1

A Quick Refresher on Analysis and Probability

Most of the material we recall in this chapter is taught in basic analysis and probability courses. You may wish to skip this chapter altogether, or skim through it, if you are well prepared. Either way, try the exercises at the end of the chapter, especially the more difficult ones.

1.1 Convex Sets and Functions

A subset $K \subset \mathbb{R}^n$ is a *convex set* if, for any pair of points in K, the line segment connecting these two points is also contained in K:

$$\lambda x + (1-\lambda)y \in K \quad \forall x, y \in K,\ \lambda \in [0,1].$$

Let $K \subset \mathbb{R}^n$ be a convex subset. A function $f\colon K \to \mathbb{R}$ is a *convex function* if

$$f\left(\lambda x + (1-\lambda)y\right) \leq \lambda f(x) + (1-\lambda) f(y) \quad \forall x, y \in K,\ \lambda \in [0,1]. \tag{1.1}$$

In other words, f is convex if, when restricted to a line segment connecting any pair of points in K, the graph of f lies below the line segment connecting these two points. Figure 1.1 illustrates the definitions of convex sets and functions.

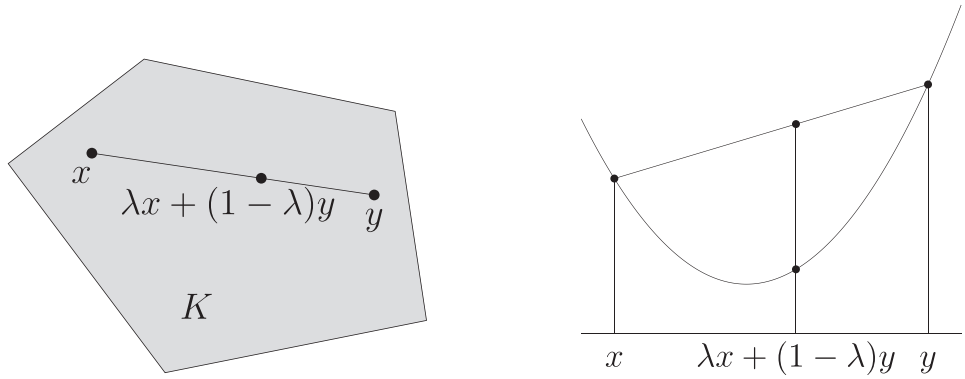


Figure 1.1 Definitions of a convex set (left) and a convex function (right).

Concave functions are defined similarly, except that the inequality in (1.1) is reversed. Equivalently, f is concave if and only if $-f$ is convex.

https://doi.org/10.1017/9781009490672.004 Published online by Cambridge University Press

The *maximum principle* states that a convex function defined on the convex set $K = \operatorname{conv}(x_1, \dots, x_n)$ attains its maximum at some point x_i (try Exercise 1.4).

1.2 Norms and Inner Products

You should already be familiar with the definitions of *metrics*, *norms*, and *inner products*. To see if you remember them well, take a minute to prove that (a) a norm is a convex function and (b) the unit ball of a normed space is a convex set.

The most popular example of a norm and an inner product on $\mathbb{R}^n$ is the Euclidean norm and the dot product, defined as

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} \quad \text{and} \quad \langle x, y \rangle = x^{\mathsf{T}} y = \sum_{i=1}^{n} x_i y_i. \tag{1.2}$$

The Euclidean norm agrees with the dot product in the sense that $\|x\|_2^2 = \langle x, x \rangle$.

More generally, for any exponent $p \in [1, \infty]$, we can define the ℓ^p *norm* on $\mathbb{R}^n$ by

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \text{ for } p \in [1, \infty), \quad \text{and } \|x\|_\infty = \max_{i \le n} |x_i|.$$

The *Minkowski inequality* states that the triangle inequality holds for the ℓ^p norm: for any vectors $x, y \in \mathbb{R}^n$ we have

$$\|x + y\|_p \le \|x\|_p + \|y\|_p.$$

It follows that the ℓ^p norm indeed defines a norm on $\mathbb{R}^n$ for every $p \in [1, \infty]$. (Check!)

The best way to visualize the geometry of an ℓ^p norm is to look at the unit ball of the space $(\mathbb{R}^n, \|\cdot\|_p)$, that is,

$$B_p^n = \{ x \in \mathbb{R}^n \colon \ \|x\|_p \le 1 \}.$$

For example, B_2^n is the unit Euclidean ball – a *round ball* with unit radius, B_∞^n is a *cube*, and B_1^n a *cross-polytope* – a high-dimensional version of an octahedron:

$$B_\infty^n = [-1, 1]^n, \quad B_1^n = \operatorname{conv}\left(\{ \pm e_1, \dots, \pm e_n \} \right), \tag{1.3}$$

where $e_1, \dots, e_n$ denotes the standard basis of $\mathbb{R}^n$ (Exercise 1.6). Figure 1.2 illustrates the unit ℓ^p balls for different exponents p.

The ℓ^p norms of a given vector x are decreasing in p:

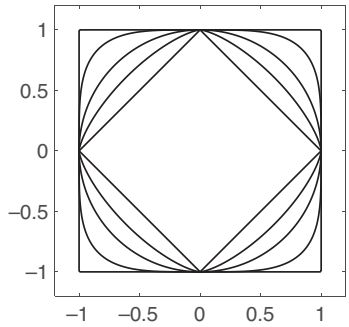
$$\|x\|_q \le \|x\|_p \quad \text{whenever } p \le q, \tag{1.4}$$

see Exercise 1.17. Equivalently, the unit ℓ^p balls are increasing in p, that is, $B_p^n \subset B_q^n$ whenever $p \le q$. (Why?)

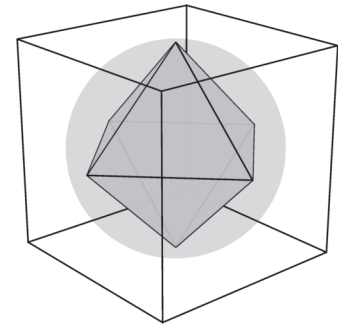
The *Cauchy–Schwarz inequality* states that for all vectors $x, y \in \mathbb{R}^n$ we have

$$|\langle x, y \rangle| \le \|x\|_2 \|y\|_2.$$

https://doi.org/10.1017/9781009490672.004 Published online by Cambridge University Press



(a) ℓ^p balls for $p = 1, 1.4, 2, 4, \infty$



(b) ℓ^p balls for $p = 1, 2, \infty$

Figure 1.2 Some unit ℓ^p balls in dimensions $n = 2$ (left) and $n = 3$ (right).

The *Hölder inequality* generalizes this result to the ℓ^p norms:

$$|\langle x, y\rangle| \le \|x\|_p \|y\|_{p'} \quad \text{if } \frac{1}{p} + \frac{1}{p'} = 1. \tag{1.5}$$

A pair of numbers $p, p' \in [1, \infty]$ satisfying the condition in (1.5) is called *conjugate exponents*.[1]

The Hölder inequality is tight. In Exercise 1.19 you will check that for any vector x one can find a vector $y \neq 0$ for which the Hölder inequality becomes an equality. In other words, the following *duality formula* holds for the ℓ^p norms:

$$\max\left\{\langle x, y\rangle \colon \ y \in B_{p'}^n\right\} = \|x\|_p. \tag{1.6}$$

1.3 Random Variables and Random Vectors

In a basic course in probability theory, we learned about the two most important quantities associated with a random variable X, namely, the *expectation*[2] (also called *mean*) and *variance*. They are denoted by[3]

$$\mathbb{E}\, X \qquad \text{and} \qquad \operatorname{Var}(X) = \mathbb{E}(X - \mathbb{E}\, X)^2. \tag{1.7}$$

The *linearity of expectation* guarantees that for any random variables $X_1, \dots, X_n$, whether independent or not, and for any fixed numbers $a_1, \dots, a_n$, we have

$$\mathbb{E}\left[a_1 X_1 + \cdots + a_n X_n\right] = a_1\, \mathbb{E}\, X_1 + \cdots + a_n\, \mathbb{E}\, X_n.$$

[1] We agree that $\frac{1}{0} = \infty$ and $\frac{1}{\infty} = 0$ to make $(1, \infty)$ and $(\infty, 1)$ pairs of conjugate exponents.

[2] If you have studied measure theory, you know that the expectation $\mathbb{E}\, X$ of a random variable X on a probability space $(\Omega, \Sigma, \mathbb{P})$ is defined as the Lebesgue integral of the function $X \colon \Omega \to \mathbb{R}$. This allows all Lebesgue integration theorems to apply for expectations of random variables.

[3] Throughout this book, we drop the brackets in the notation $\mathbb{E}[f(X)]$ and simply write $\mathbb{E}\, f(X)$ instead. Thus, nonlinear functions bind before expectation.

https://doi.org/10.1017/9781009490672.004 Published online by Cambridge University Press

Variance generally does not satisfy a similar property. However, if the random variables X_i are independent (or even uncorrelated), then

$$\operatorname{Var}(a_1X_1 + \cdots + a_nX_n) = a_1^2 \operatorname{Var}(X_1) + \cdots + a_n^2 \operatorname{Var}(X_n). \tag{1.8}$$

The simplest example of a random variable is the *indicator* of a given event E, which is denoted by $\mathbf{1}_E$. The random variable $\mathbf{1}_E$ takes value 1 if event E occurs and 0 if E does not occur. The expectation of an indicator obviously satisfies

$$\mathbb{E}\,\mathbf{1}_E = \mathbb{P}(E). \tag{1.9}$$

The *moment generating function of* X is defined as

$$M_X(t) = \mathbb{E}\,e^{tX}, \quad t \in \mathbb{R}.$$

For $p > 0$, the *pth moment* of X is defined as $\mathbb{E}\,X^p$, and the *pth absolute moment* is $\mathbb{E}|X|^p$. If we take the pth root of the absolute moment, we arrive at the notion of the L^p *norm* of a random variable:

$$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p} \text{ for } p \in (1,\infty), \quad \text{and } \|X\|_{L^\infty} = \operatorname{ess\,sup}|X|, \tag{1.10}$$

where ess sup denotes the essential supremum.

The normed space consisting of all random variables on a given probability space that have finite L^p norms is called the L^p space:

$$L^p = \{X \colon \|X\|_{L^p} < \infty\}.$$

The Minkowski inequality, which we recall in (1.21), implies that the L^p norm indeed defines a norm on the space L^p for each $p \in [1,\infty]$.

The exponent $p = 2$ is special: L^2 is not only a normed but also an inner product space. The inner product is given by

$$\langle X, Y\rangle_{L^2} = \mathbb{E}\,XY, \tag{1.11}$$

and it agrees with the L^2 norm in the sense that $\|X\|_{L^2}^2 = \langle X, X\rangle_{L^2}$.

The *standard deviation* of a random variable X is

$$\sigma(X) = \sqrt{\operatorname{Var}(X)} = \|X - \mathbb{E}\,X\|_{L^2}. \tag{1.12}$$

The *covariance* of random variables X and Y is

$$\operatorname{cov}(X,Y) = \mathbb{E}(X - \mathbb{E}\,X)(Y - \mathbb{E}\,Y) = \langle X - \mathbb{E}\,X, Y - \mathbb{E}\,Y\rangle_{L^2}. \tag{1.13}$$

The concept of a random variable can be generalized to higher dimensions. We can define a *random vector* $X = (X_1, \ldots, X_n)$ taking values in $\mathbb{R}^n$ as a vector whose all n coordinates X_i are random variables. The expected value of X is defined coordinate-wise, that is,

$$\mathbb{E}\,X = (\mathbb{E}\,X_1, \ldots, \mathbb{E}\,X_n)\,.$$

What about *variance in high dimensions*? If we want the traditional definition of variance (1.7) to make sense for random vectors, we have to decide how to square vectors $x \in \mathbb{R}^n$, or how to multiply vectors x by themselves. This can be done in two ways: as a dot product, or "inner product" $x^\mathsf{T}x = \|x\|_2^2$, and as an

https://doi.org/10.1017/9781009490672.004 Published online by Cambridge University Press

"outer product" xx^{T}. The first interpretation leads to defining the variance of X as $\mathbb{E}\|X - \mathbb{E}\,X\|_2^2$, the quantity that played a prominent role in the Appetizer. The second interpretation leads to a more informative result: it gives rise to the notion of the *covariance matrix*

$$\operatorname{cov}(X) = \mathbb{E}(X - \mathbb{E}\,X)(X - \mathbb{E}\,X)^{\mathsf{T}},$$

which is an $n \times n$ matrix whose (i,j)th entry equals $\operatorname{cov}(X_i, X_j)$. We will work with covariance matrices in Section 3.2.

1.4 Union Bound

If E_i are disjoint events, then the additivity axiom of probability tells us that

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mathbb{P}\left(E_i\right).$$

If the events E_i are not disjoint, this identity may fail: each point belonging to several events E_i is counted as one on the left-hand side, but several times on the right-hand side. Thus, instead of an equality, we get an inequality:

Lemma 1.4.1 (Union bound) *For any events $E_1, \ldots, E_n$, we have*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) \le \sum_{i=1}^{n} \mathbb{P}\left(E_i\right).$$

Proof If the event $\bigcup_{i=1}^{n} E_i$ occurs, then at least one of the events E_i occurs. Thus, the indicators satisfy

$$\mathbf{1}_{\bigcup_{i=1}^{n} E_i} \le \sum_{i=1}^{n} \mathbf{1}_{E_i}. \tag{1.14}$$

(The right-hand side counts how many E_i occur.) Now take expectations and use (1.9) and linearity of expectation to complete the proof. □

The union bound is typically used to bound the probability of a bad event.

Example 1.4.2 (Dense random graphs have no isolated vertices) Imagine that $n \ge 2$ freshmen arrive on campus. Friendships form randomly, with each pair of students becoming friends with probability p independent of all other pairs. Let us show that if $p \ge 4\ln(n)/n$ then there are no friendless students with probability at least $1 - 1/n$.

Proof Call the students $1, \ldots, n$, and let E_i denote the event that student i is friendless. This means that none of the other $n-1$ students are friends with i, and these $n-1$ sad events are independent and have probability $1-p$ each. Thus $\mathbb{P}\left(E_i\right) = (1-p)^{n-1}$.

We would like to bound the probability of the bad event

$$B = \{\text{there exists a friendless student}\} = \bigcup_{i=1}^{n} E_i.$$

https://doi.org/10.1017/9781009490672.004 Published online by Cambridge University Press

The union bound gives

$$\mathbb{P}(B) \le \sum_{i=1}^{n} \mathbb{P}\left(E_i\right) = n(1-p)^{n-1}.$$

To simplify this bound, use the inequality $1-p \le e^{-p}$ and recall the assumptions on n and p. A little computation yields $n(1-p)^{n-1} \le 1/n$. Thus, the complement of the bad event B happens with probability at least $1 - 1/n$. □

1.5 Conditioning

The conditioning trick often helps us to calculate probabilities and expectations. This method is based on the concept of *conditional probability* of an event E given event F, denoted by

$$\mathbb{P}(E \mid F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)},$$

and the concept of *conditional expectation* of a random variable X given another random variable Y, denoted by $\mathbb{E}[X|Y]$.

The *law of total expectation* says that

$$\mathbb{E}\, X = \mathbb{E}[\mathbb{E}[X|Y]]. \tag{1.15}$$

Informally, this identity says that in order to compute the expectation of X, we can first average with respect to X while keeping the value of Y fixed, and then average the result with respect to Y. This is the essence of the conditioning trick.

The conditioning trick can also help us compute probabilities. If E is any event and Y is a random variable, we can define the conditional probability of E given Y as

$$\mathbb{P}(E \mid Y) \coloneqq \mathbb{E}[\mathbf{1}_E \mid Y],$$

where $\mathbf{1}_E$ denotes the indicator of the event E. Since $\mathbb{E}\, \mathbf{1}_E = \mathbb{P}(E)$, the law of total expectation in this case reads as

$$\mathbb{P}(E) = \mathbb{E}[\mathbb{P}(E \mid Y)]. \tag{1.16}$$

Thus, in order to compute probabilities using the conditioning trick, we first need to find the probability while keeping the value of Y fixed, and then average the result with respect to Y.

Suppose we have a decomposition of the sample space Ω into mutually disjoint events $F_1, F_2, \ldots$ In other words, suppose that each outcome belongs to one, and only one, event F_i. Then the *law of total probability* allows us to compute the probability of any event E as follows:

$$\mathbb{P}(E) = \sum_i \mathbb{P}(E \mid F_i)\mathbb{P}(F_i). \tag{1.17}$$

The law of total probability can be quickly deduced from the law of total expectation. To do so, apply (1.16) for the random variable Y that takes value i if event F_i

https://doi.org/10.1017/9781009490672.004 Published online by Cambridge University Press

occurs. Note that the random variable $\mathbb{P}(E|Y)$ takes value $\mathbb{P}(E|Y=i) = \mathbb{P}(E|F_i)$ with probability $\mathbb{P}\{Y=i\} = \mathbb{P}(F_i)$.

Everything we said here about random variables X and Y is also true for random vectors.

Example 1.5.1 (Probability of a perfect cancelation) Let $a_1, \dots, a_n$ be real numbers, not all of which are zero. What is the probability that

$$\pm a_1 \pm \cdots \pm a_n = 0,$$

where the signs are chosen at random? Let us show that this probability is always bounded by $1/2$. To state this problem rigorously, we can model the random signs as independent *Rademacher random variables* $X_1, \dots, X_n$, i.e. random variables that take values -1 and 1 with probability $1/2$ each. We claim that

$$\mathbb{P}\{S_n = 0\} \le \frac{1}{2} \quad \text{where} \quad S_n = \sum_{i=1}^{n} a_i X_i.$$

Proof We can assume without loss of generality that $a_n \neq 0$ by rearranging the terms if necessary. The proof is by exposing the last term $a_n X_n$ of the sum, while keeping all the previous terms fixed by conditioning. So let us condition on the random variables $X_1, \dots, X_{n-1}$, or to be pedantic, on the random vector $(X_1, \dots, X_{n-1})$. The conditioning fixes the values of $X_1, \dots, X_{n-1}$, and therefore it fixes the value of $S_{n-1} = \sum_{i=1}^{n-1} a_i X_i$. All randomness is now left with X_n. Since $S_n = S_{n-1} + a_n X_n$, we have

$$\mathbb{P}\{S_n = 0 \,|\, X_1, \dots, X_{n-1}\} = \mathbb{P}\left\{ X_n = -\frac{S_{n-1}}{a_n} \,\middle|\, X_1, \dots, X_{n-1} \right\} \le \frac{1}{2}.$$

The inequality holds because X_n is independent of $X_1, \dots, X_{n-1}$, the value $u = -S_{n-1}/a_n$ is fixed by conditioning, and the definition of the Rademacher distribution implies that $\mathbb{P}\{X_n = u\} \le 1/2$ for any fixed value $u \in \mathbb{R}$.

We can finish the proof by applying the law of total expectation (1.16):

$$\mathbb{P}\{S_n = 0\} = \mathbb{E}[\mathbb{P}\{S_n = 0 \,|\, X_1, \dots, X_{n-1}\}] \le \mathbb{E}\,\frac{1}{2} = \frac{1}{2}. \qquad \square$$

By the way, the result in Example 1.5.1 is sharp. If there are exactly two nonzero coefficients a_i and they are equal to each other, then we have $\mathbb{P}\{S_n = 0\} = 1/2$. (Check!)

1.6 Probabilistic Inequalities

The *Jensen inequality* states that for any random variable X and a convex function $f\colon \mathbb{R} \to \mathbb{R}$, we have

$$f(\mathbb{E}\,X) \le \mathbb{E}\, f(X). \tag{1.18}$$

More generally, (1.18) holds for any random vector X taking values in $\mathbb{R}^n$ and any convex function $f\colon \mathbb{R}^n \to \mathbb{R}$. In Exercise 1.3, you will prove this for random vectors taking finitely many values; the general case can be deduced by approximation.

https://doi.org/10.1017/9781009490672.004 Published online by Cambridge University Press

In particular, since any norm on $\mathbb{R}^n$ is a convex function, the Jensen inequality yields

$$\|\mathbb{E}\, X\| \le \mathbb{E}\|X\|. \tag{1.19}$$

The L^p norms of a given random variable X are increasing[4] in p:

$$\|X\|_{L^p} \le \|X\|_{L^q} \quad \text{whenever } p \le q, \tag{1.20}$$

see Exercise 1.11.

The *Minkowski inequality* states that for any $p \in [1, \infty]$ and any random variables $X, Y \in L^p$, we have

$$\|X + Y\|_{L^p} \le \|X\|_{L^p} + \|Y\|_{L^p}. \tag{1.21}$$

In other words, the L^p norm satisfies the triangle inequality.

The *Cauchy–Schwarz inequality* states that for any random variables $X, Y \in L^2$, we have

$$\|XY\|_{L^1} \le \|X\|_{L^2} \|Y\|_{L^2}.$$

The *Hölder inequality* generalizes this result to the L^p norms. For any pair of conjugate exponents $p, p' \in [1, \infty]$ (introduced in (1.5)) and any pair of random variables $X \in L^p$ and $Y \in L^{p'}$, we have

$$\|XY\|_{L^1} \le \|X\|_{L^p} \|Y\|_{L^{p'}}. \tag{1.22}$$

As we recall from a basic probability course, the *distribution* of a random variable X is, intuitively, the information about what values X takes with what probabilities. More rigorously, the distribution of X is determined by the *cumulative distribution function* (CDF) of X, defined as

$$F_X(t) = \mathbb{P}\{X \le t\}, \quad t \in \mathbb{R}.$$

It is often more convenient to work with *tails* of random variables, namely, with

$$\mathbb{P}\{X > t\} = 1 - F_X(t).$$

The following result allows us to compute expectation in terms of the tails.

Lemma 1.6.1 (Integrated tail formula) *Any nonnegative random variable X satisfies*

$$\mathbb{E}\, X = \int_0^\infty \mathbb{P}\{X > t\}\, dt.$$

The two sides of this identity are either finite or infinite simultaneously.

[4] This might sound a little bit confusing when we recall from (1.4) that the ℓ^p norms are *decreasing* in p. There is no contradiction because the ℓ^p and L^p norms are normalized differently. The ℓ^p norm of a vector $x = (x_1, \dots, x_n)$ is $\|x\|_{\ell^p} = \big(\sum_{i=1}^n |x_i|^p\big)^{1/p}$, while the L^p norm of a random variable X that takes values $x_1, \dots, x_n$ with probabilities $1/n$ each is $\|X\|_{L^p} = \big(\frac{1}{n}\sum_{i=1}^n |x_i|^p\big)^{1/p}$.

https://doi.org/10.1017/9781009490672.004 Published online by Cambridge University Press

Proof We can represent any nonnegative real number x via the identity

$$x = \int_0^x 1\,dt = \int_0^\infty \mathbf{1}_{\{t<x\}}\,dt.$$

Replace x with the random variable X and take the expectation on both sides. This gives

$$\mathbb{E}\,X = \mathbb{E}\int_0^\infty \mathbf{1}_{\{t<X\}}\,dt = \int_0^\infty \mathbb{E}\,\mathbf{1}_{\{t<X\}}\,dt = \int_0^\infty \mathbb{P}\{t<X\}\,dt.$$

To change the order of expectation and integration in the second equality, we used the Fubini–Tonelli theorem. The proof is complete. □

Conversely, the Markov inequality bounds the tails in terms of expectation:

Proposition 1.6.2 (Markov inequality) *For any nonnegative random variable X and $t>0$, we have*

$$\mathbb{P}\{X \ge t\} \le \frac{\mathbb{E}\,X}{t}.$$

Proof Fix $t>0$. We can represent any real number x via the identity

$$x = x\mathbf{1}_{\{x\ge t\}} + x\mathbf{1}_{\{x<t\}}.$$

Replace x with the random variable X and take expectation:

$$\mathbb{E}\,X = \mathbb{E}\,X\mathbf{1}_{\{X\ge t\}} + \mathbb{E}\,X\mathbf{1}_{\{X<t\}} \ge \mathbb{E}\,t\mathbf{1}_{\{X\ge t\}} + 0 = t\cdot\mathbb{P}\{X\ge t\}.$$

Divide both sides by t to complete the proof. □

Markov's inequality gives the best possible tail bound if all we know is $\mathbb{E}\,X$. But if we also know the variance, we get a better bound that decays quadratically in t – and learn how tightly X concentrates around its mean:

Corollary 1.6.3 (Chebyshev inequality) *Let X be a random variable with mean μ and variance σ^2. Then, for any $t>0$, we have*

$$\mathbb{P}\{|X-\mu|\ge t\} \le \frac{\sigma^2}{t^2}.$$

Proof Square both sides of the bound $|X-\mu|\ge t$ and apply the Markov inequality for the random variable $(X-\mu)^2$. □

1.7 Limit Theorems

The study of *sums of independent random variables* lies at the heart of classical probability theory. Let $X_1,\dots,X_N$ be independent and identically distributed (i.i.d.) random variables with mean μ and variance σ^2. Then the formula for the variance of a sum (1.8) gives

$$\operatorname{Var}\left(\frac{1}{N}\sum_{i=1}^N X_i\right) = \frac{\sigma^2}{N}. \tag{1.23}$$

https://doi.org/10.1017/9781009490672.004 Published online by Cambridge University Press

Thus, the variance of the *sample mean* $\frac{1}{N}\sum_{i=1}^{N} X_i$ of the sample $\{X_1, \ldots, X_N\}$ shrinks to zero as the size of the sample N increases to infinity. This indicates that for large N, we should expect that the sample mean concentrates tightly about its expectation μ. One of the most important results in probability theory – the law of large numbers – states precisely this.

Theorem 1.7.1 (Strong law of large numbers) *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean μ. Consider the sum*

$$S_N = X_1 + \cdots X_N.$$

Then, as $N \to \infty$,

$$\frac{S_N}{N} \to \mu \quad \textit{almost surely.}$$

The central limit theorem makes one step further. It identifies the limiting distribution of the (properly scaled) sum S_N as a *normal* distribution, sometimes also called a *Gaussian* distribution.

Definition 1.7.2 (Normal distribution) A random variable X is a standard normal random variable, denoted by

$$X \sim N(0,1)$$

if X has density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}. \tag{1.24}$$

The random variable X has mean zero and variance one.

More generally, a random variable X has a normal distribution

$$X \sim N(\mu, \sigma^2)$$

if X can be expressed as $X = \mu + \sigma Z$ for some fixed numbers $\mu \in \mathbb{R}$ and $\sigma > 0$, and where $Z \sim N(0,1)$. The density of X equals

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

The random variable X has mean μ and variance σ^2.

Theorem 1.7.3 (Lindeberg–Lévy central limit theorem) *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Consider the sum*

$$S_N = X_1 + \cdots + X_N$$

and normalize it to obtain a random variable with zero mean and unit variance:

$$Z_N := \frac{S_N - \mathbb{E}\, S_N}{\sqrt{\mathrm{Var}(S_N)}} = \frac{1}{\sigma\sqrt{N}} \sum_{i=1}^{N} (X_i - \mu).$$

Then, as $N \to \infty$,

$$Z_N \to N(0,1) \quad \textit{in distribution.}$$

https://doi.org/10.1017/9781009490672.004 Published online by Cambridge University Press

Convergence in distribution means that the CDF of Z_N converges pointwise to the CDF of $g \sim N(0,1)$. In terms of tails, this means that for every $t \in \mathbb{R}$,

$$\mathbb{P}\{Z_N > t\} \to \mathbb{P}\{g > t\} = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\, dx \quad \text{as } N \to \infty.$$

Example 1.7.4 (Bernoulli and binomial distributions) One remarkable special case of the central limit theorem is where X_i are Bernoulli random variables with some fixed parameter $p \in (0,1)$, denoted

$$X_i \sim \mathrm{Ber}(p).$$

This means that X_i take values 1 and 0 with probabilities p and $1-p$, respectively. One can easily check that

$$\mathbb{E}\, X_i = p \text{ and } \mathrm{Var}(X_i) = p(1-p).$$

The sum $S_N := X_1 + \cdots + X_N$ of independent $\mathrm{Ber}(p)$ random variables X_i is said to have the *binomial distribution*, denoted

$$S_N \sim \mathrm{Binom}(N, p).$$

The central limit theorem (Theorem 1.7.3) yields that as $N \to \infty$,

$$\frac{S_N - Np}{\sqrt{Np(1-p)}} \to N(0,1) \quad \text{in distribution.} \tag{1.25}$$

This special case of the central limit theorem is called the *de Moivre–Laplace theorem*.

Now consider independent random variables $X_i \sim \mathrm{Ber}(p_i)$ whose parameters p_i *decay to zero* as $N \to \infty$ so fast that the sum S_N has mean $O(1)$ instead of being proportional to N. The central limit theorem fails in this regime. A different result we are about to state says that S_N still converges, but to the *Poisson distribution* instead of the normal distribution.

Definition 1.7.5 (Poisson distribution) A random variable Z has *Poisson distribution* with parameter $\lambda > 0$, denoted

$$Z \sim \mathrm{Pois}(\lambda),$$

if Z takes values in $\{0, 1, 2, \ldots\}$ with probabilities

$$\mathbb{P}\{Z = k\} = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots \tag{1.26}$$

Theorem 1.7.6 (Poisson limit theorem) *Consider independent random variables* $X_{N,i} \sim \mathrm{Ber}(p_{N,i})$ *for* $N = 1, 2, \ldots$ *and* $1 \le i \le N$. *Let*

$$S_N = X_{N,1} + \cdots + X_{N,N}.$$

Assume that, as $N \to \infty$,

$$\max_{i \le N} p_{N,i} \to 0 \quad \textit{and} \quad \mathbb{E}\, S_N = p_{N,1} + \cdots + p_{N,N} \to \lambda < \infty.$$

https://doi.org/10.1017/9781009490672.004 Published online by Cambridge University Press

Then, as $N \to \infty$,

$$S_N \to \text{Pois}(\lambda) \quad \textit{in distribution.}$$

The probability mass function (1.26) of the Poisson distribution includes the factorial $k!$, which is not an easy function to work with. For large k, we can simplify it using the Stirling approximation:

Lemma 1.7.7 (Stirling approximation) *We have*

$$n! = \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n (1 + o(1)) \quad \textit{as } n \to \infty.$$

Using the Stirling approximation in (1.26), we see that for any fixed parameter $\lambda > 0$, a Possion random variable $Z \sim \text{Pois}(\lambda)$ satisfies

$$\mathbb{P}\{Z = k\} = \frac{e^{-\lambda}}{\sqrt{2\pi k}} \left(\frac{e\lambda}{k}\right)^k (1 + o(1)) \quad \text{as } k \to \infty, \tag{1.27}$$

so this decays roughly as k^{-k} – slightly faster than exponentially.

There are non-asymptotic versions of the Stirling approximation which hold for any given value of n as opposed to the limit as $n \to \infty$. Here is one of them:

Lemma 1.7.8 (Bounds on the factorial) *For any* $n \in \mathbb{N}$, *we have*

$$\Big(\frac{n}{e}\Big)^n \le n! \le en\Big(\frac{n}{e}\Big)^n. \tag{1.28}$$

Proof If we recall the Taylor series for e^x and drop all terms except the nth, we get $e^x \ge x^n/n!$. Substitute $x = n$ and rearrange the terms to obtain the lower bound in (1.28). To prove the upper bound, note that

$$\ln(n!) = \sum_{k=1}^{n} \ln k \le \int_1^n \ln x \, dx + \ln n = n(\ln n - 1) + 1 + \ln n. \tag{1.29}$$

(The inequality in (1.29) follows by comparing the areas as in the usual proof of the integral test – do it!) Exponentiating and rearranging the terms yields the upper bound in (1.28). □

Remark 1.7.9 (Gamma function) The *gamma function* extends the notion of factorial for all real numbers, and even for all complex numbers z whose real part is positive. It is defined as follows:

$$\Gamma(z) := \int_0^\infty t^{z-1} e^{-t} dt. \tag{1.30}$$

Repeated integration by parts (do it!) reveals that

$$\Gamma(n+1) = n! \quad \text{for all } n = 0, 1, 2, \ldots$$

The Stirling approximation (Lemma 1.7.7) remains valid for the gamma function:

$$\Gamma(z+1) = \sqrt{2\pi z}\Big(\frac{z}{e}\Big)^z (1 + o(1)) \quad \text{as } \mathbb{R} \ni z \to \infty. \tag{1.31}$$

https://doi.org/10.1017/9781009490672.004 Published online by Cambridge University Press

1.8 Notes

The question we raise in Example 1.5.1 is known as the *Littlewood–Offord problem*. Originally considered by Littlewood and Offord [217] and Erdős [121], this problem and its variants was studied extensively since then; see the surveys [318, 291, 258].

Proofs of the strong law of large numbers (Theorem 1.7.1) and the Lindeberg–Lévy central limit theorem (Theorem 1.7.3) can be found, e.g. in [116, Sections 1.7 and 2.4] and [42, Sections 6 and 27].

Both Proposition 1.6.2 and Corollary 1.6.3 are due to Chebyshev. However, following the established tradition, we call Proposition 1.6.2 the Markov inequality.

In modern language, Example 1.4.2, Exercise 1.9, and Exercise 1.10 identify the threshold for the existence of isolated vertices in the Erdős–Rényi model $G(n,p)$. This problem was first studied in the seminal paper of Erdős and Rényi [122]. There are now more general and more precise results about the degrees of random graphs, see [47, Chapter 3] and [131, Chapter 3]. The second moment method showcased in Exercise 1.10 is ubiquitous in combinatorics and theoretical computer science, see [17, Chapter 4]. The result of Exercise 1.8 is asymptotically optimal: the largest cardinality of an independent set in an Erdős–Rényi random graph $G \sim G(n,p)$ is approximately $2\log_b n$ where $b = 1/(1-p)$, see [131, Section 7.2].

A short proof of the Stirling approximation (Lemma 1.7.7) can be found in [286] and [123, II.9]. This proof yields the following non-asymptotic result, which holds for any $n \in \mathbb{N}$:

$$\sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n e^{\frac{1}{12n+1}} \le n! \le \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n e^{\frac{1}{12n}}.$$

This result implies the asymptotic form of the Stirling approximation (Lemma 1.7.8) and also improves upon Lemma 1.7.8. Short proofs of the Stirling approximation for the gamma function (1.31) can be found in [44, 101]; for non-asymptotic versions, see [173].

1.9 Exercises

1.1 ☕ Consider any subset $T \subset \mathbb{R}^n$. Check that $\operatorname{conv}(T)$ is a convex set.

1.2 ☕ Check that the pointwise maximum of a finite number of convex functions is a convex function.

1.3 ☕☕ (Jensen inequality)

(a) The definition of a convex function (1.1) involves convex combinations of two points x and y. Let us extend it to arbitrarily many points. Let $K \subset \mathbb{R}^n$ be a convex subset. Prove that a function $f\colon K \to \mathbb{R}$ is convex if and only if the following holds. For any $m \in \mathbb{N}$, any vectors $x_i \in K$, and any numbers $\lambda_i \ge 0$ with $\sum_{i=1}^m \lambda_i = 1$, we have

$$f\left(\sum_{i=1}^m \lambda_i x_i\right) \le \sum_{i=1}^m \lambda_i f(x_i).$$

(b) Let X be a random vector in $\mathbb{R}^n$ that takes finitely many values, and let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a convex function. Deduce from part (a) the Jensen inequality:

$$f(\mathbb{E}\, X) \le \mathbb{E}\, f(X).$$

1.4 ☕ (A maximum principle) Prove for any convex function f and a subset $T \subset \mathbb{R}^n$:

$$\sup_{x \in \operatorname{conv}(T)} f(x) = \sup_{x \in T} f(x).$$

https://doi.org/10.1017/9781009490672.004 Published online by Cambridge University Press

1.5 ☕☕ (Expressing a cube as a convex hull of its vertices) It seems almost obvious that the cube is the convex hull of its vertices:

$$[-1,1]^n = \operatorname{conv}\left(\{-1,1\}^n\right).$$

Prove this by expressing any point in the cube as a convex combination of the vertices.

1.6 ☕☕ (Expressing a cross-polytope as a convex hull of its vertices) Check that the unit ball corresponding to the ℓ^1 norm in $\mathbb{R}^n$ is the *absolute convex hull* of the standard basis $e_1, \dots, e_n$ in $\mathbb{R}^n$, that is,

$$B_1^n = \operatorname{conv}\left(\{\pm e_1, \dots, \pm e_n\}\right).$$

Write down a formula that expresses any point $x \in B_1^n$ as a convex combination of the vectors $\pm e_1, \dots, \pm e_n$.

1.7 ☕☕ (Random graphs with random number of vertices) Suppose that in Example 1.4.2, the number n of freshmen who arrive on campus is a random variable that has Poisson distribution with mean λ. As before, each pair of students becomes friends with probability p independent of all other pairs. Show that if $p \geq 2\ln(\lambda)/\lambda$ then there are no friendless students with probability at least $1 - 1/\lambda$.

1.8 ☕☕ (Independent sets in random graphs) Call a group of people *independent* if no two members are friends. Suppose $n \geq 7$ students enroll in a class on high-dimensional probability, with each pair becoming friends independently with probability $1/2$. Show that, with probability at least $1 - 1/n$, this class has no independent subsets of more than $2\log_2 n$ students.

1.9 ☕ (Dense random graphs have no isolated vertices) Let us refine the result of Example 1.4.2. Suppose n freshmen arrive on campus, with each pair becoming friends independently with probability p_n. Fix any $\varepsilon > 0$ and assume that

$$p_n > \frac{(1+\varepsilon)\ln n}{n} \quad \text{for every } n \in \mathbb{N}.$$

Prove that there are no friendless students with probability that converges to 1 as $n \to \infty$.

1.10 ☕☕☕☕ (Sparse random graphs have isolated vertices) Let us prove a converse to Exercise 1.9. Fix any $\varepsilon > 0$ and assume that

$$p_n < \frac{(1-\varepsilon)\ln n}{n} \quad \text{for every } n \in \mathbb{N}.$$

Then there exists at least one friendless student with probability that converges to 1 as $n \to \infty$. You will prove this result using the so-called *second moment method*:

(a) Denote the number of friendless students by S_n and express it as $S_n = X_1 + \cdots + X_n$ where X_i is the indicator of the event that student i is friendless. Show that

$$\mu_n = \mathbb{E}\, S_n \to \infty.$$

https://doi.org/10.1017/9781009490672.004 Published online by Cambridge University Press

Thus, the expected number of friendless students is large. But this does not automatically imply that there exists even one friendless student with high probability! (Why?)

(b) Compute the second moment $\mathbb{E}\, S_n^2$ by expanding the square. Conclude that

$$\frac{\mathrm{Var}(S_n)}{\mu_n^2} \to 0.$$

(c) Use the Chebyshev inequality to complete the proof.

1.11 ☕☕ (Monotonicity of the L^p norm)

(a) Let X be a random variable. Show that $\|X\|_{L^p}$ is an increasing function in p:

$$\|X\|_{L^p} \le \|X\|_{L^q} \quad \text{for any } 0 \le p \le q \le \infty.$$

(b) Demonstrate that the inequality in part (a) cannot be reversed: for any $0 \le p < q \le \infty$, find an example of a random variable X with $\|X\|_{L^p} < \infty$ and $\|X\|_{L^q} = \infty$.

1.12 ☕ (Interpolation between L^1 and L^∞) We know the L^p norm of any random variable X is bounded by the L^∞ norm. We can get an even better bound if we also know that the L^1 norm of X is small. Show that

$$\|X\|_{L^p} \le \|X\|_{L^1}^{\frac{1}{p}} \|X\|_{L^\infty}^{1-\frac{1}{p}} \quad \text{for any } 1 < p < \infty.$$

1.13 ☕☕☕ (Expectation of a maximum) Let $X_1, \dots, X_n$ be nonnegative random variables.

(a) Prove that

$$\max_{i \le n} \mathbb{E}\, X_i \le \mathbb{E} \max_{i \le n} X_i \le n \cdot \max_{i \le n} \mathbb{E}\, X_i.$$

(b) Demonstrate that both inequalities in part (a) may be optimal. Specifically, find random variables $X_1, \dots, X_n$ satisfying $\max_i \mathbb{E}\, X_i = \mathbb{E} \max_i X_i > 0$ and random variables Y_i satisfying $\mathbb{E} \max_i Y_i = n \cdot \max_i \mathbb{E}\, Y_i > 0$.

(c) Demonstrate that the upper bound in part (a) may be approximately optimal even for independent random variables. Specifically, find independent random variables $X_1, \dots, X_n$ satisfying $\mathbb{E} \max_i X_i > cn \cdot \max_i \mathbb{E}\, X_i$, where $c > 0$ is an absolute constant.[5]

1.14 ☕☕ Let $X_1, \dots, X_n$ be nonnegative random variables. Prove that for any $1 \le p < \infty$, we have

$$\left(\sum_{i=1}^n (\mathbb{E}\, X_i)^p \right)^{1/p} \le \mathbb{E} \left(\sum_{i=1}^n X_i^p \right)^{1/p} \le \left(\sum_{i=1}^n \mathbb{E}\left(X_i^p\right) \right)^{1/p}.$$

1.15 ☕ (Integrated tail formulas) Prove the following more general versions of Lemma 1.6.1.

[5] You are free to choose any value for $c > 0$ as long as it is an *absolute constant*, which means that c must not depend on anything, in particular on the distribution of the random variables Y_i or their number n. From the context it should be clear that choosing a small value such as $c = 0.001$ would make your job easier than choosing a large value such as $c = 1000$.

https://doi.org/10.1017/9781009490672.004 Published online by Cambridge University Press

(a) Let X be any random variable, not necessarily nonnegative. Then

$$\mathbb{E}\, X = \int_0^\infty \mathbb{P}\{X > t\}\, dt - \int_{-\infty}^0 \mathbb{P}\{X < t\}\, dt.$$

(b) Let X be a nonnegative random variable. Let $f\colon \mathbb{R}^+ \to \mathbb{R}^+$ be an increasing, differentiable function satisfying $f(0) = 0$. Then

$$\mathbb{E}\, f(X) = \int_0^\infty \mathbb{P}\{X > t\} f'(t)\, dt.$$

(c) Let X be a random variable, not necessarily nonnegative. Deduce that for every $p \in (0, \infty)$ we have

$$\mathbb{E}|X|^p = \int_0^\infty \mathbb{P}\{|X| > t\} p t^{p-1}\, dt.$$

1.16 ☕☕ (Paley–Zygmund inequality) The Markov inequality says a random variable is unlikely to be much bigger than its expectation. But what about the reverse? Can a nonnegative random variable be much smaller than its expectation with high probability? In general, yes (example?), but not if the second moment isn't too large. Let X be a nonnegative random variable with finite variance. Show that for any $\varepsilon \in [0, 1]$,

$$\mathbb{P}\{X > \varepsilon\, \mathbb{E}\, X\} \geq (1 - \varepsilon)^2 \frac{(\mathbb{E}\, X)^2}{\mathbb{E}[X^2]}.$$

1.17 ☕☕☕ (Comparison of the ℓ^p norms) Let $0 \leq p \leq q \leq \infty$.

(a) Prove that for any vector $x \in \mathbb{R}^n$ we have

$$\|x\|_q \leq \|x\|_p \leq n^{\frac{1}{p} - \frac{1}{q}} \|x\|_q.$$

(b) Demonstrate that both inequalities in part (a) can be optimal. Specifically, find nonzero vectors $x, y \in \mathbb{R}^n$ satisfying $\|x\|_p = \|x\|_q$ and $\|y\|_p = n^{\frac{1}{p} - \frac{1}{q}} \|y\|_q$.

1.18 ☕ (The ℓ^∞ norm is the limit of the ℓ^p norms) Consider any vector $x \in \mathbb{R}^n$.

(a) Prove that

$$\|x\|_p \to \|x\|_\infty \text{ as } p \to \infty.$$

(b) In fact, p does not need to be too large for the ℓ^p norm to get reasonably close to the ℓ^∞ norm. Show that if $p \geq \ln n$, then

$$\|x\|_\infty \leq \|x\|_p \leq e\|x\|_\infty.$$

1.19 ☕☕☕ (Duality of the ℓ_p norms) Let $p, p' \in [1, \infty]$ be conjugate exponents.

(a) Show that the Hölder inequality is tight: for any vector x there exists a vector $y \neq 0$ for which $\langle x, y\rangle = \|x\|_p \|y\|_{p'}$.

(b) Conclude that for every vector $x \in \mathbb{R}^n$, we have

$$\max\left\{\langle x, y\rangle \colon y \in B_{p'}^n\right\} = \|x\|_p.$$

https://doi.org/10.1017/9781009490672.004 Published online by Cambridge University Press