = \Theta(g)$ but $2f = o(2^g),$ or conversely, functions such that $f = o(g)$ but $\log f = \Theta(\log g).$

Exercise A.8 For each pair of functions, state whether their relationship is $f = o(g),$ $f = \Theta(g),$ or $f = \omega(g)$.

1. $f(n) = \log \sqrt{n}, \ g(n) = \log(n^2)$
2. $f(n) = 3^{n/2}, \ g(n) = 2^n$
3. $f(n) = 2^n, \ g(n) = n \log n$
4. $f(n) = 2^{n+\log n}, \ g(n) = 2^n$
approximations and inequalities

<table>
<thead>
<tr>
<th>class</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>polylogarithmic</td>
<td>( f = O(\log^c n) ) for some constant ( c )</td>
</tr>
<tr>
<td>polynomial</td>
<td>( f = O(n^c) ) for some constant ( c ), or ( n^{O(1)} )</td>
</tr>
<tr>
<td>superpolynomial</td>
<td>( f = \omega(n^c) ) for every constant ( c ), or ( n^{\omega(1)} )</td>
</tr>
<tr>
<td>subexponential</td>
<td>( f = o(2^{n^c}) ) for every ( c &gt; 0 )</td>
</tr>
<tr>
<td>exponential</td>
<td>( f = \Omega(2^{n^c}) ) for some ( c &gt; 0 ), or ( f = 2^{\text{poly}(n)} )</td>
</tr>
</tbody>
</table>

Table A.1: Coarse grained classes in asymptotic analysis

We will often give coarser classifications of functions by grouping various classes of functions together as in Table A.1. For instance, the class poly\((n)\) of polynomial functions is the union, over all constant \( c \), of the class \( O(n^c) \), and we call a function exponential if it is \( 2^{\text{poly}(n)} \).

**Exercise A.9** Give several examples of functions \( f(n) \) which are superpolynomial but subexponential.

### A.2 Approximations and Inequalities

#### A.2.1 Norms

Given a real number \( r > 0 \), the \( r \)-norm of a vector \( v \) is

\[
\|v\|_r = \left( \sum_i |v_i|^r \right)^{1/r}.
\]

When \( r = 2 \), this is the standard Euclidean length. When \( r = 1 \), it becomes the “Manhattan metric” \( \sum_i |v_i| \), so called because it’s the total number of blocks we have to travel North, South, East, or West to reach our destination. In the limit \( r \to \infty \), the \( r \)-norm becomes the max norm,

\[
\|v\|_{\text{max}} = \max_i |v_i|.
\]

For any \( r \geq 1 \) the \( r \)-norm is convex in the sense that interpolating between two points doesn’t take you farther away from the origin. That is, for any \( 0 \leq \lambda \leq 1 \),

\[
\|\lambda x + (1 - \lambda)y\|_r \leq \max(\|x\|_r, \|y\|_r).
\]

The “unit disk” with respect to the \( r \)-norm, i.e., the set of vectors \( v \) such that \( \|v\|_r \leq 1 \), is a diamond for \( r = 1 \), a circle for \( r = 2 \), and a square for \( r = \infty \). The Danish mathematician and poet Piet Hein was particularly fond of the case \( r = 5/2 \). He felt that a “superellipse,” the set of points \( (x, y) \) such that \( ax^r + by^r = 1 \), was a good shape for tables and town squares [307]. See Figure A.1.

#### A.2.2 The triangle inequality

Suppose I have two vectors \( x, y \) as shown in Figure A.2. The triangle inequality is the fact that the length of their sum is at most the sum of their lengths,

\[
\|x + y\|_2 \leq \|x\|_2 + \|y\|_2.
\]
where this holds with equality if and only if $x$ and $y$ are parallel. More generally, for any set of vectors $v_i$,

$$\left\| \sum v_i \right\|_2 \leq \sum\|v\|_2.$$ 

The triangle inequality holds for the $r$-norm for any $r \geq 1$. If these quantities are simply numbers rather than vectors, it holds if $|v|$ denotes the absolute value.

### A.2.3 The Cauchy–Schwarz inequality

Another geometrical fact is that the inner product, or dot product, of two vectors is at most the product of their lengths:

$$x^T y = \sum_i x_i y_i \leq \|x\|_2 \|y\|_2.$$ 

Using Pythagoras' theorem and squaring both sides gives the Cauchy–Schwarz inequality,

$$\left( \sum_i x_i y_i \right)^2 \leq \left( \sum_i x_i^2 \right) \left( \sum_i y_i^2 \right),$$  

(A.5)
where equality holds only if $x$ and $y$ are parallel.

The Cauchy–Schwarz inequality holds for generalized inner products as well. If we have a set of non-negative coefficients $a_i$, then

$$\left( \sum_i a_i x_i y_i \right)^2 \leq \left( \sum_i a_i x_i^2 \right) \left( \sum_i a_i y_i^2 \right). \quad \text{(A.6)}$$

To see this, just write $x_i' = \sqrt{a_i} x_i$ and $y_i' = \sqrt{a_i} y_i$ and apply (A.5) to $x'$ and $y'$.

The Cauchy–Schwarz inequality is often useful even for sums that don't look like inner products. For instance,

$$\left( \sum_{i=1}^N v_i \right)^2 \leq N \sum_{i=1}^N v_i^2,$$

since we can think of the sum on the left-hand side as the inner product of $v$ with an $N$-dimensional vector $(1, \ldots, 1)$. This lets us bound the 1-norm of an $N$-dimensional vector in terms of its 2-norm,

$$\|v\|_1 = \sum_{i=1}^N |v_i| \leq \sqrt{N} \sum_{i=1}^N v_i^2 = \sqrt{N} \|v\|_2. \quad \text{(A.7)}$$

On the other hand, we also have

$$\|v\|_2 \leq \|v\|_1. \quad \text{(A.8)}$$

**Exercise A.10** Show that the extreme cases of the inequalities (A.7) and (A.8) occur when $v$ is a basis vector or when all its components are equal. Hint: consider the inscribed and circumscribed spheres of an octahedron.

### A.2.4 Taylor Series

Given an analytic function $f(x)$ we can approximate it by a Taylor series,

$$f(x) = f(0) + f'(0) x + \frac{1}{2} f''(0) x^2 + \frac{1}{6} f'''(0) x^3 + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!} f^{(k)}(0),$$

where $f^{(k)} = \frac{d^k f}{dx^k}$ is the $k$th derivative.

Since $e^x$ is its own derivative this gives the following series for the exponential,

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Another helpful series is

$$-\ln(1-x) = x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k},$$
which gives us, for instance,
\[(1 + x)^y = e^{y\ln(1 + x)} = e^{y(x - O(x^2))} = e^{xy} (1 - O(x^2y)).\]

In some cases the Taylor series gives a real inequality. For instance, for all \(x\) we have
\[1 + x \leq e^x.\]

For \(x > 0\), this follows since all the higher-order terms in \(e^x\) are positive. For \(-1 < x < 0\) their signs alternate, but they decrease in size geometrically.

### A.2.5 Stirling’s Approximation

Stirling gave a very useful approximation for the factorial,
\[n! = (1 + O(1/n)) \sqrt{2\pi n} n^n e^{-n}. \tag{A.9}\]

We prove this in Problem A.28. Some simpler approximations can be useful as well. For instance,
\[n! > n^n e^{-n},\]
and this gives us a useful upper bound on binomial coefficients \(\binom{n}{k}\),
\[\binom{n}{k} \leq \frac{n^k}{k!} \leq \left(\frac{en}{k}\right)^k. \tag{A.10}\]

As Problem A.13 shows, this approximation is tight as long as \(k = o(\sqrt{n})\).

### A.3 Chance and Necessity

The most important questions in life are, for the most part, really only problems of probability.

Pierre-Simon Laplace

Here we collect some simple techniques in discrete probability which we use in Chapter 10 and beyond. Most of the reasoning we present here is quite elementary, but even these techniques have some surprisingly powerful applications.

#### A.3.1 ANDs and ORs

The **union bound** is the following extremely simple observation. If \(A\) and \(B\) are events, the probability that one or the other of them occurs is bounded by
\[\Pr[A \cup B] \leq \Pr[A] + \Pr[B]. \tag{A.11}\]
More generally, if we have a set of possible events \( A_i \), the probability that at least one occurs is bounded by

\[
\Pr \left( \bigvee_i A_i \right) \leq \sum_i \Pr[A_i].
\]

Clearly the union bound (A.11) holds with equality only when \( A \) and \( B \) are disjoint, i.e., when \( \Pr[A \land B] = 0 \) so \( A \) and \( B \) never occur simultaneously. We can correct the union bound by taking this possibility into account, and this gives the inclusion–exclusion principle:

\[
\Pr[A \lor B] = \Pr[A] + \Pr[B] - \Pr[A \land B]. \quad (A.12)
\]

Think of this as calculating the area \( A \cup B \) in a Venn diagram. Adding the areas of \( A \) and \( B \) counts the area of \( A \cap B \) twice, and we correct this by subtracting it once.

In general, if we have a set \( S \) of events \( E_1, \ldots, E_n \), the probability that none of them occurs is an alternating sum over all subsets \( T \subseteq S \) of the probability that every event in \( T \) holds:

\[
\Pr \left[ \bigwedge_{i=1}^{n} E_i \right] = \sum_{T \subseteq \{1, \ldots, n\}} (-1)^{|T|} \Pr \left[ \bigwedge_{i \in T} E_i \right]. \quad (A.13)
\]

There are many ways to prove this—see Problem A.1 for one. Note that if \( T = \emptyset \) then \( \Pr[\bigcap_{i \in T} E_i] = 1 \).

### A.3.2 Expectations and Markov’s Inequality

Even when a random variable \( X \) is complicated, subtle, and full of correlations, it is often quite easy to calculate its average, or expectation, which we denote \( \mathbb{E}[X] \). The reason for this is the so-called linearity of expectation, the fact that the expectation of a sum is the sum of the expectations:

\[
\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].
\]

This is true even if \( X \) and \( Y \) are correlated. For instance, if \( h \) is the average height of a child, the average of the sum of two children’s heights is \( 2h \) even if they are siblings.

The expectation of the product, on the other hand, is not generally the product of the expectations: \( \mathbb{E}[XY] \) is larger or smaller than \( \mathbb{E}[X] \mathbb{E}[Y] \) if their correlation is positive or negative respectively. For instance, the average of the product of two siblings’ heights is greater than \( h^2 \).

Once we know the expectation of a random variable, a concentration inequality seeks to show that is it probably not too far from its expectation. The simplest and weakest of these is Markov’s inequality. Suppose \( X \) is a nonnegative random variable. Then for any \( \lambda \), the probability that \( X \) is at least \( \lambda \) times its expectation is bounded by

\[
\Pr[ X \geq \lambda \mathbb{E}[X] ] \leq 1/\lambda. \quad (A.14)
\]

The proof of this is very easy. If \( X \geq \tau \) with probability at least \( p \) and \( X \) is never negative, then \( \mathbb{E}[X] \geq p \tau \).

The variance of a random variable \( X \) is the expected squared difference between \( X \) and its expectation, \( \text{Var } X = \mathbb{E}[(X - \mathbb{E}[X])^2] \). Some simple things you should know about the variance are given in the following two exercises.
**Exercise A.11** Show that $\text{Var} X = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

**Exercise A.12** Two random variables $X$, $Y$ are independent if the joint probability distribution $P[X, Y]$ is the product $P[X]P[Y]$. Show that if $X$ and $Y$ are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ and $\text{Var}(X + Y) = \text{Var} X + \text{Var} Y$.

Markov's inequality can be used to prove another simple inequality, showing that if $\text{Var} X$ is small then $X$ is probably close to its expectation:

**Exercise A.13** Derive Chebyshev’s inequality,

$$\Pr[|X - \mathbb{E}[X]| > \delta] \leq \frac{\text{Var} X}{\delta^2}. \quad (A.15)$$

*Hint: apply Markov’s inequality to the random variable $(X - \mathbb{E}[X])^2$.*

Chebyshev’s inequality is a good example of a fact we will see many times—that half the battle is choosing which random variable to analyze.

Now suppose that $X$ is a nonnegative integer which counts the number of objects of some kind. Then Markov’s inequality implies

$$\Pr[X > 0] = \Pr[X \geq 1] \leq \mathbb{E}[X]. \quad (A.16)$$

An even easier way to see this is

$$\Pr[X > 0] = \sum_{x=1}^{\infty} \Pr[X = x] \leq \sum_{x=1}^{\infty} x \Pr[X = x] = \mathbb{E}[X].$$

We can use this to prove that a given type of object probably doesn’t exist. If we can show that the expected number of such objects is $o(1)$, then (A.16) implies that $\Pr[X > 0] = o(1)$. Then, with probability $1 - o(1)$ we have $X = 0$ and none of these objects exist. This is often called the *first moment method*, since the $k$th moment of $X$ is $\mathbb{E}[X^k]$.

### A.3.3 The Birthday Problem

As an application of some of these techniques, consider the so-called Birthday Problem. How many people do we need to have in a room before it becomes likely that two of them have the same birthday? Assume that each person’s birthday is uniformly random, i.e., that it is chosen from the 365 possibilities (ignoring leap years) with equal probability. Most people guess that the answer is roughly half of 365, but this is very far off the mark.

If there are $n$ people and $y$ days in the year, there are \(\binom{n}{2}\) possible pairs of people. For each pair, the probability that they have the same birthday is $1/y$. Thus by the union bound, the probability that at least one pair of people have the same birthday is at most

$$\binom{n}{2} \frac{1}{y} \approx \frac{n^2}{2y}.$$

This is small until $n \approx \sqrt{2y}$, or roughly 27 in the case $y = 365$. This is just an upper bound on this probability, but it turns out that $\sqrt{2y}$ is essentially the right answer.
Let’s repeat this calculation using the first moment method. Let $B$ denote the number of pairs of people with the same birthday. By linearity of expectation, $E[B]$ is the sum over all pairs $i, j$ of the probability $p_{ij}$ that $i$ and $j$ have the same birthday. Once again, there are $\binom{n}{2}$ pairs and $p_{ij} = 1/y$ for all of them, so

$$E[B] = \binom{n}{2} \frac{1}{y} \approx \frac{n^2}{2y}. \quad (A.17)$$

This calculation may seem exactly the same as our union bound above, but there is an important difference. That was an upper bound on the probability that any pair share a birthday, while this is an exact calculation of the expected number of pairs that do.

**Exercise A.14** Suppose there are three people, Albus, Barholomew and Cornelius, and consider the possibility that one or more pairs were born on the same day of the week. These events are positively correlated: for instance, if this is true of two pairs then it is true of all three, so we never have $B = 2$. Nevertheless, confirm that (A.17) with $n = 3$ and $y = 7$ gives the expectation of $B$ exactly.

A slightly more formal way to write this is to define an indicator random variable for each pair of people, associated with the event that they have the same birthday:

$$X_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ have the same birthday} \\ 0 & \text{if they don’t.} \end{cases}$$

Then

$$B = \sum_{i<j} X_{ij}.$$ 

Since $E[X_{ij}] = p_{ij}$, linearity of expectation gives

$$E[B] = \sum_{i<j} E[X_{ij}] = \sum_{i<j} p_{ij} = \binom{n}{2} \frac{1}{y}.$$ 

Now that we have computed $E[B]$, we can use the first moment method. If $n = o(\sqrt{y})$ then $E[B] = o(1)$, so the probability that there is a pair of people with the same birthday is $o(1)$. Thus we can say that with high probability, i.e., with probability $1 - o(1)$, all the birthdays are different.

It’s important to realize that even if the expected number of pairs is large, that doesn’t prove that there probably is one. In other words, a lower bound on $E[B]$ does not, in and of itself, give a lower bound on $Pr[B > 0]$. The reason is that the distribution of $B$ could have a very heavy tail, in which it is zero almost all the time, but is occasionally enormous. Such a variable has a very high variance, so we can eliminate this possibility by placing an upper bound on $E[B^2]$. This is the point of the second moment method outlined below. The Birthday Problem is simple enough, however, that we can prove that a pair with the same birthday really does appear with high probability when $n \approx \sqrt{2y}$; see Problem A.6.
A.3.4 Coupon Collecting

Another classic problem in discrete probability is the Coupon Collector’s Problem. In order to induce children to demand that their parents buy a certain kind of cereal, the cereal company includes a randomly chosen toy at the bottom of each box. If there are $n$ different toys and they are equally likely, how many boxes of cereal do I have to buy for my daughter before she almost certainly has one of each? Clearly I need to buy more than $n$ boxes—but how many more?

Suppose I have bought $b$ boxes of cereal so far, and let $T$ be the number of toys I still don’t have. The probability I am still missing a particular toy is $(1 - 1/n)^b$, since I get that toy with probability $1/n$ each time I buy a box. By linearity of expectation, the expected number of toys I am missing is

$$E[T] = n(1 - 1/n)^b < ne^{-b/n}.$$  

This is 1 when $b = n \ln n$. Moreover, if $b = (1 + \epsilon)n \ln n$ for some $\epsilon > 0$, then $E[T] < n^{-\alpha} = o(1)$, and by Markov’s inequality, with high probability my daughter has a complete collection.

Another type of result we can obtain is the expected number of boxes it takes to obtain a complete collection. First consider the following exercise:

**Exercise A.15** Suppose I have a biased coin which comes up heads with probability $p$. Let $t$ be the number of times I need to flip it to get the first head. Show that $E[t] = 1/p$. First do this the hard way, by showing that the probability the first head comes on the $t$th flip is $P(t) = (1 - p)^{t-1}p$ and summing the series $\sum P(t) t$. Then do it the easy way, by showing that $E[t] = p + (1 - p)(E[t] + 1)$.

Now suppose that I am missing $i$ of the toys. The probability that the next box of cereal adds a new toy to my collection is $i/n$, so the average number of boxes I need to buy to get a new toy is $n/i$. For instance, if I am a new collector I gain a new toy in the very first box, while if my collection is complete except for one last toy, it will take me an average of $n$ boxes to get it.

By linearity of expectation, the expected number of boxes it takes to get a complete collection is then the sum of $n/i$ over all $i$,

$$E[b] = \frac{n}{1} + \frac{n}{2} + \frac{n}{3} + \cdots + \frac{n}{n} = n \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right) = nH_n \approx n \ln n.$$  

Here $H_n = \sum_{i=1}^{n} 1/i$ is the $n$ harmonic number, which diverges like $\ln n$ as $n$ grows.

Note the difference between this result, which gives the expected number of boxes I need, and the previous one, which gave a number of boxes after which I have all the toys with high probability.

A.3.5 Bounding the Variance: the Second Moment Method

The first moment method is an excellent way to prove that certain things probably don’t exist. If $X$ is the number of some kind of thing then $\Pr[X > 0] \leq E[X]$, so if $E[X]$ is small then probably $X = 0$ and none of these things exist.

How can we derive a lower bound on $\Pr[X > 0]$, and thus prove that one of these things probably does exist? As mentioned above, it is not enough to show that $E[X]$ is large, since this could be due to $X$ occasionally being enormous. To exclude this possibility, we can bound $X$’s variance, or equivalently its second moment $E[X^2]$.
One way to do this is to apply Chebyshev’s inequality from Exercise A.13, as the following exercise suggests:

**Exercise A.16** Suppose \( \mathbb{E}[X] > 0 \). Show that

\[
\Pr[X > 0] \geq 1 - \frac{\text{Var } X}{\mathbb{E}[X]^2} = 2 - \frac{\mathbb{E}[X^2]}{\mathbb{E}[X]^2}.
\]

However, we will derive the following inequality, which is often much stronger:

\[
\Pr[X > 0] \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}.
\]

This is called the second moment method.

We prove (A.18) using the Cauchy–Schwarz inequality from Appendix A.2. If \( X \) and \( Y \) depend on some random variable \( i \) with a probability distribution \( p_i \), we can think of \( \mathbb{E}[XY] \) as an inner product,

\[
\mathbb{E}[XY] = \sum_i p_i X_i Y_i
\]

and using (A.6) then gives

\[
\mathbb{E}[XY]^2 \leq \left( \sum_i p_i X_i^2 \right) \left( \sum_i p_i Y_i^2 \right) = \mathbb{E}[X^2] \mathbb{E}[Y^2].
\]

Now let \( Y \) be the indicator random variable for the event \( X > 0 \), i.e., \( Y = 1 \) if \( X > 0 \) and \( Y = 0 \) if \( X = 0 \). Using the facts that \( X = XY \), \( Y^2 = Y \), and \( \mathbb{E}[Y] = \Pr[X > 0] \), the Cauchy–Schwarz inequality (A.19) gives

\[
\mathbb{E}[X]^2 = \mathbb{E}[XY]^2 \leq \mathbb{E}[X^2] \mathbb{E}[Y^2] = \mathbb{E}[X^2] \mathbb{E}[Y] = \mathbb{E}[X^2] \Pr[X > 0].
\]

Dividing both sides by \( \mathbb{E}[X^2] \) completes the proof of (A.18).

How do we calculate the second moment? Suppose that \( X \) is a sum of indicator random variables \( X_i \), each of which is associated with an event \( i \). Then

\[
\mathbb{E}[X^2] = \mathbb{E} \left[ \sum_{i,j} X_i X_j \right] = \sum_{i,j} \mathbb{E}[X_i X_j] = \sum_{i,j} \Pr[i \land j].
\]

In other words, \( \mathbb{E}[X^2] \) is the sum over all ordered pairs of events \( i, j \) of the probability that they both occur. We can write this as the probability of \( i \) times the conditional probability of \( j \) given \( i \),

\[
\Pr[i \land j] = \Pr[i] \Pr[j | i]
\]

which gives

\[
\mathbb{E}[X^2] = \sum_i \Pr[i] \sum_j \Pr[j | i].
\]

**A.20**
In order to calculate the conditional probability \( \Pr[j | i] \), we need to know how these events are correlated. In particular, if knowing that \( i \) occurs makes it more likely that \( j \) occurs, then \( \Pr[j | i] > \Pr[j] \). Pairs of events for which this effect is strong make a larger contribution to the second moment. In many applications, these events are only weakly correlated. Then \( \mathbb{E}[X^2] = (1 + o(1)) \mathbb{E}[X]^2 \), and \( X \) is concentrated around its expectation.

However, in Chapter 14 we consider events which are strongly correlated, such as the events that two truth assignments both satisfy the same random formula. In such cases, we can often prove that \( \mathbb{E}[X^2] \) is at most \( C \mathbb{E}[X]^2 \) for some constant \( C \), and therefore that \( \Pr[X > 0] \geq 1/C \).

### A.3.6 Jensen’s inequality

Let \( x \) be a random variable, and let \( f \) be a function. In general, \( \mathbb{E}[f(x)] = f(\mathbb{E}[x]) \) only if \( x \)'s probability distribution is concentrated at a single value or if \( f \) is a straight line. If \( x \) has some variance around its most likely value, and if \( f \) has some curvature, then \( \mathbb{E}[f(x)] \) and \( f(\mathbb{E}[x]) \) can be quite different.

We say that \( f(x) \) is convex if, for any \( x_1, x_2 \) and any \( \lambda \in [0, 1] \) we have

\[
f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).
\]

In other words, if we draw a line segment between any two points on the graph of \( f \) as in Figure A.3, then \( f \) lies at or below this line segment. For a continuous function, we can say equivalently that the second derivative \( f'' \) is nonnegative. **Jensen’s inequality** states that for any convex function \( f \) and any probability distribution on \( x \),

\[
\mathbb{E}[f(x)] \geq f(\mathbb{E}[x]).
\]

For instance, \( \mathbb{E}[x^2] \geq \mathbb{E}[x]^2 \) and \( \mathbb{E}[e^x] \geq e^{\mathbb{E}[x]} \).

**Exercise A.17** Prove Jensen’s inequality for the case where \( x \) takes two different values, \( x_1 \) and \( x_2 \) with probability \( p \) and \( 1 - p \) respectively.

Similarly, \( \mathbb{E}[f(x)] \leq f(\mathbb{E}[x]) \) for any concave function \( f \), i.e., a function whose second derivative is less than or equal to zero. For instance, \( \mathbb{E}[\ln x] \leq \ln \mathbb{E}[x] \) and \( \mathbb{E}[\sqrt{x}] \leq \sqrt{\mathbb{E}[x]} \).
A.4 Dice and Drunkards

Many random processes generate probability distributions of numbers and positions—for instance, the number of heads in a sequence of coin flips, or the position we end up in when we take a random walk. In this section we discuss these distributions, how concentrated they are around their expectations, and how to estimate the time that it takes a random walk to travel a certain distance from its starting point.

A.4.1 The Binomial Distribution

Suppose I flip a series of \( n \) coins, each one of which comes up heads with probability \( p \) and tails with probability \( 1 - p \). The probability that exactly \( x \) of them come up heads is the number of subsets of size \( x \), times the probability that these all come up heads, times the probability that all the others come up tails:

\[
P(x) = \binom{n}{x} p^x (1 - p)^{n-x}.
\]

This is called the binomial distribution, and is often written Bin\((n, p)\). We can also think of \( x \) as the sum of \( n \) independent random variables,

\[
x = \sum_{i=1}^{n} y_i \quad \text{where} \quad y_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}
\]

By linearity of expectation (and common sense!) the expectation of the binomial distribution is \( \mathbb{E}[x] = pn \).

The next exercise asks you to calculate its variance.

**Exercise A.18** Show that the variance of the binomial distribution Bin\((n, p)\) is \( p(1 - p)n \).

A.4.2 The Poisson Distribution

Suppose that I toss \( m \) balls randomly into \( n \) bins. What is the probability that the first bin has exactly \( x \) balls in it? This is a binomial distribution where \( p = 1/n \), so we have

\[
P(x) = \binom{m}{x} \left( \frac{1}{n} \right)^x \left( 1 - \frac{1}{n} \right)^{m-x}
\]

(A.21)

Now suppose that \( m = cn \) for some constant \( c \). The average number of balls in any particular bin is \( c \), independent of \( n \), so \( P(x) \) should become independent of \( n \) in the limit \( n \to \infty \). What distribution does \( P(x) \) converge to?

First we claim that for any constant \( x \), in the limit \( m \to \infty \) we have

\[
\binom{m}{x} \approx \frac{m^x}{x!}.
\]

Then (A.21) becomes

\[
P(x) = \lim_{n \to \infty} \frac{c^x}{x!} \left( \frac{1}{n} \right)^x \left( 1 - \frac{1}{n} \right)^{n-x} = \frac{e^{-c} c^x}{x!}.
\]

(A.22)
As Problem A.13 shows, this limit holds whenever \( x = \Theta(\sqrt{n}) \), and in particular for any constant \( x \).

The distribution (A.22) is called the Poisson distribution, and you should check that it has mean \( c \). It occurs, for instance, as the degree distribution of a sparse random graph in Section 14.2.

In physics, the Poisson distribution appears in the following way. If I am using a Geiger counter to measure a radioactive material, and if in each infinitesimal time interval \( dt \) there is an probability \( p \, dt \) of hearing a click, the total number of clicks I hear in \( T \) seconds is Poisson-distributed with mean \( p \, T \). Just as in the balls-in-bins example, this is an extreme case of the binomial distribution—there are a large number of events, each one of which occurs with very small probability.

### A.4.3 Entropy and the Gaussian Approximation

Let’s rephrase the binomial distribution a bit, and ask for the probability that the fraction \( x/n \) of heads that come up is some \( a \) between 0 and 1. Applying Stirling’s approximation A.9 to the binomial and simplifying gives

\[
\binom{n}{an} \approx \frac{1}{\sqrt{2\pi a(1-a)n}} \left( \frac{1}{a^a(1-a)^{1-a}} \right)^n \approx \frac{1}{\sqrt{2\pi a(1-a)n}} e^{nh(a)}. \tag{A.23}
\]

Here \( h(a) \) is the Gibbs–Shannon entropy,

\[
h(a) = -a \ln a - (1-a) \ln(1-a).
\]

We can think of \( h(a) \) as the average amount of information generated by a random coin which comes up heads with probability \( a \). Its maximum value is \( h(1/2) = \ln 2 \), since a fair coin generates one bit of information per toss.

Applying (A.23) to the binomial distribution gives

\[
P(an) = \frac{1}{\sqrt{2\pi a(1-a)n}} e^{nh(a\|p)}, \tag{A.24}
\]

where \( h(a\|p) \) is a weighted version of the Shannon entropy,

\[
h(a\|p) = -a \ln \frac{a}{p} - (1-a) \ln \frac{1-a}{1-p}.
\]

We can think of \( -h(a\|p) \) as a kind of distance between the probability distribution \((p, 1-p)\) and the observed distribution \((a, 1-a)\). It is called the Kullback–Leibler divergence. It is minimized at \( a = p \), where \( h(a\|p) = 0 \). The second-order Taylor series around \( p \) then gives

\[
h(p+\epsilon\|p) = -\frac{1}{2} \frac{\epsilon^2}{p(1-p)} + O(\epsilon^3),
\]

and the second-order term is an upper bound.
Let’s plug this back in to (A.24), assume that the slowly-varying part $1/\sqrt{2\pi a(1-a)n}$ is constant for $a \approx p$. Then setting $\delta = \epsilon n$ gives

$$P(pn + \delta) \approx \frac{1}{\sqrt{2\pi p(1-p)n}} \exp\left(-\frac{1}{2} \frac{\delta^2}{p(1-p)n}\right).$$

The Gaussian or normal distribution with mean zero and variance $\nu$ is

$$p(y) = \frac{1}{\sqrt{2\pi \nu}} e^{-y^2/2\nu}.$$ (A.25)

Thus when $n$ is large, the deviation $\delta$ of a binomial random variable from its mean $pn$ is distributed as a Gaussian with mean zero and variance $p(1-p)n$. This is one manifestation of the Central Limit Theorem, which states that the sum of any set of independent random variables with bounded variance converges to a Gaussian. In particular, it is tightly concentrated around its mean in a sense we will see below.

### A.4.4 Random Walks

One example of a binomial distribution is a random walk on the line. I have had too much to drink. At each step, I stumble one unit to the left or right with equal probability. If I start at the origin and take $t$ steps, the number $r$ of steps in which I move to the right is binomially distributed with $p = 1/2$, and I end at a position $x$ if $r = (x+t)/2$. The probability this happens is

$$P(x) = 2^{-t} \binom{t}{(x+t)/2} \approx \frac{2}{\sqrt{2\pi t}} e^{-x^2/2t}.$$  

Except for the factor of 2, which appears since only the values of $x$ with the proper parity, odd or even, can occur, this is a Gaussian with mean zero and variance $t$. Since the width of this distribution is $O(\sqrt{T})$, intuitively it takes about $t \sim n^2$ steps for me to get $n$ steps away from the origin—or to reach the origin from an initial position $n$ steps away.

Let’s prove a precise result to this effect. Suppose we have a random walk on a finite line, ranging from $0$ to $n$. We move left or right with equal probability unless we are at the right end, where we have no choice but to move to the left. If we ever reach the origin, we stay there.

Let $t(x)$ be the expected time it takes to reach the origin, given that we start at position $x$ where $0 \leq x \leq n$. Since $t(x) = 1 + \mathbb{E}[t(x')]$, where $x'$ is wherever we’ll be on the next step, for any $0 < x < n$ we have the following recurrence:

$$t(x) = 1 + \frac{t(x-1) + t(x+1)}{2}.$$  

Combined with the boundary conditions $t(0) = 0$ (since we’re already there) and $t(n) = f(n-1) + 1$ (since we bounce off the right end), the unique solution is

$$t(x) = x(2n-x) \leq n^2.$$ (A.26)

**Exercise A.19** Prove (A.26).
Of course, knowing the expected time to reach the origin isn’t the same as knowing a time by which we will reach it with high probability. To get a result of that form, we can cleverly focus on a different random variable. While the expectation of $x$ stays the same at each step, the expectation of $\sqrt{x}$ decreases. Specifically, a little algebra gives

$$E[\sqrt{x}] = \frac{1}{2} (\sqrt{x+1} + \sqrt{x-1}) \leq \sqrt{x} \left( 1 - \frac{1}{8x^2} \right) \leq \sqrt{x} e^{-1/8x^2}.$$  

Now let’s consider a walk on the half-infinite line, where if we ever touch the origin we stay there. If our initial position is $n$, then after $t$ steps we have

$$E[\sqrt{n}] \leq \sqrt{t} e^{-1/8n^2}.$$  

The probability we have still not reached the origin is the probability that $\sqrt{x} \geq 1$, and by Markov’s inequality this is at most $E[\sqrt{x}]$. Setting $t = 8n^2 \ln n$, say, gives $E[\sqrt{x}] \leq 1/\sqrt{n}$, so we have reached the origin with high probability.

### A.5 Concentration inequalities

The Markov and Chebyshev inequalities discussed in Section A.3.2 state, in a very weak fashion, that a random variable is probably not too far from its expectation. For distributions such as the binomial or Poisson distributions we can get much tighter results. In this section, we prove several inequalities showing that a random variable is close to its expectation with high probability.

#### A.5.1 Chernoff Bounds

One classic type of concentration inequality is the Chernoff bound. It asserts that if $x$ is chosen from the binomial distribution $\text{Bin}(n, p)$, the probability that $x$ differs from its expectation by a factor $1 \pm \delta$ decreases exponentially as a function of $\delta$ and $n$.

In order to bound the probability that $x \geq t$ for a given $t$, we start with a clever choice of random variable. We apply Markov’s inequality to $e^{\lambda x}$, rather than to $x$ itself. This gives

$$\Pr[x \geq t] = \Pr[e^{\lambda x} \geq e^{\lambda t}] \leq \frac{E[e^{\lambda x}]}{e^{\lambda t}}.$$  

(A.27)

This inequality is true for any value of $\lambda$, so we are free to choose $\lambda$ however we like. Below, we will select $\lambda$ in order to make our bounds as tight as possible.

The expectation $E[e^{\lambda x}]$ is called the moment generating function of a distribution. It is especially easy to calculate for the binomial distribution $\text{Bin}(n, p)$. Recall that $x$ is the sum of $n$ independent random variables $y_i$, each of which is 1 with probability $p$, and 0 with probability $1 - p$. Thus we can write

$$e^{\lambda x} = e^{\lambda \sum_{i=1}^{n} y_i} = \prod_{i=1}^{n} e^{\lambda y_i}.$$  

The $y_i$ are independent, so the expectation of this product is the product of these expectations. Since

$$E[e^{\lambda y_i}] = p e^\lambda + (1 - p),$$

\[\]
this gives
\[ \mathbb{E}[e^{\lambda x}] = \prod_{i=1}^{n} \mathbb{E}[e^{\lambda y_i}] = (pe^{-\lambda} + (1-p))^n = (1+(e^{-\lambda} - 1)p)^n. \]
Since \(1+z \leq e^z\), we can bound this as
\[ \mathbb{E}[e^{\lambda x}] \leq e^{(e^{-\lambda} - 1)p n} = e^{(e^{-\lambda} - 1)\mathbb{E}[x]}. \tag{A.28} \]

**Exercise A.20** Show that if \(x\) is Poisson-distributed then (A.28) is exact. That is, its moment generating function is
\[ \mathbb{E}[e^{\lambda x}] = e^{(e^{-\lambda} - 1)\mathbb{E}[x]}, \]

We are interested in bounding the probability that \(x\) is a factor \(1+\delta\) larger than its expectation. So we set \(t = (1+\delta)\mathbb{E}[x]\) and plug (A.28) into (A.27), giving
\[ \Pr[x \geq (1+\delta)\mathbb{E}[x]] \leq \frac{e^{(e^{-\lambda} - 1)\mathbb{E}[x]}}{e^{(e^{-\lambda} - 1)\mathbb{E}[x]}} = \left(\frac{\mathbb{E}[x]}{1+\delta}\right)^{\mathbb{E}[x]} \tag{A.29} \]
As stated above, we can now tune \(\lambda\) to turn (A.29) into the tightest bound possible. At this point, we ask the reader to take over.

**Exercise A.21** Find the value of \(\lambda\) that minimizes the right-hand side of (A.29) and show that, for any \(\delta \geq 0\),
\[ \Pr[x \geq (1+\delta)\mathbb{E}[x]] \leq \left(\frac{\mathbb{E}[x]}{1+\delta}\right)^{\mathbb{E}[x]} \]
Then use the inequalities
\[ \delta - (1+\delta)\ln(1+\delta) \leq \begin{cases} -\delta^2/3 & \text{if } 0 \leq \delta \leq 1 \\ -\delta/3 & \text{if } \delta > 1 \end{cases} \]
to prove the following theorem:

**Theorem A.1** For any binomial random variable \(x\),
\[ \Pr[x \geq (1+\delta)\mathbb{E}[x]] \leq \begin{cases} e^{-\delta^2\mathbb{E}[x]/3} & \text{if } 0 \leq \delta \leq 1 \\ e^{-\delta\mathbb{E}[x]/3} & \text{if } \delta > 1 \end{cases} . \]
Note that when \(\delta\) is small this bound decays exponentially with \(\delta^2\), just like the Gaussian approximation for the binomial in Section A.4.3. Indeed, it is identical to (A.25) except that we replace the constant \(1/2\) in the exponent with the weaker \(1/3\) in order to obtain a valid inequality in the interval \(\delta \in [0,1]\).

Similarly, we can bound the probability that \(x\) is a factor \(1-\delta\) less than its expectation. Applying Markov’s inequality to the random variable \(e^{-\lambda x}\) gives
\[ \Pr[x \leq t] \leq e^{\lambda t} \mathbb{E}[e^{-\lambda x}]. \]
Exercise A.22 Show that
\[ \Pr[x \leq (1 - \delta) \mathbb{E}[x]] \leq \left( e^{e^\lambda (1 - \delta) - 1} \right)^{\mathbb{E}[x]} . \]
Minimize the right-hand side as a function of \( \lambda \) and use the inequality
\[ -\delta - (1 - \delta) \ln(1 - \delta) \leq \delta^2 / 2 \quad \text{if} \ 0 \leq \delta \leq 1 \]
to prove the following theorem:

Theorem A.2 For any binomial random variable \( x \), if \( \delta \geq 0 \) then
\[ \Pr[x \leq (1 - \delta) \mathbb{E}[x]] \leq e^{\mathbb{E}[x] \delta^2 / 2} . \]

A common use of the Chernoff bound is to show that, if a randomized algorithm gives the correct yes-or-no answer with probability greater than 1/2, we can get the correct answer with high probability by running it multiple times and taking the majority of its answers. Consider the following exercise:

Exercise A.23 Suppose a coin comes up heads with probability \( 1/2 + \epsilon \) for some \( \epsilon > 0 \). Show that if we flip it \( n \) times, the majority of these flips will be heads with probability \( 1 - e^{\Omega(\epsilon^2 n)} \).

A.5.2 Martingales and Azuma’s inequality

The Chernoff bounds of the previous section show that if \( x \) is the sum of \( n \) independent random variables \( y_i \) then \( x \) is probably close to its expectation. In this section, we show that under certain circumstances, this is true even if the \( y_i \) are correlated.

First, let’s prove another type of Chernoff bound. Here the \( y_i \) are still independent, but each one has its own probability distribution. All we ask is that the \( y_i \) are bounded, and that each one has zero expectation.

Theorem A.3 Let \( x = \sum_{i=1}^n y_i \) where each \( y_i \) is chosen independently from some probability distribution \( p_i \) such that \( |y_i| \leq 1 \) and \( \mathbb{E}[y_i] = 0 \). Then for any \( a \geq 0 \),
\[ \Pr[|x| > a] \leq 2e^{-a^2/(2n)} . \]

Proof We again use the moment generating function. For each \( i \), we have
\[ \mathbb{E}[e^{\lambda y_i}] = \int_{-1}^{1} p_i(y_i) e^{\lambda y_i} \, dy_i . \quad (A.30) \]
The function \( e^{\lambda y} \) is convex, as in Section A.3.6. Thus for any \( y \in [-1, 1] \) we have
\[ e^{\lambda y} \leq \frac{1 + y}{2} e^\lambda + \frac{1 - y}{2} e^{-\lambda} = \cosh \lambda + y \sinh \lambda , \]
since this is the value we would get by drawing a straight line between \( y = +1 \) and \( y = -1 \) interpolating. This lets us bound the integral in (A.30) as follows,
\[ \mathbb{E}[e^{\lambda y_i}] \leq \int_{-1}^{1} p_i(y_i) (\cosh \lambda + y_i \sinh \lambda) \, dy_i = \cosh \lambda + \mathbb{E}[y_i] \sinh \lambda = \cosh \lambda . \quad (A.31) \]
Since the $y_i$ are independent, the moment generating function of $x$ is bounded by

$$\mathbb{E}[e^{\lambda x}] = \prod_{i=1}^{n} \mathbb{E}[e^{\lambda y_i}] \leq (\cosh \lambda)^n.$$ 

Applying Markov’s inequality to the random variable $e^{\lambda x}$ then gives, for any $\lambda > 0$,

$$\Pr[x > a] = \Pr[e^{\lambda x} > e^{\lambda a}] \leq \frac{(\cosh \lambda)^n}{e^{\lambda a}}.$$

Using the inequality

$$\cosh \lambda \leq e^{\lambda^2/2},$$

we can bound this further as

$$\Pr[x > a n] \leq e^{\lambda^2 n/2 - \lambda a}.$$ 

This is minimized when $\lambda = a/n$, in which case

$$\Pr[x > a] \leq e^{-a^2/(2n)}.$$ 

By symmetry, the same bound applies to $\Pr[x < -a]$, and by the union bound we have $\Pr[|x| > a] \leq 2e^{-a^2/(2n)}$ as stated. □

Now suppose that the $y_i$ are correlated rather than independent. We can think of choosing them one at a time, where each $y_i$ has a probability distribution $p(y_i | y_1, \ldots, y_{i-1})$ conditioned on all the ones before it. Each of these conditional distributions has an expectation $\mathbb{E}[y_i | y_1, \ldots, y_{i-1}]$. But if these conditional expectations are all zero, and if we have $|y_i| \leq 1$ as before, then exactly the same proof goes through. In other words,

**Theorem A.4** Let $x = \sum_{i=1}^{n} y_i$, where the $y_i$ are chosen from some joint probability distribution such that $|y_i| \leq 1$ and $\mathbb{E}[y_i | y_1, \ldots, y_{i-1}] = 0$. Then for any $a \geq 0$,

$$\Pr[|x| > a] \leq 2e^{-a^2/(2n)}.$$ 

**Proof** Let $x_j = \sum_{i=1}^{j} y_i$. We will show by induction on $j$ that the moment generating function of $x_j$ is at most $(\cosh \lambda)^j$, just as in Theorem A.3. Then if we take $j = n$, the analytic part of the proof will work just as before.

Assume by induction that

$$\mathbb{E}[e^{\lambda x_{j-1}}] \leq (\cosh \lambda)^{j-1}.$$ 

Now, for any $y_1, \ldots, y_{j-1}$, the conditional expectation of $e^{\lambda y_j}$ is bounded by

$$\mathbb{E}[e^{\lambda y_j} | y_1, \ldots, y_{j-1}] \leq \cosh \lambda,$$

just as in (A.31). Therefore, increasing the number of variables from $j - 1$ to $j$ multiplies the moment generating function by at most a factor of $\cosh \lambda$:

$$\begin{align*}
\mathbb{E}[e^{\lambda x_j}] &= \mathbb{E}[e^{\lambda x_{j-1}} e^{\lambda y_j}] \\
&= \mathbb{E}_{y_1, \ldots, y_{j-1}} \left[ e^{\lambda x_{j-1}} \mathbb{E}[e^{\lambda y_j} | y_1, \ldots, y_{j-1}] \right] \\
&\leq (\cosh \lambda) \mathbb{E}[e^{\lambda x_{j-1}}] \\
&\leq (\cosh \lambda)^j.
\end{align*}$$
We leave the base case $j = 0$ to the reader.

The requirement that each $y_i$ have zero conditional expectation may seem artificial. However, it comes up in a large family of probabilistic processes. A martingale is a sequence of random variables $x_0, x_1, x_2, \ldots$ such that, while each $x_t$ can depend on all the previous ones, its expectation is equal to the previous one:

$$
\mathbb{E}[x_t \mid x_0, \ldots, x_{t-1}] = x_{t-1}.
$$

If we define $y_i = x_i - x_{i-1}$ then $x_n = x_0 + \sum_{i=1}^n y_i$, and each $y_i$ has zero conditional expectation. Thus we can restate Theorem A.4 as follows:

**Theorem A.5 (Azuma's inequality)** Let the sequence $x_0, x_1, \ldots$ be a martingale with the property that $|x_t - x_{t-1}| \leq 1$ for all $i$. Then for any $a \geq 0$,

$$
\Pr[|x_n - x_0| > a] \leq 2e^{-a^2/(2n)}.
$$

One example of a martingale is the position $x_t$ of a random walk after $t$ steps, where $x_t - x_{t-1}$ is $+1$ or $-1$ with equal probability. If the random walk is biased, so that $x_t = x_{t-1} + 1$ or $x_{t-1} - 1$ with probability $p$ or $1 - p$ respectively, we can turn it into a martingale by subtracting away the expected change. Thus $z_t$ is a martingale where

$$
z_t = x_t - (2p - 1)t.
$$

Of course, $x_t$ and $z_t$ are sums of independent random variables, so we could show that they are concentrated around their expectations using Chernoff bounds instead. The power of Azuma's inequality is that it proves concentration even when these variables are based on correlated events, such as the steps of the algorithms for 3-SAT we analyze in Section 14.3.

### A.6 Asymptotic Integrals

There are two chapters in this book where we need to understand the asymptotic behavior of certain integrals. In Chapter 14, these integrals give the second moment for the number of satisfying assignments of a random formula. In Chapter 15, they give the probability distribution of a quantum walk. Even though the first kind of integral is real and the second is complex, their analysis is similar.

#### A.6.1 Laplace's Method

In Chapter 14, our calculations of the second moment of the number of solutions of a random SAT formula involve integrals of the form

$$
I = \int_a^b f(x)e^{n\phi(x)}dx,
$$

where $f(x)$ and $\phi(x)$ are smooth functions. In the limit $n \to \infty$, this integral is dominated by values of $x$ close to the $x_{\text{max}}$ that maximizes $\phi$, and the contributions from all other $x$ are exponentially smaller. Moreover, we can calculate the width of the interval which makes a significant contribution to the integral.
Assume for simplicity that \( \phi \) has a unique maximum \( \phi_{\text{max}} = \phi(x_{\text{max}}) \) in the interval \([a, b]\) and that \( a < x_{\text{max}} < b \). Assume also that this maximum is quadratic, i.e., that \( \phi''(x_{\text{max}}) < 0 \). Using the second-order Taylor series for \( \phi(x) \) near \( x_{\text{max}} \) and writing \( \phi'' \) for \( \phi''(x_{\text{max}}) \) to save ink, we have

\[
\phi(x_{\text{max}} + y) = \phi(x_{\text{max}}) - \frac{1}{2} |\phi''| y^2 + O(y^3).
\]

Comparing with (A.25), we see that \( e^{n\phi} \) is proportional to a Gaussian with mean \( x_{\text{max}} \) and variance \( 1/(n |\phi''|) \). The integral is dominated by contributions from the interval where this Gaussian is large. Its width scales as \( 1/\sqrt{n |\phi''|} \sim 1/\sqrt{n} \), and its height is \( e^{n\phi_{\text{max}}} \).

Since \( f \) is smooth, as \( n \) increases and the width narrows, \( f = f(x_{\text{max}}) \) becomes constant on this interval, giving

\[
I \approx f(x_{\text{max}}) e^{n\phi_{\text{max}}} \int_{-\infty}^{\infty} e^{-(1/2)n|\phi''|y^2} \, dy = \sqrt{\frac{2\pi}{n |\phi''|}} f(x_{\text{max}}) e^{n\phi_{\text{max}}}.
\]

This approximation is called Laplace's method. Its multiplicative error is \( 1 + O(1/n) \). If there are multiple maxima where \( \phi \) is equally large, \( I \) receives a contribution from each one.

In Section 14.5 we use another convenient form of this approximation. Suppose \( \phi(x) = \ln g(x) \). Since \( g' = 0 \) at a maximum, we have

\[
\phi'' = \frac{g''}{g_{\text{max}}}
\]

So we can also write

\[
\int_{a}^{b} f(x) g(x)^n \, dx \approx \sqrt{\frac{2\pi}{n g''/g_{\text{max}}}} f(x_{\text{max}}) g_{\text{max}}^{-n}.
\]

Laplace's method generalizes easily to higher-dimensional integrals. Suppose \( x \) is a \( d \)-dimensional vector and \( \phi \) has a unique maximum \( \phi_{\text{max}} = \phi(x_{\text{max}}) \). If the Hessian, i.e., the matrix of second derivatives

\[
(\phi'')_{ij} = \frac{\partial^2 \phi}{\partial x_i \partial x_j},
\]

is nonsingular, then

\[
\int f(x) e^{n\phi(x)} \, dx \approx \sqrt{\frac{(2\pi)^d}{n^d |\det \phi''|}} f(x_{\text{max}}) e^{n\phi_{\text{max}}},
\]

and so

\[
\int f(x) g(x)^n \, dx \approx \sqrt{\frac{(2\pi)^d}{n^d |\det (g''/g_{\text{max}})|}} f(x_{\text{max}}) g_{\text{max}}^{-n}.
\]

We can also consider the case where \( \phi''(x_{\text{max}}) = 0 \) but some higher-order derivative is nonzero. We say that \( x_{\text{max}} \) is a \( p \)-th-order stationary point if the \( p \)th derivative \( \phi^{(p)}(x_{\text{max}}) \) is nonzero but all lower derivatives are zero. In that case, the integral is dominated by an interval of width \( n^{-1/p} \) around \( x_{\text{max}} \) instead of \( n^{-1/2} \), so

\[
I \sim \frac{1}{(n \phi^{(p)}(x_{\text{max}}))^{1/p}}.
\]

Since a maximum on the real line often becomes a saddle point in the complex plane, physicists call Laplace's method the saddle-point method.
A.6.2 The method of stationary phase

In our study of quantum walks in Chapter 15, we use a generalization of Laplace’s method which applies to complex-valued integrals where the integrand oscillates rapidly in the complex plane. Consider an integral of the form

\[ I = \int_a^b e^{i n \phi(x)} \, dx. \]

In the limit \( n \to \infty \), the contributions from \( x \) with \( \phi(x) \neq 0 \) are exponentially small, since the phase oscillations cause destructive interference and the integral cancels out. Since we get constructive interference from any interval where \( \phi(x) \) is roughly constant, \( I \) is dominated by values of \( x \) near stationary points \( x_0 \) where \( \phi'(x_0) = 0 \). The width of this interval is roughly \( 1/\sqrt{n |\phi''(x_0)|} \), just as in Laplace’s method.

Suppose there is a unique stationary point \( x_0 \) in the interval \([a, b]\), and write \( \phi_0 = \phi(x_0) \). Using the second-order Taylor series for \( \phi \) just as we did in the Laplace method gives

\[ I \approx e^{i n \phi_0} \int_{a}^{b} e^{i(1/2)n \phi''(x)} x^2 \, dx = \sqrt{\frac{2i \pi}{n \phi''}} e^{i n \phi_0}. \tag{A.36} \]

If there are multiple stationary points in the interval \([a, b]\) then \( I \) receives a contribution from each one. Since each one has a phase \( e^{i n \phi_0} \), these contributions can interfere constructively or destructively. Indeed, as the parameters of the integral vary, there are often rapid oscillations, or moiré patterns, in its overall value. You can see these patterns in Figures 15.12 and 15.13.

As for the real-valued Laplace approximation, if \( x_0 \) is a \( p \)-th order stationary point then \( I \sim n^{-1/p} \). This happens, for instance, at the extreme points of the one-dimensional quantum walk in Problem 15.50.

A.7 Groups, Rings, and Fields

The familiar operations of arithmetic, addition and multiplication, have certain properties in common. First, they are associative: for any \( a, b, c \in \mathbb{R} \), we have \((a + b) + c = a + (b + c)\) and \((ab)c = a(bc)\). Second, they have an identity, something which leaves other things unchanged: \( a + 0 = 0 + a = a \) and \( 1 \cdot a = a \cdot 1 = a \). Finally, each \( a \) has an inverse, which combined with \( a \) gives the identity: \( a + (-a) = 0 \) and \( aa^{-1} = 1 \).

A general structure of this kind is called a group. To be precise, a group \( G \) is a set with a binary operation, a function from \( G \times G \) to \( G \), which we write \( a \cdot b \). This binary operation obeys the following three axioms:

1. (Associativity) for all \( a, b, c \in G \), \((a \cdot b) \cdot c = a \cdot (b \cdot c)\).
2. (Identity) there exists an element \( 1 \in G \) such that, for all \( a \in G \), \( a \cdot 1 = 1 \cdot a = a \).
3. (Inverses) for each \( a \in G \), there is an \( a^{-1} \in G \) such that \( a \cdot a^{-1} = a^{-1} \cdot a = 1 \).
There is another common property which addition and multiplication possess, namely *commutativity*:

\[ a + b = b + a \quad \text{and} \quad ab = ba. \]

Groups of this kind are called *abelian*, after the mathematician Niels Henrik Abel. (Having something named after you in mathematics is a great honor, but the greatest honor is having it in lower case.) However, there are many interesting *nonabelian* groups as well, in which the order of multiplication matters.

Here are some common examples of groups. Which ones are abelian?

1. \( (\mathbb{Z}, +) \), the integers with addition
2. \( (\mathbb{Z}_n, +) \), the integers mod \( n \) with addition
3. \( (\mathbb{Z}_n^*, \times) \), the integers mod \( n \) which are mutually prime to \( n \), with multiplication
4. \( (\mathbb{C}, +) \), the complex numbers with addition
5. \( (\mathbb{C}, \times) \), the nonzero complex numbers with multiplication
6. \( U_d \), the unitary \( d \times d \) matrices: that is, those for which \( U^{-1} = U^\dagger \), where as described in Chapter 15, \( U^\dagger \) is the complex conjugate of \( U \)'s transpose
7. \( D_4 \), the dihedral group shown in Figure 15.6, consisting of the 8 rotations and reflections of a square
8. \( S_n \), the group of \( n! \) permutations of \( n \) objects, where we multiply two permutations by composing them
9. \( A_p \), the set of functions \( f(x) = ax + b \) where \( a \in \mathbb{Z}_p^* \) and \( b \in \mathbb{Z}_p \), where we multiply two functions by composing them

### A.7.1 Subgroups, Lagrange, and Fermat

A *subgroup* of a group \( G \) is a subset \( H \subseteq G \) which is closed under multiplication. That is, if \( a, b \in H \) then \( ab \in H \). For infinite groups, we also need to require that if \( a \in H \) then \( a^{-1} \in H \). For example, for any integer \( r \), the multiples of \( r \) form a subgroup \( r\mathbb{Z} = \{ \ldots, -2r, -r, 0, r, 2r, \ldots \} \) of the integers \( \mathbb{Z} \).

A *coset* of a subgroup is a copy of it which has been shifted by multiplying it by some element—in other words, a set of the form \( aH = \{ ah : h \in H \} \) for some \( a \in G \). For instance, if \( G = \mathbb{Z} \) and \( H = r\mathbb{Z} \), the coset \( aH \) is the set of integers equivalent to \( a \) mod \( r \),

\[ a + r\mathbb{Z} = \{ a + tr : t \in \mathbb{Z} \} = \{ b \in \mathbb{Z} : b \equiv_r a \}. \]

We call \( aH \) a *left coset*. In an abelian group, \( aH = Ha \), but in a nonabelian group the right cosets \( Ha \) will typically be different. For another example, in the group \( \mathbb{R}^2 \) of two-dimensional vectors with addition, each straight line passing through the origin is a subgroup and its cosets are the other lines with the same slope.

Now let’s prove something interesting. First, we recommend the following two exercises, which you should prove using nothing but the axioms of group theory:

*Exercise A.24* Show that all the cosets of a subgroup have the same size, i.e., that \( |aH| = |H| \) for all \( a \in G \).
Exercise A.25 Show that two cosets are either identical or disjoint, i.e., for any $a, b$ we have $aH = bH$ or $aH \cap bH = \emptyset$. Hint: consider whether the "difference" $a^{-1}b$ between $a$ and $b$ is in $H$.

These two results imply that there are $|G|/|H|$ cosets of $H$. In particular, they imply Lagrange's Theorem:

Theorem A.6 If $G$ is a finite group and $H \subseteq G$ is a subgroup, then $|H|$ divides $|G|$.

Given an element $a$, the cyclic subgroup generated by $a$ is $\langle a \rangle = \{1, a, a^2, \ldots \}$, the set of all powers of $a$. Define the order of $a$ as the smallest $r$ such that $a^r = 1$. The following exercise shows that $r$ exists if $G$ is finite:

Exercise A.26 Let $G$ be a finite group and let $a \in G$. Show that there is an $r$ such that $a^r = 1$.

Thus in a finite group, $\langle a \rangle = \{1, a, a^2, \ldots, a^{r-1}\}$. Since $|\langle a \rangle| = r$, we get the following corollary to Theorem A.6:

Corollary A.7 If $G$ is a finite group, the order of any element $a$ divides $|G|$.

If $a$ has order $r$ then $a^{kr} = 1^k = 1$ for any $k$, so this implies that $a^{|G|} = 1$. In particular, Euler's totient function $\varphi(n) = |\mathbb{Z}_n^*|$ is the number of integers mod $n$ which are mutually prime to $n$. Then for any $a$ we have $a^{\varphi(n)} \equiv 1 \pmod{n}$, giving us the generalization of Fermat's Little Theorem which we use in Section 15.5.1. If $n$ is prime, we have $\varphi(n) = n - 1$, giving the original version of Fermat's Little Theorem we use in Section 4.4.1.

If $G$ is cyclic, an element $a \in G$ such that $G = \langle a \rangle$ is called a generator. For historical reasons, a generator of $\mathbb{Z}_n^*$ is also called a primitive root mod $n$.

A.7.2 Homomorphisms and isomorphisms

A homomorphism is a map from one group to another that preserves the structure of multiplication. In other words, $\phi : G \to H$ is a homomorphism if and only if $\phi(ab) = \phi(a) \cdot \phi(b)$ for all $a, b \in G$. For instance, the function $\phi(x) = e^x$ is a homomorphism from $(\mathbb{C}, \times)$ to $(\mathbb{C}, \times)$ since $e^{x+y} = e^x \times e^y$.

For another example, recall that the parity of a permutation is even or odd depending on how many times we need to swap a pair of objects to perform it. For instance, a permutation which rotates three objects $1 \to 2 \to 3 \to 1$ has even parity, since we can perform it by first swapping 2 and 3 and then swapping 1 and 2. Now define $\pi(\sigma)$ as +1 if $\sigma$ is even, and −1 if $\sigma$ is odd. Then $\pi$ is a homomorphism from $S_n$ to $(\mathbb{C}, \times)$.

Another important example, mentioned in Section 15.6, is the Fourier basis function $\phi_k(x) = e^{2\pi i k x}$ where $\omega = e^{2\pi i/n}$ is the $n$th root of 1 in the complex plane. For any integer $k$, this is a homomorphism from $\mathbb{Z}_n$ to $(\mathbb{C}, \times)$.

A subgroup $H$ is normal if for all $a \in G$ and all $h \in H$, we have $a^{-1}ha \in H$, or more compactly, if $a^{-1}Ha = H$ for all $a$. Although we don't use it in the book, the following is a nice exercise:

Exercise A.27 The kernel of a homomorphism $\phi$ is the set $K = \{a \in G : \phi(a) = 1\}$. Show that $K$ is a normal subgroup.
An isomorphism is a homomorphism which is one-to-one. As an example, consider the cyclic group $G = \{1, a, a^2, \ldots, a^{n-1}\}$ generated by an element of order $n$. This group is isomorphic to $\mathbb{Z}_n$, since the map $a^x$ is an isomorphism from $\mathbb{Z}_n$ to $G$. If two groups $G, H$ are isomorphic, we write $G \cong H$.

### A.7.3 The Chinese Remainder Theorem

Suppose we have an unknown number of objects. When counted in threes, two are left over; when counted in fives, three are left over; and when counted in sevens, two are left over. How many objects are there?

Sun Zi, ca. 400 A.D.

In this section, we discuss one of the oldest theorems in mathematics. In honor of its origins in China, it is known as the Chinese Remainder Theorem. We will state it here using modern language.

Given two groups $G_1, G_2$, their Cartesian product $G_1 \times G_2$ is a group whose elements are ordered pairs $(a_1, a_2)$ with $a_1 \in G_1$ and $a_2 \in G_2$. We define multiplication componentwise, $(a_1, a_2) \cdot (b_1, b_2) = (a_1 b_1, a_2 b_2)$.

Consider the following theorem:

**Theorem A.8** Suppose $p$ and $q$ are mutually prime. Then

$$\mathbb{Z}_{pq} \cong \mathbb{Z}_p \times \mathbb{Z}_q.$$ 

By applying Theorem A.8 inductively, we get the full version:

**Theorem A.9 (Chinese Remainder Theorem)** Let $q_1, q_2, \ldots, q_\ell$ be mutually prime and let $n = \prod_{i=1}^\ell q_i$. Then

$$\mathbb{Z}_n \cong \mathbb{Z}_{q_1} \times \mathbb{Z}_{q_2} \times \cdots \times \mathbb{Z}_{q_\ell}.$$ 

In particular, let $n = p_1^{t_1} p_2^{t_2} \cdots p_\ell^{t_\ell}$ be the prime factorization of $n$. Then

$$\mathbb{Z}_n \cong \mathbb{Z}_{p_1^{t_1}} \times \mathbb{Z}_{p_2^{t_2}} \times \cdots \times \mathbb{Z}_{p_\ell^{t_\ell}}.$$ 

Here is a more traditional restatement of Theorem A.9. It tells us that Sun Zi's riddle has a solution $x$, where $0 \leq x < 3 \times 5 \times 7$:

**Theorem A.10 (Chinese Remainder Theorem, traditional version)** Suppose that $q_1, q_2, \ldots, q_\ell$ are mutually prime, let $n = \prod_{i=1}^\ell q_i$, and let $r_1, r_2, \ldots, r_\ell$ be integers such that $0 \leq r_i < q_i$ for each $i$. Then there exists a unique integer $x$ with $0 \leq x < n$ such that $x \equiv r_i \mod q_i$ for each $i$.

It follows that for any integer $x$, the value of $x \mod n$ is determined uniquely by $x \mod p_i$ for each of the prime power factors $p_i^t$ of $n$. For instance, the following table shows how each element $x \in \mathbb{Z}_{12}$ is determined by $x \mod 3$ and $x \mod 4$:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>1</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>5</td>
<td>2</td>
<td>11</td>
</tr>
</tbody>
</table>
Notice how we move diagonally through the table as we increment \( x \), winding around when we hit the bottom or right edge. Since 3 and 4 are mutually prime, these windings cover the entire table before we return to the origin \( x = 0 \).

To prove Theorem A.8, we start by noting that the function \( f : \mathbb{Z}_{pq} \rightarrow \mathbb{Z}_p \times \mathbb{Z}_q \) defined by
\[
f(x) = (x \mod p, x \mod q),
\]
is a homomorphism. To prove that it is an isomorphism, we need to show that there is an \( x \) such that \( f(x) = (1, 0) \), and another such that \( f(x) = (0, 1) \)—in this example, 4 and 9—since these two elements generate \( \mathbb{Z}_p \times \mathbb{Z}_q \). In other words, we need to complete the following exercise, which is the two-factor case of Problem 5.26:

**Exercise A.28** Suppose that \( n = pq \) where \( p \) and \( q \) are mutually prime. Show that there is an \( x \) such that \( x \equiv_p 1 \) and \( x \equiv_q 0 \), and another \( x \) such that \( x \equiv_p 0 \) and \( x \equiv_q 1 \). Hint: recall that \( p \) and \( q \) are mutually prime if and only if there are integers \( a, b \) such that
\[
ap + bq = 1.
\]

Once Theorem A.8 is proved, Theorem A.9 follows by induction on the number of factors \( \ell \).

### A.7.4 Rings and Fields

Addition and multiplication have another important property, the *distributive law*: \( a(b + c) = ab + ac \). A ring \( R \) is a set with two binary operations, + and \( \cdot \), such that

1. both + and \( \cdot \) are associative,
2. \( (R, +) \) is an Abelian group, whose identity we denote 0, and
3. the distributive law \( a \cdot (b + c) = (a \cdot b) + (a \cdot c) \) holds.

Examples of rings include \( \mathbb{R} \) or \( \mathbb{C} \) with the usual addition and multiplication, \( \mathbb{Z}_n \) with addition and multiplication mod \( n \), and the set of polynomials over a variable \( x \) with integer or rational coefficients.

The following exercise gives us the familiar fact that, if adding 0 leaves things alone, then multiplying by 0 annihilates them:

**Exercise A.29** Show that, in any ring, we have \( 0 \cdot a = a \cdot 0 = 0 \) for all \( a \).

Therefore, if \( R \) has more than one element, 0 cannot have a multiplicative inverse.

A **field** is a ring with the additional property that all elements other than 0 have inverses. Equivalently, the nonzero elements form a group under multiplication. Infinite examples include \( \mathbb{R} \), \( \mathbb{C} \), the rationals \( \mathbb{Q} \) with the usual operations, and the set of rational functions—that is, functions of the form \( f(x) = g(x)/h(x) \) where \( g \) and \( h \) are polynomials.

The only type of finite field that we meet in this book is \( \mathbb{F}_p \), the set of integers mod \( p \) where \( p \) is a prime, with addition and multiplication mod \( p \). It is a field since every \( a \in \mathbb{F}_p^* = \{1, 2, \ldots, p - 1\} \) is mutually prime to \( p \), and so has a multiplicative inverse \( b \) such that \( ab \equiv_p 1 \).
There is also a finite field \( \mathbb{F}_q \) of size \( q \) for each prime power \( q = p^t \). Addition in \( \mathbb{F}_q \) is isomorphic to \( \mathbb{Z}_p^t \), or to addition of \( t \)-dimensional vectors mod \( p \). But if \( t > 1 \), defining multiplication is a little more complicated—it is not simply multiplication mod \( q \). Instead, it is defined in terms of polynomials of degree \( t \) with coefficients in \( \mathbb{F}_p \). For instance, we can define the finite field \( \mathbb{F}_2 = \mathbb{F}_2^3 \) as the set of polynomials \( f(x) = a_2x^2 + a_1x + a_0 \) where \( a_0, a_1, a_2 \in \mathbb{F}_2 \). When we add two polynomials their coefficients add, so \( (\mathbb{F}_2,+) \cong \mathbb{Z}_2^3 \). But when we multiply them, we take the result modulo a polynomial \( Q \), such as \( Q = x^3 + x + 1 \). Then \( x^{-1} = x^2 + 1 \), since

\[
x(x^2 + 1) = x^3 + x \equiv 1.
\]

**Exercise A.30** Compute the powers of \( x \) in this presentation of \( \mathbb{F}_8 \), and show that \( x^7 = 1 \).

Since \( Q \) is irreducible mod 2—that is, it cannot be written as a product of polynomials of lower degree with coefficients in \( \mathbb{F}_2 \)—the resulting ring is a field, just as \( \mathbb{Z}_2 \) is a field if and only if \( p \) is prime. Surprisingly, all finite fields of size \( q \) are isomorphic, where of course we demand that an isomorphism between fields preserves both the additive and the multiplicative structure.

### A.7.5 Polynomials and roots

For any ring \( R \) we can define \( R[x] \), the ring of polynomials over a variable \( x \) with coefficients in \( R \). Any polynomial \( f(x) \in R[X] \) then defines a function from \( R \) to \( R \) in the obvious way. We say \( f(x) \) has degree \( d \) if

\[
f(x) = a_dx^d + a_{d-1}x^{d-1} + \cdots + a_1x + a_0. \tag{A.37}
\]

for some coefficients \( a_0, \ldots, a_d \in R \).

Now let’s prove the following classic fact:

**Theorem A.11** If \( R \) is a field, a polynomial \( f \in R[x] \) of degree \( d \) can have at most \( d \) roots.

**Proof** First note that for any \( r \in R \), there is set of coefficients \( b_0, \ldots, b_d \) such that

\[
f(x) = b_d(x-r)^d + b_{d-1}(x-r)^{d-1} + \cdots + b_1(x-r) + b_0.
\]

To see this, start with (A.37) and set \( b_d \) equal to the leading coefficient \( a_d \). Then \( f(x) - b_d(x-r)^d \) is a polynomial of degree \( d-1 \). We set \( b_{d-1} \) equal to its leading coefficient, and so on. We continue decreasing the degree until we are left with a constant function \( b_0 \).

Now suppose that \( r \) is a root of \( f(x) \), i.e., that \( f(r) = 0 \). In that case \( b_0 = 0 \). If \( R \) is a field, we can then divide \( f(x) \) by \( x-r \), factoring it as

\[
f(x) = (x-r)\left(b_dx^d + b_{d-1}(x-r)^d + \cdots + b_1\right) = (x-r)f'(x).
\]

Since \( f'(x) \) is a polynomial of degree \( d-1 \), it follows by induction that \( f(x) \) can have at most \( d \) roots.

\( \square \)

Note that if \( R \) is a ring but not a field, Theorem A.11 does not necessarily hold. For instance, in \( \mathbb{Z}_6 \) the polynomial \( f(x) = 2x \) has two roots, \( x = 0 \) and \( x = 3 \).
Problems

A mathematical problem should be difficult in order to entice us, yet not completely inaccessible, lest it mock at our efforts. It should be to us a guidepost on the mazy paths to hidden truths, and ultimately a reminder of our pleasure in the successful solution.

David Hilbert

A.1 Inclusion and exclusion. Here is a combinatorial proof of the inclusion–exclusion principle (A.13). First, let’s separate it into one term for each of the $2^n$ possible cases of which $E_i$ are true and which are false—that is, one term for each piece of the Venn diagram. Specifically, for a subset $V \subseteq \{1, \ldots, n\}$, let

$$P_V = \Pr \left[ \left( \bigwedge_{i \in V} E_i \right) \wedge \left( \bigwedge_{i \notin V} \overline{E_i} \right) \right].$$

Then show that (A.13) can be written

$$\Pr \left[ \bigvee_{i=1}^n E_i \right] = \sum_{V \subseteq \{1, \ldots, n\}} P_V \sum_{T \subseteq V} (-1)^{|T|} \binom{|V|}{|T|}. \quad \text{(A.38)}$$

Finally, show that

$$\sum_{j=0}^m (-1)^j \binom{m}{j} = \begin{cases} 1 & \text{if } m = 0 \\ 0 & \text{if } m > 0 \end{cases},$$

so that we can rewrite (A.38) as

$$\Pr \left[ \bigvee_{i=1}^n E_i \right] = P_{\emptyset}.$$

A.2 A night at the opera. A classic puzzle describes $n$ opera-goers, whose manteaux become randomly permuted by an incompetent coat checker, and asks for the probability that not a single one of them receives their own manteau. Equivalently, if we choose randomly from among the $n!$ permutations, what is the probability that we obtain a derangement, i.e., a permutation with no fixed points? Use the inclusion–exclusion principle (A.13) to show that this probability is exactly

$$P = 1 - 1 + \frac{1}{2} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!} = \sum_{j=0}^n \frac{(-1)^j}{j!},$$

which converges quickly to $1/e$. Hint: given a set of $i$ opera-goers, what is the probability that those $i$ (and possibly some others) all get back their manteaux?

A.3 Fixed points are Poisson. Using the previous problem, show that for any constant $x$ the probability that there are exactly $x$ fixed points in a random permutation of $n$ objects approaches $P(x) = 1/(ex!)$, the Poisson distribution with mean $1$, as $n$ tends to infinity. Therefore, the number of people who get their manteaux back is distributed almost as if each of these events occurred independently with probability $1/n$. 
A.4 **The first cycle.** Given a permutation $\sigma$, a $k$-cycle is a set of $k$ objects each of which is mapped to the next by $\sigma$: that is, a set $t_1, t_2, \ldots, t_k$ such that $\sigma(t_i) = (i + 1) \mod k$. Given a random permutation of $n$ objects, show that the length of the cycle containing the first object is uniformly distributed from 1 to $n$.

A.5 **Harmonic cycles.** Show that the expected number of cycles in a random permutation of $n$ objects is exactly the $n$th harmonic number,

$$ H_n = \sum_{k=1}^{n} \frac{1}{k} \approx \ln n. $$

For instance, of the 6 permutations of 3 objects, two of them consist of a single cycle of all 3, three of them consist of a 2-cycle and a 1-cycle (i.e., a fixed point), and the identity consists of three 1-cycles. Thus the average is $(2/6) \cdot 1 + (3/6) \cdot 2 + (1/6) \cdot 3 = 11/6 = H_3$. Hint: use linearity of expectation. What is the expected number of $k$-cycles for each $k$?

A.6 **Different birthdays.** If there are $n$ people in a room and there are $y$ days in the year, show that the probability that everyone has a different birthday is at most $e^{-n(n-1)/2y}$. Combined with the discussion in Sections A.3.1 and A.3.2, we then have

$$ 1 - \frac{n(n-1)}{2y} \leq \Pr[\text{everyone has a different birthday}] \leq e^{-n(n-1)/2y}. $$

Hint: imagine assigning a birthday to one person at a time. Write the probability that each one avoids all the previous birthdays as a product, and use the inequality $1 - x \leq e^{-x}$. If $y = 365$, for what value of $n$ is this upper bound roughly $1/2$?

A.7 **A birthday triple.** If there are $n$ people in a room and $y$ days in the year, how large can $n$ be before there is a good chance that some group of three people all have the same birthday? Answer within $\Theta$.

A.8 **Rules of engagement.** There are $n$ suitors waiting outside your door. Each one, when you let them in, will make a proposal of marriage. Your goal is to accept the best one. You adopt the following strategy: you invite them in in random order. After interviewing and rejecting the first $r$ suitors for some $r \leq n$, you accept the first one which is better than any of those $r$.

Show that the probability you nab the best one is exactly

$$ P = \frac{1}{n} \sum_{i=r+1}^{n} \frac{r}{i - 1}. $$

Hint: if the best suitor is the $i$th one you invite in, you will accept them if the best among the first $i - 1$ was among the first $r$. Then replace this sum with an integral to get

$$ P \approx \int_{r/n}^{1} \frac{a}{x} \, dx = a \ln(1/a). $$

where $a = r/n$. By maximizing $P$ as a function of $a$, conclude that the best strategy is to reject the first $n/e$ suitors and that this strategy succeeds with probability $1/e$.

A.9 **A lark ascending.** Consider the definition of an increasing subsequence of a permutation given in Problem 3.21. Now consider a random permutation of $n$ objects. Show that there is some constant $A$ such that, with high probability, the longest increasing subsequence is of size $k < A \sqrt{n}$.

Hint: one way to generate a random permutation is to choose $n$ real numbers independently from the unit interval $[0, 1]$ and rank them in order. Given a set of $k$ such numbers, what is the probability that they are in increasing order? You might find the bound (A.10) on the binomial useful.

It turns out that the likely length of the longest increasing subsequence in a random permutation converges to $2\sqrt{n}$. This is known as Ulam's problem, and its exact solution involves some very beautiful mathematics [33].
A.10 Pairwise independence. Suppose that $X$ is a sum of indicator random variables $X_i$ which are pairwise independent: that is, $\mathbb{E}[X_i X_j] = \mathbb{E}[X_i] \mathbb{E}[X_j]$ for all $i \neq j$. Show that in this case the second moment method gives

$$\Pr[X > 0] \geq \frac{\mathbb{E}[X]}{1 + \mathbb{E}[X]} \geq 1 - \frac{1}{\mathbb{E}[X]},$$

so that $\Pr[X > 0]$ is close to 1 whenever $\mathbb{E}[X]$ is large.

Show that this is the case for the Birthday Problem, where $i$ and $j$ represent different pairs of people which may or may not overlap. In other words, show that the positive correlations described in the text don't appear until we consider sets of three or more pairs. Conclude that the probability that some pair of people has the same birthday is at least 1/2 when $n = \sqrt{2y} + 1$.

A.11 The case of the missing coupon. Repeat the previous problem in the case where the $X_i$ are negatively correlated, so that $\mathbb{E}[X_i X_j] \leq \mathbb{E}[X_i] \mathbb{E}[X_j]$. Use this to show that in the Coupon Collector's problem, with high probability we are still missing at least one toy when $b = (1 - \epsilon)n / n$ for any $\epsilon > 0$. Combined with the discussion in Section A.3.4, this shows that the probability we have a complete collection makes a sharp phase transition from 0 to 1 when $b$ passes $n / \log n$.

A.12 The black pearls. Suppose I have a necklace with $n$ beads on it, where each bead is randomly chosen to be black or white with probability 1/2. I would like to know how likely it is that somewhere on the necklace there is a string of $k$ consecutive black beads. Let $X_i$ be the indicator random variable for the event that there is a string of $k$ black beads starting at bead $i$. Then the number of strings is $X = \sum X_i$, where we count overlapping strings separately. For instance, a string of 5 black beads contains 3 strings of length 3.

By calculating $\mathbb{E}[X_i]$, show that with high probability there are no strings of length $(1 + \epsilon) \log n$ for any $\epsilon > 0$. Then calculate the second moment $\mathbb{E}[X^2]$ by considering the correlations between overlapping strings. Show that strings of length $k = \log \log n$ exist with probability at least 1/4, and strings of length $k = (1 - \epsilon) \log \log n$ exist with high probability for any $\epsilon > 0$.

A.13 The replacements. If we have a barrel of $n$ objects, choosing $k$ of them without replacement means to take an object out of the barrel, then another, and so on until we have chosen $k$ objects. Thus we end up with one of the $n \choose k$ possible subsets of $k$ objects. Choosing with replacement, in contrast, means tossing each object back in the barrel before we choose the next one. Show that if $k = o(\sqrt{n})$,

$$\binom{n}{k} = (1 - o(1)) \frac{n^k}{k!},$$

and therefore there is little or no distinction between choosing with or without replacement. Hint: use the Birthday Problem.

A.14 Heavy bins. Suppose that I toss $m = cn$ balls randomly into $n$ bins, where $c$ is a constant. Prove that with high probability the largest number of balls in any bin is $o(\log n)$.

A.15 Convoluted sums. If $x$ and $y$ are independent random variables whose probability distributions are $P(x)$ and $Q(y)$ respectively, the probability distribution of their sum $z = x + y$ is the convolution of $P$ and $Q$:

$$R(z) = \sum_{z-x} P(x) Q(z-x).$$

We write $R = P \ast Q$ as in Problem 3.14. Show that if $P$ and $Q$ are Poisson distributions with means $c$ and $d$ respectively, then $P \ast Q$ is a Poisson distribution with mean $c + d$. Show this analytically by actually doing the sum. Then come up with a "physical" explanation that makes it obvious.
A.16 Poisson and Gauss. Consider the Poisson distribution \( P(x) \) with mean \( c \) where \( c \) is large. Show that it can be approximated by a Gaussian for \( x \approx c \). What is its variance?

A.17 Chernoff for biased coins. Show that Theorems A.1 and A.2 hold in the more general case where \( x \) is the sum of any number of independent random variables \( x_i \), each of which is 1 or 0 with probability \( p_i \) and \( 1 - p_i \) respectively. Hint: show that the moment generating function of \( x \) is

\[
E[e^{tx}] = \prod_i (1 + p_i(e^t - 1)) \leq e^{t(e^t - 1)E[x]}.
\]

A.18 Chernoff’s tail. Consider the Chernoff bound of Theorem A.1. When \( \delta \) is large, the probability that \( x \) is \( 1 + \delta \) times its expectation is not exponentially small in \( \delta^2 \). However, it falls off faster than the simple exponential of Theorem A.1. Show that for \( \delta > 1 \),

\[
Pr[x > (1 + \delta)E[x]] \leq e^{-(\ln \delta)E[x]/2},
\]

and that asymptotically the constant 1/2 can be replaced with 1 when \( \delta \) is large.

A.19 Azuma with big steps and little ones. Generalize Theorems A.3 and A.4 as follows. If each \( y_i \) is bounded by \( |y_i| \leq c_i \) for some \( c_i \), then for any \( a \geq 0 \),

\[
Pr[|x| > a] \leq 2e^{-a^2/(2\sum c_i^2)}.
\]

A.20 Concentrated cycles. Once again, consider a random permutation of \( n \) objects. Now that we know the expected number of cycles, let’s prove that the number of cycles is \( O(\log n) \) with high probability. Specifically, show that for every constant \( a \), there is a \( b \) such that the probability there are more than \( b \log n \) cycles is less than \( n^{-a} \).

Hint: imagine constructing the cycles one by one, starting with the cycle containing the first object. Use Problem A.4 to argue that with constant probability, each cycle “eats” at least 1/2 of the remaining objects. Then use Chernoff bounds to argue that after \( b \log n \) cycles, with high probability there are no objects left.

A.21 Visits to the origin. Suppose I take a random walk on the infinite line where I start at the origin and move left or right with equal probability at each step. Let \( x_n \) be my position after \( n \) steps. First, show that \( E[|x_n|] \) is the expected number of times I visited the origin in the previous \( n - 1 \) steps. Then, show that \( E[|x_n|] \leq \sqrt{\text{Var } x_n} = \sqrt{n} \). Conclude that the expected number of times we visit the origin in \( n \) steps is \( O(\sqrt{n}) \). What does this say about the average number of steps we take in between visits to the origin?

A.22 Sticky ends. Suppose I perform a random walk on a finite line ranging from 0 to \( n \). At each step I move left or right with equal probability, but if I reach either end I stick there forever. Show that my initial position is \( x \), the probability that I will end up stuck at \( n \) instead of at 0 is exactly \( x/n \).

Prove this in two ways. First, write a recurrence similar to (A.26) for this probability \( p(x) \) as a function of \( x \), and solve it with the boundary conditions \( p(0) = 0 \) and \( p(n) = 1 \). Second, use the fact that while my position changes, its expectation does not. In other words, it is a martingale as discussed in Section A.5.2.

A.23 Distance squared minus time. Let \( x_t \) be the position of a random walk where \( x_t - x_{t-1} = +1 \) or \(-1 \) with equal probability. Show that the sequence

\[
y_t = x_t^2 - t
\]

is a martingale.
A.24 Scaling away the bias. Let $x_t$ be the position of a biased random walk where $x_t - x_{t-1} = +1$ or $-1$ with probability $p$ or $1-p$ respectively. We saw in Section A.5.2 one way to turn $x_t$ into a martingale. Here is another: show that the sequence

$$y_t = \left( \frac{1-p}{p} \right)^{x_t}$$

is a martingale.

A.25 Pólya's urn. There are some balls in an urn. Some are red, and some are blue. At each step, I pick a ball uniformly at random from the urn. I then place it back in the urn, along with a new ball the same color as the one I picked. Show that the fraction red balls in the urn is a martingale.

A.26 Resisting infinity. Roughly speaking, a random walk on a $d$-dimensional lattice consists of $d$ independent one-dimensional random walks. Argue that the probability we return to the origin after $n$ steps is $\Theta(n^{-d/2})$. Then, justify the following claim: if $d = 1$ or $d = 2$, a random walk returns to the origin an infinite number of times with probability 1, but if $d \geq 3$, with probability 1 it escapes to infinity after only a finite number of visits to the origin. Doyle and Snell [240] have written a beautiful monograph on the following analogy: if I have a $d$-dimensional lattice with a 1 Ohm resistor in each edge, the total resistance between the origin and infinity is infinite if $d < 3$ and finite if $d \geq 3$. If you know how to calculate resistances in serial and parallel, show this in a model where we replace the lattice with a series of concentric spheres.

A.27 Fibonacci, Pascal's Triangle, and Laplace. Show that the $n$th Fibonacci number can be written

$$F_n = \sum_{j=0}^{\lceil n/2 \rceil} \binom{n-j}{j}.$$  

Then use Stirling's approximation and the Laplace method to recover the result

$$F_n = \Theta(\varphi^n)$$

where $\varphi = (1 + \sqrt{5})/2$ is the golden ratio. This is probably the most roundabout way to prove this, but it's a good exercise.

A.28 Laplace and Stirling. One nice application of Laplace's method is to derive the precise version of Stirling's approximation for the factorial (A.9). We can do this using the Gamma function,

$$n! = \Gamma(n+1) = \int_0^\infty x^n e^{-x} \, dx$$

and noticing that the integrand $x^n e^{-x}$ is tightly peaked for large $n$. Write the integrand in the form

$$x^n e^{-x} = e^{n \phi(x)},$$

Taking derivatives, show that $\phi(x)$ is maximized at $x_{\text{max}} = n$, where $e^{\phi} = n/e$ and $\phi'' = -n^2$. In this case $\phi$ depends weakly on $n$, but the method of approximating the integrand with a Gaussian is otherwise the same. Then apply (A.32) and show that

$$n! = (1 + O(1/n)) \sqrt{\frac{2\pi}{n \phi''(x_{\text{max}})}} e^{n \phi(x_{\text{max}})} = (1 + O(1/n)) \sqrt{2\pi n} n^e e^{-n}.$$