# Probabilities, Vectors, and Matrices in Computing

A Textbook for CS 4850 and 5850

**Robert Kleinberg**

Copyright © 2026 Robert Kleinberg

CORNELL UNIVERSITY COURSE MATERIAL

Licensed under the Creative Commons Attribution-NonCommercial 4.0 License (the "License"). You may not use this file except in compliance with the License. You may obtain a copy of the License at https://creativecommons.org/licenses/by-nc-sa/4.0. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

*This edition: Version 1.0 — 25 January 2026*

# Contents

# Discrete Probability and Algorithms

# 1. Balls and Bins

We begin our study of randomized algorithms by analyzing one of the most elementary random sampling processes: drawing a sequence of independent, uniformly random elements of the set $[n] = \{1, 2, \ldots, n\}$. The colloquial expression "balls and bins" is often applied to such processes, because of the metaphor of a sequence of balls being thrown at random into $n$ bins.

Balls and bins are used to model computing phenomena such as hashing in data structures and load balancing in distributed systems. (We'll say much more about hashing later in this course.) A key intuition is that the process tends to lead to a fairly even distribution of balls among the bins, since all of the bins are treated symmetrically. Some of the most basic questions about balls and bins concern the extent to which the loads of the different bins (i.e. the numbers of balls they each contain) deviate from a perfectly uniform distribution. In all of the following problems, assume $m$ balls are thrown into $n$ bins.

**Birthday Paradox** How likely is it that at least one bin contains more than one ball?

**Coupon Collector Problem** How likely is it that no bin is empty?

**Load Balancing** How likely is it that the maximum and minimum bin loads differ by a factor less than $1 + \varepsilon$?

The third question illustrates a notational convention that we will use throughout this course: unless otherwise specified, the notation $\varepsilon$ represents a number in the range $0 < \varepsilon < 1$, whose value does not vary with the other parameters of the problem. (In this case, that means $\varepsilon$ does not depend on the number of balls or bins.)

## 1.1 The Birthday Paradox

Let's make the unrealistic assumption that birthdays are uniformly distributed[1] over the 365 days of the calendar. How large must a group of people be, in order for the probability that at least two of them share the same birthday to exceed $\frac{1}{2}$? Counterintuitively, this happens as long as there are at least 23 people in the group. This is generally known as the "birthday paradox", though it would be more accurate to describe the phenomenon as *surprising* rather than *paradoxical*.

The birthday paradox is a thinly disguised question about balls and bins: people's birthdays are the balls, and dates on the calendar are the bins. Before delving into the analysis that formally justifies the birthday paradox, the following intuition is helpful. The probability that any two given people have the same birthday (under our simplifying assumption of uniformly-distributed birthdays) is $\frac{1}{365}$. In a group of 23 people the number of *pairs* of people is $\binom{23}{2} = 253$. By linearity of expectation, the expected number of pairs of people who share a birthday is $\frac{253}{365} = 0.693\ldots$. In light of this calculation, it become a bit less surprising that there is a significant probability that the group contains at least one pair of people who share a birthday.

Let's consider, more generally, the *collision probability* when throwing $m$ balls into $n$ bins — that is, the probability that at least one bin is occupied by more than one ball. It turns out that the best way to calculate the collision probability is to calculate the probability of *no collisions* and then subtract from 1. An exact formula for the probability of no collisions can easily be obtained by thinking about throwing the balls sequentially. When the first ball is thrown, there cannot be a collision. When the second ball is thrown, the probability that it doesn't collide with the first is $\frac{n-1}{n}$. More generally, if $k \leq n$ balls have been thrown and all of them occupy distinct bins, then there are $n-k$ remaining unoccupied bins, so the probability that the $(k+1)^{\text{th}}$ ball does not collide with any of its predecessors is $\frac{n-k}{n}$. Multiplying all of these probabilities together, we find that

$$\Pr(\text{no collision}) = \prod_{k=1}^{m-1} \frac{n-k}{n} = \prod_{k=1}^{m-1} \left(1 - \frac{k}{n}\right). \tag{1.1}$$

Expressions like the right side of equation (1.1) are painful to deal with because if you expand out the product using the distributive law, you get a sum of exponentially many terms. (In this case, $2^{m-1}$ terms.) To make progress, it's time to apply *the most useful inequality in the analysis of randomized algorithms.*

**Fact 1.1** The inequality

$$1 + x \leq e^x$$

holds for all $x \in \mathbb{R}$, and the inequality is strict except when $x = 0$.

*Proof.* The proof of the inequality is an application of the Mean Value Theorem and case analysis. When $x = 0$ the two sides of the inequality are both equal to 1. When $x > 0$

[1] In fact, the distribution of birthdays in the United States is quite far from uniform. There are almost twice as many people born on the most common birthday, September 9, than on the least common annually-occurring one, December 25. Of course, it is even less common for someone to have the birthday February 29 because that date only occurs once every four years.

the strict inequality asserts that $\frac{e^x-1}{x} > 1$. This can be seen by applying the Mean Value Theorem to the function $f(x) = e^x$, to conclude that there exists some $y$ in the interval $(0,x)$ such that

$$\frac{f(x)-f(0)}{x} = f'(y).$$

The left side equals $\frac{e^x-1}{x}$, while the right side equals $e^y$, which is greater than 1 because $y > 0$. Similarly, when $x < 0$ the strict inequality asserts $\frac{e^x-1}{x} < 1$. (Dividing by $x$ reverses the direction of the inequality because $x$ is negative.) This holds because, again by the Mean Value Theorem, $\frac{e^x-1}{x}$ is equal to $e^y$ for some $y$ in the interval $(x,0)$. As $y$ is strictly negative, $e^y < 1$ and the inequality follows. ■

Now, applying Fact 1.1 to the birthday paradox calculation, we find that

$$\Pr(\text{no collision}) < \prod_{k=1}^{m-1} e^{-k/n} = \exp\left(-\frac{1}{n}\sum_{k=1}^{m-1} k\right) = \exp\left(-\frac{m(m-1)}{2n}\right). \tag{1.2}$$

The right side becomes less than $\frac{1}{2}$ when $\frac{m(m-1)}{2n} > \ln(2)$, or equivalently when

$$\left(m-\tfrac{1}{2}\right)^2 > 2n\ln(2)+\tfrac{1}{4}.$$

**Lemma 1.2** If $m > \sqrt{2n\ln(2)+\frac{1}{4}}+\frac{1}{2}$ and $m$ balls are thrown randomly into $n$ bins then the collision probability is greater than $\frac{1}{2}$.

When $n = 365$, the expression $\sqrt{2n\ln(2)+\frac{1}{4}}+\frac{1}{2}$ evaluates to 22.99994..., a shockingly close approximation to 23. For most applications of the birthday paradox, it is acceptable to overestimate the required value of $m$, using the function $\sqrt{2n}$ rather than the more cumbersome formula specified in the lemma.

You probably noticed that our analysis of the collision probability involved *overestimating* the probability of having no collisions, by approximating each factor $1-\frac{k}{n}$ with the overestimate $e^{-k/n}$. With a little more work, we can quantify the amount of error due to this approximation and verify that the error is very small.

**Fact 1.3** If $0 < x < \frac{1}{2}$ then $e^{-x-x^2} < 1-x$.

*Proof.* By the Taylor series for the natural logarithm,

$$-\ln(1-x) = \sum_{k=1}^{\infty}\frac{x^k}{k} < x+\sum_{k=2}^{\infty}\frac{x^k}{2} = x+\frac{x^2}{2}\sum_{j=0}^{\infty}x^j = x+\frac{x^2}{2}\cdot\frac{1}{1-x} < x+x^2$$

where the last inequality follows from our assumption that $x < \frac{1}{2}$. The inequality $e^{-x-x^2} < 1-x$ now follows by negating both sides and exponentiating. ■

Applying this fact to the birthday paradox, we see that the probability of having no collisions can be bounded *from below* as follows.

$$\Pr(\text{no collision}) > \prod_{k=1}^{m-1} \exp\left(-\frac{k}{n} - \frac{k^2}{n^2}\right) = \exp\left(-\sum_{k=1}^{m-1}\frac{k}{n}\right)\cdot\exp\left(-\sum_{k=1}^{m-1}\frac{k^2}{n^2}\right)$$
$$= \exp\left(-\frac{m(m-1)}{2n}\right)\cdot\exp\left(-\frac{m(m-1)(2m-1)}{6n^2}\right).$$

If we choose $m$ such that $\frac{m(m-1)}{2n} \approx \ln(2)$ then

$$\frac{m(m-1)(2m-1)}{6n^2} \approx \frac{\ln(2)\cdot(2m-1)}{3n} \approx \frac{\ln(2)\cdot 2\cdot\sqrt{2n\ln(2)}}{3n} = \frac{(2\ln 2)^{3/2}\cdot\sqrt{n}}{3n} < \frac{1}{\sqrt{n}}$$
$$\exp\left(-\frac{m(m-1)(2m-1)}{6n^2}\right) > 1 - \frac{m(m-1)(2m-1)}{6n^2} > 1 - \frac{1}{\sqrt{n}}.$$

Hence, the formula $\exp\left(-\frac{m(m-1)}{2n}\right)$ overestimates the probability of no collision, but only by a factor less than $(1-\frac{1}{\sqrt{n}})^{-1}$. With a little more effort, one can use this estimate to show that the minimum number of balls necessary to ensure $\Pr(\text{collision}) \geq \frac{1}{2}$ is very close to the formula given in Lemma 1.2; that formula overestimates the minimum number of balls by at most 2.

### 1.1.1 Application: Insecurity of cryptographic hash functions

A cryptographic hash function is a function $h$ that takes a long string and compresses it to a "message digest" of some fixed length. For example, the SHA-1 hash function that was in use until around 2017 had a 160-bit output. The SHA-2 family that replaces it consists of six hash functions with output lengths ranging from 224 to 512 bits.

To be considered secure, a cryptographic hash function should be *collision-resistant*, meaning that it is computationally infeasible to find two distinct inputs $x, y$ such that $h(x) = h(y)$. A simple method for attempting to "break" a collision-resistant hash function is the Birthday Attack, which consists of evaluating the hash function on uniformly-random inputs until two of them yield the same output. This is analogous to throwing balls into $n = 2^k$ bins, until there is a bin containing two balls. From Lemma 1.2 we know that the Birthday Attack is likely to succeed after roughly $\sqrt{2n} = 2^{(k+1)/2}$ attempts. However, cryptographic hash functions used in practice are not truly random functions, they are only conjectured to be computationally indistinguishable from random functions. Sometimes these conjectures are false: if the hash function has some structure that makes it distinguishable from a truly random function, it may be possible to exploit that structure to find a hash collision much more rapidly than the Birthday Attack. That is exactly what happened to SHA-1 in 2005, when Xiaoyun Wang (and two students) succeeded in finding a collision using about $2^{69}$ attempts, about 2000 times faster than the $2^{80}$ attempts predicted by the birthday paradox. This was later improved to $2^{63}$, which is more than 100,000 times faster than the Birthday Attack. The success of these attacks prompted the transition from SHA-1 to SHA-2.

### 1.1.2 Application: Sample complexity of uniformity testing

Hypothesis testing is a common task in statistics: given $m$ samples, and given a candidate distribution $p$, determine whether or not it is likely that the samples were drawn from

$p$. Uniformity testing is the special case when $p$ is the uniform distribution on the set $[n] = \{1, 2, \ldots, n\}$.

Here's one way to formalize the objective of uniformity testing. We would like to design an algorithm, parameterized by a pair of positive constants $\varepsilon, \delta$, that takes a sequence of $m$ samples drawn from some distribution on the set $[n]$, and it either outputs "uniform" or "not uniform." The algorithm will be considered *probably approximately correct (PAC)* if it satisfies the following two properties.

- If the samples are drawn from the uniform distribution, then with probability at least $1 - \delta$ the algorithm outputs "uniform".
- If the samples are drawn from a distribution $q$ that is $\varepsilon$-far from uniform, in the sense that there is a subset $S \subseteq [n]$ with $q(S) > \frac{|S|}{n} + \varepsilon$, then with probability at least $1 - \delta$ the algorithm outputs "not uniform".

The *sample complexity of* $(\varepsilon, \delta)$*-PAC uniformity testing* refers to the minimum $m = m(n)$ for which such an algorithm exists.

It is possible to use the birthday paradox to prove a simple lower bound on the sample complexity of uniformity testing. To see this, consider the problem of distinguishing between two data generating processes.

1. Samples $x_1, \ldots, x_m$ are drawn from the uniform distribution on $[n]$.
2. A random subset $T \subset [n]$ of size $n/2$ is drawn. Then, samples $x_1, \ldots, x_m$ are drawn from the uniform distribution on $T$.

If $q$ denotes the distribution from which the samples $x_1, \ldots, x_m$ are drawn, then $q$ equals the uniform distribution in Case 1, whereas $q$ is $\varepsilon$-far from the uniform distribution (for any $\varepsilon < \frac{1}{2}$) in Case 2. This is because in Case 2, $q(T) = 1$ whereas $|T|/n = \frac{1}{2}$.

If $m \leq \sqrt{n}$ then the uniformity tester faces an insurmountable dilemma: in both Case 1 and Case 2, its input sequence is probably just a random sequence of $m$ distinct elements of $[n]$. Hence, it has no useful signal for distinguishing Case 1 from Case 2. However, probable approximate correctness requires distinguishing Case 1 from Case 2, since in one case $q$ is uniform whereas in the other case it is $\varepsilon$-far from uniform.

Here's how to make this reasoning precise. Let $\mathcal{E}$ denote the event that the input sequence consists of $m$ distinct elements of $[n]$. Let $\Pr_1(\mathcal{E})$ and $\Pr_2(\mathcal{E})$ denote the probabilities of event $\mathcal{E}$ when the process generating the sequence $x_1, \ldots, x_m$ is as described in Case 1 and Case 2, respectively. In both cases, since the data generating process is invariant under permutations of the set $[n]$, the output distribution conditional on event $\mathcal{E}$ must be the uniform distribution over the set $\Sigma_{n,m}$ of length-$m$ sequences of distinct elements of $[n]$. Let $\sigma$ denote the probability that the uniformity testing algorithm outputs "uniform" when given an input sequence drawn uniformly at random from $\Sigma_{n,m}$. Assuming the uniformity tester is $(\varepsilon, \delta)$-PAC:

$$\begin{aligned}
\delta &\geq \Pr_1(\text{output "not uniform"}) \geq \Pr_1(\text{output "not uniform"} \,|\, \mathcal{E}) \cdot \Pr_1(\mathcal{E}) = (1-\sigma) \cdot \Pr_1(\mathcal{E}) \\
\delta &\geq \Pr_2(\text{output "uniform"}) \geq \Pr_2(\text{output "uniform"} \,|\, \mathcal{E}) \cdot \Pr_2(\mathcal{E}) = \sigma \cdot \Pr_2(\mathcal{E}) \\
2\delta &\geq (1-\sigma)\Pr_1(\mathcal{E}) + \sigma \Pr_2(\mathcal{E}) \geq \Pr_2(\mathcal{E}),
\end{aligned}$$

where the final inequality holds because the data generating process is less likely to generate $m$ distinct samples in Case 2 than in Case 1.

Using our analysis of the birthday paradox, we know that

$$\Pr_2(\mathcal{E}) \geq \exp\left(-\frac{m(m-1)}{2|T|}\right) \cdot \left(1 - \frac{1}{\sqrt{|T|}}\right) = \exp\left(-\frac{m(m-1)}{n}\right) \cdot \left(1 - \sqrt{\tfrac{2}{n}}\right).$$

When $m < \sqrt{n}$ and $n \geq 8$, the right side is at least $\frac{1}{2e}$. Hence, $(\varepsilon, \delta)$-PAC uniformity testing with $m < \sqrt{n}$ samples is not possible for any $\varepsilon < \frac{1}{2}$ and $\delta < \frac{1}{4e}$.

Is $m = \Theta(\sqrt{n})$ the correct sample complexity bound for uniformity testing? It turns out that the answer is yes. One might find this result surprising, because $O(\sqrt{n})$ samples constitute only a *tiny fraction of the support set of the distribution*, when $n$ is large. It is amazing that such a small set of samples contains enough information to reliably distinguish uniform distributions from non-uniform ones. Not surprisingly, the the algorithm for uniformity testing using $O(\sqrt{n})$ samples relies heavily on the birthday paradox. In fact, it works by simply counting collisions: it outputs "uniform" if the number of pairs $i \neq j$ with $x_i = x_j$ falls below a carefully-designed threshold, and "not uniform" otherwise.

## 1.2 The Coupon Collector Problem

Continuing with our analysis of balls and bins, we turn to the question: how many balls must we throw to ensure that, with probability at least $\frac{1}{2}$, every bin contains at least one ball?

### 1.2.1 Reviewing the geometric distribution

To begin solving the coupon collector problem, we must recall the definition and some basic properties of the geometric distribution.

**Definition 1.4** If $X$ is a random variable taking values in the positive integers, with the probability distribution

$$\Pr(X = n) = p \cdot (1-p)^{n-1},$$

then we say $X$ is *geometrically distributed with parameter $p$.*

If one takes a biased coin with $\Pr(\text{heads}) = p$ and tosses it until the first coin-toss that yields heads, the total number of tosses is geometrically distributed with parameter $p$.

**Expectation and variance of a geometric random variable.** It's not too hard to compute the expected value of a geometric random variable using the definition of the expectation:

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} n \cdot \Pr(X = n).$$

However, it's even easier to use the following lemma, which frequently furnishes a very useful method for computing expected values or for bounding them from above or below.

**Lemma 1.5** If $X$ is a random variable taking values in $\mathbb{N}$ then

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} \Pr(X > k).$$

*Proof.* Using the identity $n = \sum_{k=0}^{n-1} 1$, we find that

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} n \cdot \Pr(X = n) = \sum_{n=0}^{\infty} \sum_{k=0}^{n-1} \Pr(X = n) = \sum_{k=0}^{\infty} \sum_{n=k+1}^{\infty} \Pr(X = n) = \sum_{k=0}^{\infty} \Pr(X > k).$$

■

For the geometric distribution, we have

$$\Pr(X > k) = \sum_{n=k+1}^{\infty} \Pr(X = n) = \sum_{n=k+1}^{\infty} p(1-p)^{n-1} = \sum_{n=k+1}^{\infty} (1-p)^{n-1} - (1-p)^n = (1-p)^k$$

since the sum telescopes. Hence,

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} \Pr(X > k) = \sum_{k=0}^{\infty} (1-p)^k = \frac{1}{p}.$$

We can also use Lemma 1.5 to compute the variance of a geometric random variable. The key is to group the values of $k$ into intervals between consecutive squares. If $j$ is an integer and $j^2 \le k < (j+1)^2$ then $\Pr(X^2 > k) = \Pr(X > j) = (1-p)^j$. Consequently, the second moment of the geometric distribution satisfies

$$\begin{aligned}
\mathbb{E}[X^2] = \sum_{k=0}^{\infty} \Pr(X^2 > k) &= \sum_{j=0}^{\infty} \sum_{j^2 \le k < (j+1)^2} \Pr(X^2 > k) \\
&= \sum_{j=0}^{\infty} (2j+1)(1-p)^j \\
&= \sum_{i=1}^{\infty} (2i-1)(1-p)^{i-1} \\
&= \sum_{i=1}^{\infty} \frac{2i-1}{p} \cdot \Pr(X = i) = \mathbb{E}\left[\frac{2X-1}{p}\right] = \frac{2}{p^2} - \frac{1}{p}
\end{aligned}$$

where the last equation follows by linearity of expectation, substituting the formula $\mathbb{E}[X] = \frac{1}{p}$ that we already derived.

Recalling now that $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, we find that

$$\mathrm{Var}(X) = \frac{1}{p^2} - \frac{1}{p} = \frac{1-p}{p^2}.$$

### 1.2.2 Coupon collector: distribution of stopping time

Consider throwing a sequence of balls into independent, uniformly random bins until the load in each bin is positive. Let $\tau$ denote the time when this stopping condition is met. The aim of this section is to compute the distribution of the random variable $\tau$. To do so, it is useful to define random variables $\tau_1 < \tau_2 < \cdots < \tau_n = \tau$ by specifying that $\tau_k$ is the first time when there are $k$ bins each containing at least one ball. Note that $\tau_1$ is deterministically equal to 1. For the remaining random variables in this sequence, we have the following extremely useful observation.

**Lemma 1.6** For $1 \le k < n$, the random variable $Y_k = \tau_{k+1} - \tau_k$ is geometrically distributed with parameter $\frac{n-k}{n}$. These random variables $Y_1, Y_2, \ldots, Y_{n-1}$ are mutually independent.

*Proof.* By definition, $\tau_{k+1}$ is the first time after $\tau_k$ that a ball is thrown into an unoccupied bin. For any $t$ such that $\tau_k \le t < \tau_{k+1}$, the number of occupied bins at time $t$ equals $k$. Hence, the probability that that ball thrown at time $t$ lands in an unoccupied bin is $\frac{n-k}{n}$. It follows that the number of balls thrown after $\tau_k$ until one of them lands in an unoccupied bin — that is, the random variable $Y_k = \tau_{k+1} - \tau_k$ — follows a geometric distribution with parameter $\frac{n-k}{n}$. Since this distribution has no dependence on the history of the balls-and-bins preceding $\tau_k$ or following $\tau_{k+1}$, we may conclude that $Y_1, \ldots, Y_{n-1}$ are mutually independent. ■

If we define $Y_0$ to be a random variable that is deterministic equal to 1, then we have shown that the random stopping time $\tau = \tau_n$ can be represented as a sum of independent random variables $Y_0 + Y_1 + \cdots + Y_{n-1}$, where $Y_k$ is geometrically distributed with parameter $\frac{n-k}{n}$. Consequently,

$$\mathbb{E}[\tau] = \sum_{k=0}^{n-1} \mathbb{E}[Y_k] = \sum_{k=0}^{n-1} \frac{n}{n-k} = n \cdot \left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right) = n \cdot H_n$$

The number of $H_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$ is called the $n^{\text{th}}$ *harmonic number* and lies between $\ln n$ and $1 + \ln n$, by the integral test:

$$\ln n = \int_1^n \frac{dx}{x} < 1 + \frac{1}{2} + \cdots + \frac{1}{n-1} < H_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n} \le 1 + \int_1^n \frac{dx}{x} = 1 + \ln n. \quad (1.3)$$

Finally, for the variance of $\tau$, we have the following bound.

$$\mathrm{Var}(\tau) = \sum_{k=0}^{n-1} \mathrm{Var}(Y_k) = \sum_{k=0}^{n-1} \frac{k/n}{(n-k)^2/n^2} = n \sum_{k=1}^{n-1} \frac{k}{(n-k)^2} = n \sum_{j=1}^{n-1} \frac{n-j}{j^2} < n^2 \sum_{j=1}^{n-1} \frac{1}{j^2} < 2n^2.$$

The last inequality follows from a telescoping sum argument: $\sum_{j=1}^{n-1} \frac{1}{j^2} < \sum_{j=1}^{n-1} \left(\frac{2}{j} - \frac{2}{j+1}\right) = 2 - \frac{2}{n}$.

In the following section we'll see how to combine these estimates of the expectation and variance of $\tau$ to derive asymptotically tight bounds for the coupon collector problem.

### 1.2.3 Coupon collector: Tail bounds on stopping time

Let $m_{\text{coupon}}(n)$ denote the least value of $m$ such that, when $m$ balls are thrown into $n$ bins, with probability at least $\frac{1}{2}$ every bin is occupied by at least one ball. We can relate $m_{\text{coupon}}(n)$ to the sequential balls-and-bins process analyzed in the preceding section: it is the least value of $m$ such that $\Pr(\tau \le m) \ge \frac{1}{2}$.

To find an upper bound for $m_{\text{coupon}}(n)$, the easiest approach is to use Markov's inequality, which asserts that for any non-negative random variable $X$ and any factor $c \ge 1$,

$$\Pr(X \ge c \cdot \mathbb{E}X) \le \frac{1}{c}.$$

Applying this inequality with $X = \tau$ and $c = 2$, we find that

$$\Pr(\tau \geq 2nH_n) \leq \frac{1}{2}$$

and consequently $m_{\text{coupon}}(n) \leq 2nH_n$.

Markov's inequality is quite a weak inequality, so applications of Markov's inequality rarely give tight or nearly-tight bounds on the quantity of interest. This case is no exception: the estimate $m_{\text{coupon}}(n) \leq 2nH_n \approx 2n\ln n$ is off by about a factor of 2.

To obtain a tighter bound, we use Chebyshev's inequality, which is simply Markov's inequality applied to the random variable $Z = (\tau - \mathbb{E}[\tau])^2$. By our calculation of $\text{Var}(\tau)$, we know that $\mathbb{E}[Z] < 2n^2$. Hence,

$$\Pr\left(|\tau - \mathbb{E}[\tau]| \geq 2n\right) \;=\; \Pr(Z \geq 4n^2) \;\leq\; \Pr(Z \geq 2\mathbb{E}[Z]) \;\leq\; \frac{1}{2}. \tag{1.4}$$

A simple consequence of Inequality (1.4) is that

$$\mathbb{E}[\tau] - 2n \leq m_{\text{coupon}}(n) \leq \mathbb{E}[\tau] + 2n$$

which we can rewrite, using $\mathbb{E}[\tau] = nH_n$ and $\ln n < H_n \leq \ln n + 1$, as

$$n(\ln n - 2) \leq m_{\text{coupon}}(n) \leq n(\ln n + 3).$$

This improves, by approximately a factor of 2, our previous bound $m_{\text{coupon}}(n) \leq 2\mathbb{E}[\tau] = 2nH_n$. Importantly, the improved bound is asymptotically tight, i.e. the upper and lower bounds differ by a factor that converges to 1 as $n \to \infty$. To see why, note that $\frac{n(\ln n+3)}{n(\ln n-2)} = 1 + \frac{5}{\ln n-2}$.

## 1.3 Balls and bins: the heavily loaded case

We now turn to investigating the balls-and-bins problem from the standpoint of load balancing. Let us say the bins are $\beta$-balanced if the maximum bin load is no more than $\beta$ times the minimum bin load. In order for the bins to be $\beta$-balanced, the minimum bin load must be strictly positive. In the analysis of the coupon collector problem, we saw that this happens for the first time at $\tau \approx n\ln n$. At the coupon-collector stopping time, the minimum bin load is equal to 1 (by definition of $\tau$) and the maximum bin load is at least $\frac{\tau}{n}$, (by the pigeonhole principle), so the so bins are not $\beta$-balanced for any $\beta < \frac{\tau}{n}$, which is (probably) approximately $\ln n$.

As we throw even more balls into the bins, we expect the bins to become $\beta$-balanced for ever-smaller load factors $\beta$, with $\beta \to 1$ as the number of balls approaches $\infty$. Our objective in this section is to analyze how rapidly the load factor approaches 1. What is the smallest $m = m_\varepsilon(n)$ that ensures $\beta \leq 1 + \varepsilon$ with probability at least $\frac{1}{2}$?

### 1.3.1 The tail-bound-plus-union-bound method

Our strategy will exemplify a very commonly adopted method for analyzing randomized algorithms: the tail-bound-plus-union-bound method. The union bound is the following extremely simple yet useful probabilistic inequality.

**Lemma 1.7** If $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$ is a finite collection of events in a probability space, then

$$\Pr(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \cdots \cup \mathcal{E}_n) \leq \Pr(\mathcal{E}_1) + \Pr(\mathcal{E}_2) + \cdots + \Pr(\mathcal{E}_n). \tag{1.5}$$

Armed with the union bound, we will execute the following strategy for analyzing the load balancing factor of the balls-and-bins process.

1. Let $B_i$ denote the (random) number of balls in bin $i$, after $m$ balls have been thrown.
2. Observe that $\mathbb{E}[B_i] = m/n$.
3. For any $\delta$ such that $0 < \delta < 1$, let $\mathcal{E}_{i,\delta}$ denote the event that $B_i \notin \left[(1-\delta)\frac{m}{n}, (1+\delta)\frac{m}{n}\right]$.
4. *(Tail bound step.)* Choose $m$ large enough, as a function of $n$ and $\delta$, to ensure that
$$\forall i \quad \Pr(\mathcal{E}_{i,\delta}) \leq \tfrac{1}{2n}.$$
5. *(Union bound step.)* By the union bound, $\Pr\left(\bigcup_{i=1}^{n} \mathcal{E}_{i,\delta}\right) \leq \left(\frac{1}{2n}\right) \cdot n = \frac{1}{2}$.
6. The complementary event $\bigcap_{i=1}^{n} \overline{\mathcal{E}_{i,\delta}}$ has probablity at least $\frac{1}{2}$. In other words, with probability at least $\frac{1}{2}$ the load in each bin satisfies $(1-\delta)\frac{m}{n} \leq B_i \leq (1+\delta)\frac{m}{n}$.
7. Now, choose $\delta$ small enough that when $B_i$ is between $(1-\delta)\frac{m}{n}$ and $(1+\delta)\frac{m}{n}$ for each $i$, it implies the bins are $(1+\varepsilon)$-balanced.

The last step is easy to accomplish: as long as $\varepsilon \leq 1$, we can set $\delta = \varepsilon/3$ and observe that

$$(1-\delta)(1+\varepsilon) = 1 + \frac{2\varepsilon}{3} - \frac{\varepsilon^2}{3} \geq 1 + \frac{\varepsilon}{3} = 1 + \delta,$$

which implies

$$\frac{1+\delta}{1-\delta} \leq 1 + \varepsilon.$$

To complete the strategy specified above we need to fill in the details of the tail bound step. First we'll see how to do this using Chebyshev's Inequality. Then we'll see a different tail bound, the Chernoff Bound, that is quantitatively much stronger, leading to a much tighter upper bound on $m_\varepsilon(n)$.

### 1.3.2 Tail bound using Chebyshev's inequality

Our strategy for bounding the probability that $B_i$ lies outside the interval $\left[(1-\delta)\frac{m}{n}, (1+\delta)\frac{m}{n}\right]$ will be to express $B_i$ as a sum of independent random variables. Specifically, let $X_{it} = 1$ if ball $t$ is thrown into bin $i$, and $X_{it} = 0$ otherwise. The variables $X_{it}$ are independent Bernoulli random variables, each with expected value $\frac{1}{n}$ and variance $\frac{1}{n}(1-\frac{1}{n})$. Their sum, $B_i$, has expected value $\frac{m}{n}$ and variance $\frac{m}{n}(1-\frac{1}{n})$, since the variance of a sum of independent random variables is the sum of their variances.

Now, using Chebyshev's inequality,

$$\Pr(\mathcal{E}_{i,\delta}) = \Pr\left(|B_i - \mathbb{E}[B_i]| > \frac{\delta m}{n}\right) \leq \frac{\mathrm{Var}(B_i)}{(\delta m/n)^2} < \frac{m/n}{(\delta m/n)^2} = \frac{n}{\delta^2 m}.$$

To make the right side less than or equal to $\frac{1}{2n}$ we may choose $m \geq 2(n/\delta)^2$. By our choice of $\delta = \varepsilon/3$, this yields the bound

$$m_\varepsilon(n) \leq \frac{18n^2}{\varepsilon^2}.$$

In the next two sections, we'll see how to improve the dependence on $n$ from quadratic to quasi-linear using a stronger tail bound.

### 1.3.3 Introducing the Chernoff Bound

Let $X_1, X_2, \ldots, X_m$ be independent (not necessarily identically distributed) random variables taking values in $[0,1]$. In this section we derive the *Chernoff bound*, which bounds the probability that $X_1 + \cdots + X_m$ differs from its expectation by a factor lying outside the interval $[1-\varepsilon, 1+\varepsilon]$. We will assume throughout this section that $0 < \varepsilon < 1$.

The Chernoff bound is proven using the same strategy as Chebyshev's inequality.

1. Find a useful non-linear function of the random variable $X = X_1 + \cdots + X_m$.
   - In Chebyshev's inequality this function was $f(X) = (X - \mathbb{E}[X])^2$.
   - In the Chernoff bound it will be $f(x) = e^{tX}$ for a carefully-chosen value of $t$.
2. Calculate an upper bound on $\mathbb{E}[f(X)]$.
3. Show that when $X$ is far from $\mathbb{E}[X]$, the value of $f(X)$ exceeds this upper bound by a large factor.
4. Finish up by using Markov's inequality to assert that $f(X)$ is unlikely to exceed $\mathbb{E}[f(X)]$ by a large factor.

**Definition 1.8** If $X$ is a random variable, its *moment generating function* $M_X(t)$ and its *cumulant generating function* $K_X(t)$ are the functions defined by

$$M_X(t) = \mathbb{E}[e^{tX}]$$
$$K_X(t) = \ln M_X(t).$$

The following two lemmas present some useful properties of $M_X(t)$ and $K_X(t)$.

**Lemma 1.9** If $X_1, \ldots, X_m$ are independent random variables then

$$M_X(t) = \prod_{i=1}^{m} M_{X_i}(t), \qquad K_X(t) = \sum_{i=1}^{m} K_{X_i}(t).$$

*Proof.* The formula for $M_X(t)$ follows from the fact that $e^{tX} = \prod_{i=1}^{m} e^{tX_i}$ and that the expectation of the product of independent random variables equals the product of their expectations. *(This is why it's crucial, in the Chernoff bound, that the variables must be independent.)* The formula for $K_X(t)$ follows from the one for $M_X(t)$, using the product rule for logarithms. ■

▪ **Remark 1.10** When $M_X(t)$ and $K_X(t)$ are expanded as power series in $t$,

$$M_X(t) = \sum_{n=0}^{\infty} \frac{m_n(X)}{n!} \cdot t^n, \qquad K_X(t) = \sum_{n=0}^{\infty} \frac{\kappa_n(X)}{n!} \cdot t^n,$$

the power series coefficients $m_n(X)$ and $\kappa_n(X)$ are the expected values of degree-$n$ polynomials in $X$. Both of these sequences of coefficients represent important statistics about the distribution of $X$.

- $m_n(X) = \mathbb{E}[X^n]$ is the $n^{\text{th}}$ moment of $X$.

- $\kappa_n(X)$ is called the $n^{\text{th}}$ *cumulant* of $X$. The first two cumulants are well known statistics: $\kappa_1(X)$ is the expectation and $\kappa_2(X)$ is the variance. All of the cumulants satisfy the *cumulative property for independent sums:*

$$\kappa_n(X) = \kappa_n(X_1) + \kappa_n(X_2) + \cdots + \kappa_n(X_m)$$

when $X = X_1 + \cdots + X_m$ is a sum of independent random variables. This follows from the fact that $K_X(t) = \sum_{i=1}^m K_{X_i}(t)$.

■

**Lemma 1.11** For any random variable $X$ taking values in $[0,1]$, the moment generating function $M_X$ satisfies

$$M_X(t) \leq \exp\left((e^t - 1)\mathbb{E}[X]\right)$$

for all $t \in \mathbb{R}$.

*Proof.* For all $x \in [0,1]$ and all $t \in \mathbb{R}$ the inequality

$$e^{tx} \leq 1 + (e^t - 1)x$$

holds because the left side is a convex function of $x$, the right side is a linear function of $x$, and the left and right sides are equal at the endpoints $x = 0$ and $x = 1$. Applying this inequality along with linearity of expectation, we find that

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] \leq 1 + (e^t - 1)\mathbb{E}[X].$$

The lemma follows by combining this inequality with Fact 1.1. ■

**Theorem 1.12 — Chernoff bound.** If $X_1, X_2, \ldots, X_m$ are independent random variables taking values in $[0,1]$ and $X = X_1 + \cdots + X_m$, then for $0 < \varepsilon \leq 1$ we have

$$\begin{aligned}\Pr(X \geq (1+\varepsilon)\mathbb{E}[X]) &< e^{-\frac{1}{3}\varepsilon^2\mathbb{E}[X]} \\ \Pr(X \leq (1-\varepsilon)\mathbb{E}[X]) &< e^{-\frac{1}{2}\varepsilon^2\mathbb{E}[X]}\end{aligned}$$

*Proof.* Using Lemmas 1.9 and 1.11, together with $\mathbb{E}[X] = \sum_{i=1}^m \mathbb{E}[X_i]$, we find that $M_X(t) \leq \exp\left((e^t - 1)\mathbb{E}[X]\right)$ for all $t \in \mathbb{R}$. Now, from Markov's inequality we have

$$\Pr(X \geq (1+\varepsilon)\mathbb{E}[X]) = \Pr\left(e^{tX} \geq e^{(1+\varepsilon)t\mathbb{E}[X]}\right) < e^{(e^t - 1 - (1+\varepsilon)t)\mathbb{E}[X]}$$

for all $t \geq 0$. To minimize the right side, set $t = \ln(1+\varepsilon)$. Then $e^t - 1 - (1+\varepsilon)t = \varepsilon - (1+\varepsilon)\ln(1+\varepsilon)$. Using the Taylor series

$$(1+\varepsilon)\ln(1+\varepsilon) = (1+\varepsilon)\left(\varepsilon - \tfrac{1}{2}\varepsilon^2 + \tfrac{1}{3}\varepsilon^3 - \cdots\right) = \varepsilon + \tfrac{1}{2}\varepsilon^2 - \tfrac{1}{6}\varepsilon^3 + \cdots > \varepsilon + \tfrac{1}{3}\varepsilon^2$$

we find that $\varepsilon - (1+\varepsilon)\ln(1+\varepsilon) < -\frac{1}{3}\varepsilon^2$ and the upper bound on $\Pr(X \geq (1+\varepsilon)\mathbb{E}[X])$ follows.

For $t \geq 0$ another application of Markov's inequality yields

$$\Pr(X \leq (1-\varepsilon)\mathbb{E}[X]) = \Pr\left(e^{-tX} \geq e^{-(1-\varepsilon)t\mathbb{E}[X]}\right) < e^{(e^{-t} - 1 + (1-\varepsilon)t)\mathbb{E}[X]}.$$

To minimize the right side we set $t = -\ln(1-\varepsilon)$ and then $e^{-t} - 1 + (1-\varepsilon)t = -\varepsilon - (1-\varepsilon)\ln(1-\varepsilon)$. Using the Taylor series

$$-(1-\varepsilon)\ln(1-\varepsilon) = (1-\varepsilon)(\varepsilon + \tfrac{1}{2}\varepsilon^2 + \tfrac{1}{3}\varepsilon^3 + \cdots) = \varepsilon - \tfrac{1}{2}\varepsilon^2 - \tfrac{1}{6}\varepsilon^3 - \cdots < \varepsilon - \tfrac{1}{2}\varepsilon^2$$

we find that $-\varepsilon - (1-\varepsilon)\ln(1-\varepsilon) < -\frac{1}{2}\varepsilon^2$ and the upper bound on $\Pr(X \leq (1+\varepsilon)\mathbb{E}[X])$ follows. ■

A few features of the Chernoff bound are worth noting.

1. Theorem 1.12 bounds the probability of $X$ deviating from $\mathbb{E}[X]$ by a large amount. Inequalities of this type are called *large deviation inequalities* or *tail bounds*, since they quantify the amount of probability in the "tail" of the distribution of $X$.

2. The probability of a large deviation tends to zero *exponentially fast* as $\mathbb{E}[X]$ grows large. Inequalities of this type are called *exponential tail bounds*.

3. The probability of large deviation is exponentially small as a function of $\mathbb{E}[X]$, not as a function of the number of random variables being summed, $m$. Even if $m$ is very large, it's possible that the distribution of $X$ is not very concentrated around its expected value. For example, if $X_1, \ldots, X_{m-1}$ are deterministically equal to 0, and $X_m$ is equal to 0 or 1, each with probability $\frac{1}{2}$, then $X$ is equal to 0 or 1, each with probability $\frac{1}{2}$, so the event that $X$ is between $(1-\varepsilon)\mathbb{E}[X]$ and $(1+\varepsilon)\mathbb{E}[X]$ has probability zero! This is consistent with the Chernoff bound, which only says that $\Pr(X \geq (1+\varepsilon)\mathbb{E}[X])$ is small when $\mathbb{E}[X]$ is large.

4. In the exponential function on the right side of the Chernoff bound, the dependence on $\varepsilon$ is quadratic. This is typical of exponential tail bounds. In order for a deviation such as $X \geq (1+\varepsilon)\mathbb{E}[X]$ to be unlikely, the expected value of $X$ must be greater than $1/\varepsilon^2$ times the maximum value of any individual $X_i$. A useful way of summarizing this observation is, "*To estimate the frequency of an event within a factor of* $1 \pm \varepsilon$*, you must wait until you have observed the event at least* $1/\varepsilon^2$ *times.*"

### 1.3.4 Using the Chernoff Bound to analyze balls and bins

Let us now return to analyzing the load factor in the balls-and-bins process: the ratio of the maximum and minimum bin loads after $m$ balls have been thrown into $n$ bins. We previously saw that, with probability at least $\frac{1}{2}$, the load factor is less than $1+\varepsilon$ once $m \geq \frac{18n^2}{\varepsilon^2}$ balls have been thrown. We will now show that this bound can be significantly improved using the Chernoff bound.

Recall the tail-bound-plus-union-bound strategy from Sections 1.3.1 and 1.3.2. The load in bin $i$ after $m$ balls have been thrown is a random variable $B_i = X_{i1} + \cdots + X_{im}$ where $X_{i1}, \ldots, X_{im}$ are independent Bernoulli random variables. If $m$ is large enough that the event has probability less than $\frac{1}{2n}$, then the union bound implies that with probability at least $\frac{1}{2}$, the ratio of the maximum and minimum bin loads is less than $1+\varepsilon$.

The Chernoff bound constrains the probability that $B_i > \left(1+\frac{\varepsilon}{3}\right)\frac{m}{n}$ or $B_i < \left(1-\frac{\varepsilon}{3}\right)\frac{m}{n}$. The first of these probabilities is bounded above by $\exp\left(-\frac{\varepsilon^2}{27}\cdot\frac{m}{n}\right)$ while the second is bounded above by $\exp\left(-\frac{\varepsilon^2}{18}\cdot\frac{m}{n}\right)$. Hence, the sum of the two probabilities is less than

$2\exp\left(-\frac{\varepsilon^2}{27}\cdot\frac{m}{n}\right)$. We seek a value of $m$ that satisfies the inequality

$$2\exp\left(-\frac{\varepsilon^2}{27}\cdot\frac{m}{n}\right)\leq\frac{1}{2n}.$$

After taking the logarithm of both sides and doing some algebra, we discover that this is equivalent to

$$m>\frac{27n\ln(4n)}{\varepsilon^2}.$$

Compared to the bound of $\frac{18n^2}{\varepsilon^2}$ that we derived using Chebyshev's inequality, the improvement here is that the dependence on $n$ is quasi-linear, i.e. $O(n\log n)$, rather than quadratic. In fact, the quasi-linear dependence on $n$ is the best possible, because we know from our analysis of the coupon collector problem that when $m=o(n\log n)$, it is unlikely that every bin is occupied.

### 1.3.5 The Hoeffding Bound

In this section we derive a different exponential tail bound in which we once again have independent random variables $X_1,\ldots,X_n$, each taking values in a bounded interval, and their sum is denoted by $X$. This time, rather than proving that the ratio $X/\mathbb{E}[X]$ is unlikely to be far from 1, we wish to prove that the absolute difference $|X-\mathbb{E}[X]|$ is unlikely to be far from 0. In other words, whereas the Chernoff bound provides conditions under which $\mathbb{E}[X]$ is likely to be a good multiplicative approximation to $X$, we wish to understand conditions under which $\mathbb{E}[X]$ is likely to be a good additive approximation to $X$. The Hoeffding bound answers this question.

As before, the exponential tail bound will be proven by using generating functions to transform the stated inequality into an application of Markov's inequality. This time, it will be more convenient to work with the cumulant generating function rather than the moment generating function. The following lemma summarizes the implications of Markov's inequality when working with cumulant generating functions.

**Lemma 1.13** Let $X$ be a random variable with cumulant generating function $K_X(t)$, and suppose $\lambda>0$. For any $t>0$,

$$\Pr(X\geq\mathbb{E}[X]+\lambda)\leq e^{K_X(t)-t(\mathbb{E}[X]+\lambda)}$$
$$\Pr(X\leq\mathbb{E}[X]-\lambda)\leq e^{K_X(-t)+t(\mathbb{E}[X]-\lambda)}.$$

*Proof.* To derive the bound on $\Pr(X\geq\mathbb{E}[X]+\lambda)$, observe that the inequality $X\geq\mathbb{E}[X]+\lambda$ holds if and only if $e^{tX}\geq e^{t(\mathbb{E}[X]+\lambda)}$ and apply Markov's inequality. To derive the bound on $\Pr(X\leq\mathbb{E}[X]-\lambda)$, observe that the inequality $X\leq\mathbb{E}[X]-\lambda$ holds if and only if $e^{-tX}\geq e^{-t(\mathbb{E}[X]-\lambda)}$ and again apply Markov's inequality. ■

The key new ingredient in the proof of Hoeffding's Inequality is the following lemma that furnishes an upper bound on the cumulant generating function of a random variable.

**Lemma 1.14 — Hoeffding's Lemma.** If $X$ is a random variable supported on an interval $[a,b]$, with expected value $\mu$, then the cumulant generating function $K_X(t)$ satisfies

$$K_X(t) - \mu t \leq \frac{(b-a)^2 t^2}{8}.$$

*Proof.* The left side is the cumulant generating function of the random variable $X - \mu$, which has expected value zero, so we may replace $X$ with $X - \mu$ if necessary and assume henceforth, without loss of generality, that $\mathbb{E}[X] = 0$. The lemma then asserts the inequality $K_X(t) \leq \frac{1}{8}(b-a)^2 t^2$. To prove this inequality, we will use Taylor's Theorem. We know $K_X(0) = 0$ from the definition of the cumulant generating function, and we know $K_X'(0) = 0$ since the derivative of $K_X$ at 0 is the expectation of $X$. Hence, by Taylor's Theorem with Lagrange's remainder term, $K_X(t) = \frac{1}{2}K_X''(u)t^2$ for some $u$.

To conclude the proof, we need to prove that $K_X''(u) \leq \frac{1}{4}(b-a)^2$ for all $u$, when $X$ is a random variable supported on $[a,b]$. We will prove this bound by constructing a new random variable $Y$ supported on $[a,b]$ whose cumulant generating function $K_Y(t)$ satisfies

$$K_Y(t) = K_X(u+t) - K_X(u)$$

for all $t$. Then, taking the second derivative of both sides with respect to $t$, we will obtain $K_Y''(0) = K_X''(u)$. Recalling that $K_Y''(0)$ is equal to the variance of $Y$, we will be left with showing that the variance of any random variable supported on $[a,b]$ is less than or equal to $\frac{1}{4}(b-a)^2$. It will turn out that this inequality is quite easy to prove.

Let $Y$ be a random variable obtained from $X$ by "reweighting the probability of each support point $z$ by the factor $e^{uz}$." If $X$ has probability density function $f_X(z)$ this means that $Y$ has probability density function $f_Y(z) = \frac{1}{Z}e^{uz}f_X(z)$, where the normalization factor $Z = \int_{-\infty}^{\infty} e^{uz}f_X(z)$ is chosen so that the equation $\int_{-\infty}^{\infty} f_Y(z)\,dz = 1$ holds, as required for a probability density function. More generally, i.e. whether or not $X$ has a probability density function, as long as $\mathbb{E}[e^{uX}] < \infty$ we can define $Y$ by specifying its cumulative distribution function:

$$\Pr(Y \leq y) = \frac{\mathbb{E}[e^{uX} \cdot \mathbf{1}_{X \leq y}]}{\mathbb{E}[e^{uX}]}.$$

Then, we can compute $M_Y(t)$ for a given $t$ by making use of the following identity which is valid for any non-negative random variable $Z$ and whose proof is basically the same as the proof of Lemma 1.5, substituting integrals in place of sums.

$$\mathbb{E}[Z] = \int_0^{\infty} \Pr(Z > z)\,dz. \tag{1.6}$$

Setting $Z = e^{tY}$ and using the substitution $z = e^{tw}$, we find that

$$\begin{aligned} M_Y(t) = \mathbb{E}[e^{tY}] = \mathbb{E}[Z] &= \int_0^\infty \Pr(Z > z)\,dz \\ &= \int_{-\infty}^\infty \Pr(e^{tY} > e^{tw}) \cdot te^{tw}\,dw \\ &= \int_{-\infty}^\infty \Pr(Y > w) \cdot te^{tw}\,dw \\ &= \int_{-\infty}^\infty \frac{\mathbb{E}[e^{uX} \cdot \mathbf{1}_{X>w}]}{\mathbb{E}[e^{uX}]} \cdot te^{tw}\,dw \\ &= \frac{1}{M_X(u)} \int_{-\infty}^\infty \mathbb{E}[e^{uX} \cdot \mathbf{1}_{X>w}] \cdot te^{tw}\,dw \\ &= \frac{1}{M_X(u)} \mathbb{E}_X\left[e^{uX} \int_{-\infty}^X te^{tw}\,dw\right] = \frac{1}{M_X(u)} \mathbb{E}_X\left[e^{uX} e^{tX}\right] = \frac{M_X(u+t)}{M_X(u)} \end{aligned}$$

The identity $K_Y(y) = K_X(u+t) - K_X(u)$ follows by taking the logarithm of both sides.

As observed earlier, to conclude the proof of the lemma we need only show that a random variable $Y$ supported on the interval $[a,b]$ has variance at most $\frac{1}{4}(b-a)^2$. The validity of the inequality $\mathrm{Var}(Y) \le \frac{1}{4}(b-a)^2$ is unaffected if we apply an affine transformation to $Y$ and we apply the same affine transformation to the interval $[a,b]$. In other words, if we replace $Y$ with $cY+d$ and we replace $[a,b]$ with $[ca+d, cb+d]$, the validity of the inequality is unaffected because the variance of $Y$ is scaled by $c^2$, and the squared-length of the interval is also scaled by $c^2$. Hence, without loss of generality (applying an affine transformation to $Y$ and to $[a,b]$ if necessary) we can assume $[a,b] = [-1,1]$ and $\frac{1}{4}(b-a)^2 = 1$. The variance of $Y$ is $\mathbb{E}[Y^2] - \mathbb{E}[Y]^2$. The first term on the right side is clearly no greater than 1 because $Y$ is supported on $[-1,1]$. Since $\mathbb{E}[Y]^2 \ge 0$, it follows that $\mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \le 1$, as desired. ■

**Theorem 1.15 — Hoeffding's Inequality.** Suppose $X_1, X_2, \ldots, X_n$ are independent random variables and that for each $i$, the support of $X_i$ is contained in a bounded interval $[a_i, b_i]$. Let $X = X_1 + \cdots + X_n$. For any $\lambda > 0$,

$$\Pr(X \ge \mathbb{E}[X] + \lambda) \le \exp\left(-\frac{2\lambda^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$
$$\Pr(X \le \mathbb{E}[X] - \lambda) \le \exp\left(-\frac{2\lambda^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

*Proof.* Let $\mu_i = \mathbb{E}[X_i]$ for each $i$, and let $\mu = \sum_{i=1}^n \mu_i = \mathbb{E}[X]$. By Hoeffding's Lemma, $K_{X_i}(t) - \mu_i t \le \frac{1}{8}(b_i - a_i)^2 t^2$ for all $t$ and all $i$. Summing over $i$,

$$K_X(t) - \mu t \le \frac{1}{8} \sum_{i=1}^n (b_i - a_i)^2 t^2.$$

Let $c = \frac{1}{8}\sum_{i=1}^n (b_i - a_i)^2$. By Lemma 1.13,

$$\Pr(X \ge \mathbb{E}[X] + \lambda) \le e^{K_X(t) - \mu t - \lambda t} \le e^{ct^2 - \lambda t}.$$

The proof concludes by setting $t = \lambda/(2c)$, so that $ct^2 - \lambda t = -\frac{\lambda^2}{4c} = -\frac{2\lambda^2}{\sum_{i=1}^n (b_i - a_i)^2}$. ■

## 1.4 Further applications of the Chernoff and Hoeffding bounds

The Chernoff and Hoeffding bounds are some of the most versatile tools in the analysis of randomized algorithms and the average-case analysis of algorithms. In this section we will present a number of applications of both.

### 1.4.1 Estimating the expected value of a distribution

Suppose $Y$ is a random variable taking values in an interval $[0,M]$ whose expected value we wish to estimate. Let $Y_1, Y_2, \ldots$ be a sequence of independent random variables, each having the same distribution as $Y$. One way to estimate $\mathbb{E}[Y]$ is to simply take the unweighted average of the first $N$ samples,

$$\hat{Y} = \frac{1}{N}(Y_1 + \cdots + Y_N).$$

We wish to determine a value of $N$ such that the error of the estimate is very unlikely to exceed $\varepsilon$:

$$\Pr(|\hat{Y} - \mathbb{E}[Y]| > \varepsilon) < \delta.$$

This type of guarantee is summarized by saying that the estimator $\hat{Y}$ is "probability approximately correct," often abbreviated as PAC.

By Hoeffding's Inequality,

$$\Pr(|\hat{Y} - \mathbb{E}[Y]| > \varepsilon) = \Pr(|Y_1 + \ldots + Y_N - N\mathbb{E}[Y]| > N\varepsilon) \leq 2\exp\left(-\frac{2N^2\varepsilon^2}{NM^2}\right) = 2\exp\left(-\frac{2N\varepsilon^2}{M^2}\right).$$

To make this less than $\delta$, we require

$$\begin{aligned}
\exp\left(-\frac{2N\varepsilon^2}{M^2}\right) &< \frac{\delta}{2} \\
\exp\left(\frac{2N\varepsilon^2}{M^2}\right) &> \frac{2}{\delta} \\
\frac{2N\varepsilon^2}{M^2} &> \ln\left(\frac{2}{\delta}\right) \\
N &> \frac{M^2}{2\varepsilon^2}\ln\left(\frac{2}{\delta}\right).
\end{aligned}$$

This sample complexity bound has several features that are typical for estimation procedures that use independent, identically distributed samples to estimate a scalar quantity. The number of samples required depends inverse-quadratically on the tolerable level of "relative error;" in this example the tolerable relative error is $\varepsilon/M$ because we are trying to estimate a quantity belonging to an interval of length $M$, and we tolerate additive error up to $\varepsilon$. On the other hand, the number of samples depends on logarithmically on the inverse of the "confidence parameter," $\delta$, which governs the maximum failure probability that is deemed tolerable.

### 1.4.2 Generalization error of empirical risk minimization

We now show how the Hoeffding bound can be applied to the important problem of *generalization error* in machine learning. To keep the analysis as simple as possible, we will focus on the task of *hypothesis selection*, where there is a finite set of hypotheses and the learner aims to use a set of training data to choose a hypothesis that generalizes to unseen data.

We can model the hypothesis selection problem as follows. We have:

- a random variable $Z$ taking values in a set $\mathcal{Z}$;
- a finite set of hypotheses, $\mathcal{H} = \{h_1, \ldots, h_m\}$;
- independent random variables $Z_1, Z_2, \ldots, Z_N$, each identically distributed to $Z$, collectively called the *training set*.
- a loss function $L : \mathcal{H} \times \mathcal{Z} \to [0,1]$. The value $L(h,z)$ indicates how poorly hypothesis $h$ fits data point $z$.

We assume that the learner is given the training set $\{Z_1, \ldots, Z_N\}$ but does not know the distribution from which the $Z_i$'s were sampled. There are two important ways of evaluating a hypothesis $h$.

1. The *population loss* is $\overline{L}(h) = \mathbb{E}[L(h,Z)]$. This measures how well hypothesis $h$ performs on *the actual distribution* from which the data is sampled, including data points that were not present in the training data.
2. The *empirical loss* is $\frac{1}{N}\sum_{i=1}^{N} L(h,Z_i)$. This has the advantage that it can be computed from the training data, unlike the population loss which can only be computed if one knows the data distribution.

*Empirical risk minimization* is the algorithm that selects the hypothesis $h_{ERM}$ that minimizes empirical loss on the training set. The hope is that if the training set is a representative sample of the data distribution, then $h_{ERM}$ will also perform near-optimally when evaluated according to population loss, even though it was selected to minimize empirical loss rather than population loss.

**Theorem 1.16** Let $h_*$ denote the element of $\mathcal{H}$ that minimizes population loss. For any $0 < \varepsilon, \delta < 1$, if the number of data samples, $N$, satisfies $N > \frac{2}{\varepsilon^2}\ln(2m/\delta)$ then with probability at least $1-\delta$, $\overline{L}(h_{ERM}) \leq \varepsilon + \overline{L}(h_*)$.

*Proof.* Let $\phi(Z_1, \ldots, Z_N)$ denote the Boolean predicate:

$$\forall h \in \mathcal{H} \quad \left| \overline{L}(h) - \frac{1}{N}\sum_{i=1} L(h,Z_i) \right| < \frac{\varepsilon}{2}.$$

When $Z_1, \ldots, Z_n$ satisfy property $\phi$, it implies that $\overline{L}(h_{ERM}) \leq \varepsilon + \overline{L}(h_*)$ because

$$\overline{L}(h_{ERM}) \leq \frac{\varepsilon}{2} + \frac{1}{N}\sum_{i=1} L(h_{ERM}, Z_i) \leq \frac{\varepsilon}{2} + \frac{1}{N}\sum_{i=1} L(h_*, Z_i) \leq \varepsilon + \overline{L}(h_*).$$

The first and third inequalities are applications of property $\phi$, the second inequality follows from the definition of $h_{ERM}$.

To complete the proof we just need to show that $\Pr(\phi(Z_1,\ldots,Z_n)) \geq 1-\delta$. For $j = 1,2,\ldots,m$ let $\psi_j(Z_1,\ldots,Z_n)$ denote the predicate $\left|\overline{L}(h_j) - \frac{1}{N}\sum_{i=1}^{N} L(h_j,Z_i)\right| \geq \frac{\varepsilon}{2}$. The random variables $X_i = \frac{1}{N}L(h_j,Z_i)$ take values in $[0,\frac{1}{N}]$ and the expectation of their sum is $\overline{L}(h_j)$, so applying Hoeffding's inequality with $\lambda = \varepsilon/2$ yields

$$\Pr(\psi_j(Z_1,\ldots,Z_n)) \leq 2\exp\left(-\frac{\varepsilon^2 N}{2}\right) \leq \frac{\delta}{m},$$

by our assumption that $N > \frac{2}{\varepsilon^2}\ln(2m/\delta)$. The Union Bound (Lemma 1.7) implies that

$$\Pr(\vee_{j=1}^{m}\psi_j(Z_1,\ldots,Z_n)) \leq \sum_{j=1}^{m}\Pr(\psi_j(Z_1,\ldots,Z_n)) \leq \delta.$$

Since $\vee_{j=1}^{m}\psi_j(Z_1,\ldots,Z_n)$ is the negation of $\phi(Z_1,\ldots,Z_n)$, it follows that $\Pr(\phi(Z_1,\ldots,Z_n)) \geq 1-\delta$ as claimed. ■

### 1.4.3 Reducing error rate of randomized algorithms

Our last application of the Chernoff bound comes from the theory of randomized algorithms for decision problems. A decision problem is a problem whose output is an element of $\{0,1\}$, with 0 representing "no" and 1 representing "yes." A decision problem belongs to the complexity class P if there is a deterministic polynomial-time algorithm — i.e., an algorithm running in time $O(n^c)$ where $n$ is the input size (in bits) and $c$ is a constant — that answers the decision problem correctly on every possible input. The complexity class BPP consists of decision problems $\Pi$ having a randomized polynomial-time algorithm $A$ that satisfies the following guarantee, where $x$ denotes the problem input and $r$ denotes the random string used by $A$.

$$\forall x \;\; \Pr(A(x,r) \neq \Pi(x)) \leq \frac{1}{3}. \tag{1.7}$$

The random string $r$ is assumed to be a uniformly random binary string whose length, $L(n)$, is bounded by a polnomial function of the input size, $n$. Property (1.7) is often stated equivalently as follows: if $\Pi(x) = 1$ then $\Pr(A(x,r) = 1) \geq \frac{2}{3}$, while if $\Pi(x) = 0$ then $\Pr(A(x,r) = 1) \leq \frac{1}{3}$.

The error rate of a BPP algorithm can be reduced by running it repeatedly using independent random strings, and taking a majority vote of the outcomes. The following algorithm uses a random string $R = r_1 : r_2 : r_3 : \cdots : r_m$ of length $m \cdot L(n)$, for some specified $m \in \mathbb{N}$.

**Algorithm 1** Algorithm $B_m(x,R)$

1: Let $n$ denote the number of bits in $x$.
2: Break $R$ into strings $r_1, r_2, \ldots, r_m$, each of length $L(n)$.
3: Let $a = \frac{1}{m}\sum_{i=1}^{m} A(x,r_i)$.
4: If $a \geq \frac{1}{2}$, output 1. Else, output 0.

**Lemma 1.17** If randomized algorithm $A(x,r)$ satisfies $\Pr(A(x,r) \neq \Pi(x)) \leq \frac{1}{3}$ for all $x$, then for any $\delta > 0$, randomized algorithm $B_m(x,R)$ with $m > 18\ln(1/\delta)$ satisfies $\Pr(B_m(x,R) \neq \Pi(x)) \leq \delta$ for all $x$.

*Proof.* Since algorithm $A$ satisfies the BPP property (1.7), when $\Pi(x) = 1$ we have $\mathbb{E}[A(x,r_i)] \geq \frac{2}{3}$ and when $\Pi(x) = 0$ we have $\mathbb{E}[A(x,r_i)] \leq \frac{1}{3}$. If $B_m(x,R) \neq \Pi(x)$ then either $\Pi(x) = 0$ and $\sum_{i=1}^m A(x,r_i) \geq \frac{m}{2}$, or $\Pi(x) = 1$ and $\sum_{i=1}^m A(x,r_i) < \frac{m}{2}$. In the former case, $\sum_{i=1}^m A(x,r_i)$ exceeds its expected value by at least $\frac{m}{6}$, while in the latter case it falls short of its expected value by at least the same amount. In both cases, Hoeffding's Inequality ensures that the probability of this occurring is no greater than

$$e^{-2(m/6)^2/m} = e^{-m/18} < e^{\ln(\delta)} = \delta.$$

■

Lemma 1.17 has the following consequence for complexity theory. A decision problem $\Pi$ is said to belong to the complexity class P/poly if there is a family of deterministic algorithms $\{B_n \mid n \in \mathbb{N}\}$ such that:

1. for every input $x$ of size $n$, $B_n(x) = \Pi(x)$;
2. for some constant $c < \infty$ and every $n \in \mathbb{N}$, the worst-case running time of $B_n$ on inputs of size $n$ is bounded by $O(n^c)$.

This is summarized by saying that the decision problem $\Pi$ has a *non-uniform family of polynomial-time algorithms*: it can be solved deterministically in polynomial time for all input sizes, but the choice of algorithm depends on the input size.

**Theorem 1.18** If $\Pi$ is a decision problem in BPP then $\Pi$ belongs to P/poly.

*Proof.* Let $A$ be a randomized polynomial-time algorithm for $\Pi$ that satisfies property (1.7). For any $n \in \mathbb{N}$ let $m = \lceil 18\ln(2)\cdot n \rceil = \lceil 18\ln(2^n) \rceil$ and consider the randomized algorithm $B_m$. According to Lemma 1.17, for all $x \in \{0,1\}^n$, $\Pr(B_m(x,R) \neq \Pi(x)) < 2^{-n}$. By the union bound, $\Pr(\exists x \in \{0,1\}^n\ B_m(x,R) \neq \Pi(x)) < 1$. Hence, it is not the case that for all $R \in \{0,1\}^{m\cdot L(n)}$, there exists an $x \in \{0,1\}^n$ such that $B_m(x,R) \neq \Pi(x)$. In other words, there exists some $R_n \in \{0,1\}^{m\cdot L(n)}$ such that for all $x \in \{0,1\}^n$, $B_m(x,R_n) = \Pi(x)$. Let $B_n$ be the algorithm that on input $x$, computes $B_m(x,R_n)$. Then the family $\{B_n \mid n \in \mathbb{N}\}$ constitutes a non-uniform family of polynomial-time algorithms for $\Pi$. ■

A very natural and worthy goal is to eliminate the non-uniformity in Theorem 1.18 and prove that BPP = P. This would show that giving algorithms access to random bits does not affect the set of decision problems that can be solved in polynomial time, or equivalently, that every polynomial-time randomized algorithm for a decision problem can be efficiently "derandomized" to yield a deterministic polynomial-time algorithm for the same problem. Most complexity theorists believe such a derandomization of BPP is possible. The effort to derandomize BPP and other complexity classes is currently one of the most active research areas in complexity theory.

# 2. Hashing

In Section 1.3 we saw that throwing $m$ balls at random into $n$ bins is an excellent way to balance load. However, in CS the metaphorical "balls" — often pieces of data or tasks — may have *identifiers* or *keys*, and the event of throwing a ball with identifier $x$ into bin $b$ needs to be *reproducible*. In other words, the bin number $b$ must be bound to the identifier $x$ in such a way that the system can later fulfill a request to retrieve the ball with identifier $x$, or to remove it from its bin, or to throw another ball with the same identifier into the same bin.

The word for this type of "random but reproducible" mapping of balls to bins is *hashing*. In this section we will introduce an extremely important application of hashing in CS, namely dictionary data structures. In Section 2.1 we'll begin by defining the dictionary as an abstract data type. We'll describe a simple deterministic implementation of a dictionary and analyze its space and time efficiency. It turns out that all of the most efficient known implementations of dictionaries use randomization in the form of *hash functions*. In Section 2.2 we'll introduce the abstraction of hash functions. Then in Section 2.3 we'll introduce the hash table, a randomized data structure that uses hash functions to implement a dictionary. We'll first analyze the hash table's efficiency under an unrealistically strong assumption about the randomness of the hash function. Then in Section 2.4 we'll see how to substitute a weaker randomness assumption called *pairwise independence* without hurting the expected efficiency of the hash table. Furthermore, we'll present some efficient constructions of pairwise independent families of hash functions.

## 2.1 Introducing the dictionary data structure

The quintessential application of this capability is the *dictionary* abstract data type, also known as an associative array or key-value store. A data structure that implements a dictionary stores a set of key-value pairs, with at most one value per key. We will use $\mathcal{X}$

to denote the set of potential keys, and $x \in \mathcal{X}$ to denote one such key. Similarly we will use $\mathcal{V}$ to denote the set of potential values, and $v$ to denote one such value. The dictionary supports (at least) the following three operations.

1. LOOKUP$(x)$ returns the value $v$ stored with key $x$, or it returns a special "not found" symbol, $\perp$, if $x$ is not in the dictionary.
2. INSERT$(x,v)$ inserts the pair $(x,v)$ into the dictionary. If another pair $(x,v')$ was already stored in the dictionary, it is overwritten. (The value $v'$ is replaced with $v$.)
3. DELETE$(x)$ removes the (unique) pair $(x,v)$ with key $x$ from the dictionary, if any such pair exists.

Data structures implementing this functionality in common programming languages include the Python dictionary, the Java HashMap, and the C++ unordered_map. These implementations often provide other functionality (e.g., iterators) but in these notes we will focus on the three basic operations listed above, which are the three essential operations a dictionary must support.

A reasonably efficient deterministic implementation of the dictionary data type uses a balanced binary search tree, such as a red-black tree, to store the keys and pointers to the values. Then, all three of the dictionary operations can be implemented to run in $O(\log_2 m)$ time. As for the space efficiency of this dictionary implementation, if we assume each key and value occupies $O(1)$ space, then the storage required for this data structure is $O(m)$. More generally, if $S$ is the total amount of storage occupied by all the keys and values (which might asymptotically exceed $m$, for example if the values stored in the dictionary are large objects that occupy a super-constant amount of space) then the balanced binary search tree, and the accompanying storage space for storing the values that the tree nodes point to, will occupy $O(S)$ space.

To sum up, the balanced binary search tree is already quite a space-efficient and time-efficient implementation of a dictionary: its space usage is within a constant factor of optimal, and the time complexity of each operation is $O(\log m)$. However, dictionaries are so widely used in computing, and their efficiency is so highly prized, that the running time of $O(\log m)$ per operation is considered slow. We will see that there is a randomized implementation with $O(1)$ running time per operation.

## 2.2 Hash functions

Hash tables are a randomized implementation of dictionaries in which the LOOKUP, INSERT, and DELETEoperations are faster than in the balanced binary search tree. They speed up the (expected) running time of these operations from $O(\log m)$ to $O(1)$ by "flattening" the data structure: instead of searching for keys stored in a hierarchical structure of depth $\log_2(m)$, the keys are stored in a one-dimensional array of $n = O(m)$ *hash buckets*, and a *hash function* applied to any key directs the user to the appropriate bucket.

A *hash function* needs to support two operations, generally with the aid of a source of random bits.

- INITHASH$(\mathcal{X},\mathcal{B})$ initializes a hash function with $\mathcal{X}$ as the set of potential keys and $\mathcal{B}$ as the set of bins. The operation is called only once, when the hash function is initialized.
- HASH$(x)$ evaluates the hash function on key $x \in \mathcal{X}$, returning bin $b \in \mathcal{B}$.

Implementations of hash functions strive for an unattainable goal of achieving three properties simultaneously.

**Uniform randomness** If we call INITHASH$(\mathcal{X},\mathcal{B})$ followed by HASH$(x_1)$, HASH$(x_2)$, ..., HASH$(x_m)$ for any distinct keys $x_1, x_2, \ldots, x_m \in \mathcal{X}$, the values returned are independent, uniformly random elements of $\mathcal{B}$.

**Reproducibility** For any given $x \in \mathcal{X}$, all calls to HASH$(x)$ must return the same value.

**Space and time efficiency** The hash function should be stored using a small amount of space, and evaluating HASH$(x)$ should be fast. If $N = |\mathcal{X}|$, $n = |\mathcal{B}|$, and locations in memory can store $\log(n)$ bits, then the space required to store the representation of the hash function should ideally be $O(\log_2 N)$ and the time required to evaluate HASH$(x)$ should ideally be $O(\log_n N)$.

It's easy to achieve any two of these properties while sacrificing the third one. For example, the second and third properties are satisfied by any deterministic function that is easy to store and evaluate, for example a constant function that always outputs 1. The first and third properties are satisfied if we implement HASH$(x)$ by calling a random number generator to draw a uniformly random sample from $\mathcal{B}$ every time HASH is called, but of course this violates the second property (reproducibility). The first and second properties are achieved by implementing INITHASH$(\mathcal{X},\mathcal{B})$ to sample a uniformly random function $h : \mathcal{X} \to \mathcal{B}$ and store its values in a giant array of size $N$, but this is definitely not space-efficient, and the time-efficiency of evaluating HASH$(x)$ also suffers if $N$ is large enough that an array of size $N$ doesn't fit in RAM.

However, it is impossible to implement a hash function that achieves all three properties simultaneously. To see why, consider initializing a hash function with key set $\mathcal{X}$ and bucket set $\mathcal{B}$ and then calling

$$\text{HASH}(x_1), \text{HASH}(x_2), \ldots, \text{HASH}(x_m), \text{HASH}(x_1), \text{HASH}(x_2), \ldots, \text{HASH}(x_m)$$

where the keys $x_1, x_2, \ldots, x_m$ are any $m$ distinct elements of $\mathcal{X}$. The implementation of the hash function has an internal state, and we will use $s$ to represent the state at the end of the $m^{\text{th}}$ call to HASH. For any $\mathrm{b} = (b_1, b_2, \ldots, b_m) \in [n]^m$ let $S(b)$ denote the set of internal states that may potentially be reached in an execution of the above sequence of HASH operations when the first $m$ HASH operations have outputs $b_1, \ldots, b_m$ respectively. We now make the following observations.

1. By the uniform randomness property, every $\mathrm{b} \in [n]^m$ has a positive probability of being realized as the outputs of the first $m$ HASH operations. Hence, $S(\mathrm{b})$ is non-empty for every $\mathrm{b} \in [n]^m$.
2. If $\mathrm{b} \neq \mathrm{b}'$ then $S(b)$ and $S(b')$ must be disjoint. To see why, for any internal state $s$ and consider the outcome of calling HASH$(x_1)$, HASH$(x_2)$, ..., HASH$(x_m)$ starting from internal state $s$. By the reproducibility property, if $s \in S(\mathrm{b})$ then HASH$(x_1)$, ..., HASH$(x_m)$ must return b, whereas if $s \in S(\mathrm{b}')$ then HASH$(x_1)$, ..., HASH$(x_m)$ must return $\mathrm{b}'$. Since $\mathrm{b} \neq \mathrm{b}'$, $s$ cannot belong to both $S(\mathrm{b})$ and $S(\mathrm{b}')$, so the two sets are disjoint as claimed.
3. The set $\bigcup_{\mathrm{b} \in [n]^m} S(\mathrm{b})$ is a union of $n^m$ disjoint non-empty sets, so its cardinality is at least $n^m$.
4. To encode each internal state using a distinct string of bits, we must use at least $\log_2(n^m) = m \log_2(n)$ bits to encode the internal state.

5. A single memory location holds at most $\log_2(n)$ bits. Hence, storing the internal state requires at least $m$ memory locations. When $m \gg \log(N)$ this violates space-efficiency.

## 2.3 Hash tables

In applications of hashing, reproducibility is usually treated as a hard constraint, and space and time efficiency are strongly desired. (The more efficient, the better.) On the other hand, the property of uniform randomness is often stronger than what's really needed. A useful paradigm for designing algorithms that make use of hashing is to start by fantasizing that all three of the properties we desire in a hash function can be fulfilled. Under this unrealistic assumption, we design and analyze an algorithm. Then, we scrutinize the analysis of the algorithm to identify a weakening of the uniform randomness property that suffices for the analysis. Lastly, we try to design a hash function implementation that satisfies the weaker property while maintaining space and time efficiency.

An important and illustrative case study is the *hash table with chain hashing*, which implements a dictionary using a randomized hash function to map keys to buckets. Suppose that, when initializing the dictionary, we are given the space of potential keys, $\mathcal{X}$, the space of potential values, $\mathcal{V}$, and an upper bound on the maximum number of key-value pairs[1] that will ever be stored in the dictionary at once, $m$. Given these parameters, the hash table uses a set $\mathcal{B} = [n]$ of $n = O(m)$ hash buckets, each with an associated linked list where key-value pairs are stored. A hash function is used to map keys to bins. The hash table operations are implemented as follows.

1. To initialize the hash table, one calls INITHASH$(\mathcal{X}, \mathcal{B})$, allocates an array of size $n$ for the bins, and populates each cell of the array with a pointers to an empty linked list.
2. LOOKUP$(x)$ calls HASH$(x)$ to obtain a bucket $b$, then searches the linked list stored in bucket $b$ to see if a pair $(x, v)$ is found.
3. INSERT$(x, v)$ first calls LOOKUP$(x)$. If a pair $(x, v')$ is found, the value stored is modified from $v'$ to $v$. If $x$ is not found in the hash table, the pair $(x, v)$ is appended to the linked list stored in bucket HASH$(x)$.
4. DELETE$(x)$ scans the linked list stored in bucket HASH$(x)$ and, if a pair with key $x$ is found, deletes that pair from the linked list.

On an architecture where keys and values can be stored in constant space, the space required for the hash table is $O(m+n)$. The time required for LOOKUP, INSERT, and DELETEoperations is linear in the length of the linked list stored in the bucket HASH$(x)$, where $x$ is the key associated with the operation.

Suppose for a moment that the hash function is a uniformly random function $h : \mathcal{X} \to \mathcal{B}$. Then, we could bound the expected length of the linked list stored in the bucket

[1] For the sake of convenience, we are assuming that an upper bound on $m$ is known at initialization time. If this assumption is not satisfied, there are "doubling tricks" that involve making an initial guess that $m = O(1)$, and then every time the number of key-value pairs stored in the hash table exceeds the current guess, the guess is doubled and the hash table is resized accordingly. One then needs to analyze the time complexity of these resizing operations. It is not hard to show that the amortized cost of resizing the hash table is bounded by the total cost of the insertion operations. We omit the amortized analysis from these notes, but it can found in textbooks on data structures.

$b = \text{HASH}(x)$ as follows. If $x$ itself is already stored in the hash table, then $b$ definitely contains one element. There are at most $m-1$ other elements $y \neq x$ stored in the hash table, and each of them has probability $1/n$ of being stored in bucket $b$, so the expected length of the linked list in that bucket is $1+(m-1)/n$. If $x$ is not stored in the hash table, then there are at most $m$ elements stored in it, each of them has $1/n$ probability of belonging to bucket $b$, and hence the expected length of the linked list is $m/n$. The ratio $m/n$ is called the *load factor* of the hash table. Our analysis has shown that *if the hash function is uniformly random*, the expected running time of hash table operations is $O(1+m/n)$. Typically one chooses the parameters of the hash table to make the load factor a constant less than 1, resulting in $O(1)$ running time for lookup, insertion, and deletion.

It would appear that this entire analysis rests on the assumption that the hash function is uniformly random, an assumption which unfortunately is incompatible with space and time efficiency of the hash function operations. Fortunately, upon closer examination, our analysis only made use of the randomness of $h$ in one step, when bounding the expected number of elements other than $x$ that occupy bucket $b = \text{HASH}(x)$. By linearity of expectation, this quantity can be calculated as

$$\sum_{b \in \mathcal{B}} \sum_{y \neq x} \Pr(\text{HASH}(x) = \text{HASH}(y) = b).$$

The event $\text{HASH}(x) = \text{HASH}(y) = b$ is called a *hash collision* of keys $x$ and $y$ at bucket $b$. For any specific $x$, $y$, and $b$, the probability of the event $\text{HASH}(x) = \text{HASH}(y) = b$ is $1/n^2$. The double-sum above has at most $mn$ terms, each equal to $1/n^2$, so the expected number of $y \neq x$ occupying bucket $b = \text{HASH}(x)$ is bounded above by $(mn)/n^2 = m/n$, the load factor of the hash table.

## 2.4 Pairwise independence

Reviewing the analysis of the hash table with chain hashing, we see that the hash function need not be uniformly random, it only needs to satsfy the much weaker property that for all $b \in \mathcal{B}$ and all $x, y \in \mathcal{X}$ such that $x \neq y$, we have $\Pr(\text{HASH}(x) = \text{HASH}(y) = b) = 1/n^2$.

**Definition 2.1** A *2-universal hash family* is a set of functions $\mathcal{H}$, with each $h \in \mathcal{H}$ being a function from $\mathcal{X}$ to $\mathcal{B}$, such that when $h$ is sampled uniformly at random from $\mathcal{H}$, for every distinct $x, y \in \mathcal{X}$ the ordered pair of values $(h(x), h(y))$ is uniformly distributed over $\mathcal{B}^2$.

A 2-universal hash family is sometimes also called a *pairwise independent hash family*, though that term is a bit of a misnomer because the definition requires not only that pair of the hash values $h(x)$ and $h(y)$ are independent, but also that each of them is uniformly distributed.

Since the analysis of chain hashing in Section 2.3 relied only on the 2-universality of the hash family, we have established the following.

**Proposition 2.2** Consider a hash table with $n$ buckets that uses chain hashing with a hash function drawn uniformly from a 2-universal hash family. Let $s_{\text{hash}}(n)$ denote the space complexity of storing a representation of the hash function, and let $t_{\text{hash}}(n)$ denote the time complexity of evaluating $\text{HASH}(x)$ on any given key $x$. For any sequence of hash table operations with at most $m \leq cn$ key-value pairs stored in the table at any time,

the expected time per operation is $O(c+1+t_{\text{hash}})$. The space complexity of the hash table is $O(m+n+s_{\text{hash}})$.

Of course, in order for a 2-universal hash family to be computationally useful, it is necessary to be able to efficiently store and evaluate the hash functions in the family. The remainder of this section provides some useful constructions of 2-universal hash families.

### 2.4.1 Linear congruential hashing

The simplest construction of a 2-universal hash function works when the number of hash buckets is a prime number, $p$, and the space of potential keys, $\mathcal{X}$, has $p$ or fewer[2] elements.

Let $\mathbb{F}_p$ denote the set $\{0,1,\ldots,p-1\}$ under the operations of addition and multiplication modulo $p$. In other words, to add or multiply two elements of $\mathbb{F}_p$, one adds them or multiplies them as ordinary integers, and then if the result is greater than or equal to $p$, one divides it by $p$ and outputs the remainder. This structure is called the *prime field of order p*, but for present purposes it is not necessary to know what a field is in order to analyze the hash function family we now present.

The *linear congruential hash function family modulo p* is the family of functions $h:\mathbb{F}_p\to\mathbb{F}_p$ defined by

$$h(x)=ax+b$$

where the coefficients $a$ and $b$ are allowed to be any elements of $\mathbb{F}_p$. Denote this family of functions by $\mathcal{H}_p$.

**Lemma 2.3** For any prime number $p$, the linear congruential hash function family $\mathcal{H}_p$ is a 2-universal hash family.

*Proof.* The hash function family $\mathcal{H}_p$ has $p^2$ elements, one for each pair of coefficients $a,b\in\mathbb{F}_p$. We intend to argue that for any $x\neq y$, the function $\phi_{x,y}:\mathcal{H}_p\to\mathbb{F}_p^2$ defined by

$$\phi_{x,y}(h)=(h(x),h(y))$$

is a bijection. If so, then it follows that for $h$ drawn uniformly at random from $\mathcal{H}_p$, the pair $(h(x),h(y))=\phi_{x,y}(h)$ will be drawn uniformly at random from $\mathbb{F}_p^2$ as required by the definition of a 2-universal hash family.

Since $\mathcal{H}_p$ and $\mathbb{F}_p^2$ both have exactly $p^2$ elements, the assertion that $\phi_{x,y}$ is a bijection is equivalent to the assertion it is one-to-one. In other words, if $h,h'$ are two (possibly identical) elements of $\mathcal{H}_p$ such that $\phi_{x,y}(h)=\phi_{x,y}(h')$, we are accountable for proving that $h=h'$. Let $a,b$ and $a',b'$ be the coefficient pairs for $h$ and $h'$, respectively. In other words, for all $z$, $h(z)=az+b$ and $h'(z)=a'z+b'$. Then, rewriting the equation $\phi_{x,y}(h)=\phi_{x,y}(h')$ as $h(x)=h'(x)$ and $h(y)=h'(y)$ we find that

$$\begin{aligned}(a-a')x+(b-b')&=h(x)-h'(x)=0\\(a-a')y+(b-b')&=h(y)-h'(y)=0\\(a-a')(x-y)&=0\end{aligned}$$

[2] In typical hash tables, the number of potential keys is *much* greater than the number of buckets, so the assumption that $|\mathcal{X}|\leq p$ is quite limiting. The hash function family that we construct and analyze in this section nevertheless has other very useful applications. One such application will be presented later in Section 3.2.

where all addition and multiplication operations are interpreted modulo $p$, and the equation on the last line is obtained by subtracting both sides of the equations on the two lines above. The equation $(a-a')(x-y) = 0$ (modulo $p$) means that $p$ is a divisor of the product $(a-a')(x-y)$. Since $p$ is prime, it must divide at least one of the factors, $a-a'$ or $x-y$. However, since $a, a', x, y$ all belong to the set $\{0, 1, \ldots, p-1\}$, the differences $a-a'$ and $x-y$ belong to the interval $[-(p-1), p-1]$, and 0 is the only multiple of $p$ in that interval. Hence, either $a = a'$ or $x = y$. By assumption $x \neq y$, so we have shown $a = a'$. Now, rewriting the equation $h(x) = h'(x)$ as $ax+b = a'x+b'$ and using the equation $a = a'$, we find that $b = b'$ as well. Thus, $h = h'$ as desired. ■

To evaluate the space and time efficiency of linear congruential hashing, recall our standing assumption that one memory location can store $O(\log n) = O(\log p)$ bits, which is sufficient to store the value of one element of $\mathbb{F}_p$. To store a representation of a hash function $h \in \mathcal{H}_p$ one only needs to store the coefficients $a$ and $b$, hence the representation of the function $h$ requires only $O(1)$ space. Evaluating $h$ requires only 2 arithmetic operations (mod $p$), so it takes $O(1)$ time.

### 2.4.2 Inner product hashing

Linear congruential hashing can be generalized to allow for applications in which the number of buckets is still a prime number, $p$, but the number of potential keys, $N$, may be much greater than $p$. In that case, we will let $d = \lceil \log_p(N) \rceil$ and we will identify the space of $N$ keys, $\mathcal{X}$, with a subset of the set $\mathbb{F}_p^d$ of $d$-tuples of elements of $\mathbb{F}_p$. The set $\mathbb{F}_p^d$ constitutes a *$d$-dimensional vector space over the prime field of order $p$*. Analogous to the dot-product operation on vectors in $\mathbb{R}^d$ we have the following *inner product* operation which is defined for any two elements $\mathrm{w} = (w_1, \ldots, w_d)$ and $\mathrm{x} = (x_1, \ldots, x_d)$ in $\mathbb{F}_p^d$.

$$\langle \mathrm{w}, \mathrm{x} \rangle = \sum_{i=1}^{d} w_i x_i.$$

As usual, the addition and multiplication operations on the right side are interpreted modulo $p$. Now, let $\mathcal{H}_p^d$ denote the set of hash functions $h : \mathbb{F}_p^d \to \mathbb{F}_p$ of the form

$$h(\mathrm{x}) = \langle \mathrm{a}, \mathrm{x} \rangle + b$$

where $\mathrm{a} \in \mathbb{F}_p^d$ and $b \in \mathbb{F}_p$. Storing the representation of a function $h \in \mathcal{H}_p^d$ requires storing $d+1$ elements of $\mathbb{F}_p$, which takes $O(d)$ space. Evaluating $h(\mathrm{x})$ takes $O(d)$ time, since it involves performing $d$ multiplication and $d$ addition operations in $\mathbb{F}_p$.

**Lemma 2.4** For any prime number $p$, the linear congruential hash function family $\mathcal{H}_p$ is a 2-universal hash family.

*Proof.* We can prove that the hash family $\mathcal{H}_p^d$ is 2-universal by mimicking the proof in the $d = 1$ case, Lemma 2.4. Consider any $\mathrm{x}, \mathrm{y} \in \mathbb{F}_p^d$. If $\mathrm{x} \neq \mathrm{y}$ it means there is a coordinate $j$ such that $x_j \neq y_j$. Without loss of generality assume $j = 1$. Then, we can break down the process of sampling $h \in \mathcal{H}_p^d$ into two steps:

1. sample $a_2, a_3, \ldots, a_d \in \mathbb{F}_p$ independently and uniformly at random;
2. sample $a_1, b \in \mathbb{F}_p$ independently and uniformly at random.

For any fixed choice of $a_2, \ldots, a_d$ in the first sampling step, there are $p^2$ choices of $a_1$ and $b$ in the second step and corresponding to each of them there is an ordered pair of hash values $(h(\mathrm{x}), h(\mathrm{y})) \in \mathbb{F}_p^2$. If we can prove that the mapping $(a_1, b) \mapsto (h(\mathrm{x}), h(\mathrm{y}))$ is one-to-one, then it must be bijective, from which it follows that $(h(\mathrm{x}), h(\mathrm{y}))$ is uniformly distributed over $\mathbb{F}_p^2$, conditional on our fixed choice of $a_2, \ldots, a_d$. Since this holds for any fixed choice of $a_2, \ldots, a_d$, it then follows (by averaging over $a_2, \ldots, a_d$) that the unconditional distribution of $(h(\mathrm{x}), h(\mathrm{y}))$ is uniform over $\mathbb{F}_p^2$, as desired.

To prove that the function $(a_1, b) \mapsto (h(\mathrm{x}), h(\mathrm{y}))$ is one-to-one, consider any coefficient pairs $(a_1, b)$ and $(a_1', b')$ with associated hash functions $h$ and $h'$, respectively, and assume $h(\mathrm{x}) = h'(\mathrm{x})$ and $h(\mathrm{y}) = h'(\mathrm{y})$. Define $\mathrm{a} = (a_1, a_2, \ldots, a_d)$ and $\mathrm{a}' = (a_1', a_2, \ldots, a_d)$; note that these two vectors differ only in their first coordinate. We have

$$
\begin{aligned}
0 &= h(\mathrm{x}) - h'(\mathrm{x}) = \langle \mathrm{a}, \mathrm{x} \rangle + b - \langle \mathrm{a}', \mathrm{x} \rangle - b' = \langle \mathrm{a} - \mathrm{a}', \mathrm{x} \rangle + (b - b') = (a_1 - a_1')x_1 + (b - b') \\
0 &= h(\mathrm{y}) - h'(\mathrm{y}) = \langle \mathrm{a}, \mathrm{y} \rangle + b - \langle \mathrm{a}', \mathrm{y} \rangle - b' = \langle \mathrm{a} - \mathrm{a}', \mathrm{y} \rangle + (b - b') = (a_1 - a_1')y_1 + (b - b') \\
0 &= (a_1 - a_1')(x_1 - y_1).
\end{aligned}
$$

By assumption, $x_1 \neq y_1$ so, as in the proof of Lemma 2.4, it follows that $a_1 = a_1'$. In that case $\mathrm{a} = \mathrm{a}'$, so rewriting the equation $h(\mathrm{x}) = h'(\mathrm{x})$ as $\langle \mathrm{a}, \mathrm{x} \rangle + b = \langle \mathrm{a}, \mathrm{x} \rangle + b'$, we see that it implies $b = b'$ and hence $h = h'$, as desired. ■

# 3. Data Streaming and Sketching

In this section we survey some of the applications of hashing to the analysis of datasets that are too large to fit in the computer's memory all at once.

In the *streaming* model of computation, an algorithm observes a sequence $a_1, a_2, \ldots, a_n$ of data items, each represented by at most $b$ bits. Thus, the set of potential data items (called "tokens" henceforth) has size $m = 2^b$. The algorithm has a working memory of size $s$, where each memory location is assumed to be capable of storing $O(\log n)$ bits. Typically we require $s$ to have sublinear dependence on $n$ (the length of the stream) and at most linear dependence on $b$ (the number of bits representing each element). Hence it is infeasible to store each data item, which would require space $s \geq n$ even if $b$ were $O(\log n)$, and it's also infeasible to store a count of how many times each token was seen in the data stream, which could require space $s \geq 2^b$ if every element of $\{0,1\}^b$ were observed at least once in the stream.

Some of the typical objectives of streaming algorithms are to find the most frequently occurring element (or elements) in the data stream, approximate the number of distinct elements, or approximate the $p^{\text{th}}$ frequency moment, $\sum_j f_j^p$, where $f_j$ denotes the number of occurrences of the token $j$ in the stream.

## 3.1 Finding frequent elements

To illustrate the model, we begin by presenting an example of a non-trivial streaming algorithm that makes no use of hashing and is, in fact, completely deterministic. This is an algorithm of Misra and Gries that uses space $s = O(k(b + \log n))$ to find every token that occurs more than $n/(k+1)$ times in the stream. The algorithm allocates its storage space for a $k$-tuple of tokens $b_1, \ldots, b_k$, and a $k$-tuple of counters, $c_1, \ldots, c_k$. Initially each pair $(b_j, c_j)$ is initialized to $(\perp, 0)$, where $\perp$ denotes a null symbol that doesn't belong to the set of tokens. While the algorithm is processing the stream, if it sees one of the tokens

$b_1, \ldots, b_k$ then it increments the corresponding counter. Otherwise, if one of the counters $c_j$ is equal to zero, it stores the new element as $b_j$ and sets $c_j$ to 1. Otherwise, if all of the counters are strictly positive, it decrements each of them. When the algorithm finishes processing the stream, it outputs the set of all tokens that have positive counters.

---

Initialize $(b_j, c_j) = (\perp, 0)$ for $j = 1, 2, \ldots, k$.
**for** $i = 1, 2, \ldots, n$ **do**
    **if** $a_i = b_j$ for some $j \in [k]$ **then**
        $c_j \leftarrow c_j + 1$
    **else if** $c_j = 0$ for some $j \in [k]$ **then**
        $b_j \leftarrow a_i$
        $c_j \leftarrow 1$
    **else**
        Decrement $c_j$ to $c_j - 1$ for each $j \in [k]$.
    **end if**
**end for**
Output $\{b_j \mid c_j > 0\}$.

---

**Proposition 3.1** The output of the Misra-Gries algorithm contains every token that occurs more than $n/(k+1)$ times in the data stream (and potentially some tokens that occur fewer than $n/(k+1)$ times).

*Proof.* Picture marking elements of the sequence $a_1, a_2, \ldots, a_n$ as follows. Initially all elements are unmarked. At the start of the loop iteration that processes element $a_i$, it becomes marked. There are three cases for what could happen during the loop iteration. In the first two cases, if $a_i \in \{b_1, \ldots, b_k\}$ or if $a_i \notin \{b_1, \ldots, b_k\}$ but $c_j = 0$ for some $j$, then $a_i$ remains marked. In the third case, if $a_i \notin \{b_1, \ldots, b_k\}$ and $c_j > 0$ for all $j$, then we remove the mark from $a_i$, and we also remove red marks from the earliest marked copy of each of the tokens $b_1, \ldots, b_k$.

We claim that at all times, there are $c_j$ marked copies of $b_j$ for each $j \in [k]$, and no token other than $b_1, \ldots, b_k$ is marked. The proof is by induction on $i$. In the base case $i = 0$, no tokens are marked and $c_j = 0$ for all $j$. For the induction step, if $a_i$ belongs to the set $\{b_1, \ldots, b_k\}$ or is inserted into that set, then it remains marked at the end of the loop iteration and the corresponding counter $c_j$ is incremented. If $a_i$ doesn't belong to the set $\{b_1, \ldots, b_k\}$ and $c_j > 0$ for all $j$, then the mark is removed from $a_i$ and (by the induction hypothesis) there is at least one marked copy of $b_j$ for every $j \in [k]$, so a mark is removed from one copy of each $b_j$ as $c_j$ is decremented.

Each time a loop iteration removes any marks, it removes $k+1$ of them. Since an element of the sequence is only marked once and its mark is removed at most once, there are at most $n/(k+1)$ loop iterations in which marks are removed. If a token appears strictly more than $n/(k+1)$ times in the sequence, then some copies of that token are marked at the end of the final loop iteration, so that token must be one of $b_1, \ldots, b_k$. ■

## 3.2 Estimating the number of distinct elements

The Misra-Gries algorithm is atypical of streaming algorithms because it's deterministic. Generally a streaming algorithm's objective can't be achieved deterministically within

the given space bound, so these algorithms use randomness and are usually evaluated according to the PAC (probably approximately correct) objective: one wants to show that with probability at least $1-\delta$, the algorithm's output approximates the target quantity with relative error $\varepsilon$ or less.

Here's a famous example due to Flajolet and Martin. The algorithm estimates the number of distinct tokens in the data stream. Note that this number might be as large as $m = 2^b$, but we aim to estimate the number of distinct token in space $s = \text{poly}(b, \log n)$, so keeping a list of every distinct token encountered in the stream is not an option. Instead, we will use a hash function $h : [m] \to [M]$, for some large integer $M$. The function $h$ will be drawn at random from a 2-universal hash family.

The key observation is that if there are $d$ distinct tokens in the stream, then the random variable $Z = \min\{h(a_i) \mid 1 \leq i \leq n\}$ is on the order of $\frac{M}{d}$. In fact, if we assume without loss of generality that the $d$ distinct tokens belonging to the stream are $a_1, \ldots, a_d$, then for any $k \in [M]$ we can define the random variables

$$X_{ik} = \begin{cases} 1 & \text{if } h(a_i) \leq k \\ 0 & \text{otherwise} \end{cases}$$

$$Y_k = \sum_{i=1}^{d} X_{ik} = \text{number of distinct tokens whose hash value is } \leq k.$$

Then, we make the following observations.

1. $\mathbb{E}[X_{ik}] = \frac{k}{M}$.
2. $\mathbb{E}[Y_k] = \frac{dk}{M}$.
3. $\text{Var}[Y_k] = \frac{dk(M-k)}{M^2} < \frac{dk}{M}$. This is because

$$\begin{aligned} \text{Var}[Y_k] = \mathbb{E}[Y_k^2] - \mathbb{E}[Y_k]^2 &= \sum_{i=1}^{d}\sum_{j=1}^{d} \mathbb{E}[X_{ik}X_{jk}] - \frac{d^2k^2}{M^2} \\ &= \sum_{i=1}^{d} \mathbb{E}[X_{ik}] + \sum_{i=1}^{d}\sum_{j\neq i} \mathbb{E}[X_{ik}X_{jk}] - \frac{d^2k^2}{M^2} \\ &= d \cdot \frac{k}{M} + d(d-1) \cdot \frac{k^2}{M^2} - - \frac{d^2k^2}{M^2} \\ &= d \cdot \frac{k}{M} - d \cdot \frac{k^2}{M^2} = \frac{dk(M-k)}{M^2}. \end{aligned}$$

In these observations, we made use of the assumption that the hash function $h$ is sampled from a 2-universal hash family. The first such usage was when we asserted $\mathbb{E}[X_{ik}] = \frac{k}{M}$, which requires $h(a_i)$ to be uniformly distributed. The second such usage was in the calculation of $\text{Var}[Y_k]$: the derivation of the third line uses the fact that $\mathbb{E}[X_{ik}X_{jk}] = \frac{k^2}{M^2}$, which follows from the pair $(h(a_i), h(a_j))$ being uniformly distributed in $[M]^2$.

Recall that $Z = \min\{h(a_i) \mid i \in [n]\}$. As a first attempt at estimating $d$, we can approximate it with the quantity $M/Z$. By Markov's Inequality, if $k = \lfloor M/6d \rfloor$, then

$$\Pr\left(\tfrac{M}{Z} > 6d\right) = \Pr\left(\tfrac{M}{6d} > Z\right) = \Pr\left(Y_k \geq 1\right) \leq \mathbb{E}[Y_k] = \tfrac{dk}{M} \leq \tfrac{1}{6}. \tag{3.1}$$

On the other hand, by Chebyshev's Inequality, if $\ell = \lfloor 6M/d \rfloor$,

$$\begin{aligned}\Pr\left(\tfrac{M}{Z} < \tfrac{d}{6}\right) &= \Pr\left(\tfrac{6M}{d} \le Z\right) = \Pr(Y_\ell = 0) \le \Pr(|Y_\ell - \mathbb{E}Y_\ell| \ge \mathbb{E}Y_\ell) \\ &\le \frac{\mathrm{Var}(Y_\ell)}{(\mathbb{E}Y_\ell)^2} \\ &< \frac{\mathbb{E}Y_\ell}{(\mathbb{E}Y_\ell)^2} = \frac{1}{\mathbb{E}Y_\ell} = \frac{M}{d\ell} \le \frac{1}{6} + \frac{1}{6M-5}. \end{aligned} \tag{3.2}$$

Hence, the probability that the estimate $M/Z$ lies outside the interval $[d/6, 6d]$ is at most $\frac{1}{3} + \frac{1}{6M-5}$.

We can obtain a better estimate of $d$ using $Z_t$, the $t^{\text{th}}$ smallest of the values $\{h(a_i)\}_{i=1}^n$, for a suitable choice of the parameter $t > 1$. Intuitively, the reason is that $Z_t$ "aggregates a greater amount of randomness", namely the randomness in the positions of the $t$ smallest elements rather than just the smallest one. To make this intuition a bit more precise, if we set $k = \lfloor tM/d \rfloor$ such that the expected number of elements that hash into the set $[k]$ is $\mathbb{E}[Y_k] = dk/M \approx t$, then the variance $\mathrm{Var}[Y_k]$ is less than $t$, so the probability that $Y_k$ differs from its expected value by more than $\varepsilon t$ is at most $\frac{1}{\varepsilon^2 t}$ by Chebyshev's Inequality. For $t > \frac{1}{\varepsilon^2 \delta}$, this probability will be less than $\delta$. This argument doesn't directly lead to the conclusion that $tM/Z_t$ approximates $d$ within $\varepsilon$, but a variation on the argument — using random variables $Y_q$ and $Y_r$ for values $q$ and $r$ differing from $k$ by $1 \pm \varepsilon$ factors, and accounting more carefully for hash collisions — does the trick.

**Algorithm 2** Algorithm for estimating distinct elements

1: Set $t = \lceil \frac{12}{\varepsilon^2 \delta} \rceil$.
2: Choose $M \ge 6m/(\varepsilon\delta)$ and randomly sample $h : [m] \to [M]$ from a 2-universal hash family.
3: Initialize $(Z_1, Z_2, \ldots, Z_t) = \perp^t$.
4: **for** $i = 1, \ldots, n$ **do**
5: Observe $a_i$ and calculate $z = h(a_i)$.
6: **if** $z < Z_t$ **then**
7: Update $Z_1, \ldots, Z_t$ to be the $t$ smallest hash values yet seen, in increasing order.
8: **end if**
9: **end for**
10: Output $tM/Z_t$.

Note that the space required by the algorithm is equal to $O(t \log M)$ plus the amount of space required to store $h$. In Section 2.4 we saw that when $M$ is a prime number, we can sample $h$ from the linear congruential hash family and the space required for storing $h$ will be $O(\log M)$ bits. Since there is always a prime number between $6m/(\varepsilon\delta)$ and $12m/(\varepsilon\delta)$, we can ensure $\log(M) \le \log(m) - \log(\varepsilon\delta) + O(1)$. Also $t \le \frac{12}{\varepsilon^2\delta} + 1$, so the space required by the algorithm is $s = O\left(\frac{\log m - \log(\varepsilon\delta)}{\varepsilon^2 \delta}\right)$ bits.

**Proposition 3.2** When Algorithm 2 is run on a stream with $d$ distinct elements, the probability that it outputs an answer in the range $[(1-\varepsilon)d, (1+\varepsilon)d]$ is at least $1 - \delta$.

*Proof.* If the output, $tM/Z_t$, lies outside the range $[(1-\varepsilon)d, (1+\varepsilon)d]$, it means that $Z_t$ lies outside the range $\left[\frac{tM}{(1+\varepsilon)d}, \frac{tM}{(1-\varepsilon)d}\right]$. To reason about the circumstances under which $Z_t$

could lie outside the range $\left[\frac{tM}{(1+\varepsilon)d}, \frac{tM}{(1-\varepsilon)d}\right]$, we let

$$q = \left\lceil \frac{tM}{(1+\varepsilon)d} \right\rceil - 1, \qquad r = \left\lfloor \frac{tM}{(1-\varepsilon)d} \right\rfloor$$

and observe that there are three cases in which $Z_t$ lies outside $\left[\frac{tM}{(1+\varepsilon)d}, \frac{tM}{(1-\varepsilon)d}\right]$.

1. If $Z_t < \frac{tM}{(1+\varepsilon)d}$ it means that at least $t$ distinct tokens in the stream are mapped to hash buckets in the range $[q] = \{1, 2, \ldots, q\}$. Denote this event by $\mathcal{E}_1$.
2. If $Z_t > \frac{tM}{(1-\varepsilon)d}$ it means that one of the following two cases must occur.
   a. Fewer than $\left(1 + \frac{\varepsilon}{2-2\varepsilon}\right)t$ distinct tokens in the stream are mapped to hash buckets in the range $[r]$. Denote this event by $\mathcal{E}_{2a}$.
   b. At least $\left(1 + \frac{\varepsilon}{2-2\varepsilon}\right)t$ distinct tokens in the stream are mapped to hash buckets in $[r]$, but these tokens occupy fewer than $t$ distinct buckets because there are more than $\frac{\varepsilon t}{2-2\varepsilon}$ hash collisions in buckets whose number belongs to $[r]$. Denote this event by $\mathcal{E}_{2b}$.

By the union bound, to prove $\Pr\left(Z_t \in \left[\frac{tM}{(1+\varepsilon)d}, \frac{tM}{(1-\varepsilon)d}\right]\right) \geq 1 - \delta$, it suffices to prove $\Pr(\mathcal{E}_1) + \Pr(\mathcal{E}_{2a}) + \Pr(\mathcal{E}_{2b}) \leq \delta$. We will do so by defining some random variables, reasoning about their expectations and variances using the 2-universality of the random hash function $h$, and applying Markov's and Chebyshev's Inequalities.

Assume without loss of generality that $a_1, a_2, \ldots, a_d$ are the distinct tokens in the stream and for $i \in [d]$ and $\ell \in [M]$ let

$$X_{i\ell} = \begin{cases} 1 & \text{if } h(a_i) = \ell \\ 0 & \text{otherwise.} \end{cases}$$

The random variable $X_{i\ell}$ indicates whether $a_i$ hashes into bucket $\ell$, and for $i \neq j$ the random variable $X_{i\ell}X_{j\ell}$ indicates whether $a_i$ and $a_j$ collide in bucket $\ell$. By 2-universality,

$$\mathbb{E}[X_{i\ell}] = \tfrac{1}{M} \tag{3.3}$$

$$\mathbb{E}[X_{i\ell}X_{j\ell}] = \tfrac{1}{M^2} \qquad \text{if } i \neq j. \tag{3.4}$$

For any $k \in [M]$, the random variable

$$Y_k = \sum_{i=1}^{d} \sum_{\ell=1}^{k} X_{i\ell}$$

counts the number of distinct tokens that hash to buckets numbered in the range $[k]$. Its expectation and variance can be calculated using equations (3.3) and (3.4).

$$\mathbb{E}[Y_k] = \frac{dk}{M} \tag{3.5}$$

$$\mathrm{Var}[Y_k] = dk\left(\frac{1}{M} - \frac{1}{M^2}\right) < \frac{dk}{M}. \tag{3.6}$$

We can use these calculations to bound the probabilities of $\mathcal{E}_1$ and $\mathcal{E}_{2a}$.

$$\begin{aligned}
\Pr(\mathcal{E}_1) &= \Pr(Y_q \geq t) \leq \Pr\left(|Y_q - \mathbb{E}[Y_q]| \geq t - \mathbb{E}[Y_q]\right) \leq \frac{\mathrm{Var}[Y_q]}{(t-\mathbb{E}[Y_q])^2} < \frac{dq/M}{(t-dq/M)^2} \\
\Pr(\mathcal{E}_2) &= \Pr\left(Y_r < \left(1+\tfrac{\varepsilon}{2-2\varepsilon}\right)t\right) \leq \Pr\left(|Y_r - \mathbb{E}[Y_r]| > \mathbb{E}[Y_r] - \left(1+\tfrac{\varepsilon}{2-2\varepsilon}\right)t\right) \\
&\leq \frac{\mathrm{Var}[Y_r]}{\left(\mathbb{E}[Y_r] - \left(1+\frac{\varepsilon}{2-2\varepsilon}\right)t\right)^2} < \frac{dr/M}{\left(dr/M - \left(1+\frac{\varepsilon}{2-2\varepsilon}\right)t\right)^2}.
\end{aligned}$$

By our choice of $q$ and $t$ we have

$$\begin{aligned}
\tfrac{dq}{M} &< \tfrac{t}{1+\varepsilon} \\
t - \tfrac{dq}{M} &> \tfrac{\varepsilon t}{1+\varepsilon} \\
\frac{dq/M}{(t-dq/M)^2} &< \frac{t/(1+\varepsilon)}{\varepsilon^2 t^2/(1+\varepsilon)^2} = \frac{1+\varepsilon}{\varepsilon^2 t} < \frac{\delta}{3}.
\end{aligned}$$

By our choice of $r$ and $t$ we have

$$\begin{aligned}
\tfrac{t}{1-\varepsilon} \leq \tfrac{dr}{M} &< \tfrac{t}{1-\varepsilon} + 1 \\
\tfrac{dr}{M} - \left(1+\tfrac{\varepsilon}{2-2\varepsilon}\right)t &\geq \tfrac{t}{1-\varepsilon} - t - \tfrac{\varepsilon t}{2-2\varepsilon} = \tfrac{\varepsilon t}{2-2\varepsilon} \\
\frac{dr/M}{\left(dr/M - \left(1+\frac{\varepsilon}{2}\right)t\right)^2} &< \frac{t/(1-\varepsilon)+1}{\varepsilon^2 t^2/4(1-\varepsilon)^2} = \frac{4(1-\varepsilon)+4(1-\varepsilon)^2/t}{\varepsilon^2 t} < \frac{\delta}{3}
\end{aligned}$$

Hence, $\Pr(\mathcal{E}_1)$ and $\Pr(\mathcal{E}_{2a})$ are both less than $\delta/3$. To deal with $\Pr(\mathcal{E}_{2b})$ we observe that the number of pairs of distinct stream tokens whose hash values collide in bucket set $[r]$ is equal to the random variable $W = \sum_{\ell=1}^{r} \sum_{1 \leq i < j \leq d} X_{i\ell} X_{j\ell}$. By equation (3.4) and our choice of $M \geq \frac{6m}{\varepsilon\delta}$ we have

$$\begin{aligned}
\mathbb{E}[W] &= \frac{r \cdot \sharp\{(i,j) \mid 1 \leq i < j \leq d\}}{M^2} = \frac{rd(d-1)}{2M^2} < \left(\frac{dr}{M}\right)\left(\frac{d}{2M}\right) < \left(\frac{t+1}{1-\varepsilon}\right)\left(\frac{m}{2M}\right) \\
\Pr(\mathcal{E}_3) &= \Pr\left(W > \frac{\varepsilon t}{2(1-\varepsilon)}\right) < \frac{\frac{t+1}{1-\varepsilon} \cdot \frac{m}{2M}}{\varepsilon t/2(1-\varepsilon)} = \left(1+\frac{1}{t}\right)\frac{m}{\varepsilon M} < \frac{\delta}{3}.
\end{aligned}$$

We have shown that each of the events $\mathcal{E}_1, \mathcal{E}_{2a}, \mathcal{E}_{2b}$ has probability less than $\delta/3$, and that the union of these three events covers all cases in which $tM/Z_t$ lies outside the interval $[(1-\varepsilon)d, (1+\varepsilon)d]$. Hence, with probability at least $1-\delta$, the algorithm's estimate $tM/Z_t$ belongs to the interval $[(1-\varepsilon)d, (1+\varepsilon)d]$ as claimed. ■

## 3.3 Sketching token frequencies

Data sketching is an algorithmic paradigm that combines streaming with data structures. As before, an algorithm processes a stream of tokens, $a_1, \ldots, a_n$, taking values in $[m]$, and it is allowed to store $s = O(\mathrm{poly}(\log n, \log m))$ bits of information about the stream. However, rather than wanting to estimate a single attribute of the stream, such as the number of distinct elements, the algorithm designer's objective is to be able to answer queries about the stream afterward. In this setting, the $s$-bit internal representation of the stream is called a *sketch* of the data.

Consider the task of sketching the frequency of each token in the data stream. In other words, the algorithm will be asked to answer queries of the form, "How many times did $x$ occur in the stream?" and the goal will be to output an approximately correct answer with probability $1-\delta$. In this section we will present two different algorithms for this task. The algorithms have different benefits and drawbacks. The first algorithm has smaller space complexity and only suffers from one-sided error, i.e. it can overestimate the number of occurrences of $x$ but it never underestimates. The second algorithm requires more space and suffers from two-sided error, but it satisfies a significantly stronger approximate-correctness property.

**Algorithm 3** Count-Min Sketch

Given positive integers $B, t$ ...
Sample $h_1, \ldots, h_t : [m] \to [B]$ independently from a 2-universal hash family.
Initialize a two-dimensional array $C$ of dimensions $B \times t$, setting $C[k,\ell] = 0$ for each $k, \ell$.
**for** each $i \in [n]$ **do**
    Observe $a_i$.
    **for** each $\ell \in [t]$ **do**
        Compute $k = h_\ell(a_i)$.
        Increment $C[k,\ell]$ by 1.
    **end for**
**end for**
When queried about frequency of token $x$, return $\min_{\ell \in [t]} \{C[h_\ell(x), \ell]\}$.

The first algorithm we'll analyze, called the Count-Min Sketch, is based on a hashing scheme presented in Algorithm 3. The idea behind the algorithm is simple: we choose $t$ independent random hash functions $h_1, \ldots, h_t$, with range $[B]$ for some moderately large $B$, and for each "hash bucket" $k \in [B]$ we count how many elements of the stream are hashed to $k$ by each of the $t$ functions. If $h$ is a hash function and $x$ is a token appearing $r$ times in the stream, then the counter for bucket $h(x)$ will reach a value which is at least $r$. To the extent that the counter exceeds $r$, the difference is due to hash collisions — other elements of the stream that hash to the same bucket as $x$. For large $B$, this will typically be only a small fraction of the stream. By repeating this counting procedure in parallel using $t$ different hash functions, we minimize the probability of getting an anomalously large number of hash collisions.

**Lemma 3.3** The CountMin sketch uses space $s = O(Bt \log(mn))$ and satisfies the following guarantee for every $x \in [m]$: if the true frequency of $x$ in the stream is denoted by $f_x$, the sketch's estimate $\hat{f}_x$ satisfies $f_x \leq \hat{f}_x$ with probability 1 and $\hat{f}_x \leq f_x + \frac{2n}{B}$ with probability at least $1 - 2^{-t}$.

*Proof.* The space complexity bound follows from the observation that the algorithm only needs to store an array of dimensions $B \times t$, with each element of the array being an integer in the range $0, 1, \ldots, n$, plus descriptions of $t$ hash functions each requiring space $O(\log m)$.

For each $\ell \in [t]$, the counter $C[h_\ell(x), \ell]$ is incremented each time $x$ appears in the stream — $f_x$ times in total — and it is also incremented each time another token $y \neq x$ appears in the stream and satisfies $h_\ell(y) = h_\ell(x)$. There are $n - f_x$ tokens other than $x$ in the stream, and

for each of them the probability that $h_\ell(y) = h_\ell(x)$ is $1/B$, so by linearity of expectation we have $\mathbb{E}[C[h_\ell(x),\ell] - f_x] = (n - f_x)/B$. Then, by Markov's Inequality,

$$\Pr\left(C[h_\ell(x),\ell] - f_x > \tfrac{2n}{B}\right) \leq \tfrac{1}{2}.$$

Since the hash function $\{h_1,\ldots,h_t\}$ are mutually independent,

$$\Pr\left(\forall \ell \in [t]\, C[h_\ell(x),\ell] - f_x > \tfrac{2n}{B}\right) \leq \left(\tfrac{1}{2}\right)^t,$$

and the lemma follows. ■

**Corollary 3.4** For any $\varepsilon, \delta > 0$ the Count-Min Sketch with parameters $B = \lceil \frac{2}{\varepsilon} \rceil$ and $t = \lceil \log_2(1/\delta) \rceil$ achieves the following guarantee: for any token $x$, with probability at least $1 - \delta$ the estimated frequency of $x$ differs from the true frequency by no more than $\varepsilon n$. The space complexity of the sketch with these parameters is $O\left(\log(mn)\log(1/\delta)/\varepsilon\right)$.

The second algorithm we'll analyze uses more space, namely $O\left(\log n \log(1/\delta)/\varepsilon^2\right)$, but achieves a stronger approximate-correctness guarantee: with probability at least $1 - \delta$, the estimate of $f_x$ differs from the true value by at most $\varepsilon \|\mathrm{f}\|_2$. Here, f denotes the "frequency vector" of the stream, an $m$-dimensional vector whose $x^{\text{th}}$ component $f_x$ is the frequency of token $x$ in the stream. Since the sum of frequencies of all tokens is $n$, we have $\mathrm{f}_1 = n$. Note that $\mathrm{f}_2 \leq \mathrm{f}_1$ for any vector f, so the error bound of $\varepsilon \|\mathrm{f}\|_2$ is never worse than the $\varepsilon n$ error bound of the Count-Min Sketch. However, $\|\mathrm{f}\|_2$ can be much smaller than $n$; for example, when the tokens are uniformly distributed we have $\|\mathrm{f}\|_2 \approx \frac{n}{\min\{\sqrt{m},\sqrt{n}\}}$.

**Algorithm 4** Count Sketch

Given positive integers $B, t$ ...
Sample $h_1,\ldots,h_t : [m] \to [B]$ independently from a 2-universal hash family.
Sample $g_1,\ldots,g_t : [m] \to \{\pm 1\}$ independently from a 2-universal hash family.
Initialize a two-dimensional array $C$ of dimensions $B \times t$, setting $C[k,\ell] = 0$ for each $k,\ell$.
**for** each $i \in [n]$ **do**
    Observe $a_i$.
    **for** each $\ell \in [t]$ **do**
        Compute $k = h_\ell(a_i)$.
        $C[k,\ell] \leftarrow C[k,\ell] + g_\ell(a_i)$.
    **end for**
**end for**
When queried about frequency of token $x$, return the median of the multiset $\{g_\ell(x) \cdot C[h_\ell(x),\ell]\}$.

The intuition for the Count Sketch is similar to that for the Count-Min Sketch with one important difference. As before, if $x$ occurs $f_x$ times in the stream, then with each occurrence we add $g_\ell(x)$ to $C[h_\ell(x),\ell]$, resulting in a total of $g_\ell(x) \cdot f_x$. Since $g_\ell(x)^2 = 1$, this means that the random variable $g_\ell(x) \cdot C[h_\ell(x),\ell]$ equals $f_x + Z$, where the random variable $Z$ accounts for the "noise" due to other tokens $y \neq x$ that are hashed by $h_\ell$ to the same bucket as $x$, similarly to the analysis of the Count-Min Sketch. However, the

key difference is that the noise variable $Z$ in the Count Sketch is a sum of randomly-signed contributions from the various tokens that occupy the same hash bucket as $x$. In aggregate we can expect some of these noise terms to cancel each other out because they are oppositely signed. Hence, we might hope that the Count Sketch suffers from less additive error when estimating the frequency $f_x$. The following analysis substantiates that hope.

**Lemma 3.5** The Count Sketch uses space $s = O(Bt\log(mn))$ and satisfies the following guarantee for every $x \in [m]$: if the true frequency of $x$ in the stream is denoted by $f_x$, the sketch's estimate $\hat{f}_x$ satisfies $|\hat{f}_x - f_x| \leq \sqrt{\frac{3}{B}}\|\mathrm{f}\|_2$ with probability at least $1 - e^{-t/18}$.

*Proof.* Fix $x \in [m]$. For any $y \in [m]$ and $\ell \in [t]$ define random variables $X_{y\ell}$ and $Z_{y\ell}$ by

$$X_{y\ell} = \begin{cases} 1 & \text{if } h_\ell(y) = h_\ell(x) \\ 0 & \text{if } h_\ell(y) \neq h_\ell(x) \end{cases}$$
$$Z_{y\ell} = g_\ell(x)g_\ell(y)X_{y\ell}f_y.$$

In words, $X_{y\ell}$ equals 1 or 0 depending whether or not $h_\ell$ has a hash collision between $y$ and $x$, and $Z_{y\ell}$ is a random variable representing the amount (positive or negative) that occurrences of token $y$ in the stream contribute to the value of $g_\ell(x) \cdot C[h_\ell(x), \ell]$. To substantiate the latter interpretation, observe that

$$C[h_\ell(x), \ell] = \sum_{y=1}^{m} g_y(\ell)X_{y\ell}f_y$$

because token $y$ occurs $f_y$ times in the stream, and each of these occurrences contribute $g_y(\ell)$ to the counter $C[h_\ell(x), \ell]$ if and only if $X_{y\ell} = 1$, otherwise each occurrence of $y$ in the stream has zero contribution to $C[h_\ell(x), \ell]$.

The random variable $Z_{x\ell}$ is deterministically equal to $f_x$ because $g_\ell(x)^2 = 1$ and $X_{x\ell} = 1$. As for $Z_{y\ell}$ when $y \neq x$, we have

$$\mathbb{E}[Z_{y\ell}] = \mathbb{E}[g_\ell(x)g_\ell(y)X_{y\ell}f_y] = \mathbb{E}[g_\ell(x)] \cdot \mathbb{E}[g_\ell(y)] \cdot \mathbb{E}[X_{y\ell}] \cdot f_y = 0, \tag{3.7}$$

where we have used the fact that $g_\ell(x)$, $g_\ell(y)$, and $X_{y\ell}$ are mutually independent, and that $\mathbb{E}[g_\ell(x)] = \mathbb{E}[g_\ell(y)] = 0$. To verify the mutual independence, observe that $X_{y\ell}$ depends only on the hash function $h_\ell$ which is independent of $g_\ell$, and the values $g_\ell(x), g_\ell(y)$ are independent of one another by the pairwise-independence property of $g_\ell$.

Using linearity of expectation we have

$$\mathbb{E}[g_\ell(x) \cdot C[h_\ell(x), \ell]] = \sum_{y=1}^{m} \mathbb{E}[Z_{y\ell}] = f_x + \sum_{y \neq x} \mathbb{E}[Z_{y\ell}] = f_x. \tag{3.8}$$

To continue with the analysis of the Count Sketch, the next step is to analyze the variance

of $g_\ell(x) \cdot C[h_\ell(x), \ell]$ and apply Chebyshev's Inequality. We have

$$\begin{aligned}\mathrm{Var}[g_\ell(x) \cdot C[h_\ell(x), \ell]] &= \mathrm{Var}[f_x + \sum_{y \neq x} Z_{y\ell}] = \mathrm{Var}[\sum_{y \neq x} Z_{y\ell}] \\ &= \mathbb{E}\left[\left(\sum_{y \neq x} Z_{y\ell}\right)^2\right] \\ &= \sum_{y \neq x} \sum_{w \neq x} \mathbb{E}[Z_{y\ell} Z_{w\ell}] = \sum_{y \neq x} \mathbb{E}[Z_{y\ell}^2] + \sum_{y \neq x} \sum_{w \notin \{x,y\}} \mathbb{E}[Z_{y\ell} Z_{w\ell}].\end{aligned}$$

Now,

$$\mathbb{E}[Z_{y\ell}^2] = \mathbb{E}[X_{y\ell}^2 f_y^2] = \mathbb{E}[X_{y\ell} f_y^2] = \tfrac{1}{B} f_y^2,$$

since $X_{y\ell} = 1$ with probability $\frac{1}{B}$ and $X_{y\ell} = 0$ otherwise. (Here we have used the fact that $h_\ell$ is drawn from a 2-universal hash family, so for any $y \neq x$ the probability of $h_\ell(y) = h_\ell(x)$ is $1/B$.) Furthermore, if $w \notin \{x, y\}$ then

$$\mathbb{E}[Z_{y\ell} Z_{w\ell}] = \mathbb{E}[g_\ell(y) g_\ell(w) X_{y\ell} X_{w\ell} f_y f_w] = \mathbb{E}[g_\ell(y)] \cdot \mathbb{E}[g_\ell(w)] \cdot \mathbb{E}[X_{y\ell} X_{w\ell}] \cdot f_y f_w = 0,$$

where we have again used the mutual independence of the random variables $g_\ell(y), g_\ell(w)$, and $X_{y\ell} X_{w\ell}$. (Note that $X_{y\ell}$ and $X_{w\ell}$ may be correlated with one another, we only need to use the fact that their product is independent of $g_\ell(y)$ and $g_\ell(w)$, which holds because $X_{y\ell} X_{w\ell}$ depends only on the hash function $h_\ell$, which is independent of $g_\ell$.) Substituting the calculated values of $\mathbb{E}[Z_{y\ell}^2]$ and $\mathbb{E}[Z_{y\ell} Z_{w\ell}]$ into the variance calculation, we find that

$$\mathrm{Var}[g_\ell(x) \cdot C[h_\ell(x), \ell]] = \frac{1}{B} \sum_{y \neq x} f_y^2 \leq \frac{1}{B} \|\mathrm{f}\|_2^2.$$

By Chebyshev's Inequality,

$$\Pr\left(|g_\ell(x) \cdot C[h_\ell(x), \ell] - f_x| \geq \sqrt{\tfrac{3}{B}}\, \|\mathrm{f}\|_2\right) \leq \frac{\mathrm{Var}[g_\ell(x) \cdot C[h_\ell(x), \ell]]}{\frac{3}{B} \|\mathrm{f}\|_2^2} = \frac{1}{3}. \quad (3.9)$$

We have shown that each of the individual estimates $g_\ell(x) \cdot C[h_\ell(x), \ell]$ has probability at most $\frac{1}{3}$ of differing from the target value $f_x$ by more than $\sqrt{3/B} \cdot \|\mathrm{f}\|_2$. There are $t$ such estimates, one for each $\ell \in [t]$, and they are independent random variables. In order for their *median* to be less than $f_x - \sqrt{3/B} \cdot \|\mathrm{f}\|_2$ or greater than $f_x + \sqrt{3/B} \cdot \|\mathrm{f}\|_2$, at least $t/2$ of the estimates must differ from $f_x$ by more than $\sqrt{3/B} \cdot \|\mathrm{f}\|_2$. To finish up, we use the Hoeffding Bound to show that the probability of this happening is less than $e^{-t/18}$. In more detail, let $W_\ell$ be a random variable which equals 1 if $|g_\ell(x) \cdot C[h_\ell(x), \ell] - f_x| \geq \sqrt{3/B} \cdot \|\mathrm{f}\|_2$, otherwise $W_\ell = 0$. Inequality (3.9) says that $\mathbb{E}[W_\ell] \leq \frac{1}{3}$. Since the random variables $\{W_\ell : \ell \in [t]\}$ are mutually independent, Hoeffding's Inequality says that

$$\Pr\left(W_1 + \cdots + W_t \geq \tfrac{t}{2}\right) = \Pr\left(W_1 + \cdots + W_t \geq \mathbb{E}[W_1 + \cdots + W_t] + \tfrac{t}{6}\right) \leq e^{-2(t/6)^2/t} = e^{-t/18}.$$

■

**Corollary 3.6** For any $\varepsilon, \delta > 0$ the Count-Min Sketch with parameters $B = \lceil \frac{3}{\varepsilon^2} \rceil$ and $t = \lceil 18 \ln(1/\delta) \rceil$ achieves the following guarantee: for any token $x$, with probability at least $1 - \delta$ the estimated frequency of $x$ differs from the true frequency by no more than $\varepsilon \|\mathrm{f}\|_2$. The space complexity of the sketch with these parameters is $O\left(\log(mn)\log(1/\delta)/\varepsilon^2\right)$.

## 3.4 Quantile estimation

The last streaming algorithm we'll present in these notes is applicable when the tokens in the stream come from an ordered set. As in previous sections, we'll assume the tokens come from the set $[m] = \{1, 2, \ldots, m\}$, ordered by the $<$ relation. For convenience, we'll assume in this section that a single token can be stored in $O(1)$ space, i.e. that a single memory location can store $\log(m)$ bits. All of the algorithms we present here can be applied even when this assumption is violated; the space complexity bounds presented here would just need to be scaled by the amount of space required to store one token from the stream.

For a stream $a_1, \ldots, a_n$ and a token $a \in [m]$, we'll say the quantile of token $a$ relative to the stream is

$$q(a) = \frac{\#\{i \mid a_i \leq a\}}{n}$$

and we'll say that $\hat{q}(a)$ is an $\varepsilon$-approximate quantile of $a$ if $|\hat{q}(a) - q(a)| \leq \varepsilon$. Algorithms for *quantile estimation* maintain a sketch that enables them to estimate an $\varepsilon$-approximate quantile of any element.

In this section we'll present a simple algorithm for quantile estimation using $s = O\left(\frac{1}{\varepsilon^2} + \log n\right)$ bits of space. The design of the algorithm illustrates an important technique called *reservoir sampling* for maintaining a random sample of elements of a data stream. To analyze the algorithm we'll introduce and apply the Glivenko-Cantelli Theorem, an important theorem in statistics about estimating a univariate distribution from random samples.

### 3.4.1 Reservoir sampling

One of the most basic tasks in data streaming is downsampling: for a specific integer $t > 0$, draw $t$ tokens uniformly at random from the multiset of $n$ tokens in the stream. To state the goal of downsampling more precisely, the algorithm should output a sequence of $t$ tokens, and the distribution of its output as an unordered multiset (i.e., ignoring the ordering of tokens in the output) should match the output distribution of a hypothetical algorithm that stores the entire stream, samples $t$ distinct indices $1 \leq i(1) < i(2) < \cdots < i(t) \leq n$ uniformly at random from among all $\binom{n}{t}$ such $t$-tuples of indices, and outputs the multiset $\{a_{i(1)}, a_{i(2)}, \ldots, a_{i(t)}\}$.

Reservoir sampling is a simple and widely used downsampling procedure that works by maintaining a buffer $\mathrm{b} = (b_1, \ldots, b_t)$ of size $t$ containing the random samples, and a counter that keeps track of the number of stream tokens seen so far. After an initialization phase that fills the buffer with the first $t$ tokens of the stream, the token observed at time $s > t$ is discarded with probability $1 - \frac{t}{s}$, and otherwise a random element of the buffer is overwritten with the new token.

**Algorithm 5** Reservoir sampling

Given positive integer $t$
**for** each $s \in [n]$ **do**
    Observe $a_s$.
    **if** $s \leq t$ **then**
        Set $b_s = a_s$.
    **else**
        With probability $\frac{t}{s}$, sample index $i \in [t]$ uniformly at random and set $b_i = a_i$.
    **end if**
**end for**
Output $(b_1, \ldots, b_t)$.

**Proposition 3.7** When Algorithm 5 is used to process a stream of $n \geq t$ tokens, the indices of the $t$ tokens it outputs constitute a uniformly random $t$-element subset of $[n]$.

*Proof.* We use induction on $n$. In the base case $n = t$ the algorithm is guaranteed to output tokens $\{a_i \mid i \in [t]\}$, and $[t]$ is the only $t$-element subset of $[t]$.

When $n > t$, consider any $t$-tuple of distinct indices $1 \leq i(1) < i(2) < \cdots < i(t) \leq n$. If $i(t) < n$, then by the induction hypothesis the probability that the buffer stores elements $a_{i(1)}, \ldots, a_{i(t)}$ at the *start* of the final loop iteration is $1 \Big/ \binom{n-1}{t}$. With probability $1 - \frac{t}{n}$ the final loop iteration does not change the buffer, so the probability of outputting $\{a_{i(1)}, \ldots, a_{i(t)}\}$ is

$$\frac{1 - t/n}{\binom{n-1}{t}} = \frac{n-t}{n} \cdot \frac{t!(n-1-t)!}{(n-1)!} = \frac{t!(n-t)!}{n!} = \frac{1}{\binom{n}{t}}.$$

On the other hand, if $i(t) = n$, then the algorithm outputs $\{a_{i(1)}, \ldots, a_{i(t)}\}$ if and only if its buffer contents at the start of the final loop iteration are indexed by $t$-tuple $1 \leq j(1) < j(2) < \cdots < j(t) \leq n-1$ such that the set $J_t = \{j(1), \ldots, j(t)\}$ contains $I_{t-1} = \{i(1), \ldots, i(t-1)\}$ as a subset. The number of $t$-element sets $J_t \subseteq [n]$ containing $I_{t-1}$ as a subset is $n - t$, because $J_t \setminus I_{t-1}$ can be any of the $n - t$ elements of $[n-1] \setminus I_{t-1}$. For each such set $J_t$, the probability of overwriting the unique element of $J_t \setminus I_{t-1}$ in the final iteration is $\frac{t}{n} \cdot \frac{1}{t} = \frac{1}{n}$. Since there are $n - t$ potential sets $J_t$, each having $1 \Big/ \binom{n-1}{t}$ of being in the buffer at the start of the final loop iteration, and overwriting $J_t \setminus I_{t-1}$ with $i(t) = n$ in the final iteration has probability $1/n$, we find that the probability of outputting $\{a_{i(1)}, a_{i(2)}, \ldots, a_{i(t)}\}$ is

$$(n-t) \cdot \frac{1}{\binom{n-1}{t}} \cdot \frac{1}{n} = \frac{n-t}{n} \cdot \frac{t!(n-1-t)!}{(n-1)!} = \frac{t!(n-t)!}{n!} = \frac{1}{\binom{n}{t}}.$$

■

### 3.4.2 Quantile estimation via reservoir sampling

Reservoir sampling leads to a very natural idea for quantile estimation: if the buffer $(b_1, \ldots, b_t)$ is a uniformly random $t$-element subset of the full stream, then it should constitute a representative sample of the full stream. If so, a good way to estimate the

quantile of any token $a \in [m]$ with respect to the full stream is to calculate its quantile with respect to the tokens contained in the buffer.

The analysis of the quantile estimation algorithm will go more smoothly if we make a small modification to it. Rather than maintaining one reservoir sample of size $t$, our algorithm will maintain $t$ independent reservoir samples, each of size 1. Reservoir sampling with $t = 1$ requires $O(1)$ space for the buffer and $O(\log n)$ space for the counter, so $t$ independent reservoir sampling algorithms with a shared counter use $O(t + \log n)$ space.

Given $t$ independent samples $b_1, \ldots, b_t$, each uniformly distributed over the multiset of tokens in the stream, the quantile estimate for any token $a \in [m]$ is

$$\hat{q}(a) = \frac{\#\{j \mid b_j \leq a\}}{t}. \tag{3.10}$$

**Proposition 3.8** If $b_1, \ldots, b_t$ are independent random samples from the stream $a_1, \ldots, a_n$, each uniformly distributed over its $n$ tokens, then for any $a \in [m]$ and $\varepsilon > 0$, the quantile estimate $\hat{q}(a)$ computed by Equation (3.10) satisfies $|\hat{q}(a) - q(a)| \leq \varepsilon$ with probability at least $1 - 2e^{-2\varepsilon^2 t}$.

*Proof.* We can bound the probability that the estimate differs from $q(a)$ by more than $\varepsilon$ using Hoeffding's Inequality. For $1 \leq j \leq t$, let $X_j = 1$ if $b_j \leq a$ and $X_j = 0$ otherwise. The random variables $\{X_j\}$ are independent and $\{0,1\}$-valued, and each has expected value $\mathbb{E}[X_j] = q(a)$. The event that $|\hat{q}(a) - q(a)| > \varepsilon$ is equivalent to the event that the sum $X = X_1 + \cdots + X_t$ satisfies $|X - \mathbb{E}[X]| > \varepsilon t$. Hence,

$$\Pr\left(|\hat{q}(a) - q(a)| > \varepsilon\right) = \Pr\left(|X - \mathbb{E}[X]| > \varepsilon t\right) < 2\exp\left(-\frac{2\varepsilon^2 t^2}{t}\right) = 2e^{-2\varepsilon^2 t}.$$

■

As a corollary, if we want each quantile query to have an $(\varepsilon, \delta)$-PAC answer, then $t \geq \frac{1}{2}\varepsilon^{-2}\ln(2/\delta)$ independent uniformly random samples are sufficient.

### 3.4.3 Uniformly accurate quantile sketches via Glivenko-Cantelli

One can interpret the buffer of samples $(b_1, \ldots, b_t)$ used in Section 3.4.2 as a sketch that supports answering quantile queries via the formula (3.10). In light of that interpretation, Proposition 3.8 says that for any *particular* quantile query $q(a)$, the response $\hat{q}(a)$ is $(\varepsilon, \delta)$-PAC when $t \geq \frac{1}{2}\varepsilon^{-2}\ln(2/\delta)$. However, we can ask for more: the sketch $(b_1, \ldots, b_t)$ is *uniformly $\varepsilon$-accurate for quantile queries* if $|\hat{q} - q|_\infty \leq \varepsilon$. In other words, a uniformly $\varepsilon$-accurate sketch supplies quantile estimates $\hat{q}$ that satisfy $|\hat{q}(a) - q(a)| \leq \varepsilon$ *simultaneously* for all $a \in [m]$.

Let us say an algorithm for quantile sketching is *uniformly $(\varepsilon, \delta)$-PAC* if, for all input streams, the probability that the algorithm produces a uniformly $\varepsilon$-accurate sketch is at least $1 - \delta$.

A simple application of the union bound and Proposition 3.8 leads to the conclusion that $t \geq \frac{1}{2}\varepsilon^{-2}\ln(2m/\delta)$ independent uniformly random samples suffice for a uniformly $(\varepsilon, \delta)$-PAC quantile sketch. That's because when $t$ satisfies the specified lower bound, for any $a \in [m]$ an application of Proposition 3.8 ensures that

$$\Pr\left(|\hat{q}(a) - q(a)| > \varepsilon\right) \leq \frac{\delta}{m}.$$

Summing over $a \in [m]$ and applying the union bound, we find that the quantile sketch obtained from $(b_1, \ldots, b_t)$ is uniformly $(\varepsilon, \delta)$-PAC. However, setting $t \geq \frac{1}{2}\varepsilon^{-2}\ln(2m/\delta)$ results in a sketch whose space complexity is logarithmic in $m$. A more careful analysis will justify using a number of samples with *no dependence on m at all*. The key is the following inequality for independent samples drawn from any univariate distribution.

**Lemma 3.9** Let $b_1, \ldots, b_t$ be $t$ independent, identically distributed random numbers each with cumulative distribution function $F$. The function $\hat{q}$ defined by formula (3.10) satisfies

$$\forall \varepsilon > 0 \quad \Pr\left(|\hat{q} - F|_\infty > \varepsilon\right) < \frac{4}{\varepsilon} e^{-\varepsilon^2 t/2}. \tag{3.11}$$

*Proof.* Consider any $\varepsilon > 0$, and let $k = \lceil 2/\varepsilon \rceil$. For $1 \leq i \leq k$ let

$$a^{(i)} = \sup\left\{a \mid F(a) < \tfrac{i}{k}\right\}.$$

By the definition of $a^{(i)}$, we have

$$\forall a < a^{(i)} \quad F(a) < \tfrac{i}{k}. \tag{3.12}$$

Furthermore, since cumulative distribution functions are right-continuous, we have

$$F(a^{(i)}) = \inf\{F(a) \mid a > a^{(i)}\} \geq \frac{i}{k}. \tag{3.13}$$

For each $a \in \mathbb{R}$ define $\hat{q}(a)$ as in Equation (3.10) and define

$$\check{q}(a) = \frac{\#\{j \mid b_j < a\}}{t}. \tag{3.14}$$

Let $F(a^{(i)}-)$ denote $\sup\{F(a) \mid a < a^{(i)}\}$, the probability of sampling a value strictly less than $a^{(i)}$ under the common distribution of $b_1, \ldots, b_t$. We claim that whenever $\|\hat{q} - F\|_\infty > \varepsilon$ there exists some $i \in [k-1]$ such that either $F(a^{(i)}) - \hat{q}(a^{(i)}) > \varepsilon/2$ or $\check{q}(a^{(i)}) - F(a^{(i)}-) > \varepsilon/2$. To prove the claim, we suppose $|\hat{q}(a) - F(a)| > \varepsilon$ for some $a \in \mathbb{R}$ and perform the following case analysis.

1. If $F(a) - \hat{q}(a) > \varepsilon$ then let $i/k$ denote the greatest multiple of $1/k$ less than $F(a)$. Note that $i \leq k-1$ because $\frac{i}{k} < F(a) \leq 1$. We have $F(a^{(i)}) \geq \frac{i}{k}$ by (3.13) whereas $\hat{q}(a^{(i)}) \leq \hat{q}(a)$ by monotonicity of $\hat{q}$. Hence, subtracting the two inequalities,

$$F(a^{(i)}) - \hat{q}(a^{(i)}) \geq \tfrac{i}{k} - \hat{q}(a) \geq F(a) - \tfrac{1}{k} - \hat{q}(a) > \varepsilon - \tfrac{1}{k} > \tfrac{\varepsilon}{2}.$$

2. If $\hat{q}(a) - F(a) > \varepsilon$ then let $i/k$ denote the least multiple of $1/k$ greater than $F(a)$. Note that $i \leq k-1$ because

$$\frac{i+1}{k} \leq F(a) + \frac{2}{k} \leq F(a) + \varepsilon < \hat{q}(a) \leq 1.$$

Since $F(a) < \frac{i}{k}$ we have $a < a^{(i)}$ and hence $\hat{q}(a) \leq \check{q}(a^{(i)})$. Then,

$$\check{q}(a^{(i)}) - F(a^{(i)}-) \geq \hat{q}(a) - F(a^{(i)}-) \geq \hat{q}(a) - \tfrac{i}{k} \geq \hat{q}(a) - F(a) - \tfrac{1}{k} > \varepsilon - \tfrac{1}{k} > \tfrac{\varepsilon}{2}.$$

For each $i \in [k]$ an application of Hoeffding's inequality as in Proposition 3.8 ensures that

$$\Pr\left(F(a^{(i)}) - \hat{q}(a^{(i)}) > \varepsilon/2\right) < e^{-\varepsilon^2 t/2}$$

and

$$\Pr\left(\check{q}(a^{(i)}) - F(a^{(i)}-) > \varepsilon/2\right) < e^{-\varepsilon^2 t/2}$$

By the union bound,

$$\Pr\left(\exists i \in [k-1]\, F(a^{(i)}) - \hat{q}(a^{(i)}) > \varepsilon/2 \text{ or } \check{q}(a^{(i)}) - F(a^{(i)}-) > \varepsilon/2\right) < 2(k-1)e^{-\varepsilon^2 t/2} < \frac{4}{\varepsilon} e^{-\varepsilon^2 t/2} \tag{3.15}$$

where the second inequality follows because $k - 1 < 2/\varepsilon$, by our choice of $k$. ■

Lemma 3.9 has two corollaries that are worthy of note. The first is an upper bound with on the number of samples needed for uniformly $(\varepsilon, \delta)$-PAC quantile sketching. As promised, the bound has no dependence on $m$, the size of the set from which tokens are drawn.

**Corollary 3.10** If $t \geq 2\varepsilon^{-2} \ln\left(\frac{4}{\varepsilon\delta}\right)$ then the quantile sketch obtained from $t$ independent uniformly-distributed stream tokens is uniformly $(\varepsilon, \delta)$-PAC.

The second corollary of Lemma 3.9 is the Glivenko-Cantelli Theorem, an important theorem in statistics asserting that the empirical distribution of $n$ independent, identically distributed random variables converges uniformly to the distributon from which the variables were drawn, as $n$ tends to infinity.

**Theorem 3.11** If $X_1, X_2, \ldots$ is an infinite sequence of independent, identically distributed random variables, each with cumulative distribution function $F$, and if $F_t$ denotes the empirical cumulative distribution function

$$F_t(a) = \frac{\sharp\{i \mid X_i \leq a \text{ and } 1 \leq i \leq t\}}{t}$$

then $\lim_{t\to\infty} \|F_t - F\|_\infty = 0$ almost surely.

*Proof.* The function $F_t$ is identical to the quantile estimate $\hat{q}$ determined by the buffer $(b_1, \ldots, b_t) = (X_1, \ldots, X_t)$. Hence, for any $\varepsilon > 0$ we can apply Lemma 3.9 and linearity of expectation to conclude that the expected number of $t$ such that $\|F_t - F\|_\infty > \varepsilon$ is less than

$$\frac{4}{\varepsilon} \sum_{t=1}^{\infty} e^{-\varepsilon^2 t/2} = \frac{4}{\varepsilon} \cdot \frac{1}{1 - e^{-\varepsilon^2/2}} < \infty.$$

By the Borel-Cantelli Lemma, the number of $t$ such that $\|F_t - F\|_\infty > \varepsilon$ is almost surely finite. In other words,

$$\Pr\left(\limsup_{t\to\infty} \|F_t - F\|_\infty > \varepsilon\right) = 0.$$

Since the above probability equals zero for every $\varepsilon$, it follows that $\lim_{t\to\infty} \|F_t - F\|_\infty = 0$ almost surely. ■

One concluding remark is that the bound in Lemma 3.9 can be tightened further. The sharp bound is called the Dvoretzky-Kiefer-Wolfowitz Inequality (or DKW Inequality) and takes the following form.

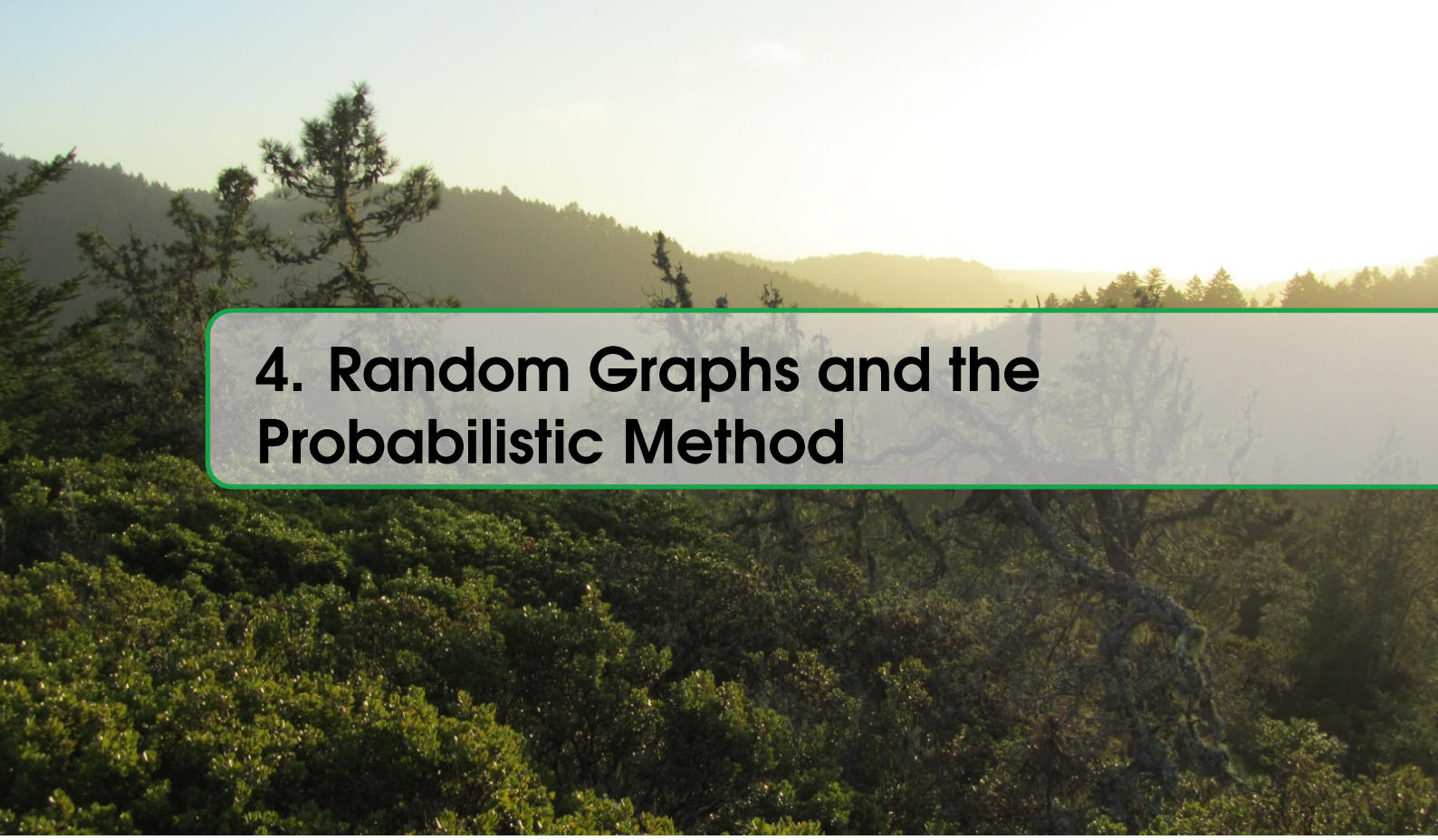
$$\forall \varepsilon > 0 \quad \Pr\left(\|\hat{q} - F\|_\infty > \varepsilon\right) < 2e^{-2\varepsilon^2 t}. \tag{3.16}$$

The inequality is quite remarkable, because the right-hand side is the same as the upper bound on $\Pr\left(|\hat{q}(a) - F(a)| > \varepsilon\right)$ for a *single* value of $a$ derived using Hoeffding's Inequality in Proposition 3.8. The DKW Inequality pertains of to the union of the events $\{|\hat{q}(a) - F(a)| > \varepsilon\}$ for *uncountably many* values of $a$, yet somehow taking the union of all these events comes at no cost in terms of the probability bound.

The DKW inequality justifies that the reservoir-sampling-based quantile sketch is uniformly $(\varepsilon, \delta)$-PAC as long as $t \geq \frac{1}{2}\varepsilon^{-2}\ln(2/\delta)$. The sketching algorithm runs in

$$s = O(\varepsilon^{-2}\log(1/\delta) + \log_m(n))$$

bits of space, with the $\log_m(n)$ bits being needed for the counter in the reservoir sampling algorithm. A different quantile estimation algorithm, due to Greenwald and Khanna, runs in $O(\varepsilon^{-1}\log(\varepsilon n))$ space, improving the dependence on $1/\varepsilon$ from quadratic to quasi-linear at the cost of slightly worse dependence on $\log(n)$.



# 4. Random Graphs and the Probabilistic Method

The study of random graphs began in the middle of the twentieth century, and it led to a flood of important results in combinatorics and, later on, became a benchmark model for average-case analysis of algorithms. These notes introduce random graphs and present some of their basic combinatorial and algorithmic properties.

## 4.1 The Erdős-Rényi Models

The two most fundamental models of random graphs are denoted by $G(n,p)$ and $G(n,m)$. Both are named *Erdős-Rényi Models*, after the two Hungarian mathematicians who founded and popularied random graph theory. They were the first to write about $G(n,m)$. Interestingly, it turns out that the first person to write about $G(n,p)$ was neither Erdős nor Rényi, but another mathematician named Edgar Gilbert. (He is also a namesake of the Gilbert-Varshamov bound in coding theory.)

The models are very similar. The $G(n,m)$ model denotes the uniform distribution on undirected graphs with $n$ vertices and $m$ edges, whereas $G(n,p)$ is a random graph with $n$ vertices, where $p$ is the probability that any of the $\binom{n}{2}$ pairs of vertices is present in the edge set of the graph, and the presence or absence of different edges are mutually independent random variables. In other words, to randomly sample a graph from the $G(n,p)$ model one samples $\binom{n}{2}$ independent Bernoulli random variables $X_{\{u,v\}}$, each with expected value $p$, and the edge set of the graph is defined to be the set of vertex pairs $\{u,v\}$ such that $X_{\{u,v\}} = 1$.

The expected number of edges in $G(n,p)$ is $p\binom{n}{2}$. As long as $p$ is not very close to zero, the ratio of the sampled number of edges to its expected value is very unlikely to lie outside of $[1-\varepsilon, 1+\varepsilon]$. By the Chernoff bound, the probability of this event is exponentially small in $n$, as long as $p > 1/n$. So the $G(n,p)$ and $G(n,m)$ models tend to behave very similarly when $m = p\binom{n}{2}$. It's usually more convenient to work with $G(n,p)$, so we will focus on

that model in these notes.

Closely related to both $G(n,p)$ and $G(n,m)$ is the *evolving random network process*, which is the random sequence of $n$-vertex graphs obtained by starting with an empty graph and introducing edges one by one, in randomly-permuted order, until all $\binom{n}{2}$ vertex pairs have been connected by edges. If the graphs in this random sequence are numbered $G_0, G_1, \ldots, G_{n(n-1)/2}$, then the $m^{\text{th}}$ graph in the sequence is a random sample from the $G(n,m)$ distribution.

## 4.2 Connectivity, diameter, and expansion

Recall that a graph $G$ is called *connected* if every two vertices can be joined by a path in $G$. The *diameter* of a connected graph is the smallest $d$ such that every two vertices can be joined by a path made up of $d$ or fewer edges. A first question one can ask about $G(n,p)$ is: *what is the probability that a random sample from the $G(n,p)$ distribution is a connected graph?* When $p$ is large enough that $G(n,p)$ is connected with high probability, one can also ask about the distribution of the random graph's diameter.

We will see that there exist constants $0 < a < b < \infty$ such that when $p < \frac{a \log n}{n}$, the graph $G(n,p)$ is disconnected with probability very close to 1, while for $p > \frac{b \log n}{n}$, with probability very close to 1, $G(n,p)$ is connected and has diameter $O(\log n)$.

### 4.2.1 Isolated vertices in $G(n,p)$

It turns out that isolated vertices constitute the main obstruction to $G(n,p)$ being connected. This statement can be given a precise interpretation in terms of the evolving random network process $G_0, G_1, \ldots, G_{n(n-1)/2}$. One can define $m_{\text{conn}}(n)$ to be the least $m$ such that $G_m$ is connected, and $m_{\text{isol}}(n)$ to be the least $m$ such that $G_m$ has no isolated vertices. It has been proven that $m_{\text{isol}}(n) = m_{\text{conn}}(n)$ with probability tending to 1 as $n \to \infty$. The proof of this result is beyond the scope of these notes, but we will present a simpler analysis that gives less precise bounds on the probability that $G(n,p)$ has an isolated vertex.

For any vertex $v$ of $G(n,p)$ let $X_v$ denote the Bernoulli random variable that equals 1 if and only if $v$ is isolated. Since $v$ has $n-1$ potential neighbors, we have

$$\mathbb{E}[X_v] = (1-p)^{n-1}.$$

If $u$ and $v$ are any two distinct vertices, there are $2n-3$ vertex pairs that intersect $\{u,v\}$: one of them is $\{u,v\}$ itself, and the other $2n-4$ are the pairs composed of one vertex in $\{u,v\}$ and one chosen from among the $n-2$ other vertices. Hence, the probability that both $u$ and $v$ are isolated is

$$\mathbb{E}[X_u X_v] = (1-p)^{2n-3}.$$

For the number of isolated vertices, $X = \sum_{v \in V} X_v$, we have

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{v \in V} \mathbb{E}[X_v] = n(1-p)^{n-1} \\
\mathrm{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \sum_{v \in V} \sum_{u \in V} \mathbb{E}[X_u X_v] - \mathbb{E}[X_u]\mathbb{E}[X_v] \\
&= \sum_{v \in V} \mathrm{Var}[X_v] + \sum_{v \in V} \sum_{u \in V \setminus \{v\}} (1-p)^{2n-3} - (1-p)^{2n-2} \\
&= n(1-p)^{n-1}(1-(1-p)^{n-1}) + pn(n-1)(1-p)^{2n-3} \\
&= (1-(1-p)^{n-1} + p(n-1)(1-p)^{n-2})\mathbb{E}[X] = (1+(pn-1)(1-p)^{n-2})\mathbb{E}[X] \leq pn\,\mathbb{E}[X].
\end{aligned}$$

When $p \geq \frac{\ln(n)+c}{n-1}$ we have

$$\Pr(X > 0) \leq \mathbb{E}[X] = n(1-p)^{n-1} < ne^{-p(n-1)} = e^{-c}, \tag{4.1}$$

so the probability of $G(n,p)$ having an isolated vertex converges to zero exponentially fast as $p$ exceeds $\frac{\ln(n)}{n-1}$. On the other hand, when $p \leq \frac{\ln(n)}{2n-2}$ and $n \geq 2$, then $p < \frac{1}{2}$ so $1-p \geq 4^{-p}$. Now, by Chebyshev's Inequality,

$$\Pr(X = 0) \leq \frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2} \leq \frac{pn\,\mathbb{E}[X]}{\mathbb{E}[X]^2} = \frac{pn}{n(1-p)^{n-1}} = \frac{p}{(1-p)^{n-1}} \leq p \cdot 4^{p(n-1)} \leq p \cdot n^{\ln(2)} \leq n^{-0.3}, \tag{4.2}$$

so the probability of $G(n,p)$ having an isolated vertex converges to 1 as $n \to \infty$ and $p \leq \frac{\ln n}{2n-2}$.

An immediate corollary of our analysis of isolated vertices is the following proposition about the probability that $G(n,p)$ is disconnected.

**Proposition 4.1** When $p \leq \frac{\ln n}{2n-2}$, the probability that $G(n,p)$ is disconnected converges to 1 as $n$ tends to infinity.

*Proof.* We have shown that the probability of $G(n,p)$ containing an isolated vertex converges to 1 as $n \to \infty$. A graph with $n$ vertices that has at least one isolated vertex is disconnected as long as $n > 1$. ■

### 4.2.2 Connectedness of $G(n,p)$

In this section we prove that for $p \geq \frac{3\ln n}{n}$, with probability $1 - o(1)$ the graph $G(n,p)$ is connected. The proof is based on the observation that a graph is disconnected if and only if its vertex set can be partitioned into two non-empty subsets, $A$ and $B$, such that the graph has no edges joining $A$ to $B$. Call such a partition a "disconnecting partition". When $p \geq \frac{3\ln n}{n}$ we can exhaustively consider all partitions of the vertex set into two non-empty subsets. Although there are exponentially many such partitions, we can use linearity of expectation to show that the expected number of disconnecting partitions is much less than 1.

Suppose $A, B$ is a partition of $G$ into two non-empty vertex sets with $|A| \leq |B|$. Let $k = |A|$ and observe that, since $k \leq n-k$, we must have $k \leq n/2$ and $n-k \geq n/2$. Since

there are $k(n-k)$ vertex pairs $(u,v)$ with $u \in A$, $v \in B$, the probability that $A$ and $B$ form a disconnecting partition is

$$(1-p)^{k(n-k)} \le (1-p)^{kn/2} \le \left(1 - \frac{3\ln n}{n}\right)^{kn/2} < \exp\left(-\frac{3\ln n}{n} \cdot \frac{kn}{2}\right) = e^{-\frac{3}{2}k\ln n} = n^{-3k/2}.$$

Summing over all partitions with $1 \le |A| = k \le n/2$, we find that

$$\begin{aligned}
\Pr(G(n,p) \text{ is disconnected}) &\le \mathbb{E}[\text{number of disconnecting partitions}] \\
&= \sum_{k=1}^{n/2} \binom{n}{k} (1-p)^{k(n-k)} \\
&< \sum_{k=1}^{n/2} \binom{n}{k} n^{-3k/2} \\
&< -1 + \sum_{k=0}^{n} \binom{n}{k} n^{-3k/2} \\
&= -1 + (1+n^{-3/2})^n < -1 + \left(e^{n^{-3/2}}\right)^n = e^{n^{-1/2}} - 1 < \frac{e-1}{n^{1/2}}
\end{aligned}$$

where the last inequality follows from the fact that $e^x - 1 < (e-1)x$ whenever $0 < x < 1$.

To sum up, when $p \ge \frac{3\ln n}{n}$, the probability that $G(n,p)$ is disconnected is $O(n^{-1/2})$. Combining this with the observation from Section 4.2.1 that $G(n,p)$ contains isolated vertices with probability $1-o(1)$ when $p \le \frac{\ln n}{2n-2}$, we have shown that the transition from disconnectivity to connectivity in $G(n,p)$ occurs when $p$ is somewhere in the range $\left[\frac{1}{2} \cdot \frac{\ln n}{n}, \frac{3\ln n}{n}\right]$. Using more sophisticated methods, mathematicians have shown that the connectivity transition occurs in the range $\left[\frac{(1-\varepsilon)\ln n}{n}, \frac{(1+\varepsilon)\ln n}{n}\right]$ for any $\varepsilon > 0$. In other words, when $p$ is below this range $G(n,p)$ is disconnected with probability $1-o(1)$ as $n \to \infty$, whereas when $p$ is above this range $G(n,p)$ is connected with probability $1-o(1)$ as $n \to \infty$.

### 4.2.3 Diameter and expansion of $G(n,p)$

In this section we prove that for $p \ge \frac{7\ln n}{n}$, with probability $1-o(1)$ the graph $G(n,p)$ is connected and has diameter $O(\log n)$. The method of proof will in fact show that an even stronger property holds with probability $1-o(1)$, namely that $G(n,p)$ is a *vertex expander*.

**Definition 4.2** A graph $G$ with $n$ vertices is an $\alpha$-expander if for every non-empty vertex set $S$ with $|S| \le n/2$, the set

$$\partial S = \{v \notin S \mid \exists \text{edge } (u,v) \text{ with } u \in S\}$$

has at least $\alpha|S|$ elements.

**Lemma 4.3** If $G$ is an $\alpha$-expander with $n$ vertices then its diameter is at most $\frac{2\ln(n)}{\ln(1+\alpha)}$.

*Proof.* For any vertex $s \in V(G)$ and any integer $r \ge 0$ let $B_r(s)$ denote the set of vertices that can be joined to $s$ by a path composed of $r$ or fewer edges. Trivially, $B_0(s) = \{s\}$. Furthermore, for each $r$ such that $|B_r(s)| \le n/2$, the relation $|B_{r+1}(s)| \ge (1+\alpha)|B_r(s)|$

holds by the $\alpha$-expansion property, applied to the set $S = B_r(s)$. This is because for each $v \notin S$ such that there exists an edge $(u,v)$ with $u \in S$, we can form a path from $s$ to $u$ composed of at most $r$ edges, and then we can append the edge $(u,v)$ to the end of this path, to obtain a path from $s$ to $v$ composed of at most $r+1$ edges.

Applying the inequality $|B_{r+1}(s)| \geq (1+\alpha)|B_r(s)|$ inductively starting from the base case $r = 0$, we find that $|B_{r+1}(s)| \geq (1+\alpha)^{r+1}$ for all $r$ such that $|B_r(s)| \leq n/2$. Contrapositively, if $|B_{r+1}(s)| < (1+\alpha)^{r+1}$ then $|B_r(s)| > n/2$. Now, when $r = \left\lfloor \frac{\ln(n)}{\ln(1+\alpha)} \right\rfloor$ we have $(1+\alpha)^{r+1} > n$, so clearly $|B_{r+1}(s)| < (1+\alpha)^{r+1}$. Consequently, $|B_r(s)| > n/2$.

To deduce the bound on the diameter of $G$, consider any two vertices, $s$ and $t$. For $r = \left\lfloor \frac{\ln(n)}{\ln(1+\alpha)} \right\rfloor$, we have shown that the sets $B_r(s)$ and $B_r(t)$ each have strictly more than $n/2$ elements. Since their union has at most $n$ elements, their intersection must be non-empty. That means there is a vertex $u$ that can be joined to each of $s$ and $t$ by a path composed of at most $r$ edges. Concatenating these two paths together (and removing loops if necessary) we obtain a path from $s$ to $t$ composed of at most $2r$ edges, which confirms the stated bound on the diameter. ■

**Proposition 4.4** If $p \geq \frac{7\ln(n)}{n}$ then $G(n,p)$ is a $\frac{1}{2}$-expander with probability $1 - o(1)$.

*Proof.* If $G$ is any graph with $n$ vertices, we will say that a pair of vertex sets $S \subseteq T$ are "bad" if the following two conditions are satisfied.

1. $G$ has no edges from $S$ to the complement of $T$
2. $\frac{2}{3}|T| < |S| \leq \frac{n}{2}$

If $S \subseteq T$ is a bad pair then $G$ is not a $\frac{1}{2}$-expander: the first condition ensures that $\partial S \subseteq T \setminus S$, while the second condition implies that $|T \setminus S| < \frac{1}{2}|S|$ while $|S| \leq \frac{n}{2}$. Conversely, if $G$ is not a $\frac{1}{2}$-expander, then there must exist at least one bad pair: take any $S$ such that $|S| \leq n/2$ and $|\partial S| < \frac{1}{2}|S|$, and let $T = S \cup \partial S$.

Since $\frac{1}{2}$-expanders are *precisely* the graphs that have no bad pairs, the proposition can be restated as asserting that when $p \geq \frac{7\ln(n)}{n}$ the probability that $G(n,p)$ has a bad pair of vertex sets is $o(1)$. We will prove this by calculating the expected number of bad pairs and showing that it is $o(1)$.

For any specific pair of vertex sets $S \subseteq T$ with $|T| = k$ and $|S| = \ell$, the probability that $G(n,p)$ has no edges from $S$ to the complement of $T$ is $(1-p)^{\ell(n-k)}$. The cardinality constraints on bad pairs stipulate that $\frac{2}{3}k < \ell \leq \frac{1}{2}n$, so $k < \frac{3}{4}n$ and

$$\ell(n-k) > \frac{2k}{3} \cdot \frac{n}{4} = \frac{kn}{6}$$
$$(1-p)^{\ell(n-k)} < (1-p)^{kn/6} < \exp\left(-\frac{7\ln(n)}{n} \cdot \frac{kn}{6}\right) = n^{-7k/6}.$$

For any $k < \frac{3}{4}n$, there are $\binom{n}{k}$ sets $T$ of size $k$, and for each such $T$ we can bound the number of subsets of size greater than $\frac{2}{3}k$ by the total number of subsets of $T$, which is $2^k$.

Hence,

$$\mathbb{E}[\sharp \text{ bad pairs}] \leq \sum_{k=1}^{3n/4} \binom{n}{k} 2^k n^{-7k/6} < -1 + \sum_{k=0}^{n} \binom{n}{k} 2^k n^{-7k/6}$$
$$= -1 + \left(1 + \frac{2}{n^{7/6}}\right)^n < -1 + \left(e^{2/n^{7/6}}\right)^n = e^{2/n^{1/6}} - 1 < \frac{2(e-1)}{n^{1/6}}$$

where, in the final step, we made use of the inequality $e^x - 1 < (e-1)x$ which is valid whenever $0 < x < 1$. We have shown that the expected number of bad pairs is $o(1)$ which concludes the proof that $G(n,p)$ is a $\frac{1}{2}$-expander with probability $1 - o(1)$. ■

In presenting Proposition 4.4 we favored the succinctness of the proof over the sharpness of the constants. In particular, the constant 7 in our lower bound for $p$ is far from the best possible: for any constant $c > 1$ there exists an $\alpha > 0$ such that when $p = \frac{c\ln(n)}{n}$ the graph $G(n,p)$ is an $\alpha$-expander with probability tending to 1 as $n \to \infty$.

## 4.3 Ramsey's Theorem and the Probabilistic Method

In analyzing any large dataset, one hopes to find patterns reflecting meaningful and generalizable properties of the population or process from which the data originated. However, one must be careful to distinguish the "signal" from the "noise": patterns that are present in the data either because of random coincidence or because their presence is mathematically inevitable. Ramsey Theory is the branch of mathematics that seeks to understand which patterns are mathematically inevitable. The following story, related by Noga Alon and Michael Krivelevich in *The Princeton Companion to Mathematics*, illustrates the point nicely.

> In the course of an examination of friendship between children some fifty years ago, the Hungarian sociologist Sandor Szalai observed that among any group of about twenty children he checked he could always find four children any two of whom were friends, or else four children no two of whom were friends. Despite the temptation to try to draw sociological conclusions, Szalai realized that this might well be a mathematical phenomenon rather than sociological one. Indeed, a brief discussion with the mathematicians Erdős, Turán, and Sós convinced him this was the case.

To discuss the pattern Szalai noted, it is helpful to define the following terminology.

**Definition 4.5** In an undirected graph $G$, a set of vertices $S$ is called a *clique* if every two elements of $S$ are joined by an edge in $G$. The set $S$ is called an *independent set* if no two elements of $S$ are joined by an edge in $G$.

If one wants to prove that any group of twenty children either contains four children any two of whom are friends, or four children no two of whom are friends, the logic turns out to be quite messy. The following simpler proposition illustrates the type of reasoning involved.

**Proposition 4.6** In any graph with six vertices, there are three vertices that form a clique or an independent set.

*Proof.* Consider any vertex, $u$. Of the five other vertices, either a majority of them are neighbors of $u$ or a majority of them are not neighbors of $u$.

Suppose a majority of the five other vertices are neighbors of $u$. That means $u$ has at least three distinct neighbors, $v_1, v_2, v_3$. If any two of $v_1, v_2, v_3$ are joined to each other by an edge, then those two vertices together with $u$ form a clique. Otherwise, $\{v_1, v_2, v_3\}$ is an independent set.

Now suppose a majority of the five other vertices are not neighbors of $u$. That means there are at least three distinct vertices — call them $w_1, w_2, w_3$ — none of whom are joined to $u$ by an edge. If any two of these three vertices are not connected to each other, then those two vertices together with $u$ form an independent set. Otherwise, $\{w_1, w_2, w_3\}$ is a clique. ■

Generalizing Proposition 4.6, one can define $R(k,\ell)$ to be the minimum $n$ such that every undirected graph with $n$ or more vertices has either a clique of size $k$ or an independent set of size $\ell$, and one can ask whether $R(k,\ell)$ is finite for every $k$ and $\ell$, and if so, how large can it be? The finiteness of $R(k,\ell)$ is called *Ramsey's Theorem*.

**Theorem 4.7 — Ramsey's Theorem.** If $n \geq 2^{k+\ell-3}$ then every graph on $n$ vertices must either contain a clique of size $k$ or an independent set of size $\ell$. In other words, $R(k,\ell) \leq 2^{k+\ell-3}$.

*Proof.* Let $t = k+\ell-3$. We will iteratively construct a sequence of vertices $v_0, v_1, \ldots, v_t$ and vertex sets $S_0, S_1, \ldots, S_t$, with the following properties.

1. $v_i \in S_i$ for $0 \leq i \leq t$.
2. $v_i \notin S_{i+1}$ for $0 \leq i < t$.
3. $S_i \supset S_{i+1}$ for $0 \leq i < t$.
4. For $0 \leq i < t$, either every element of $S_{i+1}$ is a neighbor of $v_i$ in $G$, or none of the elements of $S_{i+1}$ are neighbors of $v_i$ in $G$.
5. $S_i$ contains at least $2^{t-i}$ vertices, for $0 \leq i \leq t$.

The sequences are constructed as follows. First let $v_0$ be an arbitrary vertex of $G$ and let $S_0 = V(G)$ be the set of all vertices of $G$. Now, for $i = 0, 1, \ldots, t-1$, assume $v_0, \ldots, v_i$ and $S_0, \ldots, S_i$ have already been constructed. Then, the set $S_i \setminus \{v_i\}$ has at least $2^{t-i}-1$ elements. If at least half of its elements are neighbors of $v_i$, then let $S_{i+1}$ be the set of all neighbors of $v_i$ in $S_i \setminus \{v_i\}$. Otherwise, let $S_{i+1}$ be the set of all elements of $S_i \setminus \{v_i\}$ that are not neighbors of $v_i$. In either case, $S_{i+1}$ has at least $2^{t-i-1}$ elements, which confirms that $S_{i+1}$ satisfies the fifth property listed above. By construction, it also satisfies the third and fourth properties. Since $2^{t-i-1} \geq 1$, we know $S_{i+1}$ is non-empty, so we can let $v_{i+1}$ be any vertex of $S_{i+1}$. Then, by construction, $v_{i+1}$ belongs to $S_{i+1}$ but $v_i$ does not, confirming the first two properties listed above.

Now, observe that $v_0, v_1, \ldots, v_t$ must all be distinct from one another: if $i < j$ then $v_i$ cannot equal $v_j$ because $v_i \notin S_{i+1}$ whereas $v_j \in S_j \subseteq S_{i+1}$. Assign a color to each vertex in the sequence except $v_t$, according to the following rule. If $v_i$ is adjacent to every element of $S_{i+1}$ then color $v_i$ blue; if $v_i$ is adjacent to none of the elements of $S_{i+1}$ then color $v_i$

red. Of the $t = k+\ell-3$ vertices in the sequence $v_0, \ldots, v_{t-1}$, either the number of blue vertices exceeds $k-2$ or the number of red vertices exceeds $\ell-2$. If there are $k-1$ blue vertices in the sequence. Then those $k-1$ blue vertices, together with $v_t$, form a clique in $G$ of size $k$. If there are $\ell-1$ red vertices in the sequence, then those $\ell-1$ red vertices, together with $v_t$, form a clique in $G$ of size $\ell$. ■

Ramsey's Theorem has been known since the 1930's, but there are very few pairs $(k,\ell)$ for which the exact value of $R(k,\ell)$ is known. It is easy to work out the value of $R(k,\ell)$ when $\min\{k,\ell\}$ is equal to 1 or 2. Apart from those trivial cases, the values of $R(3,\ell)$ for $\ell \leq 9$ and the values of $R(4,\ell)$ for $\ell \leq 5$ are known and all other Ramsey numbers $R(k,\ell)$ with $k \leq \ell$ are unknown.

Given the difficulty of exactly computing Ramsey numbers, attention has focused on order-of-growth estimates. Theorem 4.7 shows that $R(k,\ell)$ is at most exponential in $k+\ell$. Is this the correct order of growth, or is $R(k,\ell)$ bounded above by a sub-exponential function of $k+\ell$? In 1947 Erdős answered this question by showing that $R(k,k) > 2^{k/2}$. His method of proof was revolutionary: rather than directly constructing a graph on $n = 2^{k/2}$ vertices with no clique or independent set of size $k$, he proved the existence of such a *k-Ramsey graph* non-constructively, by showing that a random sample from $G(n,1/2)$ is a $k$-Ramsey graph with positive probability. This paradigm of proving that objects with certain properties exist by showing that a random object has the specified properties with positive probability is called *the probabilistic method*, and it has become an influential and widely used principle in discrete math and theoretical computer science.

**Definition 4.8** An undirected graph $G$ is called a $k$-Ramsey graph if no $k$ of its vertices form a clique or independent set.

**Theorem 4.9** If $k \geq 3$ and $n \leq 2^{(k+1)/2}$, then a random sample from $G(n,1/2)$ has positive probability of being a $k$-Ramsey graph. Consequently, $R(k,k) > 2^{(k+1)/2}$.

*Proof.* If $k = 3$ or $k = 4$ then $2^{(k+1)/2} < 2k-1$ and one can verify that a path graph composed of $n < 2k-1$ vertices is always a $k$-Ramsey graph. (In a path graph, there are no cliques of size greater than 2, and the largest independent set has $\lceil n/2 \rceil$ vertices.) For the remainder of the proof, we will assume $k \geq 5$. This assumption will be helpful because every $k \geq 5$ satisfies the inequality $2^{k+1} < k!$ which implies the following binomial coefficient inequality that is used below.

$$\binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} < \frac{n^k}{k!} < \frac{1}{2}\left(\frac{n}{2}\right)^k. \tag{4.3}$$

We now proceed to calculate the expected number of cliques of size $k$ in $G(n,1/2)$. For each $k$-element set $S$, the probability that $S$ forms a $k$-clique is $2^{-k(k-1)/2}$ because the $k$ vertices of $S$ form $\binom{k}{2} = \frac{k(k-1)}{2}$ pairs, each of which is an edge of $G(n,1/2)$ independently with probability $1/2$. Since there are $\binom{n}{k}$ vertex sets of size $k$, we find that

$$\mathbb{E}[\text{number of } k\text{-cliques}] = \binom{n}{k}2^{-k(k-1)/2} < \frac{1}{2}\left(\frac{n}{2}\right)^k 2^{-k(k-1)/2} = \frac{1}{2}\left(\frac{n}{2^{(k+1)/2}}\right)^k.$$

Hence, if $n \leq 2^{(k+1)/2}$, the expected number of $k$-cliques is less than $\frac{1}{2}$. By Markov's inequality, the probability that $G(n,1/2)$ contains a $k$-clique is less than $\frac{1}{2}$. An identical

calculation shows that the expected number of independent sets of size $k$ is also less than $\frac{1}{2}$, hence the probability that $G(n,1/2)$ contains an independent set of size $k$ is less than $\frac{1}{2}$. By the union bound, the probability that $G(n,1/2)$ is not a $k$-Ramsey graph is less than 1. ■

With a little bit more care, one can show that for $n \leq 2^{(k+1)/2}$, the probability that $G(n,1/2)$ is not a $k$-Ramsey graph converges to zero super-exponentially fast as $k \to \infty$; in other words, the convergence to zero happens at a faster rate than $c^k$ for any $0 < c < 1$. This is because the inequality $2^{k+1} < k!$ that we used in one step of the proof has a super-exponential amount of slack as $k \to \infty$: the factorial function grows strictly more rapidly than any exponential function. Hence, $k$-Ramsey graphs are incredibly abundant among the $n$-vertex graphs when $k$ is large and $n \leq 2^{(k+1)/2}$. Despite their abundance, no explicit construction of a $k$-Ramsey graph is known when $n$ is exponential in $k$. People have jokingly likened the problem of explicitly constructing Ramsey graphs to the problem of "finding hay in a haystack." To date, the best known explicit construction of $k$-Ramsey graphs yields graphs with $2^{k^c}$ vertices, where $c$ is a (small) positive constant.

The application of the probabilistic method to prove existence of Ramsey graphs is not an isolated example. Here is another illustrative example from graph theory.

**Definition 4.10** The *girth* of an undirected graph is the length of the shortest cycle contained in the graph. (A graph with no cycles has infinite girth.)

**Definition 4.11** A *$k$-coloring* of a graph $G = (V,E)$ consists of a $k$-element set $C$, called the set of colors, and a function $h : V \to C$ that assigns a color to each vertex. A $k$-coloring is *proper* if the endpoints of every edge are assigned distinct colors. The chromatic number of $G$ is the minimum $k$ such that $G$ has a proper $k$-coloring.

If a graph $G$ has infinite girth, then every connected component of $G$ is a tree and can be 2-colored. More generally, if $G$ has girth greater than $g$, then every subgraph of $g$ or fewer vertices is acyclic, hence 2-colorable, so in some sense $G$ is "locally 2-colorable". On the other hand, a graph of high girth may not be globally 2-colorable, i.e., it may not have a proper 2-coloring. For example, when $n \geq 3$ is odd, an $n$-cycle has girth $n$ and chromatic number 3. It's less clear, however, how to construct graphs with large girth having no proper 3-coloring. For example, for girth 5 the smallest such graph has 21 vertices.

In general, given lower bounds on the girth and chromatic number of a graph, can we always find a finite graph that satisfies the bounds? A famous application of the probabilistic method, again by Erdős, answers this question affirmatively.

**Theorem 4.12** For any finite $g$ and $k$, there exist finite undirected graphs with girth greater than $g$ and chromatic number greater than $k$.

*Proof.* Consider a random sample $G$ drawn from the $G(n,p)$ distribution, where the value of $p$ will be determined later. The first part of the proof consists of calculating the expected number of cycles of length less than or equal to $g$. To estimate the expected number of cycles of length $\ell \leq g$, we can reason as follows. The vertices of such a cycle constitute a sequence $v_1, \ldots, v_{\ell-1}, v_\ell$, such that for all $i \in [\ell]$, $(v_{i-1}, v_i)$ belongs to the edge set of $G$. (Here, we are interpreting $v_0$ as being equal to $v_\ell$.) For a given such sequence $v_1, \ldots, v_\ell$, the probability that all of the required edges are present is $p^\ell$. The number of such sequences is less than $n^\ell$. Hence, the expected number of $\ell$-cycles is bounded above by $(pn)^\ell$. Summing over all cycle lengths from 3 up to $g$, the expected number of cycles of length less than or

equal to $g$ satisfies the bound

$$\mathbb{E}[\text{number of cycles of length } \leq g] \leq \sum_{\ell=3}^{g} (pn)^{\ell}.$$

If we choose $p$ satisfying

$$\frac{1}{n} \leq pn \leq \left(\frac{n}{4g}\right)^{1/g}$$

then $(pn)^{\ell}$ is an increasing function of $\ell$ so

$$\sum_{\ell=3}^{g} (pn)^{\ell} < g(pn)^{g} \leq \frac{n}{4}.$$

By Markov's inequality,

$$\Pr\left(G \text{ has more than } \tfrac{n}{2} \text{ cycles of length } \leq g\right) \leq \frac{1}{2}. \tag{4.4}$$

The second part of the proof examines the probability that $G$ contains an independent set of size $t = \lceil n/(2k) \rceil$. For a given vertex set $S$ of size $t$, the probability that $S$ is an independent set in $G(n,p)$ is

$$(1-p)^{t(t-1)/2} < e^{-pt(t-1)/2}.$$

The number of vertex sets of size $t$ is less than $n^t$, so

$$\Pr(G \text{ has an independent set of size } t) < n^t \cdot e^{-pt(t-1)/2} = \left(\frac{n}{e^{p(t-1)/2}}\right)^t.$$

If $p \geq \frac{4k\ln(2n)}{n-2k}$ then

$$\begin{aligned} \frac{p(t-1)}{2} &\geq \frac{p\left(\frac{n}{2k}-1\right)}{2} = \frac{p(n-2k)}{4k} = \ln(2n) \\ e^{p(t-1)/2} &\geq 2n \\ \frac{n}{e^{p(t-1)/2}} &\leq \frac{1}{2} \end{aligned}$$

$$\Pr(G \text{ has an independent set of size } t) < 2^{-t}.$$

To summarize the proof up to this point, if $4k\ln(2n)/(n-2k) \leq p \leq n^{-1} \cdot (n/4g)^{1/g}$ then the event that $G$ has more than $n/2$ cycles of length $g$ or smaller has probability less than $\frac{1}{2}$, and the event that $G$ has an independent set of size $t$ has probability less than $2^{-t}$. By the union bound, there is a positive probability that neither of these events happens. Hence, there exists a graph on $n$ vertices with $n/2$ or fewer cycles of length $g$ or smaller, and with no independent set of size $t$. Let $G_0$ be one such graph. Let $G_1$ be a graph obtained from $G_0$ by deleting one vertex from each cycle of length $g$ or smaller. By assumption, at most $n/2$ vertices are deleted from $G_0$ when forming $G_1$, so $G_1$ has at least $n/2$ vertices. By construction, $G_1$ has girth greater than $g$. Now, if $h$ is any $k$-coloring of $G_1$ then, by the pigeonhole principle, there is some color $c \in [k]$ such that the set $S = h^{-1}(c)$ has at least

$n/(2k)$ vertices. However, since $G_1$ is a subgraph of $G_0$ it has no independent set of size $t = \lceil n/(2k) \rceil$. Consequently, $S$ cannot be an independent set in $G_1$; there must be an edge between two vertices in $S$. Since $h$ assigns color $c$ to both of these vertices, $h$ is not a proper coloring. Hence, the chromatic number of $G_1$ is greater than $k$.

To conclude the proof, we merely need to observe that the inequality $\frac{4k\ln(2n)}{n-2k} \leq n^{-1} \cdot \left(\frac{n}{4g}\right)^{1/g}$, is satisfied by all sufficiently large $n$, so when $n$ is large enough we may always choose $p$ satsfying the necessary upper and lower bounds. ■

# II Probability Meets Linear Algebra

# 5. Vector Spaces

Representing data in the form of vectors lies at the core of machine learning, data science, and scientific computing. These notes explain the basic theory of vector spaces over the real numbers. Differing from most introductory courses on linear algebra, we will adopt a "coordinate-free" viewpoint that treats vectors are an abstract data type supporting the operations of addition and scalar multiplication.

## 5.1 Algebraic definitions

**Definition 5.1** A *vector space* (over the real numbers) is a non-empty set $V$ of elements, called *vectors*, equipped with two operations, called *addition* and *scalar multiplication*.

- Addition is a binary operation of type $V \times V \to V$. In other words two vectors x and y can be added to yield another vector, $\mathrm{x}+\mathrm{y}$.
- Scalar multiplication is a binary operation of type $\mathbb{R} \times V \to V$. In other words we can scale a vector x by a real number $a$ to obtain another vector, $a\mathrm{x}$.

These operations are required to satisfy the associative, commutative, distributive, and multiplicative identity laws.

1. $\mathrm{x}+\mathrm{y} = \mathrm{y}+\mathrm{x}$.
2. $(\mathrm{x}+\mathrm{y})+\mathrm{z} = \mathrm{x}+(\mathrm{y}+\mathrm{z})$ and $(ab)\mathrm{x} = a(b\mathrm{x})$.
3. $(a+b)\mathrm{x} = a\mathrm{x}+b\mathrm{x}$ and $a(\mathrm{x}+\mathrm{y}) = a\mathrm{x}+a\mathrm{y}$.
4. $1\mathrm{x} = \mathrm{x}$.

These laws imply the existence of a vector called the *zero vector*, which we denote by 0, that satisfies $0+\mathrm{x} = \mathrm{x}$ and $0\mathrm{x} = 0$ for every $\mathrm{x} \in V$.

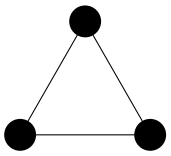
The most important and archetypical vector spaces are the spaces $\mathbb{R}^n$, defined for each

positive integer $n$. Vectors in $\mathbb{R}^n$ are $n$-tuples of real numbers. Addition and scalar multiplication are defined component-wise. In these notes we will represent elements of $\mathbb{R}^n$ by column vectors, such as $\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$.

The key distinction here is that we are not defining vectors to be $n$-tuples of real numbers, and then defining addition and scalar multiplication as operations on $n$-tuples. Rather, we are defining a vector space to be *any* structure possessing addition and scalar multiplication operations that satisfy the key properties in Definition 5.1, and then we are admitting the vector space $\mathbb{R}^n$ as an example of one such structure. This is similar to the distinction between an abstract data type, such as a list, and a data structure that implements that abstract data type, such as a doubly linked list. For the purpose of reasoning about vectors, everything we need to know about them is summarized in the abstract definition; for the purpose of calculating with them, we need to choose a specific way of representing them.

▪ **Example 5.2** For any set $S$, the set $\mathbb{R}^S$ of functions from $S$ to the real numbers, with addition and scalar multiplication defined pointwise: if x and y are two functions from $S$ to $\mathbb{R}$, $a \in \mathbb{R}$ is a scalar, and $s$ is any element of $S$, then $\mathrm{x}+\mathrm{y}$ is the function defined by $(\mathrm{x}+\mathrm{y})(s) = \mathrm{x}(s)+\mathrm{y}(s)$ and $a\mathrm{x}$ is the function defined by $(a\mathrm{x})(s) = a\mathrm{x}(s)$.

For example, if $G$ is the graph shown at right, then $\mathbb{R}^{V(G)}$ is the vector space of functions that label each vertex of $G$ with a real number. It's evident that we can represent elements of $\mathbb{R}^{V(G)}$ as ordered 3-tuples of real numbers by choosing an ordering of the vertices of $G$. However, the choice of ordering is arbitrary, so there are at least six equally reasonable ways to model the elements of $\mathbb{R}^{V(G)}$ as elements of $\mathbb{R}^3$. We describe this state of affairs by saying that the vector spaces $\mathbb{R}^{V(G)}$ and $\mathbb{R}^3$ are *isomorphic* but not equal to one another.



▪

▪ **Example 5.3** Continuing with the example above, let $Z$ denote the subset of $\mathbb{R}^{V(G)}$ consisting of functions that sum to zero. In other words x belongs to $Z$ if and only if it satisfies $\sum_{v \in V(G)} \mathrm{x}(v) = 0$. Then $Z$ is also a vector space. An element of $Z$ could be represented by an ordered triple of real numbers that sum to zero, such as the function values at the top, left, and right vertices respectively. Alternatively, we could represent an element of $Z$ by an ordered pair of numbers, such as the function values at the left and right vertices only, since the value at the top vertex is uniquely determined by the other two. The vector space $Z$ will become a running example in these notes. ▪

### 5.1.1 Linear transformations and isomorphisms

Now that we've defined vector spaces, it's time to talk about functions between vector spaces. The most important type of function between vector spaces is called a linear transformation, and it preserves all of the algebraic structure of the space.

**Definition 5.4** If $V$ and $W$ are vector spaces, a *linear transformation* from $V$ to $W$ is a function $T : V \to W$ that satisfies

$$T(a\mathrm{x}+b\mathrm{y}) = aT(\mathrm{x})+bT(\mathrm{y})$$

for all $\mathrm{x}, \mathrm{y} \in V$ and $a, b \in \mathbb{R}$.

A linear transformation is called an *isomorphism*, or equivalently *invertible*, if there

is another linear transformation $T^{-1} : W \to V$ such that $T^{-1} \circ T$ and $T \circ T^{-1}$ are the identity functions of $V$ and $W$, respectively. We then call $T^{-1}$ the *inverse* of $T$. We say $V$ and $W$ are *isomorphic* if there is an isomorphism from $V$ to $W$.

▪ **Example 5.5** When $m < n$, an important class of linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$ are the *coordinate projections*: functions that modify an $n$-tuple to an $m$-tuple by extracting a specified subset of the coordinates. For example, the coordinate projection $\pi_{13}$ from $\mathbb{R}^3$ to $\mathbb{R}^2$ deletes the middle coordinate of a 3-tuple while preserving the first and third coordinates, e.g. $\pi_{13}\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$. ▪

For any vector spaces $V$ and $W$, the set of linear transformations from $V$ to $W$ forms a vector space under pointwise addition and scalar multiplication. This vector space is denoted by $\hom(V,W)$. The isomorphisms from $V$ to $W$ don't form a vector space because, for example, when we multiply an isomorphism by the scalar 0 we obtain the function that maps every x in $V$ to 0 in $W$, which is no longer an isomorphism.

### 5.1.2 Bases and dimension

It seems self-evident that the vector spaces $\mathbb{R}^2$ and $\mathbb{R}^3$ are not isomorphic, because one of them is two-dimensional while the other is three-dimensional. How do we actually define dimension of a vector space? How do we confirm that vector spaces of different dimensions are really not isomorphic to one another? To answer these questions, we must first introduce the very important notion of a *basis* for a vector space.

**Definition 5.6** A *linear combination* of vectors $\mathrm{x}_1, \ldots, \mathrm{x}_k$ is a finite sum of the form $a_1\mathrm{x}_1 + a_2\mathrm{x}_2 + \cdots + a_k\mathrm{x}_k$. It is non-trivial if at least one of the coefficients $a_i$ is not equal to zero. A set $S$ of vectors is *linearly independent* if the zero vector cannot be expressed as a non-trivial linear combination of elements of $S$. A *basis* of a vector space is a maximal linearly independent set.

**Lemma 5.7** If $B$ is a basis of a vector space $V$, then every element of $V$ can be represented as a linear combination of elements of $B$. This representation is unique, up to a reordering of the summands.

*Proof.* If $v \in V$ cannot be represented as a linear combination of elements of $B$, then $B \cup \{v\}$ is linearly independent, contradicting the maximality of $B$. Hence, every element of $V$ can be expressed as a linear combination of elements of $B$. To see why the representation is unique, consider any $\mathrm{x} \in V$ and consider two representations

$$\mathrm{x} = a_1\mathrm{b}_1 + \cdots + a_m\mathrm{b}_m = a'_1\mathrm{b}'_1 + \cdots + a'_n\mathrm{b}'_n.$$

Subtracting these two representations of x from one another, we obtain a representation of 0 as a linear combination of elements of $B$. Since $B$ is linearly independent, all the coefficients in this linear combination must be zero. Hence, the two representations of x are identical, up to a reordering of the terms of the sum. ■

**Corollary 5.8** If $V$ is a vector space with a finite basis $B$, then the linear transformation $T : \mathbb{R}^B \to V$ defined by $T(f) = \sum_{\mathrm{b}\in B} f(\mathrm{b})\mathrm{b}$ is an isomorphism.

*Proof.* By Lemma 5.7, for every $\mathrm{x} \in V$ there is a unique representation of the form $\mathrm{x} = \sum_{\mathrm{b}\in B} a_\mathrm{b}\mathrm{b}$. Let $C(\mathrm{x})$ be the function in $\mathbb{R}^B$ defined by $C(\mathrm{x})(\mathrm{b}) = a_\mathrm{b}$. We leave it as an

exercise for the reader to verify that $C$ is a linear transformation and that $C \circ T$ and $T \circ C$ are the identity functions of their respective vector spaces. ■

The image of $\mathrm{x} \in V$ under the isomorphism $C : V \to \mathbb{R}^B$ is a $B$-tuple of real numbers. The elements of this tuple are called the *components* of x in the basis $B$. **??** shows that every finite-dimensional vector space $V$ is isomorphic to $\mathbb{R}^n$ for some value of $n$, and we will soon see this value is unique. However, there are many isomorphisms from $V$ to $\mathbb{R}^n$, corresponding to all the different ways to choose an (ordered) basis of $V$. For this reason, the components of a vector are only well-defined in contexts where an ordered basis has been specified.

**Definition 5.9** The *standard basis* of $\mathbb{R}^n$ is the basis $\{\mathrm{e}_1, \ldots, \mathrm{e}_n\}$, where $\mathrm{e}_i$ denotes a vector whose $i^{\text{th}}$ coordinate is 1 and all other coordinates are zero.

**Lemma 5.10** A set of vectors $B \subset V$ is a basis if and only if every element of $V$ can be uniquely expressed as a linear combination of elements of $B$.

*Proof.* The "only if" direction was proven in Lemma 5.7. If $B \subset V$ is a subset having the property that every element of $V$ can be uniquely expressed as a linear combination of elements of $B$, then in particular the only representation of 0 as a linear combination of elements of $B$ is the trivial representation; this verifies that $B$ is linearly independent. Furthermore, for any $\mathrm{x} \notin B$, by our assumption on $B$ we can find a representation $\mathrm{x} = a_1\mathrm{b}_1 + \cdots + a_m\mathrm{b}_m$. Then the equation $0 = a_1\mathrm{b}_1 + \cdots + a_m\mathrm{b}_m - \mathrm{x}$ shows that $B \cup \{vx\}$ is not linearly independent. Hence, $B$ is a *maximal* linearly independent set, i.e. $B$ is a basis, as claimed. ■

We will be defining the dimension of a vector space to be the cardinality of any basis. However, in order to make such a definition we need to verify that all bases have the same cardinality. This is accomplished in the following pair of lemmas.

**Lemma 5.11 — Exchange Lemma.** If $V$ is a vector space with basis $B$, then for any nonzero vector $\mathrm{x} \notin B$ we can obtain another basis from $B$ by exchanging x for one of the vectors $\mathrm{y} \in B$. In other words, $B' = (B \setminus \{\mathrm{y}\}) \cup \{\mathrm{x}\}$.

*Proof.* Using Lemma 5.7 and the fact that $\mathrm{x} \neq 0$, we know that x can be expressed as a non-trivial linear combination $\mathrm{x} = a_1\mathrm{b}_1 + \cdots + a_m\mathrm{b}_m$. Assume without loss of generality that $a_1 \neq 0$. Then

$$\mathrm{b}_1 = \mathrm{x} - \frac{a_2}{a_1}\mathrm{b}_2 - \cdots - \frac{a_m}{a_1}\mathrm{b}_m. \tag{5.1}$$

For any vector z that can be expressed as a linear combination of elements of $B$, we can substitute the right side of (5.1) in place of $\mathrm{b}_1$, to obtain a representation of z as a linear combination of elements of $B' = (B \setminus \{\mathrm{b}_1\}) \cup \{\mathrm{x}\}$. To see that this representation of z is unique, consider subtracting any two distinct representations of z as linear combinations of elements of $B'$, to obtain a nontrivial representation of 0 as a linear combination of elements of $B'$. Let $a_x$ denote the coefficient of x in this representation of 0. Our hypothesis that $B$ is linearly independent means that 0 cannot be represented as a nontrivial linear combination of elements of $B$, so we know that $a_x \neq 0$. Now if we substitute the expression $a_1\mathrm{b}_1 + \cdots + a_m\mathrm{b}_m$ in place of x, we obtain another representation of 0, this time as a linear combination of elements of $B$, in which the coefficient of $\mathrm{b}_1$ is $a_1a_x$. Since $a_1a_x \neq 0$, this contradicts the assumption that $B$ is linearly independent. ■

**Theorem 5.12** If $V$ is a vector space with a finite basis, then all bases of $V$ have the same number of elements.

*Proof.* Let $B$ and $B'$ be two bases of $V$, with $B$ finite. Denote the elements of $B$ by $\{\mathrm{b}_1, \ldots, \mathrm{b}_d\}$. Now construct a sequence of bases by the following procedure. Start with $B_0 = B'$, and repeatedly perform the exchange procedure in the proof of Lemma 5.11, inserting elements of $B$ one by one. This yields a sequence of bases $B_0, B_1, \ldots, B_d$, such that $B_0 = B'$, and for $i > 0$, $B_i = (B_{i-1} \cup \{\mathrm{b}_i\}) \setminus \mathrm{b}'_{i-1}$ for some $\mathrm{b}'_{i-1} \in B_{i-1}$. When choosing the vector $\mathrm{b}'_{i-1}$ to remove from $B_{i-1}$ while inserting $\mathrm{b}_i$, let us *never remove a vector that belongs to B*. This is possible because in the proof of Lemma 5.11, the vector that was removed from the basis when inserting x was allowed to be any vector having a nonzero coefficient when x is represented using the basis $B$. We know that when $\mathrm{b}_i$ is represented using the basis $B_{i-1}$, at least one of the basis vectors with a nonzero coefficient does not belong to $B$; this is because $B$ is linearly independent, so $\mathrm{b}_i$ cannot be represented as a non-trivial linear combination of elements of $B \setminus \{\mathrm{b}_i\}$.

By the time we reach $B_d$ in this repeated-exchange process, we have inserted each element of $B$ and have not removed any elements of $B$, so $B \subseteq B_d$. One basis cannot be a proper subset of another, since that would violate the maximality property of bases. Hence $B = B_d$. Since each two consecutive sets in the sequence of $B_0, \ldots, B_d$ have the same cardinality, we conclude that $B' = B_0$ must have the same cardinality as $B$, as claimed. ■

### 5.1.3 Inner products and the dual of a vector space

An important binary operation on $\mathbb{R}^n$ is the *dot product* operation, defined by $\mathrm{x} \cdot \mathrm{y} = \sum_{i=1}^n x_i y_i$. In the setting of abstract vector spaces, the appropriate generalization of the dot product is a structure called a *positive definite inner product*, whose essential properties are defined as follows.

**Definition 5.13** An *inner product* structure on a vector space is a function of type $V \times V \to \mathbb{R}$, with the inner product of vectors $\mathrm{x}, \mathrm{y} \in V$ being denoted by $\langle \mathrm{x}, \mathrm{y} \rangle$. An inner product is required to satisfy the following properties.

1. Bilinearity:

$$\langle a\mathrm{x} + b\mathrm{y}, \mathrm{z} \rangle = a \langle \mathrm{x}, \mathrm{z} \rangle + b \langle \mathrm{y}, \mathrm{z} \rangle \quad \text{and} \quad \langle \mathrm{x}, a\mathrm{y} + b\mathrm{z} \rangle = a \langle \mathrm{x}, \mathrm{y} \rangle + b \langle \mathrm{x}, \mathrm{z} \rangle .$$

2. Symmetry:

$$\langle \mathrm{x}, \mathrm{y} \rangle = \langle \mathrm{y}, \mathrm{x} \rangle .$$

An inner product is called *non-degenerate* if for every $\mathrm{x} \neq 0$ there exists a y such that $\langle \mathrm{x}, \mathrm{y} \rangle \neq 0$. It is called *positive semidefinite* if $\langle \mathrm{x}, \mathrm{x} \rangle \geq 0$ for all x, and *positive definite* if the inequality is strict for all $\mathrm{x} \neq 0$.

Note that a positive definite inner product is always non-degenerate: if $\mathrm{x} \neq 0$ then $\langle \mathrm{x}, \mathrm{x} \rangle \neq 0$. The dot product on $\mathbb{R}^n$ is positive definite because $\langle \mathrm{x}, \mathrm{x} \rangle = x_1^2 + \cdots + x_n^2$, which is always non-negative and equals zero only when $\mathrm{x} = 0$.

An example of a non-degenerate inner product that is *not* positive definite is the Lorentzian inner product on $\mathbb{R}^n$:

$$\langle \mathrm{x}, \mathrm{y} \rangle_L = -x_1 y_1 + x_2 y_2 + \cdots + x_n y_n .$$

This inner product plays an important role in the physics of spacetime, where the first coordinate represents the time dimension and the remaining coordinates represent the spatial dimensions. According to Einstein's theory of special relativity, the linear transformations that one should apply to shift from one observer's system of spacetime coordinates to another's are precisely the linear transformations that preserve the Lorentzian inner product of vectors.

### 5.1.4 The dual of a vector space

**Definition 5.14** The vector space $\hom(V,\mathbb{R})$ of real-valued linear functions on $V$ is called the *dual* of $V$ and is denoted by $V^*$.

**Lemma 5.15** Every finite-dimensional vector space is isomorphic to its own dual.

*Proof.* Suppose $V$ is a vector space and $B$ is a finite basis for $V$. Recall from Corollary 5.8 that $V$ is isomorphic to $\mathbb{R}^B$. The dual vector space $V^*$ is also isomorphic to $\mathbb{R}^B$, via the isomorphism that maps each linear function $\mathrm{f}: V \to \mathbb{R}$ to the function $\mathrm{f}^B : B \to \mathbb{R}$ obtained by restricting the domain of f from $V$ to $B$. To prove this is an isomorphism between $V^*$ and $B$ we need to prove it has an inverse. In other words, we need to show that for each function $\mathrm{f}^B : B \to \mathbb{R}$ there is a unique linear function $\mathrm{f}: V \to \mathbb{R}$ that restricts to $\mathrm{f}^B$. If f is any linear function that restricts to $\mathrm{f}^B$, then for any vector $\mathrm{x} = \sum_{\mathrm{b}\in B} x_{\mathrm{b}}\mathrm{b}$ the value $\mathrm{f}(\mathrm{x})$ must satisfy

$$\mathrm{f}(\mathrm{x}) = \sum_{\mathrm{b}\in B} x_{\mathrm{b}}\mathrm{f}^B(\mathrm{b}).$$

This shows there can be *at most one* linear function $\mathrm{f}: V \to \mathbb{R}$ that restricts to $\mathrm{f}^B$, since the value $\mathrm{f}(\mathrm{x})$ on any x is uniquely determined by the equation above. To verify that there is exactly one linear function that restricts to $\mathrm{f}^B$, we must check that the function f defined above is linear. Suppose $\mathrm{x},\mathrm{y} \in V$ and $r,s \in \mathbb{R}$. If $\mathrm{x} = \sum_{\mathrm{b}\in B} x_{\mathrm{b}}\mathrm{b}$ and $\mathrm{y} = \sum_{\mathrm{b}\in B} y_{\mathrm{b}}\mathrm{b}$ then

$$r\mathrm{x} + s\mathrm{y} = \sum_{\mathrm{b}\in B} (rx_{\mathrm{b}} + sy_{\mathrm{b}})\mathrm{b}$$

so

$$\mathrm{f}(r\mathrm{x}+s\mathrm{y}) = \sum_{\mathrm{b}\in B} (rx_{\mathrm{b}} + sy_{\mathrm{b}})\mathrm{f}^B(\mathrm{b}) = r\sum_{\mathrm{b}\in B} x_{\mathrm{b}}\mathrm{f}^B(\mathrm{b}) + s\sum_{\mathrm{b}\in B} y_{\mathrm{b}}\mathrm{f}^B(\mathrm{b}) = r\mathrm{f}(\mathrm{x}) + s\mathrm{f}(\mathrm{y})$$

which confirms that f is linear. ■

▪ **Example 5.16** Every real-valued linear function f on $\mathbb{R}^3$ can be represented (uniquely) by a sequence of three coefficients $a_1, a_2, a_3$ such that

$$\mathrm{f}\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) = a_1x_1 + a_2x_2 + a_3x_3.$$

The dual of $\mathbb{R}^3$ is isomorphic to $\mathbb{R}^3$, under the isomorphism that maps a linear function f to the coefficient vector $\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$. ▪

Generalizing the previous example, the dual of $\mathbb{R}^n$ is isomorphic to $\mathbb{R}^n$ via the isomorphism that maps a linear function to its coefficient vector. To facilitate distinguishing

between $\mathbb{R}^n$ and its dual, we will represent elements of $(\mathbb{R}^n)^*$ as row vectors rather than column vectors. In other words, we will prefer to represent the linear function f in Example 5.16 using the row vector $\mathrm{f} = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}$ rather than the column vector $\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$. This notation is convenient because it means that the application of the function f to the vector x can simply be written as fx, using the rules for multiplying a 1-by-$n$ matrix by an $n$-by-1 matrix.

Generalizing these examples still further, a non-degenerate inner product structure on a finite-dimensional vector space always allows one to define an isomorphism between $V$ and $V^*$. However, it's important to note that there are many isomorphisms between $V$ and $V^*$, and there's no particular way to single out one of them without singling out a non-degenerate inner product structure.

**Lemma 5.17** If $V$ is a finite dimensional vector space and $\langle \cdot, \cdot \rangle$ is a non-degenerate inner product, then there is an isomorphism $T : V \to V^*$ where $T(\mathrm{x})$ is defined to be the linear function $\mathrm{t_x}$ specified by the formula $\mathrm{t_x}(\mathrm{y}) = \langle \mathrm{x}, \mathrm{y} \rangle$.

*Proof.* The bilinearity property of inner products ensures that the function $T$ defined in the lemma statement is a linear function from $V$ to $V^*$. It is injective because if $\mathrm{x}, \mathrm{y} \in V$ satisfy $T(\mathrm{x}) = T(\mathrm{y})$, then for all $\mathrm{z} \in V$ we have $\langle \mathrm{x} - \mathrm{y}, \mathrm{z} \rangle = \langle \mathrm{x}, \mathrm{z} \rangle - \langle \mathrm{y}, \mathrm{z} \rangle = \mathrm{t_x}(\mathrm{z}) - \mathrm{t_y}(\mathrm{z}) = 0$. As $\langle \cdot, \cdot \rangle$ is non-degenerate, this implies $\mathrm{x} - \mathrm{y} = 0$ hence $\mathrm{x} = \mathrm{y}$. From Lemma 5.15, we know that $V$ and $V^*$ have the same dimension. We leave it as an exercise to the reader to verify that an injective map between finite-dimensional vector spaces of the same dimension must be an isomorphism. ■

▪ **Example 5.18** Suppose $V$ is the subspace of $\mathbb{R}^3$ consisting of vectors whose coordinates sum to zero, with the positive definite inner product structure given by

$$\left\langle \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \\ z_z \end{bmatrix} \right\rangle = x_1 x_2 + y_1 y_2 + z_1 z_2.$$

One element of $V^*$ is the linear function f that sums the first two coordinates of a vector, i.e. the function $\mathrm{f}\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}\right) = x + y$. If we are representing elements of $V$ by three-tuples $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ then f can be represented by the row vector $\begin{bmatrix} 1 & 1 & 0 \end{bmatrix}$. However, since $-z = x + y$ for every $\begin{bmatrix} x \\ y \\ z \end{bmatrix} \in V$, the function f is also expressed by the formula $\mathrm{f}\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}\right) = -z$ and can be represented by the vector $\begin{bmatrix} 0 & 0 & -1 \end{bmatrix}$.

This example underscores the importance of distinguishing between a vector space and its dual. The vector space $\mathbb{R}^3$ is isomorphic to its dual, however when we pass to a subspace of $\mathbb{R}^3$, the dual of the subspace is *not* a subspace of $(\mathbb{R}^3)^*$. Instead, it is a *quotient* of $(\mathbb{R}^3)^*$, i.e. a vector space whose elements are *equivalence classes of vectors* in $(\mathbb{R}^3)^*$. ▪

## 5.2 Convexity and norms

One the wonderful things about vector spaces is that, although they are defined by algebraic operations, we can also reason about them using geometric notions like convexity, distance, and volume. In this section we develop some basic facts about these three notions.

### 5.2.1 Convex sets and functions

A subset of a vector space is convex if it contains the line segment joining any two of its points. This informal definition is formalized as follows.

**Definition 5.19** If $F = \{x_1, \ldots, x_m\}$ is a finite subset of a vector space $V$, an *affine combination* of points of $F$ is a linear combination $a_1x_1 + \cdots + a_mx_m$ whose coefficients satisfy $a_1 + \cdots + a_m = 1$. A *convex combination* of points of $F$ is an affine combination whose coefficients are non-negative. (Another name for a convex combination of vectors is a *weighted average*.) The *affine hull* and *convex hull* of $F$ are the set of all affine combinations and all convex combinations of elements of $F$, respectively.

The affine hull of two points is the line passing through them, the affine hull of three non-collinear points is the plane passing through them, and so on. The convex hull of two points is the line segment joining them, the convex hull of three non-collinear points is the triangle joining them, and so on.

**Definition 5.20** A subset $K$ of a vector space is *convex* if every convex combination of points in $K$ belongs to $K$. Equivalently, $K$ is convex if, for every $x, y \in K$ and every $t \in [0,1]$, the vector $(1-t)x + ty$ also belongs to $K$.

A simple inductive proof verifies that the two formulations of convexity in Definition 5.20 are, indeed, equivalent. Clearly, the first definition implies the second because the expression $(1-t)x + ty$ defines a convex combination of x and y when $0 \leq t \leq 1$. Conversely, suppose $K$ satisfies the second definition. We assert that for every $m \geq 2$, every convex combination of $m$ points of $K$ belongs to $K$. The base case $m = 2$ is simply a restatement of the second definition. For the inductive step, if non-negative coefficients $a_1, a_2, \ldots, a_m$ sum up to 1, assume without loss of generality that $a_m > 0$, and let $t = 1 - a_1 = a_2 + a_3 + \cdots + a_m$. Since $t > 0$, we have

$$a_1x_1 + a_2x_2 + \cdots + a_mx_m = (1-t)x_1 + t\left(\tfrac{a_2}{t}x_2 + \cdots + \tfrac{a_m}{t}x_m\right).$$

By the induction hypothesis, the vector $x' = \left(\frac{a_2}{t}x_2 + \cdots + \frac{a_m}{t}x_m\right)$ belongs to $K$. Hence, $(1-t)x_1 + tx'$ also belongs to $K$, as desired.

An important type of convex set is a halfspace, which is a set of the form

$$H = \{x \mid f(x) \leq \theta\}, \tag{5.2}$$

for some nonzero $f \in V^*$ and some $\theta \in \mathbb{R}$. Equivalently, due to Lemma 5.15, we can define a halfspace using a non-degenerate inner product as

$$H = \{x \mid \langle w, x\rangle \leq b\}, \tag{5.3}$$

where w is a nonzero vector in $V$ called the *normal vector* to $H$, and $\theta \in \mathbb{R}$. To verify that the set $H$ defined using (5.2) is convex, observe that

$$f(a_1x_1 + \cdots + a_mx_m) = a_1f(x_1) + \cdots + a_mf(x_m).$$

If $a_1, a_2, \ldots, a_m$ are the coefficients of a convex combination, then the right side of this equation is a weighted average of the values $f(x_1), \ldots, f(x_m)$. If each of those values is less than or equal to $\theta$, then their weighted average must also be less than or equal to $\theta$.

Convexity of a closed set[1] can be equivalently defined using halfspaces.

[1] A subset $S$ of a finite-dimensional vector space is called *closed* if the limit point of every convergent sequence of vectors in $S$ is also contained in $S$.

**Lemma 5.21** If $V$ is a finite-dimensional vector space and $K$ is a closed subset of $V$, then $K$ is convex if and only if it is equal to the intersection of a set of halfspaces.

The proof of the lemma is not quite self-contained. It uses some facts from topology that we state here without proof.

1. If $V$ is a finite-dimensional vector space and $\langle \cdot, \cdot \rangle$ is a positive definite inner product, then for any x the function $q(\mathrm{y}) = \langle \mathrm{y}-\mathrm{x}, \mathrm{y}-\mathrm{x} \rangle$ is continuous.
2. If $S$ is a non-empty, closed, bounded subset of a finite-dimensional vector space and $f$ is a continuous function on $S$, then there exists a point $\mathrm{z} \in S$ such that $f(\mathrm{z}) = \inf\{f(\mathrm{y}) \mid \mathrm{y} \in S\}$.

*Proof.* From the definition of a convex set, it is clear that an intersection of convex sets is convex. Conversely, if $K$ is convex then we must prove it is the intersection of a set of halfspaces. Specifically, let $\mathcal{H}(K)$ denote the set of halfspaces that contain $K$ as a subset, and let $K' = \cap_{H \in \mathcal{H}(K)} H$. (If $\mathcal{H}(K) = \emptyset$ then interpret this intersection to be the entirety of $V$.) Then $K'$ is the intersection of a set of halfspaces, and we shall show that $K' = K$. The containment $K \subseteq K'$ is immediate from the definition of $K'$. To show that $K' \subseteq K$, we prove the reverse containment $V \setminus K \subseteq V \setminus K'$. In other words, if $\mathrm{x} \in V \setminus K$, we must find a halfspace $H$ that contains $K$ but not x. Let $\langle \cdot, \cdot \rangle$ be a positive-definite inner product on $V$, and consider the continuous function $q(\mathrm{y}) = \langle \mathrm{y}-\mathrm{x}, \mathrm{y}-\mathrm{x} \rangle$. Let $q_0 = \inf\{q(\mathrm{y}) \mid \mathrm{y} \in K\}$ and observe that $q_0 > 0$. The set $K_0 = \{\mathrm{y} \in K \mid q(\mathrm{y}) \leq q_0 + 1\}$ is non-empty, closed, and bounded, so there exists $\mathrm{z} \in K_0$ with $q(\mathrm{z}) = q_0$.

Now consider the set

$$H = \{\mathrm{y} \mid \langle \mathrm{z}-\mathrm{x}, \mathrm{y}-\mathrm{x} \rangle \geq q_0\} = \{\mathrm{y} \mid \langle \mathrm{z}-\mathrm{x}, \mathrm{y} \rangle \geq q_0 + \langle \mathrm{z}-\mathrm{x}, \mathrm{x} \rangle\}.$$

This is a halfspace, and $\mathrm{x} \notin H$ because $\langle \mathrm{z}-\mathrm{x}, \mathrm{x}-\mathrm{x} \rangle = 0 < q_0$. To conclude the proof we will show that $K \subseteq H$. For any $\mathrm{y} \in K$ consider the function

$$f(t) = q(\mathrm{z}+t(\mathrm{y}-\mathrm{z})) = \langle \mathrm{z}-\mathrm{x}+t(\mathrm{y}-\mathrm{z}), \mathrm{z}-\mathrm{x}+t(\mathrm{y}-\mathrm{z}) \rangle = q(\mathrm{z}) + 2t \langle \mathrm{z}-\mathrm{x}, \mathrm{y}-\mathrm{z} \rangle + t^2 \langle \mathrm{y}-\mathrm{z}, \mathrm{y}-\mathrm{z} \rangle .$$

For $0 \leq t \leq 1$ the vector $\mathrm{z}+t(\mathrm{y}-\mathrm{z}) = (1-t)\mathrm{z}+t\mathrm{y}$ belongs to $K$, and we know that the minimum value of $q$ on $K$ is attained at z, so the quadratic function $f(t)$ on the interval $0 \leq t \leq 1$ attains its minimum value at $t = 0$. Therefore, $f'(0) \geq 0$, which implies $\langle \mathrm{z}-\mathrm{x}, \mathrm{y}-\mathrm{z} \rangle \geq 0$. Now, we find that

$$\langle \mathrm{z}-\mathrm{x}, \mathrm{y}-\mathrm{x} \rangle = \langle \mathrm{z}-\mathrm{x}, \mathrm{y}-\mathrm{z} \rangle + \langle \mathrm{z}-\mathrm{x}, \mathrm{z}-\mathrm{x} \rangle \geq 0 + q_0$$

hence y satisfies the defining inequality of the halfspace $H$. As y was an arbitrary element of $K$, we have proven $K \subseteq H$ as desired. ■

In addition to convex sets, another important notion is that of a *convex function*.

**Definition 5.22** If $V$ is a vector space, $K \subseteq V$ is a convex set, and $h : K \to \mathbb{R}$ is a function, we say that $h$ is convex if it satisfies

$$h((1-t)\mathrm{x}+t\mathrm{y}) \leq (1-t)h(\mathrm{x}) + th(\mathrm{y}) \qquad \forall \mathrm{x}, \mathrm{y} \in K,\ 0 \leq t \leq 1$$

Analogous to the two equivalent definitions of a convex set, it is equivalent to say that $h$ is convex if and only if, for all finite sets $F = \{\mathrm{x}_1, \ldots, \mathrm{x}_m\} \subseteq K$ and convex combinations $\mathrm{x} = a_1\mathrm{x}_1 + \cdots + a_m\mathrm{x}_m$, the inequality

$$h(\mathrm{x}) \leq a_1 h(\mathrm{x}_1) + \cdots + a_m h(\mathrm{x}_m).$$

This inequality (along with its generalization to integrals rather than finite sums) goes by the name of *Jensen's convex function inequality.*

We proceed to state two more definitions related to convex functions and then a lemma providing two equivalent characterizations of convexity.

**Definition 5.23** If $V$ is a vector space, $K \subseteq V$, and $h : K \to \mathbb{R}$, then the *epigraph* of $h$ is the set of all pairs $(\mathrm{x}, y) \in V \times \mathbb{R}$ such that $y \geq h(\mathrm{x})$. For any $\mathrm{x} \in K$, the *subdifferential* of $h$ at x is defined to be the set

$$\partial h(\mathrm{x}) = \{\mathrm{f} \in V^* \mid \mathrm{f}(\mathrm{y}) - \mathrm{f}(\mathrm{x}) \leq h(\mathrm{y}) - h(\mathrm{x}) \ \forall \mathrm{y} \in K\}.$$

One can visualize the epigraph of $h$ as an infinitely tall multidimensional bowl-shaped region sitting above the graph of $h$ in $V \times \mathbb{R}$. To visualize what it means for f to belong to the subdifferential of $h$, note that the graph of the function $L_{\mathrm{f},\mathrm{x}}(\mathrm{y}) = \mathrm{f}(\mathrm{y}) - \mathrm{f}(\mathrm{x}) + h(\mathrm{x})$ is a hyperplane in $V \times \mathbb{R}$ and it touches the graph of $h$ at the point $(\mathrm{x}, h(\mathrm{x}))$. If the graph of $L_{\mathrm{f},\mathrm{x}}$ is a *supporting hyperplane* of the epigraph of $h$ (i.e., a hyperplane that touches the epigraph of $h$ at least once point and lies (weakly) below it everywhere) then f belongs to the subdifferential $\partial h(\mathrm{x})$.

If $V$ has a non-degenerate inner product, this defines an isomorphism between $V^*$ and $V$. The image of $\partial h(\mathrm{x})$ under this isomorphism is a set of vectors called the *subgradient* of $h$ at x.

To relate epigraphs and subgradients to convexity, we need to define one more notion: open subsets of a finite-dimensional vector space. Intuitively, a subset $U \subseteq V$ is *open* if every point of $U$ is completely surrounded by other points of $U$. For example, in the open-dimensional vector space $\mathbb{R}$, an open interval $(a,b) = \{x \mid a < x < b\}$ is open whereas a closed interval $[a,b] = \{x \mid a \leq x \leq b\}$ is not, because the endpoints of a closed interval are not surrounded on both sides by other points of the interval.

**Definition 5.24** If $V$ is a finite-dimensional vector space, a subset $U \subseteq V$ is called an *open set* if it satisfies the following property: for all $\mathrm{x}, \mathrm{y} \in U$ there exists some $\delta > 0$ such that for every $\varepsilon$ with $|\varepsilon| < \delta$, the vector $\mathrm{x} + \varepsilon\mathrm{y}$ belongs to $U$.

**Lemma 5.25** For a convex open subset $K$ of a finite-dimensional vector space $V$, the following properties of a function $h : K \to \mathbb{R}$ are equivalent.

1. $h$ is convex.
2. The epigraph of $h$ is a convex subset of $V \times \mathbb{R}$.
3. The subdifferential of $h$ is nonempty at every point of $K$.

*Proof.* We will prove the cycle of implications (3) $\Rightarrow$ (1) $\Rightarrow$ (2) $\Rightarrow$ (3), which suffices to prove the equivalence of the three condiions.

**(3) $\Rightarrow$ (1):** If the subdifferential of $h$ is nonempty at every point of $K$, then consider any two points $\mathrm{x}, \mathrm{x}'$ and their convex combination $\mathrm{x}'' = (1-t)\mathrm{x} + t\mathrm{x}'$. The subdifferential

$\partial h(\mathrm{x}'')$ is non-empty, so it contains some $\mathrm{f} \in V^*$ that satisfies $\mathrm{f}(\mathrm{y}) - \mathrm{f}(\mathrm{x}'') \leq h(\mathrm{y}) - h(\mathrm{x}'')$ for all $\mathrm{y} \in K$. In particular, we have the two inequalities

$$\begin{aligned} \mathrm{f}(\mathrm{x}) - \mathrm{f}(\mathrm{x}'') &\leq h(\mathrm{x}) - h(\mathrm{x}'') \\ \mathrm{f}(\mathrm{x}') - \mathrm{f}(\mathrm{x}'') &\leq h(\mathrm{x}') - h(\mathrm{x}''). \end{aligned}$$

Multiplying the first by $1-t$ and the second by $t$ we obtain

$$(1-t)\mathrm{f}(\mathrm{x}) + t\mathrm{f}(\mathrm{x}') - \mathrm{f}(\mathrm{x}'') \leq (1-t)h(\mathrm{x}) + th(\mathrm{x}') - h(\mathrm{x}'').$$

The left side is zero, because f is a linear function that $\mathrm{x}'' = (1-t)\mathrm{x} + t\mathrm{x}'$. Hence, $(1-t)h(\mathrm{x}) + th(\mathrm{x}') \geq h(\mathrm{x}'') = h((1-t)\mathrm{x} + t\mathrm{x}'')$ which confirms that $h$ is convex.

**(1) ⇒ (2):** Suppose $h$ is convex. Let $(\mathrm{x}, y)$ and $(\mathrm{x}', y')$ denote two points in the epigraph of $h$. Then $y \geq h(\mathrm{x})$ and $y' \geq h(\mathrm{x}')$ so

$$(1-t)y + ty' \geq (1-t)h(\mathrm{x}) + th(\mathrm{x}') \geq h((1-t)\mathrm{x} + t\mathrm{x}')$$

which shows that $(1-t)(\mathrm{x}, y) + t(\mathrm{x}', y')$ belongs to the epigraph of $h$ and thus confirms that the epigraph is convex.

**(2) ⇒ (3):** If the epigraph of $h$ is convex and x is a point of $K$, then for every $n > 0$ the point $(\mathrm{x}, h(\mathrm{x}) - 1/n)$ does not belong to the epigraph of $h$. The closure of the epigraph of $h$ (i.e., the set consisting of the epigraph along with every point in $V \times \mathbb{R}$ that is the limit of a sequence of points in the epigraph) is a closed, convex subset of $V \times \mathbb{R}$. By Lemma 5.21 it follows that there is a halfspace $H_n$ that contains the epigraph of $h$ but doesn't contain $(\mathrm{x}, h(\mathrm{x}) - 1/n)$. The set of points $(\mathrm{x}', y') \in H_n$ is defined by an inequality of the form $\mathrm{f}_n(\mathrm{x}') - a_n y' \leq \theta_n$, where $\mathrm{f}_n \in V^*$ and $a_n, \theta_n \in \mathbb{R}$.

Choose an isomorphism between $V^* \times \mathbb{R}$ and $\mathbb{R}^{d+1}$, where $d$ is the dimension of $V$, and let $S$ be the image of the unit sphere in $\mathbb{R}^{d+1}$ under this isomorphism. By rescaling $(\mathrm{f}_n, a_n)$ if necessary, we can assume that $(\mathrm{f}_n, a_n) \in S$ for each $n$. Since $S$ is a closed and bounded subset of $V^* \times \mathbb{R}$, and $V^* \times \mathbb{R}$ is finite dimensional, the sequence $(\mathrm{f}_n, a_n)_{n=1}^{\infty}$ has an infinite subsequence that converges to a limit point $(\mathrm{f}, a) \in S$. If we look at the values $\theta_n$ as $n$ ranges over the same subsequence, we claim that they converge to the number $\theta = \mathrm{x}(\mathrm{x}) - ah(\mathrm{x})$. To prove this, note that $(\mathrm{x}, h(\mathrm{x})) \in H_n$ but $(\mathrm{x}, h(\mathrm{x}) - 1/n) \notin H_n$, which means

$$\mathrm{f}_n(\mathrm{x}) - a_n h(\mathrm{x}) \leq \theta_n < \mathrm{f}_n(\mathrm{x}) - a_n h(\mathrm{x}) + a_n/n. \tag{5.4}$$

Passing to a subsequence on which $(\mathrm{f}_n, a_n)$ converges to $(\mathrm{f}, a)$, the left and right sides both converge to $\mathrm{f}(\mathrm{x}) - ah(\mathrm{x})$, so $\theta_n$ must converge to the same number.

Next we claim that the halfspace $H$ consisting of all pairs $(\mathrm{x}', y')$ satisfying the inequality $\mathrm{f}(\mathrm{x}') - ay' \leq \theta$ contains the epigraph of $h$. To see that this is so, assume $y' \geq h(\mathrm{x}')$ and note that for each $n$ we know that $(\mathrm{x}', y')$ belongs to $H_n$ hence it satisfies $\mathrm{f}_n(\mathrm{x}') - a_n y' \leq \theta_n$. Passing to a subsequence on which $(\mathrm{f}_n, a_n) \to (\mathrm{f}, a)$ and then taking the liminf of both sides, we find that $\mathrm{f}(\mathrm{x}') - ay' \leq \theta$, as claimed.

For any $\mathrm{x}'$ in $K$, the point $(\mathrm{x}', h(\mathrm{x}'))$ belongs to the epigraph of $H$, hence it satisfies

$$\mathrm{f}(\mathrm{x}') - ah(\mathrm{x}') \leq \theta = \mathrm{f}(\mathrm{x}) - ah(\mathrm{x}).$$

Rearranging this equation we find that

$$\tfrac{1}{a}\mathrm{f}(\mathrm{x}') - \tfrac{1}{a}\mathrm{f}(\mathrm{x}) \leq h(\mathrm{x}') - h(\mathrm{x})$$

for all $\mathrm{x}' \in K$, which confirms that $\frac{1}{a}\mathrm{f}$ belongs to $\partial h$, so $\partial h$ is nonempty. ■

### 5.2.2 Norms

A *norm* on a vector space provides a way to measure the length of a vector, and hence the distance between two vectors.

**Definition 5.26** If $V$ is a vector space, a *norm* on $V$ is a function $\|\cdot\|$ from $V$ to $\mathbb{R}$ satisfying:

1. Non-negativity: $\|\mathrm{x}\| \geq 0$ for all $\mathrm{x} \in V$, with equality if and only if $\mathrm{x} = 0$.
2. Linear homogeneity: $\|a\mathrm{x}\| = |a|\|\mathrm{x}\|$ for all $a \in \mathbb{R}$ and $\mathrm{x} \in V$.
3. Subadditivity: $\|\mathrm{x}+\mathrm{y}\| \leq \|\mathrm{x}\| + \|\mathrm{y}\|$ for all $\mathrm{x}, \mathrm{y} \in V$.

Common examples of norms on $\mathbb{R}^n$ are the $L_p$ norms, defined for $1 \leq p < \infty$ by

$$\|\mathrm{x}\|_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

and for $p = \infty$ by

$$\|\mathrm{x}\|_\infty = \max_{i=1}^{n} \{|x_i|\}.$$

It is easy to check that these norms satisfy non-negativity and linear homogeneity; the proof of subadditivity is omitted from these notes but can be found in many textbooks.

**Lemma 5.27** For $\mathrm{x} \in \mathbb{R}^n$, the $p$-norm $\|\mathrm{x}\|_p$ is a non-increasing function of $p$.

*Proof.* For $\mathrm{x} = 0$ the assertion is trivial, since $\|\mathrm{x}\|_p = 0$ for all $p$. Otherwise, consider any $\mathrm{x} \neq 0$ and any $p, q$ such that $1 \leq p < q$. We wish to show that $\|\mathrm{x}\|_p \geq \|\mathrm{x}\|_q$. By rescaling x if necessary, we may assume $\|\mathrm{x}\|_q = 1$. (The rescaling doesn't affect the validity of the inequality, since the linear homogeneity property ensures both sides are scaled by the same amount.) This implies that $|x_i| \leq 1$ for all $i$, either because $q = \infty$ or because $q < \infty$, $\sum_{i=1}^{q} |x_i|^q \leq 1$, and every term in the sum is non-negative. Since $|x_i| \leq 1$ and $p < q$, we have $|x_i|^p \geq |x_i|^q$. Summing these inequalities,

$$\|\mathrm{x}\|_p^p = \sum_{i=1}^{n} |x_i|^p \geq \sum_{i=1}^{n} |x_i|^q = 1.$$

Taking the $p^{\text{th}}$ root of both sides, $\|\mathrm{x}\|_p \geq 1 = \|\mathrm{x}\|_q$. ∎

When x is a vector with just one nonzero coordinate $x_i$, the $p$-norm $\|\mathrm{x}\|_p$ is equal to $|x_i|$ for every $p$. When x has more than one nonzero coordinate, $\|\mathrm{x}\|_p$ is a strictly decreasing function of $p$: it is largest when $p = 1$ and smallest when $p = \infty$. More generally, having large 1-norm can often be interpreted as a sign of *density* (i.e., having many nonzero coordinates) while having small 1-norm is often interpreted as a sign of *sparsity*. This intuition will be put to use later in the course.

**Definition 5.28** If $V$ is a vector space and $\|\cdot\|$ is a norm, the *unit ball* of $\|\cdot\|$ is the set of all vectors in $V$ whose norm is less than or equal to 1.

**Lemma 5.29** If $V$ is a vector space and $\|\cdot\|$ is a norm, the unit ball of $\|\cdot\|$ is a closed, bounded, convex set that is *centrally symmetric*, meaning that for every vector x in the unit ball, $-$x along belongs to the unit ball. Conversely, for any closed, bounded, centrally symmetric convex set $B$, there exists a norm whose unit ball is $B$.

The following important inequality is usually called the Cauchy-Schwartz inequality.

**Lemma 5.30** If $\langle\cdot,\cdot\rangle$ is a positive definite inner product on a vector space, then for any two vectors $\mathrm{x},\mathrm{y}$ we have

$$\langle \mathrm{x},\mathrm{y}\rangle \leq \langle \mathrm{x},\mathrm{x}\rangle^{1/2} \cdot \langle \mathrm{y},\mathrm{y}\rangle^{1/2},$$

with equality if and only if x is a scalar multiple of y or vice-versa.

*Proof.* If x or y is equal to 0 then both sides of the inequality are zero, so the lemma holds. Otherwise, note that replacing x and y with $a$x and $b$y, respectively, multiplies both sides of the inequality by $ab$. Hence, we may prove the lemma in the special case when $\langle \mathrm{x},\mathrm{x}\rangle = \langle \mathrm{y},\mathrm{y}\rangle = 1$; the general case will then follow by scaling x and y suitably.

When $\langle \mathrm{x},\mathrm{x}\rangle = \langle \mathrm{y},\mathrm{y}\rangle = 1$, we have

$$0 \leq \langle \mathrm{x}-\mathrm{y},\mathrm{x}-\mathrm{y}\rangle = \langle \mathrm{x},\mathrm{x}\rangle - 2\langle \mathrm{x},\mathrm{y}\rangle + \langle \mathrm{y},\mathrm{y}\rangle = 2 - 2\langle \mathrm{x},\mathrm{y}\rangle.$$

Furthermore, the inequality is strict when $\mathrm{x}-\mathrm{y} \neq 0$. Hence, we conclude that $\langle \mathrm{x},\mathrm{y}\rangle \leq 1 = \langle \mathrm{x},\mathrm{x}\rangle^{1/2} \cdot \langle \mathrm{y},\mathrm{y}\rangle^{1/2}$ and that the inequality is strict unless $\mathrm{x} = \mathrm{y}$. ■

An easy application of the Cauchy-Schwartz inequality shows that any positive definite inner product can be used to define a norm on a vector space.

**Lemma 5.31** If $V$ is a vector space with a positive definite inner product $\langle\cdot,\cdot\rangle$, then the function defined by

$$\|\mathrm{x}\| = \langle \mathrm{x},\mathrm{x}\rangle^{1/2}$$

is a norm.

*Proof.* Non-negativity follows from positive definiteness of the inner product, and linear homogeneity follows from bilinearity. To prove subadditivity, observe that for any $\mathrm{x},\mathrm{y}$,

$$\|\mathrm{x}+\mathrm{y}\|^2 = \langle \mathrm{x}+\mathrm{y},\mathrm{x}+\mathrm{y}\rangle = \|\mathrm{x}\|^2 + 2\langle \mathrm{x},\mathrm{y}\rangle + \|\mathrm{y}\|^2.$$
$$(\|\mathrm{x}\| + \|v\mathrm{y}\|)^2 = \|\mathrm{x}\|^2 + 2\|\mathrm{x}\|\,\|\mathrm{y}\| + \|\mathrm{y}\|^2.$$

The Cauchy-Schwartz inequality implies that the right side of the first equation is less than or equal to the right side of the second equation. ■

For the standard inner product on $\mathbb{R}^n$ the norm defined in Lemma 5.31 coincides with the $L_2$ norm. For other positive definite inner products on $\mathbb{R}^n$, it constitutes a different norm whose unit ball is an ellipsoidal (egg-shaped) region.

### 5.2.3 Differentials and gradients

The gradient of a function on $\mathbb{R}^n$ is usually defined using partial derivatives. In this section we will see that a differentiable function on a vector space $V$ always has a well-defined "differential" at every point, which is an element of the dual space $V^*$. However, to define the gradient requires choosing an isomorphism between $V$ and $V^*$; hence, the gradient of a multivariate function depends on the choice of inner product structure for the vector space on which the function is defined.

**Definition 5.32** If $(V, \|\cdot\|)$ is a normed vector space, a function $g : V \to \mathbb{R}$ is said to *vanish to first order at* 0 if $\frac{g(\mathrm{x})}{\|\mathrm{x}\|} \to 0$ as $\|\mathrm{x}\| \to 0$, uniformly in x. More precisely, $g$ vanishes to first order at 0 if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $\frac{g(\mathrm{x})}{\|\mathrm{x}\|} < \varepsilon$ whenever $\|\mathrm{x}\| < \delta$.

**Definition 5.33** If $(V, \|\cdot\|)$ is a normed vector space, $S \subseteq V$, and $f : V \to \mathbb{R}$, we say that $f$ is *differentiable* at a point $\mathrm{x} \in S$ if there exists a linear function $\mathrm{df}_\mathrm{x} \in V^*$, called the *differential* of $f$ at x, such that

$$\forall \mathrm{y}\ f(\mathrm{x}+\mathrm{y}) = f(\mathrm{x}) + \mathrm{df}_\mathrm{x}(\mathrm{y}) + g(\mathrm{y}),$$

where the remainder $g(\mathrm{y})$ vanishes to first order at 0. If $f$ is differentiable at every point of $S$, we simply say that $f$ is differentiable.

The following lemma explains the relationship between differentials and subdifferentials of convex functions.

**Lemma 5.34** If $f$ is a convex function and $f$ is differentiable at x, then the subdifferential $\partial f(\mathrm{x})$ at the point x is the one-element set $\{\mathrm{df}_\mathrm{x}\}$.

*Proof.* Let $g(\mathrm{y}) = f(\mathrm{x}+\mathrm{y}) - f(\mathrm{x}) - \mathrm{df}_\mathrm{x}(\mathrm{y})$. From the Definition 5.33 we know that $g$ vanishes to first order at 0. On the other hand, $g$ is convex because it is a convex function, minus a constant, minus a linear function. To complete the proof of the lemma it suffices to prove that the subdifferential of $g$ at 0 is a singleton set consisting of $0 \in V^*$, i.e. the constant function that maps every vector in $V$ to 0. From Definition 5.22 we know that the subdifferential $\partial g(\mathrm{x})$ is a nonempty set. To prove it equals $\{0\}$, let h be any nonzero element of $V^*$ and we will show $\mathrm{h} \notin \partial g(0)$. Suppose y is a vector such that $\mathrm{h}(\mathrm{y}) \neq 0$. Replacing y with $-\mathrm{y}$ if necessary, we can assume $\mathrm{h}(\mathrm{y}) > 0$. Now, since $\mathrm{h} \in \partial g(0)$, we have $g(\mathrm{z}) \geq g(0) + \mathrm{h}(\mathrm{z}-0) = \mathrm{h}(\mathrm{z})$ for all vectors z. In particular, letting $\mathrm{z} = t\mathrm{y}$ for $t \in \mathbb{R}$, we find that $g(\mathrm{z}) = t\mathrm{h}(\mathrm{y})$ and

$$\lim_{t\to 0} \frac{g(\mathrm{z})}{\|\mathrm{z}\|} = \lim_{t\to 0} \frac{t\,h(\mathrm{y})}{t\,\|\mathrm{y}\|} = \frac{h(\mathrm{y})}{\|\mathrm{y}\|} > 0.$$

This contradicts the fact that $g$ vanishes to first order at 0. ∎

Closely related to the differential of a function is its gradient, which encodes information about the derivative of $f$ in the form of a vector in $V$ rather than $V^*$.

**Definition 5.35** If $V$ is a vector space, $\langle\cdot,\cdot\rangle$ is a non-degenerate inner product, and $f : V \to \mathbb{R}$ is a function differentiable at x, the *gradient* of $f$ at x, denoted by $\nabla \mathrm{f}_\mathrm{x}$, is the image of the differential $\mathrm{df}_\mathrm{x}$ under the isomorphism $V^* \to V$ induced by the inner

product.

When $V = \mathbb{R}^n$ with the standard inner product structure, these definitions accord with the usual definitions given using partial derivatives. The differential of $f$ is the row vector

$$\mathrm{df_x} = \left[ \frac{\partial f}{\partial x_1} \;\; \frac{\partial f}{\partial x_2} \;\cdots\; \frac{\partial f}{\partial xn} \right]$$

and the gradient $\nabla \mathrm{f_x}$ is the column vector obtained by transposing this row vector.

▪ **Example 5.36** This example illustrates the difference between the gradient with respect to the standard inner product and the gradient with respect to a non-standard inner product. Let $V = \mathbb{R}^2$ and consider the function $f : V \to \mathbb{R}$ defined by $f(x_1, x_2) = 4x_1^2 + x_2^2$.

To calculate the differential of $f$ at $\mathrm{x} = \left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]$, we expand $f(\mathrm{x}+\mathrm{y})$ in powers of $y_1$ and $y_2$:

$$f(\mathrm{x}+\mathrm{y}) = 4(x_1+y_1)^2 + (x_2+y_2)^2 = (4x_1^2+x_2^2) + (8x_1y_1+2x_2y_2) + (4y_1^2+y_2^2) = f(\mathrm{x}) + (8x_1y_1+2x_2y_2) + g$$

where the function $g(\mathrm{y}) = 4y_1^2 + y_2^2$ vanishes to first order at 0. This indicates that

$$\mathrm{df_x}(\mathrm{y}) = 8x_1y_1 + 2x_2y_2.$$

The right side of the equation is a linear function of $\mathrm{y} \in \mathbb{R}^2$. In other words, the differential of $f$ an element of $(\mathbb{R}^2)^*$, as expected.

The gradient of $f$ with respect to the standard inner product is obtained by stacking the two partial derivatives of $f$ into a vector.

$$\nabla \mathrm{f_x} = \begin{bmatrix} 8x_1 \\ 2x_2 \end{bmatrix}.$$

What about the gradient of $f$ with respect to the non-standard inner product defined by

$$\langle \mathrm{x}, \mathrm{y} \rangle = 2x_1y_1 + x_2y_2.$$

The gradient $\nabla \mathrm{f_x}$ is defined to be the image of $\mathrm{df_x}$ under the isomorphism $(\mathbb{R}^2)^* \to \mathbb{R}^2$ induced by the inner product. In other words, $\nabla \mathrm{f_x}$ is the unique vector $\mathrm{z} = \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right]$ that satisfies

$$\forall \mathrm{y} = \left[\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right] \;\; \langle \mathrm{z}, \mathrm{y} \rangle = 8x_1y_1 + 2x_2y_2.$$

Recall that the inner product $\langle \mathrm{z}, \mathrm{y} \rangle$ is defined to be $2z_1y_1 + z_2y_2$. So, for all $\mathrm{y} \in \mathbb{R}^2$, we require the equation

$$8x_1y_1 + 2x_2y_2 = 2z_1y_1 + z_2y_2$$

to hold. Equating the coefficients of $y_1$ and $y_2$, we may conclude that $z_1 = 4x_1$ and $z_2 = 2x_2$. Hence,

$$\nabla \mathrm{f_x} = \begin{bmatrix} 4x_1 \\ 2x_2 \end{bmatrix}.$$

▪

### 5.2.4 Gradient descent

Minimizing a real-valued function on a vector space is one of the most important optimization problems in Computer Science. Among other uses, it underlies the training of machine learning models: in that application, each vector in the vector space represents a different parameter setting for the model, and the function to be minimized is called a "loss function" and is interpreted as a measure of how poorly the model with those parameters fits the training data.

The most popular family of algorithms for minimizing real-valued functions on vector spaces is based on a principle called *gradient descent*. These are iterative algorithms that take a sequence of small steps, each in a direction that locally improves the function value. In this section we introduce the gradient descent algorithm and analyze its performance when minimizing a convex function. Many of the most important contemporary applications of gradient descent involve non-convex functions, but the performance guarantees for gradient descent are much weaker when the function being optimized is non-convex.

The most elementary gradient descent algorithm has a "step size" parameter, $\eta$. The algorithm is as follows.

**Algorithm 6** Gradient descent with fixed step size

**Parameters:** Starting point $\mathrm{x}_0 \in \mathbb{R}^n$, step size $\eta > 0$, number of iterations $T \in \mathbb{N}$.

1: **for** $t = 0, \ldots, T-1$ **do**
2: $\quad \mathrm{x}_{t+1} = \mathrm{x}_t - \eta \nabla \mathrm{f}_{\mathrm{x}_t}$
3: **end for**
4: Output $\hat{\mathrm{x}} = \arg\min\{f(\mathrm{x}_0), \ldots, f(\mathrm{x}_T)\}$.

We will analyze the behavior of gradient descent under the following assumptions.

1. $V$ has a positive definite inner product, $\langle \cdot, \cdot \rangle$. Gradients and norms of vectors are defined with respect to this inner product.

2. $f$ is convex.

3. For some $L < \infty$ called the *Lipschitz constant of* $f$, the following inequality is satisfied by all $\mathrm{x}, \mathrm{y} \in V$.

$$|f(\mathrm{x}) - f(\mathrm{y})| \leq L \cdot \|\mathrm{x} - \mathrm{y}\|.$$

Let $\mathrm{x}^*$ denote a point in $V$ at which $f$ is minimized. The analysis of the algorithm will show that if $\|\mathrm{x}^* - \mathrm{x}_0\| \leq D$ then gradient descent (Algorithm 6) with $\eta = \varepsilon/L^2$ finds a point $\hat{\mathrm{x}}$ where $f(\hat{\mathrm{x}}) \leq f(\mathrm{x}^*) + \varepsilon$ after $T = L^2 D^2/\varepsilon^2$ iterations. The key parameter in the analysis is the squared distance $\|\mathrm{x}_t - \mathrm{x}^*\|^2$. The following lemma does most of the work, by showing that this parameter must decrease if $f(\mathrm{x}_t)$ is sufficiently far from $f(\mathrm{x}^*)$.

**Lemma 5.37** $\|\mathrm{x}_{t+1} - \mathrm{x}^*\|^2 \leq \|\mathrm{x}_t - \mathrm{x}^*\|^2 - 2\eta(f(\mathrm{x}_t) - f(\mathrm{x}^*)) + \eta^2 L^2$.

*Proof.* Letting $\mathrm{x} = \mathrm{x}_t$ we have

$$\begin{aligned}\|\mathrm{x}_{t+1} - \mathrm{x}^*\|^2 &= \|\mathrm{x} - \mathrm{x}^* - \eta \nabla \mathrm{f}_\mathrm{x}\|^2 \\ &= \|\mathrm{x} - \mathrm{x}^*\|^2 - 2\eta \langle \nabla \mathrm{f}_\mathrm{x}, \mathrm{x} - \mathrm{x}^* \rangle + \eta^2 \|\nabla \mathrm{f}_\mathrm{x}\|^2 \\ &= \|\mathrm{x} - \mathrm{x}^*\|^2 + 2\eta \mathrm{df}_\mathrm{x}(\mathrm{x}^* - \mathrm{x}) + \eta^2 \|\nabla \mathrm{f}_\mathrm{x}\|^2 \\ &\leq \|\mathrm{x} - \mathrm{x}^*\|^2 + 2\eta (f(x^*) - f(x)) + \eta^2 \|\nabla \mathrm{f}_\mathrm{x}\|^2.\end{aligned}$$

The proof concludes by observing that the $L$-Lipschitz property of $f$ implies $\|\nabla \mathrm{f}_\mathrm{x}\| \leq L$. ∎

Now, to complete the analysis of gradient descent, let $\Phi(t) = \|\mathrm{x}^t - \mathrm{x}^*\|^2$; we will refer to $\Phi$ as the "potential function" and to $\Phi(t)$ as the "potential at time $t$". When $\eta = \varepsilon/L^2$, the lemma implies that for every $t$ such that $f(\mathrm{x}_t) > f(\mathrm{x}^*) + \varepsilon$, the decrease in potential at time $t$ is bounded below by

$$\Phi(t) - \Phi(t+1) > 2\eta\varepsilon - \eta^2 L^2 = \varepsilon^2/L^2. \tag{5.5}$$

Since $\Phi(0) \leq D$ and $\Phi(t) \geq 0$ for all $t$, the equation (5.5) cannot be satisfied for all $0 \leq t \leq L^2 D^2/\varepsilon^2$. Hence, if we run gradient descent for $T = L^2 D^2/\varepsilon^2$ iterations, at least one of the iterates $\mathrm{x}_t$ satisfies $f(\mathrm{x}_t) \leq f(\mathrm{x}^*) + \varepsilon$, and hence the algorithm will set $\hat{\mathrm{x}}$ to be a point that satisfies such that $f(\hat{\mathrm{x}}) \leq f(\mathrm{x}^*) + \varepsilon$.

A few observations about this analysis of gradient descent are in order.

1. The upper bound on the number of iterations *does not depend on the dimension of the vector space*. The bound is $L^2 D^2/\varepsilon^2$, which depends on the Lipschitz constant of the function (namely $L$) and on the distance of the starting point $\mathrm{x}_0$ from the optimal point $\mathrm{x}^*$ (namely $D$), but the number of iterations required to find an $\varepsilon$-optimal point does not tend to infinity as the dimension increases, provided those other parameters do not increase with dimension. This partially explains why gradient descent is such a useful algorithm for contemporary optimization problems with billions of parameters, such as training very large neural networks. To be honest, though, in those applications it is quite unlikely that the initial distance from optimality, $D$, would remain constant as the number of parameters tends to infinity.
2. The number of iterations depends quadratically on $1/\varepsilon$, which is quite bad. Later in the course we will see a variant of gradient descent that needs only $O(\log(1/\varepsilon))$ iterations, when the gradient $\nabla \mathrm{f}_\mathrm{x}$ is neither too rapidly nor too slowly varying as x varies.
3. As noted in Section 5.2.3, the gradient (unlike the differential) is only well-defined in the context of an inner product structure on $V$. Under a different choice of inner product, the gradient of a function would be calculated in a different way, which would cause gradient descent to behave differently. This can be seen by plotting the iterations of gradient descent when minimizing a function such as $f(x) = 4x^2 + y^2$, whose level sets are ellipses. The gradient vectors with respect to the standard inner product are perpendicular to the level sets. The negative gradient (i.e., the direction of the steps taken by the gradient descent algorithm) is directed toward a point on the major axis of the ellipse but not toward its center. Hence, gradient descent with respect to the standard inner product will tend to zig-zag back and forth across the major axis as it makes it way toward the global minimum of $f$, repeatedly

overshooting in the $x$ direction and then correcting its course, while making steady progress in the $y$ direction. If, instead of the standard inner product, one takes the gradient of $f$ with respect to the non-standard inner product

$$\langle x, y \rangle = 2x_1y_1 + x_2y_2,$$

then the gradient descent algorithm makes steady progress in both the $x$ and $y$ directions. Thus, while gradient descent using the standard inner product is adequately efficient, if one knows something about the geometry of the function being optimized then choosing an inner product adapted to the geometry of the problem can make gradient descent even more efficient.

## 5.3 Geometry in high dimensions

When visualizing high-dimensional vector spaces, it is important to keep in mind some stark quantitative differences between low-dimensional and high-dimensional geometry. In high dimensions, when we circumscribe a cube around a sphere, the cube's volume exceeds that of the sphere by a greater-than-exponential factor. (In other words, as the dimension increases, the volume ratio of the two shapes grows faster than any exponential function of the dimension.) Almost all of the volume of a high-dimensional ball is located in a thin shell near the surface. In addition, almost all of the ball's volume is located near the equator. Finally, if we sample $m$ vectors at random from a $d$-dimensional ball and $m$ is subexponential in $d$, then with high probability all of the vectors are nearly orthogonal to one another.

### 5.3.1 Preliminaries

We will derive all of the geometric facts cited above using a few basic facts from geometry and analysis.

In the vector space $\mathbb{R}^d$ there is a function denoted by $\mathrm{Vol}_d(\cdot)$ that assigns to certain subsets $S \subseteq \mathbb{R}^d$ a non-negative (possibly infinite) number $\mathrm{Vol}_d(S)$ called the $d$-dimensional volume of $S$. The sets for which $\mathrm{Vol}_d(S)$ is defined are called *measurable sets* and we will not give a definition here, but we will note that any (topologically) closed subset of $\mathbb{R}^d$ is measurable, and the collection of measurable subsets is closed under complementation and under taking unions or intersections of countably many sets. The $d$-dimensional volume of a set $S$ contained in a $d$-dimensional hyperplane in $\mathbb{R}^n$ (i.e., a set obtained from a $d$-dimensional linear subspace by translation) because

Furthermore, the $d$-dimensional volume satisfies the following properties.

1. The $d$-dimensional volume of a set is invariant under translations and rotations.
2. When we scale a set by a scale factor $\lambda > 0$, its $d$-dimensional volume is scaled by $\lambda^d$. In other words, if we define

   $$\lambda \cdot S = \{\lambda \cdot \mathrm{x} \mid \mathrm{x} \in S\}$$

   and if $S$ is measurable, then $\lambda \cdot S$ is measurable and $\mathrm{Vol}_d(\lambda \cdot S) = \lambda^d \cdot \mathrm{Vol}_d(S)$.
3. If $A$ and $B$ are disjoint measurable sets, then $\mathrm{Vol}_d(A \cup B) = \mathrm{Vol}_d(A) + \mathrm{Vol}_d(B)$. More generally, if $A_1, A_2, \ldots$ is an infinite sequence of pairwise disjoint measurable sets,

then

$$\mathrm{Vol}_d\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{Vol}_d(A_i).$$

Define a $d$-dimensional hyperplane in $\mathbb{R}^n$ to be a set obtained from a $d$-dimensional linear subspace by translation. For every $d$-dimensional hyperplane $W$ in $\mathbb{R}^n$ we can let $w : W \to \mathbb{R}^d$ be any (Euclidean) distance-preserving bijection and define $\mathrm{Vol}_d(\cdot)$ on measurable subsets of $W$ by specifying that $\mathrm{Vol}_d(S) = \mathrm{Vol}_d(w(S))$. This definition of $\mathrm{Vol}_d(S)$ doesn't depend on the choice of distance-preserving bijection, because $\mathrm{Vol}_d$ is invariant under translations and rotations.

The volumes of $d$-dimensional and $(d-1)$-dimensional sets are related by the following integral formula.

**Fact 5.38** If $S \subseteq \mathbb{R}^d$ is measurable and $W_s = \{\mathrm{x} \in \mathbb{R}^d \mid x_1 = s\}$, then

$$\mathrm{Vol}_d(S) = \int_{-\infty}^{\infty} \mathrm{Vol}_{d-1}(S \cap W_s)\, ds.$$

Using Fact 5.38 we can derive the formula for the volume of a cone. If $T$ is a subset of $\mathbb{R}^{d-1}$ and $h > 0$, then a *cone of height $h$ with base $T$* is any set congruent to the following subset of $\mathbb{R}^d = \mathbb{R} \times \mathbb{R}^{d-1}$:

$$\mathrm{Cone}(T,h) = \{\mathrm{x} = ((1-t)h, t\mathrm{y}) \mid 0 \le t \le 1, \mathrm{y} \in T\}.$$

**Fact 5.39** If $T \subseteq \mathbb{R}^{d-1}$ is measurable, the volume of $\mathrm{Cone}(T,h)$ is $\frac{h}{d}\mathrm{Vol}_{d-1}(T)$.

*Proof.* The intersection $\mathrm{Cone}(T,h) \cap W_s$ is empty unless $0 \le s \le h$, and then its $(d-1)$-dimensional volume is $t^d \cdot \mathrm{Vol}_{d-1}(T)$, where $t$ is the solution to the equation $s = (1-t)h$; in other words, $t = 1 - \frac{s}{h}$. Using Fact 5.38 and the substitution $t = 1 - \frac{s}{h}$ we obtain

$$\mathrm{Vol}_d(\mathrm{Cone}(T,h)) = \int_0^h \left(1 - \tfrac{s}{h}\right)^d \cdot \mathrm{Vol}_{d-1}(T)\, ds = \mathrm{Vol}_{d-1}(T) \cdot \int_0^1 h t^d\, dt = \tfrac{h}{d}\mathrm{Vol}_{d-1}(T),$$

as claimed. ■

Finally, in evaluating the volumes of high-dimensional sets it will be useful for us to be able to estimate the factorial function up to a constant factor. The following lemma furnishes the required estimate.

**Lemma 5.40** For any positive integer $n$,

$$\sqrt{en}\left(\frac{n}{e}\right)^n < n! < e\sqrt{n}\left(\frac{n}{e}\right)^n. \tag{5.6}$$

*Proof.* Upon taking logarithms, the inequalities stated in the lemma become equivalent to

$$n\ln(n) - n + \tfrac{1}{2}\ln(n) + \tfrac{1}{2} < \ln(n!) < n\ln(n) - n + \tfrac{1}{2}\ln(n) + 1, \tag{5.7}$$

and we will prove the stated inequalities in this equivalent form. For all $k$ and all $t \in (0,1)$ we have

$$\ln(k) + t\,(\ln(k+1) - \ln(k)) < \ln(k+t) < \ln(k) + \tfrac{t}{k},$$

where the left inequality is derived from the fact that the logarithm function is strictly concave, and the right inequality is derived from strict concavity along with the fact that the derivative of the natural logarithm at $k$ is $\frac{1}{k}$. Integrating with respect to $t$ and applying the substitution $x = k+t$, we find that

$$\ln(k) + \tfrac{1}{2}\,(\ln(k+1) - \ln(k)) < \int_k^{k+1} \ln(x)\,dx < \ln(k) + \tfrac{1}{2k}. \tag{5.8}$$

Now, summing over $k = 1, \ldots, n-1$,

$$\ln(n!) - \tfrac{1}{2}\ln(n) < \int_1^n \ln(x)\,dx < \ln(n!) - \ln(n) + \frac{1}{2}\sum_{k=1}^{n-1}\frac{1}{k} < \ln(n!) - \tfrac{1}{2}\ln(n) + \tfrac{1}{2}, \tag{5.9}$$

where we have used the fact that $\sum_{k=1}^{n-1}\frac{1}{k} < \ln(n) + 1$. (To derive that inequality, write the sum on the left as $1 + \sum_{k=2}^{n-1}\frac{1}{k}$ and note that this is bounded above by $1 + \int_1^{n-1}\frac{dt}{t}$.) Rearranging terms in (5.9) and using the fact that $\int_1^n \ln(x)\,dx = n\ln(n) - n + 1$, we derive

$$n\ln(n) - n + \tfrac{1}{2}\ln(n) + \tfrac{1}{2} < \ln(n!) < n\ln(n) - n + \tfrac{1}{2}\ln(n) + 1 \tag{5.10}$$

as claimed. ■

### 5.3.2 Volume distribution near boundary

In this section we will explore a simple consequence of the rule for how $\mathrm{Vol}_d$ transforms under scaling, $\mathrm{Vol}_d(\lambda \cdot S) = \lambda^d \cdot \mathrm{Vol}_d(S)$. We'll see that this implies almost all of a high-dimensional sphere's volume is concentrated in a thin shell near the surface of the sphere.

**Proposition 5.41** Let $B^d(r)$ denote the Euclidean ball of radius $r$ centered at $0 \in \mathbb{R}^d$, i.e. the ball of radius $r$ in the $L_2$ norm. For any $c > 0$, the set of points whose distance from the boundary of $B^d(1)$ is greater than $c/d$ constitutes less than $e^{-c}$ fraction of the ball's volume.

*Proof.* If $c \geq d$, then the set of points whose distance from the boundary of $B = B^d(1)$ is greater than $c/d$ is empty. Otherwise, the set is equal to the interior of the ball $B^d(1 - \frac{c}{d})$, so its volume is equal to

$$\left(1 - \tfrac{c}{d}\right)^d \mathrm{Vol}_d(B).$$

To finish up, we use the inequality $1 - x < e^{-x}$ which is valid for all non-zero $x \in \mathbb{R}$. Applying this inequality with $x = \frac{c}{d}$, we find that

$$\left(1 - \tfrac{c}{d}\right)^d < \left(e^{-c/d}\right)^d = e^{-c}$$

which completes the proof of the proposition. ■

### 5.3.3 Estimating the volume of the Euclidean ball

Let $B_1^d(r), B_2^d(r), B_\infty^d(r)$ denote the unit balls of radius $r$ in $\mathbb{R}^d$ under the $L_1, L_2$, and $L_\infty$ norms, respectively. In this section we will show that the volume of $B_2^d(1)$ is $d^{-d/2+o(d)}$, where the expression $o(d)$ in the exponent indicates an error term that grows sublinearly in $d$, as $d \to \infty$. To do so, we will inscribe an $L_\infty$ ball inside $B = B_2^d(1)$ and circumscribe an $L_1$ ball around it, and we'll bound the volume of $B$ from below and above by these inscribed and circumscribed shapes.

**Lemma 5.42** For any dimension $d \geq 1$, $B_\infty^d(d^{-1/2}) \subset B_2^d(1) \subset B_1^d(d^{1/2})$.

*Proof.* Every $\mathrm{x} \in B_\infty^d(d^{-1/2})$ satisfies $|x_i| \leq d^{-1/2}$ for $i = 1, 2, \ldots, d$, which implies

$$\sum_{i=1}^{d} x_i^2 \leq \sum_{i=1}^{d} \frac{1}{d} = 1,$$

hence $\mathrm{x} \in B_2^d(1)$. To prove $B_2^d(1) \subseteq B_1^d(d^{1/2})$, consider any $\mathrm{x} \in B_2^d(1)$ and let y denote a vector in $\{\pm 1\}^d$ such that $x_i y_i \geq 0$ for all $i$; in other words, $y_i = -1$ if $x_i < 0$, $y_i = 1$ if $x_i > 0$, and $y_i$ is an arbitrary element of $\{\pm 1\}$ if $x_i = 0$. We have

$$\|\mathrm{x}\|_1 = \sum_{i=1}^{d} |x_i| = \sum_{i=1}^{d} x_i y_i \leq \|\mathrm{x}\|_2 \|\mathrm{y}\|_2,$$

where the last step is the Cauchy-Schwartz Inequality. Recalling that $\|\mathrm{x}\|_2 \leq 1$ and calculating that $\|\mathrm{y}\|_2 = d^{1/2}$, we find that $\|\mathrm{x}\|_1 \leq d^{1/2}$, as claimed. ∎

**Lemma 5.43** The unit balls of the $L_1$ and $L_\infty$ norms $\mathbb{R}^d$ have volumes

$$\mathrm{Vol}_d(B_1^d(1)) = \frac{2^d}{d!}, \quad \mathrm{Vol}_d(B_\infty^d(1)) = 2^d.$$

*Proof.* The $L_\infty$ ball $B_\infty^d$ is simply the set $[-1, 1]^d$ of vectors whose coordinates are all between $-1$ and 1. This is a product of $d$ intervals of length 2, so its volume is $2^d$.

To estimate $\mathrm{Vol}_d(B_1^d)$, first dissect $B_1^d$ into two congruent pieces: one consisting of the vectors in $B_1^d$ whose first coordinate is non-negative, and the other consisting of the vectors in $B_1^d$ whose first coordinate is non-positive. (These sets have a non-empty intersection consisting of vectors whose first coordinate is zero, but the $d$-dimensional volume of this intersection is zero.) Both pieces of this dissection are congruent to $\mathrm{Cone}(B^{d-1}, 1)$. Hence,

$$\mathrm{Vol}_d(B^d) = 2\,\mathrm{Vol}_d(\mathrm{Cone}(B^{d-1}, 1)) = \frac{2}{d}\,\mathrm{Vol}_{d-1}(B^{d-1}).$$

Solving this recurrence with the base case $\mathrm{Vol}_1(B^1) = 2$, we obtain $\mathrm{Vol}_d(B^d) = \frac{2^d}{d!}$. ∎

**Proposition 5.44** The volume of the Euclidean unit ball in $\mathbb{R}^d$ satisfies

$$\left(\frac{2}{\sqrt{d}}\right)^d < \mathrm{Vol}_d(B_2^d(1)) < \left(\frac{2e}{\sqrt{d}}\right)^d. \tag{5.11}$$

*Proof.* By Lemma 5.42 we have $\mathrm{Vol}_d(B_\infty^d(d^{-1/2})) < \mathrm{Vol}_d(B_2^d(1)) < \mathrm{Vol}_d(B_1^d(d^{1/2}))$. Applying the rule that scaling a set in $\mathbb{R}^d$ scales its volume by $\lambda$ to the formulas for $\mathrm{Vol}_d(B_\infty^d(1))$ and $\mathrm{Vol}_d(B_1^d(1))$, we can calculate the volumes of $B_\infty^d(d^{-1/2})$ and $B_1^d(d^{1/2})$ exactly and conclude that

$$2^d \cdot d^{-d/2} < \mathrm{Vol}_d(B_2^d(1)) < \frac{2^d \cdot d^{d/2}}{d!}. \tag{5.12}$$

From Lemma 5.40 we know that $\frac{1}{d!} < (\frac{e}{d})^d$, and substituting this upper bound for $\frac{1}{d!}$ into inequality (5.12), we obtain inequality (5.11). ■

Above we have estimated the volume of a Euclidean unit ball by "sandwiching" it between the unit balls of the $L_\infty$ and $L_1$ norms. A slightly more complicated way to obtain qualitatively similar estimates is to sandwich the $d$-dimensional ball between a cylinder and a cone. The benefit of the latter approach is that it enables us to estimate (within a constant factor) to volume ratio of the unit balls in $d$ and $d-1$ dimensions, which will be helpful in the following section.

**Lemma 5.45** For any $\varepsilon > 0$, the Euclidean unit ball $B = B_2^d(1)$ is contained in the cone $C(\varepsilon) = \left\{ \mathrm{x} \in \mathbb{R}^d \,\middle|\, \varepsilon x_1 + \sqrt{(1-\varepsilon^2)(x_2^2 + x_3^2 + \cdots + x_d^2)} \le 1 \right\}$.

*Proof.* For any $\mathrm{x} \in B$, apply the Cauchy-Schwartz inequality to the two-dimensional vectors

$$\mathrm{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ \sqrt{x_2^2 + \cdots + x_d^2} \end{bmatrix}, \qquad \mathrm{b} = \begin{bmatrix} \varepsilon \\ \sqrt{1-\varepsilon^2} \end{bmatrix}.$$

Observe that $\|\mathrm{a}\|_2 = \|\mathrm{x}\|_2 \le 1$ since $\mathrm{x} \in B$, and that $\|\mathrm{b}\|_2 = 1$. Hence, the Cauchy-Schwartz Inequality implies $\langle \mathrm{a}, \mathrm{b} \rangle \le 1$. Expressing this inequality in terms of the coordinates of the vector x, we find that x satisfies the inequality defining $C(\varepsilon)$. ■

**Lemma 5.46** Let $B^d = B_2^d(1)$ and $B^{d-1} = B_2^{d-1}(1)$ denote the Euclidean unit balls in $d$ and $d-1$ dimensions respectively. The volumes $\mathrm{Vol}_d(B^d)$ and $\mathrm{Vol}_{d-1}(B^{d-1})$ obey the following relation:

$$\frac{2}{\sqrt{e}} \cdot \frac{\mathrm{Vol}_{d-1}(B^{d-1})}{\sqrt{d}} \le \mathrm{Vol}_d(B^d) \le 2\sqrt{e} \cdot \frac{\mathrm{Vol}_{d-1}(B^{d-1})}{\sqrt{d}}. \tag{5.13}$$

*Proof.* Let $A$ denote the cylinder

$$A = \left\{ \mathrm{x} \in \mathbb{R}^d \,\middle|\, x_1^2 \le \frac{1}{d},\ x_2^2 + x_3^2 + \cdots + x_d^2 \le \frac{d-1}{d} \right\}$$

and observe that $A \subset B^d$ since every $\mathrm{x} \in A$ satisfies $x_1^2 + x_2^2 + \cdots + x_d^2 \le \frac{1}{d} + \frac{d-1}{d} = 1$. The height of cylinder $A$ is $\frac{2}{\sqrt{d}}$ and its base is a $(d-1)$-dimensional ball of radius

$$r = (1 + \frac{1}{d-1})^{-1/2} > e^{-1/(2d-2)}$$

so its volume is

$$\mathrm{Vol}_d(A) = \frac{2}{\sqrt{d}}\, r^{d-1}\, \mathrm{Vol}_{d-1}(B^{d-1}) > \frac{2}{\sqrt{e}} \cdot \frac{\mathrm{Vol}_{d-1}(B^{d-1})}{\sqrt{d}}. \tag{5.14}$$

Recall the infinite cone $C(\varepsilon)$ defined in Lemma 5.45, and let $-C(\varepsilon)$ denote the set $\{-\mathrm{x} \mid \mathrm{x} \in C(\varepsilon)\}$. The intersection $C = C(\varepsilon) \cap -C(\varepsilon)$ is a union two cones, each of height $\frac{1}{\varepsilon}$, whose common base is a $(d-1)$-dimensional ball whose radius is $(1-\varepsilon^2)^{-1/2}$. If we set $\varepsilon = \frac{1}{\sqrt{d}}$ then

$$(1-\varepsilon^2)^{-1/2} = \left(1 - \frac{1}{d}\right)^{-1/2} = \left(1 + \frac{1}{d-1}\right)^{1/2} < e^{1/2(d-1)}.$$

$$\mathrm{Vol}_d(C) = \frac{2}{d} \cdot \frac{1}{\varepsilon} \cdot \left(1-\varepsilon^2\right)^{-(d-1)/2} \cdot \mathrm{Vol}_{d-1}(B^{d-1}) < \frac{2\sqrt{e}}{\sqrt{d}} \cdot \mathrm{Vol}_{d-1}(B^{d-1}) \tag{5.15}$$

Since Lemma 5.45 tells us that $B^d \subset C(\varepsilon)$ and $B^d \subset -C(\varepsilon)$, we have $B \subset C$. Combining the set-theoretic relations $A \subseteq B \subseteq C$ with the volume bounds derived in Inequalities (5.14) and (5.15), we obtain the relation (5.13) asserted in the lemma statement. ■

### 5.3.4 Volume distribution near equator

As one consequence of estimating the Euclidean ball's volume, we can prove that most of the volume is located in a thin layer near the equator. In fact, letting $B = B_2^d(1)$ denote the Euclidean unit ball in $\mathbb{R}^d$, if $L_i(w) = \{x \in B \mid -w \le x_i \le w\}$ denotes a layer of width $2w$ centered on the equatorial disc $\{x \in B \mid x_i = 0\}$, then we will prove that for any $c > 0$, the complement of $L_i = L_i\left(\sqrt{c/d}\right)$ contains only $2e^{-\beta c}$ fraction of the volume of $B$, for some constant $\beta > 0$.

As a warm-up before proving this exponentially small upper bound on the volume of $B \setminus L_i$, let us prove a simpler upper bound showing that for any $c > 1$, the set $C_i = B \setminus L_i(\sqrt{c/d})$ contains at most $\frac{1}{c}$ of the volume of $B$. The key observation is that every point $\mathrm{x} \in B$ belongs to fewer than $d/c$ of the sets $C_1, C_2, \ldots, C_d$. Indeed, if $\mathrm{x} \in C_i$ then $x_i^2 > c/d$, and the constraint $\sum_{i=1}^d x_i^2 \le 1$ ensures that fewer than $d/c$ indices $i$ satisfy the inequality $x_i^2 > c/d$. Since $C_1, C_2, \ldots, C_d$ are subsets of $B$ and every point of $B$ belongs to fewer than $d/c$ of them, their combined volume is less than $\frac{d}{c}\mathrm{Vol}(B)$. Since all of the sets are congruent to each other, they all have the same volume, which must therefore be less than $\frac{1}{c}\mathrm{Vol}(B)$.

**Proposition 5.47** Let $B = B_2^d(1)$ denote the Euclidean unit ball, and for some $c \ge 4$ let $L = L_i(\sqrt{c/d})$ denote the layer of width $2\sqrt{c/d}$ around the equator. The volume of $B \setminus L$ satisfies

$$\mathrm{Vol}_d(B \setminus L) < \sqrt{\frac{e}{c}} e^{-c/2}\, \mathrm{Vol}_d(B).$$

*Proof.* If $c \ge d$ then $B \setminus L$ is an empty set and there is nothing to prove. Assume henceforth that $c < d$. Then the set $B \setminus L$ consists of two congruent spherical caps. The base of each spherical cap is a $(d-1)$-dimensional ball $B_2^{d-1}(r)$ whose radius $r$ satisfies $r^2 + \frac{c}{d} = 1$.

For $c \geq 4$ this implies $r < 1 - \frac{c}{2(d-1)} < e^{-c/2(d-1)}$. Applying Lemma 5.45 with $\varepsilon = \sqrt{\frac{c}{d}}$, we know that $B$ is contained in the infinite cone $C(\varepsilon)$. The portion of this cone sitting above the hyperplane $\{x_1 = \varepsilon\}$ has base consisting of the points x such that $x_1 = \varepsilon$ and $x_2^2 + x_3^2 + \cdots + x_d^2 \leq 1 - \varepsilon^2 = 1 - \frac{c}{d} = r^2$; this matches the base of the spherical cap. Hence, the volume of the spherical cap is less than the volume of the cone, which is

$$\frac{1}{d} \cdot \left(\frac{1}{\varepsilon} - \varepsilon\right) \cdot r^{d-1} \cdot \mathrm{Vol}_{d-1}(B^{d-1}) < \sqrt{\frac{1}{cd}} \cdot e^{-c/2} \cdot \mathrm{Vol}_{d-1}(B^{d-1}) \quad < \frac{1}{2}\sqrt{\frac{e}{c}} e^{-c/2} \cdot \mathrm{Vol}_d(B^d),$$

where the last inequality follows from Lemma 5.46. ■

### 5.3.5 Random high-dimensional vectors are nearly orthogonal

In $d$ dimensions, if non-zero vectors $\mathrm{z}_1, \ldots, \mathrm{z}_k$ are pairwise orthogonal, meaning that the dot product of any two of them is zero, then the vectors are linearly independent[2] and thus $k$ must be less than or equal to $d$. In this section we will see that the situation is completely different if we require the vectors to be *nearly orthogonal*, meaning that the angle between any two of them lies in the interval from $\frac{\pi}{2} - \varepsilon$ to $\frac{\pi}{2} + \varepsilon$ radians, for some arbitrarily small $\varepsilon > 0$. We will prove that the maximum number of pairwise nearly orthogonal vectors in $d$ dimensions grows exponentially with $d$, for any fixed $\varepsilon > 0$. The proof that we present shows, in fact, that if $m < e^{-\varepsilon^2 d/16}$, then with high probability a random set of $m$ vectors sampled independently and uniformly at random from the unit ball $B^d$ in $d$-dimensional Euclidean space will be pairwise nearly orthogonal. This is an illustration of a powerful technique called the *probabilistic method* in which one proves that an object having a certain property exists by showing that a random object possesses the property with positive probability. (In the case presented here, the "object" in question is a collection of $m$ vectors in $\mathbb{R}^d$, and the property is pairwise near orthogonality.) In many cases, including this one, directly constructing an object with the required property is much more difficult than proving the existence of such an object using the probabilistic method.

At the heart of our proof that a random $m$-tuple of vectors in $B^d$ are likely to be pairwise nearly orthogonal is the following lemma concerning the probability that two random vectors form an angle that differs from $\frac{\pi}{2}$ by more than $\varepsilon$.

**Lemma 5.48** Suppose x, y are two random vectors sampled independently and uniformly at random from $B^d$. Let $\theta \in [0, \pi]$ denote the angle formed between x and y. For any $\varepsilon$ such that $0 < \varepsilon < \frac{1}{8}$ and any $d > \frac{4e}{\varepsilon^2}$, the probability that $|\frac{\pi}{2} - \theta| > \varepsilon$ is less than $2e^{-\varepsilon^2 d/8}$.

*Proof.* The joint distribution of the pair x, y is rotation-invariant, so the conditional distribution of the angle $\theta$ given that y is parallel to the standard basis vector $\mathrm{e}_1$ is the same as the unconditional distribution of $\theta$. Furthermore, since $\theta$ depends only on the orientations of x and y, not their lengths, we can condition on the event $\mathrm{y} = \mathrm{e}_1$ without affecting the distribution of $\theta$.

Recalling now that the dot product of two vectors is equal to the product of their lengths, times the cosine of the angle between them, we find that our assumption that $\mathrm{y} = \mathrm{e}_1$ allows

[2] One way to see this must be the case is to consider the linear function $\mathrm{f}_i(\mathrm{x}) = \langle \mathrm{z}_i, \mathrm{x} \rangle$ for $i = 1, 2, \ldots, k$. By assumption, $\mathrm{f}_(\mathrm{x})$ evaluates to zero at $\mathrm{z}_j$ for any $j \neq i$, hence $\mathrm{f}(\mathrm{x}) = 0$ whenever x is a linear combination of $\{\mathrm{z}_j \mid j \neq i\}$. However, $\mathrm{f}(\mathrm{z}_i) = \langle \mathrm{z}_i, \mathrm{z}_i \rangle > 0$, so $\mathrm{z}_i$ is not a linear combination of $\{\mathrm{z}_j \mid j \neq i\}$. Since this holds for every $i$, we may conclude that $\mathrm{z}_1, \ldots, \mathrm{z}_k$ are linearly independent as claimed.

us to calculate the cosine of $\theta$ as follows.

$$\cos(\theta) = \frac{\langle \mathrm{x}, \mathrm{y} \rangle}{\|\mathrm{x}\|_2 \|\mathrm{y}\|_2} = \frac{x_1}{\|\mathrm{x}\|_2}. \tag{5.16}$$

Using the identity $\sin\left(\frac{\pi}{2} - \theta\right) = \cos(\theta)$, we find that the event $|\frac{\pi}{2} - \theta| > \varepsilon$ is equivalent to the event

$$|\sin(\theta)| > \sin(\varepsilon).$$

The inequality $\sin(\varepsilon) > \varepsilon/2$ is valid whenever $0 < \varepsilon < \frac{1}{8}$, hence

$$\Pr\left(\left|\frac{\pi}{2} - \theta\right| > \varepsilon\right) \leq \Pr\left(\frac{|x_1|}{\|\mathrm{x}\|_2} > \frac{\varepsilon}{2}\right) \leq \Pr(\|\mathrm{x}\|_2 < 1 - \varepsilon) + \Pr\left(|x_1| > (1-\varepsilon) \cdot \frac{\varepsilon}{2}\right). \tag{5.17}$$

The second inequality follows because the inequality $\frac{|x_1|}{\|\mathrm{x}\|_2} > \frac{\varepsilon}{2}$ is only satisfied when at least one of the following two events happens: either $\|\mathrm{x}\|_2 < 1 - \varepsilon$ or $|x_1| > (1-\varepsilon) \cdot \frac{\varepsilon}{2}$. Therefore the event $\frac{|x_1|}{\|\mathrm{x}\|_2} > \frac{\varepsilon}{2}$ is contained in the union of the latter two events, and its probability is bounded above by the sum of their probabilities.

Proposition 5.41 implies that the probability of the event $\|\mathrm{x}\|_2 < 1 - \varepsilon$ is less than $e^{-\varepsilon d}$, which is less than $e^{-\varepsilon^2 d/8}$ due to our assumption that $\varepsilon < 1/8$. Applying Proposition 5.47 with $c = \varepsilon^2 d/4$, and hence $\sqrt{c/d} = \varepsilon/2$, we find that the probability of the event $|x_1| > \varepsilon/2$ is less than $\sqrt{4e}\varepsilon^2 d e^{-\varepsilon^2 d/8}$, which is less than $e^{-\varepsilon^2 d/8}$ due to our assumption that $d > \frac{4e}{\varepsilon^2}$. To sum up, we have shown that both probabilities on the right side of (5.17) are less than $e^{-\varepsilon^2 d/8}$, hence the probability on the left side is less than $2e^{-\varepsilon^2 d/8}$. ■

**Proposition 5.49** For every $\varepsilon, d, m$ satisfying $0 < \varepsilon < \frac{1}{8}$, $d > \frac{4e}{\varepsilon^2}$, $m < e^{\varepsilon^2 d/16}$, if vectors $\mathrm{x}_1, \ldots, \mathrm{x}_m$ are drawn independently and uniformly at random from the Euclidean unit ball $B^d \subset \mathbb{R}^d$, then with probability at least $\frac{1}{2m}$ every pair of vectors in the set $\{\mathrm{x}_1, \ldots, \mathrm{x}_m\}$ forms an angle between $\frac{\pi}{2} - \varepsilon$ and $\frac{\pi}{2} + \varepsilon$.

*Proof.* We can use Lemma 5.48 to put an upper bound on the expected number of pairs $\mathrm{x}_i, \mathrm{x}_j$ that form an angle $\theta$ such that $|\frac{\pi}{2} - \theta| > \varepsilon$. The probability that any one such pair forms such an angle is less than $2e^{-\varepsilon^2 d/8}$, which is less than $\frac{2}{m^2}$ by our assumption on $m$. The number of unordered pairs $\{\mathrm{x}_i, \mathrm{x}_j\}$ with $i \neq j$ is

$$\binom{m}{2} = \frac{m^2 - m}{2}.$$

By linearity of expectation, the expected number of (unordered) pairs $\{\mathrm{x}_i, \mathrm{x}_j\}$ that form an angle $\theta$ not lying between $\frac{\pi}{2} - \varepsilon$ and $\frac{\pi}{2} + \varepsilon$ is less than

$$\frac{2}{m^2} \cdot \frac{m^2 - m}{2} = 1 - \frac{1}{2m}.$$

The proposition follows, since a non-negative integer-valued random variable must always satisfy the inequality $\mathbb{E}[X] \geq \Pr(X > 0)$. ■

## 5.4 Matrices

A matrix is a two-dimensional array of real numbers, $M$, with entries denoted by $M_{ij}$. Here, the ranges of $i$ and $j$ are finite intervals $[m] = \{1,2,\ldots,m\}$ and $[n] = \{1,2,\ldots,n\}$, where $m$ and $n$ are the number of *rows* and *columns*, respectively, of the matrix $M$.

Matrices play at least three distinct important roles in mathematics, computer science, and data science.

1. They encode information that takes the form of a two-dimensional array. A running example in this section will be a matrix encoding course enrollments in a department, with two rows that tabulate the number of undergraduate and graduate students, respectively, and with one column for each course offered by the department. In this example, the column for course $j$ would contain entries $M_{1j}$ and $M_{2j}$ encoding the number of undergraduates and grad students, respectively, enrolled in course $j$.
2. An $m \times n$ matrix can encode a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. In this encoding, the matrix $M$ encodes the function $T : \mathbb{R}^n \to \mathbb{R}^m$ where $T(\mathrm{x})$ is the vector $\mathrm{y} \in \mathbb{R}^m$ whose coordinates are defined, for each $i \in [m]$ by the equation

$$y_i = \sum_{j=1}^{n} M_{ij} x_j.$$

3. An $m \times n$ matrix can also encode a *bilinear function* on $\mathbb{R}^m \times \mathbb{R}^n$. A function $A : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is called bilinear if it satisfies

$$\forall \mathrm{x},\mathrm{y} \in \mathbb{R}^m, \mathrm{z} \in \mathbb{R}^n \;\; A(a\mathrm{x}+b\mathrm{y},\mathrm{z}) = aA(\mathrm{x},\mathrm{z}) + bA(\mathrm{y},\mathrm{z})$$
$$\forall \mathrm{x} \in \mathbb{R}^m, \mathrm{y},\mathrm{z} \in \mathbb{R}^n \;\; A(\mathrm{x},a\mathrm{y}+b\mathrm{z}) = aA(\mathrm{x},\mathrm{y}) + bA(\mathrm{x},\mathrm{z}).$$

Equivalently, $A$ is bilinear if and only if for every $\mathrm{y} \in \mathbb{R}^n$ the function $f(\mathrm{x}) = A(\mathrm{x},\mathrm{y})$ is a linear function of x, and for every $\mathrm{x} \in \mathbb{R}^m$ the function $g(\mathrm{y}) = A(\mathrm{x},\mathrm{y})$ is a linear function of y. We say that matrix $M$ encodes the bilinear function $A : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ if

$$A(\mathrm{x},\mathrm{y}) = \sum_{i=1}^{m}\sum_{j=1}^{n} M_{ij} x_i y_j.$$

### 5.4.1 Change of basis

One of the tricky things about working with matrices is that we often want to write a matrix representing "the same thing" as $M$ using a different basis. Doing this can be confusing because the way to rewrite $M$ depends on what "thing" we are encoding using $M$.

▪ **Example 5.50** Let us return to our running example of a matrix $M$ with 2 rows and $n$ columns, representing the enrollments of $n$ courses by noting the number of undergraduate students in the first row and the number of graduate students in the second row. A different matrix representing the same information might have the total number of students in the first row and the number of graduate students in the second row. Let us call this second matrix $M'$. Its relationship to $M$ can be expressed by the formulas

$$M'_{1j} = M_{1j} + M_{2j}, \qquad M'_{2j} = M_{2j}$$

or more succinctly by the equation

$$M' = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right] M.$$

The matrix $B = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ is a "change of basis" matrix describing how the entries of $M$ transform when we rewrite the data in the format of $M'$.

To illustrate the subtlety of working with change-of-basis matrices, let us now suppose that the university's budget model credits the department with \$2 for every undergraduate student and \$1 for every graduate student. (These aren't realistic numbers, we're just using them for the sake of this example.) Consider course $j$ whose enrollment is represented in the first basis by the vector $\mathrm{m}_j = \begin{bmatrix} M_{1j} \\ M_{2j} \end{bmatrix}$ and in the second basis by the vector $\mathrm{m'}_j = \begin{bmatrix} M'_{1j} \\ M'_{2j} \end{bmatrix}$. The department's revenue from course $j$ can be calculated by the expression $\left[\begin{smallmatrix} 2 & 1 \end{smallmatrix}\right]\mathrm{m}_j$ (\$2 for every undergraduate student plus \$1 for every graduate student), but it can also be calculated by the expression $\left[\begin{smallmatrix} 2 & -1 \end{smallmatrix}\right]\mathrm{m'}_j$ (\$2 for every student, minus \$1 for every graduate student). Evidently, the change of basis which transforms $\mathrm{m}_j$ to $\mathrm{m'}_j$ also transforms the linear function represented by the row vector $\left[\begin{smallmatrix} 2 & 1 \end{smallmatrix}\right]$ to the one represented by the row vector $\left[\begin{smallmatrix} 2 & -1 \end{smallmatrix}\right]$, even though

$$\left[\begin{smallmatrix} 2 & 1 \end{smallmatrix}\right] \cdot \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right] \neq \left[\begin{smallmatrix} 2 & -1 \end{smallmatrix}\right] \qquad \text{and} \qquad \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right] \cdot \left[\begin{smallmatrix} 2 \\ 1 \end{smallmatrix}\right] \neq \left[\begin{smallmatrix} 2 \\ -1 \end{smallmatrix}\right].$$

What's going on here is that a linear function represented in the first basis by a row vector r becomes represented in the second basis by a row vector $\mathrm{r}' = \mathrm{r}B^{-1}$. A change of basis which operates on vectors via left multiplication by $B$ operates on linear functions (represented as row vectors) via right multiplication by $B^{-1}$. If we choose to represent a linear function of m as an inner product $\langle \mathrm{c}, \mathrm{m} \rangle$, where c is a column vector, then the change-of-basis formula becomes even more obscure: the change of basis that transforms m to $B\mathrm{m}$ acts on c by transforming it into $(B^{-1})^\top \mathrm{c}$. ∎

To derive the correct change-of-basis formulae for different types of vectors and matrices it is useful to introduce the notion of a *based vector space*. This is not a widely used mathematical term, but just a useful term we are using in this course to simplify the discussion of how to account for a change of basis.

**Definition 5.51** A *based vector space* is a finite-dimensional vector space $V$ together with a choice of isomorphism $\beta : \mathbb{R}^n \to V$ for some $n \in \mathbb{N}$.

Recall that $\mathbb{R}^n$ has a *standard basis* $\mathrm{e}_1, \ldots, \mathrm{e}_n$ where $\mathrm{e}_i$ has a 1 in its $i^{\text{th}}$ coordinate and 0 in every other coordinate. If $V$ is a based vector space then the vectors $\beta(\mathrm{e}_1), \ldots, \beta(\mathrm{e}_n)$ constitute a basis of $V$. Conversely, if $\mathrm{v}_1, \ldots, \mathrm{v}_n$ is an ordered $n$-tuple of vectors that form a basis of $V$, then there is a unique isomorphism $\beta : \mathbb{R}^n \to V$ such that $\beta(\mathrm{e}_i) = \mathrm{v}_i$. Thus, giving a vector space $V$ the structure of a based vector space is equivalent to choosing a basis for $V$ and arranging the elements of the basis into an ordered $n$-tuple.

For a vector space $V$ whose elements are semantically meaningful (e.g., course enrollments rather than abstract ordered pairs of numbers), giving $V$ the structure of a based vector space is tantamount to settling on a convention for how to represent elements of $V$ as $n$-tuples of numbers. This phenomenon already exists — and is well known — in the context of one-dimensional vector spaces, where the process of representing physical quantities as numbers requires choosing units. For example, it is meaningless to say, "The mass of my textbook is 2.5," whereas the sentence, "The mass of my textbook is 2.5 kilograms," is perfectly meaningful. In this case, masses of physical objects can be interpreted

as elements of an abstract one-dimensional vector space $V$ in which addition represents the operation of combining two disjoint physical objects. Two different choices of units, such as kilograms versus grams, are represented by two differed based vector space structures $\beta_{\text{kg}} : \mathbb{R} \to V$ and $\beta_{\text{g}} : \mathbb{R} \to V$ that send the element $1 \in \mathbb{R}$ to a one-kilogram mass and a one-gram mass, respectively. Choosing different based vector space structures for a higher-dimensional vector space $V$ can be interpreted as a higher-dimensional counterpart to the process of converting between two different systems of units.

**Definition 5.52** If $V$ is an $n$-dimensional vector space and $\beta_1 : \mathbb{R}^n \to V$ and $\beta_2 : \mathbb{R}^n \to V$ are two different based vector space structures on $V$, the linear transformation $\beta_2^{-1} \circ \beta_1 : \mathbb{R}^n \to \mathbb{R}^n$ is represented by an $n \times n$ matrix called the *change of basis matrix* from $\beta_1$ to $\beta_2$.

▪ **Example 5.53** Returning to our running example, course enrollments can be interpreted as elements of an abstract two-dimensional vector space $V$. When course enrollments are represented as columns of the matrix $M$ in Example 5.50, this corresponds to choosing a based vector space structure $\beta_1$ on $V$ that sends $\text{e}_1$ to the element of $V$ represented in matrix $M$ by the column vector $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$, i.e. a course with one undergraduate and no graduate students. Let us denote this element of $V$ by u, for "undergraduate". Meanwhile $\beta_1(\text{e}_2)$ is the element of $V$ represented in matrix $M$ by the column vector $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$, i.e. a course with no undergraduates and one graduate student. Let us denote this element of $V$ by g, for "graduate".

The matrix $M'$ represents course enrollments (i.e., elements of $V$) in an alternate data format that corresponds to a different based vector space structure. In this structure, $\beta_2(\text{e}_1)$ is the element of $V$ represented in matrix $M$ by the column vector $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$, i.e. a course with one student in total, but zero graduate students. This is again the vector $\text{u} \in V$. However, $\beta_2(\text{e}_2)$ is the element of $V$ represented in matrix $M'$ by the column vector $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$, i.e. a course with *zero students in total, but one graduate student!* It's a bit hard to wrap one's head around what this means, but the most natural way to interpret it is that adding this vector to a course enrollment represents the operation of one undergraduate dropping the course and being replaced by a graduate student. (That operation has zero effect on the total number of students, but it increments the number of graduate students.) In other words, $\beta_2(\text{e}_2) = \text{g} - \text{u}$.

Now, let's compute the change of basis matrix $B$. It is a two-by-two matrix whose columns are $B\text{e}_1$ and $B\text{e}_2$. We can calculate each column as follows.

$$B\text{e}_1 = \beta_2^{-1}(\beta_1(\text{e}_1)) = \beta_2^{-1}(\text{u}) = \beta_2^{-1}(\beta_2(\text{e}_1)) = \text{e}_1$$
$$B\text{e}_2 = \beta_2^{-1}(\beta_1(\text{e}_2)) = \beta_2^{-1}(\text{g}) = \beta_2^{-1}(\text{u} + (\text{g} - \text{u})) = \beta_2^{-1}(\beta_2(\text{e}_1) + \beta_2(\text{e}_2)) = \text{e}_1 + \text{e}_2.$$

Hence, $B = \left[\begin{smallmatrix}1 & 1\\0 & 1\end{smallmatrix}\right]$, consistent with the change of basis formula for converting matrix $M$ to $M'$ derived in Example 5.50. ▪

As the preceding example makes clear, when a matrix $M$ represents a data table whose columns belong to a vector space $V$, if we change from one basis of $V$ to another, the matrix $M$ is transformed to $BM$, where $B$ is the change-of-basis matrix. When $M$ represents a linear transformation from $V$ to $W$ or a bilinear function on $V \times W$, the rules for how $M$ transforms under a change of basis for $V$ or $W$ (or both) can be derived by reasoning about the equations that must be satisfied after the change of basis.

For example, suppose $V$ and $W$ are vector spaces of dimensions $n$ and $m$, respectively. Suppose $V$ and $W$ each have two different bases, represented by based vector space

structures $\beta_{V1}$ and $\beta_{V2}$ in the case of $V$, and by $\beta_{W1}$ and $\beta_{W2}$ in the case of $W$. Let $B_V$ and $B_W$ denote the respective change of basis matrices. Consider any linear transformation $T : V \to W$ and let $M_1, M_2$ be the matrices that represent $T$ with respect to the based vector space structures $\beta_{V1}, \beta_{W1}$ and $\beta_{V2}, \beta_{W2}$, respectively. Then for all $\mathrm{x} \in V$,

$$T(\mathrm{x}) = \beta_{W1}(M_1(\beta_{V1}^{-1}(\mathrm{x})))$$
$$T(\mathrm{x}) = \beta_{W2}(M_2(\beta_{V2}^{-1}(\mathrm{x})))$$

hence

$$\beta_{W1} \circ M_1 \circ \beta_{V1}^{-1} = \beta_{W2} \circ M_2 \circ \beta_{V2}^{-1}.$$

To isolate a formula for $M_2$ we multiply the last equation on the left by $\beta_{W2}^{-1}$ and on the right by $\beta_{V2}$, obtaining

$$M_2 = \beta_{W2}^{-1} \circ \beta_{W1} \circ M_1 \circ \beta_{V1}^{-1} \circ \beta_{V2} = B_W M_1 B_V^{-1}. \tag{5.18}$$

▪ **Example 5.54** We return once more to our running example of course enrollments. Recall that in Example 5.50, if $V$ is the vector space of course enrollments, we defined a linear function $V \to \mathbb{R}$ represented in basis $\beta_{V1}$ by the row vector $\mathrm{r} = [2\ 1]$. When we transform to the basis $\beta_{V2}$, the change-of-basis formula for a linear transformation specifies that we should transform r to $\mathrm{r}B_V^{-1}$. This explains the reason why right-multiplication by the inverse of the change-of-basis matrix is the appropriate way to transform the coefficient vector of a linear function. ▪

Now let us explore how the matrix representing a bilinear function transforms under change of basis. Recall that a bilinear function $A$ on $\mathbb{R}^m \times \mathbb{R}^n$ is represented by a matrix $M$ satisfying

$$A(\mathrm{x}, \mathrm{y}) = \sum_{i=1}^{m} \sum_{j=1}^{n} M_{ij} x_i y_j = \sum_{i=1}^{m} x_i \sum_{j=1}^{n} M_{ij} y_j = \langle \mathrm{x}, M\mathrm{y} \rangle .$$

More generally, if $A : V \times W \to \mathbb{R}$ is a bilinear function and $\beta_V, \beta_W$ are based vector space structures on $V$ and $W$, respectively, then the matrix $M$ representing $A$ with respect to these bases satisfies

$$\forall \mathrm{v} \in V, \mathrm{w} in W\ A(\mathrm{v}, \mathrm{w}) = \left\langle \beta_V^{-1}(\mathrm{v}), M\beta_W^{-1}(\mathrm{w}) \right\rangle .$$

As before, if $V$ and $W$ each have two based vector space structures, denoted by $\beta_{V1}, \beta_{V2}$ and $\beta_{W1}, \beta_{W2}$, and the bilinear function $A$ is represented by matrices $M_1$ and $M_2$ with respect to these two pairs of based vector space structures, then we have the equation

$$\forall \mathrm{v} in V, \mathrm{w} \in W\ \left\langle \beta_{V2}^{-1}(\mathrm{v}), M_2 \beta_{W2}^{-1}(\mathrm{w}) \right\rangle = \left\langle \beta_{V1}^{-1}(\mathrm{v}), M_1 \beta_{W1}^{-1}(\mathrm{w}) \right\rangle .$$

Let $\mathrm{v} = \beta_{V2}(\mathrm{x})$ and $\mathrm{w} = \beta_{W2}(\mathrm{y})$.

$$\forall \mathrm{x} \in \mathbb{R}^m, \mathrm{y} \in \mathbb{R}^n\ \langle \mathrm{x}, M_2 \mathrm{y} \rangle = \left\langle \beta_{V1}^{-1}(\beta_{V2}(\mathrm{x})), \beta_{W1}^{-1}(\beta_{W2}(\mathrm{y})) \right\rangle = \left\langle B_V^{-1}\mathrm{x}, B_w^{-1}\mathrm{y} \right\rangle = \left\langle \mathrm{x}, (B_V^{-1})^\top M_1 B_W^{-1} \mathrm{y} \right\rangle$$

where the last step used the identity $\langle M\mathrm{x}, \mathrm{y} \rangle = \left\langle \mathrm{x}, M^\top \mathrm{y} \right\rangle$. Hence $M_2 = (B_V^{-1})^\top M_1 B_W^{-1}$.

### 5.4.2 Adjoints and orthogonality

Taking the transpose of a matrix is an important operation in linear algebra. When the matrix represents a linear transformation between abstract vector spaces, the linear transformation that corresponds to the transpose of the matrix is called its *adjoint* and is defined in the following lemma. Before stating the lemma, we need the following definition.

**Definition 5.55** If $V$ is a finite-dimensional vector space with an inner product $\langle\cdot,\cdot\rangle_V$, a based vector space structure $\beta : \mathbb{R}^n \to V$ is said to be *compatible* with the inner product structure on $V$ if it satisfies

$$\forall \mathrm{x},\mathrm{y} \in \mathbb{R}^n \ \langle \mathrm{x},\mathrm{y}\rangle = \langle \beta(\mathrm{x}), \beta(\mathrm{y})\rangle_V$$

where $\langle \mathrm{x},\mathrm{y}\rangle$ denotes the dot product of x and y, i.e. the standard inner product structure on $\mathbb{R}^n$.

**Lemma 5.56** If $V, W$ are finite-dimensional vector spaces, each equipped with a non-degenerate inner product, and $T : V \to W$ is a linear transformation, then there is a unique linear transformation $U : W \to V$ called the *adjoint* of $T$, that satisfies

$$\forall \mathrm{v} \in V, \mathrm{w} \in W \ \langle T\mathrm{v}, \mathrm{w}\rangle = \langle \mathrm{v}, U\mathrm{w}\rangle .$$

If $\beta_V, \beta_W$ are based vector space structures on $V, W$ that are compatible with their respective inner products, and $M_T, M_U$ are the matrices representing $T$ and its adjoint $U$, respectively, then $M_U$ is the transpose of $M_T$.

*Proof.* Because the inner product structures on $V$ and $W$ are non-degenerate, there are isomorphisms $\iota_V : V \to V^*$ and $\iota_W : W \to W^*$ such that $\iota_V(\mathrm{x})$ is the linear function f that, when applied to a vector $\mathrm{y} \in V$, yields the inner product $\mathrm{f}(\mathrm{y}) = \langle \mathrm{x},\mathrm{y}\rangle$, and $\iota_W$ is defined similarly using the inner product on $W$. Let $T^* : W^* \to V^*$ denote the linear transformation such that for all $\mathrm{g} \in W^*$, $T^*(\mathrm{g})$ is the linear function $\mathrm{f} \in V^*$ defined by $\mathrm{f}(\mathrm{y}) = \mathrm{g}(T(\mathrm{y}))$. Let $U = \iota_V^{-1} \circ T^* \circ \iota_W$. Then for any $\mathrm{v} \in V, \mathrm{w} \in W$, if we let $\mathrm{f} = T^*(\iota_W(\mathrm{w}))$, then

$$\langle \mathrm{v}, U\mathrm{w}\rangle = \left\langle \mathrm{v}, \iota_V^{-1}(T^*(\iota_W(\mathrm{w})))\right\rangle = \left\langle \mathrm{v}, \iota_V^{-1}(\mathrm{f})\right\rangle = \left\langle \iota_V^{-1}(\mathrm{f}), \mathrm{v}\right\rangle = \mathrm{f}(\mathrm{v}) = \iota_W(\mathrm{w})(T\mathrm{v}) = \langle \mathrm{w}, T\mathrm{v}\rangle = \langle T\mathrm{v}, \mathrm{w}\rangle ,$$

which verifies that $U$ satisfies the equation defining the adjoint of $T$. To verify that $U$ is unique, observe that if $U'$ also satisfies the defining equation of the adjoint, then for all $\mathrm{v} \in V, \mathrm{w} \in W$,

$$\left\langle \mathrm{v}, U\mathrm{w} - U'\mathrm{w}\right\rangle = \langle \mathrm{v}, U\mathrm{w}\rangle - \left\langle \mathrm{v}, U'\mathrm{w}\right\rangle = \langle T\mathrm{v}, \mathrm{w}\rangle - \langle T\mathrm{v}, \mathrm{w}\rangle = 0.$$

Since v was an arbitrary vector in $V$ and the inner product on $V$ is non-degenerate, this implies that $U\mathrm{w} - U'\mathrm{w} = 0$. Since w was an arbitrary vector in $W$, this means $U = U'$.

Finally, the fact that $M_U$ is the transpose of $M_T$ can be checked by verifying that the standard inner product on $\mathbb{R}^n$ satisfies $\langle M\mathrm{x},\mathrm{y}\rangle = \left\langle \mathrm{x}, M^\top \mathrm{y}\right\rangle$ for all vectors $\mathrm{x},\mathrm{y} \in \mathbb{R}^n$ and matrices $M \in \mathbb{R}^{n\times n}$. ■

A matrix $M \in \mathbb{R}^{n\times n}$ is called *symmetric* if $M = M^\top$, and it is called *orthogonal* if $M^\top$ is the inverse of $M$. Based on Lemma 5.56 we can generalize the definitions of symmetric and orthogonal matrices to the setting of abstract inner product spaces as follows.

**Definition 5.57** If $V$ is a vector space with a non-degenerate inner product and $T : V \to V$ is a linear transformation, we say that $T$ is *self-adjoint* with respect to the inner product on $V$ if it equal to its own adjoint. In other words, a self-adjoint linear transformation is one that satisfies the equation

$$\langle T\mathrm{x}, \mathrm{y}\rangle = \langle \mathrm{x}, T\mathrm{y}\rangle$$

for all $\mathrm{x}, \mathrm{y} \in V$. We say that $T$ is *orthogonal* with respect to the inner product on $V$ if its adjoint is $T^{-1}$. Equivalently, an orthogonal linear transformation is one that satisfies the equation

$$\langle T\mathrm{x}, T\mathrm{y}\rangle = \langle \mathrm{x}, \mathrm{y}\rangle$$

for all $\mathrm{x}, \mathrm{y} \in V$.

### 5.4.3 Symmetric positive definite matrices

A very important set of square matrices are the symmetric positive definite matrices, i.e. the set of all matrices that represent positive definite inner products on $\mathbb{R}^n$. There are a number of equivalent characterizations of symmetric positive definite matrices, and all of them are important in different contexts. In this section we present several equivalent characterizations and prove their equivalence. A key starting point for the proof is the following observation.

**Lemma 5.58** If $V$ in a vector space of dimension $n < \infty$ with a positive definite inner product $\langle \cdot, \cdot \rangle_V$, then $V$ is isomorphic to $\mathbb{R}^n$ with the standard inner product structure. In other words there is a based vector space structure $\beta : \mathbb{R}^n \to V$ such that for all $\mathrm{x}, \mathrm{y} \in \mathbb{R}^n$,

$$\forall \mathrm{x}, \mathrm{y} \in \mathbb{R}^n \ \langle \mathrm{x}, \mathrm{y}\rangle = \langle \beta\mathrm{x}, \beta\mathrm{y}\rangle_V \,. \tag{5.19}$$

*Proof.* The proof is by induction on $n$. When $n = 0$ there is nothing to prove, since $V$ and $\mathbb{R}^n$ are both singleton sets consisting of the vector 0, whose inner product with itself is 0.

For $n > 0$, let $W$ be an $(n-1)$-dimensional subspace of $V$, equipped with the inner product structure obtained by restricting $\langle \cdot, \cdot \rangle_V$ to pairs of vectors in $W$. There is a linear transformation $T : V \to W^*$ that maps each vector $\mathrm{x} \in V$ to the linear function $f_\mathrm{x} : W \to \mathbb{R}$ defined by $f_\mathrm{w}(\mathrm{w}) = \langle \mathrm{x}, \mathrm{w}\rangle_V$. Let $\mathrm{v}_1, \ldots, \mathrm{v}_n$ be a basis of $V$. The vectors $T(\mathrm{v}_1), \ldots, T(\mathrm{v}_n) \in W^*$ must be linearly dependent, since $\dim(W^*) = \dim(W) = n - 1$. Hence we can express $0 \in W^*$ as a non-trivial linear combination

$$0 = \sum_{i=1}^{n} a_i T(\mathrm{v}_i) = T\left(\sum_{i=1}^{n} a_i \mathrm{v}_i\right)$$

where the coefficients $a_1, \ldots, a_n$ are not all equal to zero. Let $\mathrm{v} = \sum_{i=1}^n a_i \mathrm{v}_i$, which is a nonzero vector in $V$ since $\mathrm{v}_1, \ldots, \mathrm{v}_n$ is a basis and $a_1, \ldots, a_n$ are not all zero. Recalling the definition of the linear transformation $T$, we see that the equation $T(\mathrm{v}) = 0$ means

$$\forall \mathrm{w} \in W \ \langle \mathrm{v}, \mathrm{w}\rangle_V = 0. \tag{5.20}$$

Since the inner product on $V$ is positive definite and $\mathrm{v} \neq 0$, we know that $\langle \mathrm{v}, \mathrm{v}\rangle_V > 0$. Rescaling v if necessary, we can assume $\langle \mathrm{v}, \mathrm{v}\rangle = 1$. The rescaling doesn't affect the validity of (5.20).

The induction hypothesis implies there is an isomorphism $\beta_W : \mathbb{R}^{n-1} \to W$ such that $\langle \mathrm{x}, \mathrm{y} \rangle = \langle \beta_W \mathrm{x}, \beta_W \mathrm{y} \rangle_V$ for all $\mathrm{x}, \mathrm{y} \in \mathbb{R}^{n-1}$. Let us now define $\beta : \mathbb{R}^n \to V$ by specifying that

$$\beta \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \beta_W \begin{bmatrix} x_1 \\ \vdots \\ x_{n-1} \end{bmatrix} + x_n \mathrm{v}.$$

We must verify that this $\beta$ satisfies Equation (5.19). For any $\mathrm{x}, \mathrm{y} \in \mathbb{R}^n$, let $\mathrm{x}', \mathrm{y}'$ denote the vectors in $\mathbb{R}^{n-1}$ obtained by extracting the first $n-1$ coordinates of x and y, respectively. We have

$$\begin{aligned} \langle \beta \mathrm{x}, \beta \mathrm{y} \rangle_V &= \left\langle \beta_W \mathrm{x}' + x_n \mathrm{v}, \beta_W \mathrm{y}' + y_n \mathrm{v} \right\rangle_V \\ &= \left\langle \beta_W \mathrm{x}', \beta_W \mathrm{y}' \right\rangle_V + x_n \left\langle \mathrm{v}, \beta_W \mathrm{y}' \right\rangle_V + y_n \left\langle \beta_W \mathrm{x}', \mathrm{v} \right\rangle_V + x_n y_n \langle \mathrm{v}, \mathrm{v} \rangle_V \end{aligned}$$

Thinking about the four terms on the right side, the induction hypothesis implies that the first term equals $\langle \mathrm{x}', \mathrm{y}' \rangle$, the second and third terms vanish because of equation (5.20), and the four term equals $x_n y_n$ because we normalized v to ensure $\langle \mathrm{v}, \mathrm{v} \rangle = 1$. Hence,

$$\langle \beta \mathrm{x}, \beta \mathrm{y} \rangle_V = \left\langle \mathrm{x}', \mathrm{y}' \right\rangle + x_n y_n = \langle \mathrm{x}, \mathrm{y} \rangle ,$$

as desired. ■

**Proposition 5.59** For a square matrix $M \in \mathbb{R}^{n \times n}$ the following properties are equivalent.

1. The bilinear function $f(\mathrm{x}, \mathrm{y}) = \langle \mathrm{x}, M\mathrm{y} \rangle$ is a positive definite inner product.
2. $M = BB^\top$ for some invertible square matrix $B$.
3. $M = BB^\top$ for some (possibly rectangular) matrix $B$ whose column space is $\mathbb{R}^n$.
4. $M = \sum_{i=1}^m a_i \mathrm{x}_i \mathrm{x}_i^\top$ for some coefficients $a_1, \ldots, a_m > 0$ and some sequence of vectors $\mathrm{x}_1, \ldots, \mathrm{x}_m \in \mathbb{R}^n$ that contains a basis for $\mathbb{R}^n$.
5. $M = QDQ^\top$ for some orthogonal matrix $Q$ and diagonal matrix $D$ with positive diagonal entries.
6. $M$ is a symmetric matrix whose eigenvalues are all strictly positive.

# 6. Markov Chains and Sampling Algorithms

Markov chains model discrete-time random processes whose future state evolution depends only on the present state, not on the entire sequence of states leading up to the present. As such, they represent an important class of probabilistic models. However, in algorithm design they serve an important additional role: the most popular algorithmic procedure for sampling from complicated probability distributions is to design an appropriate Markov chain and simulate its state evolution. This method is known as *Markov Chain Monte Carlo (MCMC)*. In these notes we will present some aspects of the fundamental theory of Markov chains and of the MCMC paradigm for designing sampling algorithms.

Before delving into definitions, let us give some examples to illustrate what we mean by "sampling from complicated probability distributions."

▪ **Example 6.1** If $G$ is a $q$-colorable graph then the uniform distribution on proper $q$-colorings of $G$ is easy to define but potentially hard to sample. For example if $q \geq 3$ and $G$ is allowed to be an arbitrary graph, it is NP-hard to decide if *any* $q$-coloring of $G$ exists, let alone sample a uniformly random one. ▪

▪ **Example 6.2** Generalizing the preceding example, given a graph $G$ and two parameters $\beta, \gamma$, we may want to sample a random labeling of its vertices using labels in some set $\Sigma$, i.e. a random function $L : V(G) \to \Sigma$, with probability proportional to

$$w(L) = \prod_{(u,v)\in E(G)} \begin{cases} \beta & \text{if } L(u) = L(v) \\ \gamma & \text{if } L(u) \neq L(v). \end{cases}$$

The first example (sampling a random $q$-coloring) specializes this one by setting $|\Sigma| = q$, $\beta = 0$, $\gamma = 1$. ▪

▪ **Example 6.3** Given a tuple of non-negative integers $(d_1, d_2, \ldots, d_n)$, consider the set of graphs with vertex set $[n] = \{1, 2, \ldots, n\}$ such that for all $i \in [n]$ the degree of vertex $i$ is $d_i$.

When this set is non-empty, one may wish to draw random samples from it. For example, sampling graphs from this distribution may be useful for simulating the performance of algorithms or distributed protocols on networks that resemble (in terms of their size and degree distribution) observed real-world network topologies. Alternatively, the ability to draw samples from this distribution may aid a statistician in testing the hypothesis that a network topology observed in the real world has some structure that is statistically distinguishable from random graphs with the same size and degree distribution. ∎

∎ **Example 6.4** Suppose we are given:

1. a deep neural network (DNN) that generates random images by transforming an input layer of independent (Gaussian) random numbers into an output layer of pixels;
2. an image $I$ with some missing pixels.

The DNN defines a probability distribution over output images (i.e., the distribution that results from feed-forward propagation of Gaussian random numbers at the input layer), and one may wish to draw samples from the conditional distribution over output images, conditioned on the pixel values matching the data present in $I$. For example, this sampling task may form part of the pipeline in an image completion algorithm: given a DNN that models natural scenes, and an image with a natural scene in the background and an object in the foreground that occludes part of the scene, the sampling algorithm could be used to generate hypothetical completions of the background image. ∎

One can define the following class of algorithmic random sampling problems that includes all of the examples above, along with many other important and practical random sampling problems.

**Definition 6.5** An *unnormalized distribution* on a finite set $\mathbf{X}$ is a function $w : \mathbf{X} \to \mathbb{R}_{\geq 0}$ such that

$$Z_w \stackrel{\Delta}{=} \sum_{x \in \mathbf{X}} w(x) > 0.$$

The corresponding probability distribution is $p(x) = w(x)/Z_w$. Sampling from $w$ refers to the process of drawing a random sample $x \in \mathbf{X}$ with probability $p(x) = w(x)/Z_w$. Approximately sampling from $w$ refers to any process that draws a random sample $x$ from $\mathbf{X}$ such that for all $A \subseteq \mathbf{X}$,

$$|\Pr(x \in A) - \sum_{y \in A} p(y)| \leq \varepsilon$$

for some specified approximation parameter $\varepsilon > 0$.

One can often specify an unnormalized distribution $w$ by specifying an efficient algorithm to calculate $w(x)$ for every $x \in \mathbf{X}$. This brings us to the main question we address below.

> *Given an efficient algorithm for evaluating an unnormalized distribution $w(x)$, when is it possible to efficiently draw random samples from the probability distribution $p = w/Z_w$?*

Before continuing, let us pause to illustrate how the first and last examples above can be cast as special cases of this problem.

For the example of sampling a random $q$-coloring of a graph $G = (V, E)$, we can take $\mathbf{X}$ to be the set of all functions from $V$ to $[q]$ (called "labelings" henceforth), and we can take $w$ to be a function that assigns the value 1 to labelings that are proper colorings of $G$ and 0 to all other labelings. Then the probability distribution $p$ is the uniform distribution on proper colorings of $G$.

For the example of image completion, we can take $\mathbf{X}$ to be the set of all functions that label each node of the DNN with a number called the node's *activation*.[1] We can then define $w(x)$ to be zero if the node activations in $x$ don't obey the DNN's weights and activation functions, or if the values in the output layer don't match the pixel values given in the input, $I$. However, when $x$ does obey the DNN's weights and activation functions and matches the given pixel values in the output layer, we define $w(x)$ to be the product of the (Gaussian) probabilities of the input node activations. Then the distribution $p(x)$ is the conditional distribution defined in Example 6.4.

## 6.1 Markov chains and their stationary distributions

In this section we formally define Markov chains, introduce the notion of a stationary distribution, and identify conditions under which a Markov chain has a unique stationary distribution such that the marginal distribution of the time-$t$ state is guaranteed to converge to the stationary distribution as $t \to \infty$.

**Definition 6.6** A Markov chain with (finite) state set $\mathbf{X}$ is a probability distribution on infinite sequences $X_0, X_1, \ldots$ of elements of $\mathbf{X}$, satisfying the Markov property:

$$\forall t > 0 \, \forall (x_0, x_1, \ldots, x_t) \in \mathbf{X}^{t+1} \quad \Pr(X_t = x_t \,|\, X_0 = x_0, \ldots, X_{t-1} = x_{t-1}) = \Pr(X_t = x_t \,|\, X_{t-1} = x_{t-1}).$$

In other words, the conditional distribution of $X_t$ depends only on the value of $X_{t-1}$ and not on any of the values that came before time $t-1$.

A Markov chain is *time-homogeneous* if for all pairs $(x, y) \in \mathbf{X}^2$, and all $t > 0$,

$$\Pr(X_t = x | X_{t-1} = y) = \Pr(X_{t+1} = x | X_t = y).$$

For a time-homogeneous Markov chain, the matrix $P$ defined by $P_{xy} = \Pr(X_t = y | X_{t-1} = x)$ is called the *transition matrix*.

For the remainder of these lecture notes, all the Markov chains we consider will be time-homogeneous. Accordingly, when we use the term *Markov chain* below it always implicitly refers to a time-homogeneous Markov chain.

The probability distribution of a Markov chain's state at time $t$ can be represented by a row vector $\pi_t \in \mathbb{R}^{\mathbf{X}}$, whose $x^{\text{th}}$ coordinate is the probability that $X_t = x$:

$$(\pi_t)_x = \Pr(X_t = x).$$

[1] Since our formalism requires $\mathbf{X}$ to be finite, we must quantize the set of numbers that can be used as a node's label. For example, we could limit the label set to be the set of 32-bit floating point numbers, or we could quantize node activations even more aggressively. Such quantization schemes have been advocated in the neural network literature, for the sake of making the training and inference process more efficient in terms of storage space, running time, and energy consumption.

For $t > 0$ we can then calculate that

$$\begin{aligned}(\pi_t)_x = \Pr(X_t = x) &= \sum_{y \in \mathbf{X}} \Pr(X_t = x \wedge X_{t-1} = y) \\ &= \sum_{y \in \mathbf{X}} \Pr(X_t = x \,|\, X_{t-1} = y) \cdot \Pr(X_{t-1} = y) = \sum_{y \in \mathbf{X}} (\pi_{t-1})_y P_{yx}\end{aligned}$$

This can be summarized more succinctly as

$$\pi_t = \pi_{t-1} P$$

and, by induction, we obtain

$$\pi_t = \pi_0 P^t.$$

**Definition 6.7** A probability distribution $\pi$ is a stationary distribution for a Markov chain with transition matrix $P$ if it satisfies

$$\pi P = \pi.$$

A stationary distribution is thus a fixed point of the Markov chain's transition dynamics: if the initial state distribution $\pi_0$ is equal to the stationary distribution $\pi$, then every future state $\pi_t$ is also distributed according to $\pi$.

It turns out that every Markov chain with finite state set has a stationary distribution. This fact, as well as a sufficient condition for the stationary distribution to be unique, can be deduced from the Perron-Frobenius Theorem, a fundamental theorem from linear algebra that concerns the eigenvalues of square matrices with non-negative entries.

**Definition 6.8** If $A$ is an $n \times n$ square matrix with non-negative entries, let $G_A$ be the directed graph (potentially with self-loops) having vertex set $[n]$ and edge set $\{(i, j) | A_{ij} > 0\}$. We say $A$ is *irreducible* if $G_A$ is strongly connected, and we say $A$ is *aperiodic* if the cycle lengths in $G_A$ have no common divisor greater than 1.

Irreducible matrices are characterized by the property that every entry of $A + A^2 + A^3 + \ldots + A^n$ is strictly positive. Among irreducible matrices, the aperiodic ones are characterized by the property that for some positive integer $k$, every entry of $A^k$ is strictly positive.

**Theorem 6.9 — Perron-Frobenius.** If $A$ is an irreducible $n \times n$ square matrix with non-negative entries, then $A$ has a unique right eigenvector $v \in \mathbb{R}^n$ whose components are strictly positive. The eigenvalue associated to $v$, called the *Perron-Frobenius eigenvalue*, has multiplicity one, and every other (complex) eigenvalue $\lambda'$ satisfies $|\lambda'| \leq \lambda$. This inequality is strict if $A$ is aperiodic.

The proof of the Perron-Frobenius Theorem can be found in many linear algebra textbooks, for example Felix Gantmacher's *The Theory of Matrices* (AMS Chelsea Publishing, 2000). For the sake of making these lecture notes self-contained, we will prove an easier result that pertains to Markov chain transition matrices.

**Theorem 6.10** If $P$ is the transition matrix of an irreducible, aperiodic Markov chain with finite state set, then there is a unique stationary distribution $\pi$ such that $\pi P = \pi$. For any starting distribution $\pi_0$, the time-$t$ state distribution $\pi_t = \pi_0 P^t$ converges to $\pi$ as $t \to \infty$. In fact, the convergence is exponentially fast: there are constants $C < \infty$ and $\delta > 0$ such that

$$\|\pi_t - \pi\|_1 \leq C(1-\delta)^t$$

for all $t \in \mathbb{N}$.

*Proof.* Since $P$ is irreducible and aperiodic, there exists some $k$ such that all entries of $P^k$ are positive. Let $N = |\mathbf{X}|$ denote the number of states of the Markov chain, and choose $\varepsilon < 0$ such that all entries of $P^k$ are greater than or equal to $\varepsilon/N$. Let $Q = (11^\top)/N$. Then

$$P^k = \varepsilon Q + (1-\varepsilon)R$$

where $R$ is a non-negative matrix.

A *row-stochastic matrix* is a non-negative matrix whose row sums are all equal to 1. Equivalently, the non-negative matrix $A$ is called row-stochastic if $A1 = 1$; from this characterization it is evident that the set of row-stochastic matrices is closed under multiplication. Note that $Q$ is row-stochastic since $1^\top 1 = N$. Furthermore, $P$ is row-stochastic since for every $x \in \mathbf{X}$ we have $\sum_y P_{xy} = \sum_{y \in \mathbf{X}} \Pr(X_t = y \,|\, X_{t-1} = x) = 1$. Hence $P^k$ is row-stochastic, and we may conclude that $R$ is also row-stochastic using the equation

$$(1-\varepsilon)R1 = P^k 1 - \varepsilon Q 1 = 1 - \varepsilon 1 = (1-\varepsilon)1.$$

For $t \geq 0$ let $\Delta_t = \pi_{t+1} - \pi_t = \pi_0(P^{t+1} - P^t)$. We have

$$\Delta_t Q = \frac{1}{N}\pi_0(P^{t+1} - P^t)11^\top = 0,$$

since $(P^{t+1} - P^t)ones = 1 - 1 = 0$. Therefore,

$$\Delta_{t+k} = \Delta_t P^k = (1-\varepsilon)\Delta_t R.$$

The inequality $\|vR\|_1 \leq \|v\|_1$ holds for any vector $v$. To prove this, it suffices to verify it when $\|v\|_1 \leq 1$. A vector whose 1-norm is less than or equal to 1 is a convex combination of the standard basis vectors and their negations, hence we only need to check that $\|vR\|_1 \leq 1$ when $v$ is one of the standard basis vectors. In that case $vR$ is a row of $R$, i.e. a non-negative vector whose components sum up to 1, so $\|vR\| = 1$. Now, using the inequality $\|vR\|_1 \leq \|v\|_1$, we find that

$$\|\Delta_{t+k}\|_1 \leq (1-\varepsilon)\|\Delta_t\|.$$

For any $t \in \mathbb{N}$, if $q = \lfloor t/k \rfloor$, then

$$\begin{aligned}
\sum_{s=t}^{\infty} \|\Delta_s\| &\leq \sum_{r=q}^{\infty} \sum_{i=0}^{k-1} \|\Delta_{kr+i} \\
&\leq \sum_{r=q}^{\infty} \sum_{i=0}^{k-1} (1-\varepsilon)^r \|\Delta_i\| \\
&= \frac{(1-\varepsilon)^q}{\varepsilon}(\|\Delta_0\|_1 + \cdots + \|\Delta_{k-1}\|_1)
\end{aligned}$$

This confirms that the sequence $\pi_t = \pi_0 + \sum_{s=0}^{t-1} \Delta_s$ converges absolutely as $t \to \infty$ and that the rate of convergence is exponential. Denote the limit point by $\pi$. To conclude the proof we must show that $\pi$ is a stationary distribution of $P$. The equation $\pi P = \pi$ follows by observing that

$$\pi P = \lim_{t\to\infty}(\pi_t P) = \lim_{t\to\infty} \pi_{t+1} = \pi.$$

The fact that $\pi$ is a probability distribution follows from the fact that $\pi_t$ is a probability distribution for each $t$, and that the set of probability distributions on $\mathbb{R}^{\mathbf{X}}$ is topologically closed. ■

## 6.2 Reversible Markov chains and the Metropolis-Hastings algorithm

In general, computing the stationary distribution of a Markov chain requires solving a linear system, but there is one case in which the stationary distribution has a simple closed-form formula. This is the case of a reversible Markov chain.

In this section, for an unnormalized distribution $w$, we will use the notations $w(x)$ and $w_x$ interchangeably.

**Definition 6.11** A Markov chain with transition matrix $P$ is reversible with respect to (unnormalized) distribution $w$ if it satisfies

$$w_x P_{xy} = w_y P_{yx}$$

for all $x, y \in \mathbf{X}$.

**Lemma 6.12** If $P$ is reversible with respect to $w$, then $\pi = w/Z_w$, is a stationary distribution for $P$.

*Proof.* Multiplying both sides of the reversibility equation $w_x P_{xy} = w_y P_{yx}$ by the normalizing constant $Z_w^{-1} = (\sum_x w_x)^{-1}$, we find that $\pi_x P_{xy} = \pi_y P_{yx}$ for all $x, y \in \mathbf{X}$. Hence,

$$(\pi P)_x = \sum_{y\in S} \pi_y P_{yx} = \sum_{y\in S} \pi_x P_{xy} = \pi_x \left( \sum_{y\in S} P_{xy} \right) = \pi_x.$$

■

The reversibility condition can be interpreted as a type of "detailed balance" condition: at stationarity, the rate of state transitions from $x$ to $y$ equals the rate of state transitions from $y$ to $x$, for all state pairs $x$ and $y$.

The Metropolis-Hastings algorithm is a procedure that takes an unnormalized distribution $w$ and creates a Markov chain $P$ whose state transitions are computationally easy to simulate, and whose stationary distribution is $\bar{w}$. Actually the procedure makes use of an auxiliary Markov chain $K$, called the *proposal distribution*, whose stationary distribution is simple and often unrelated to $w$. In many applications the stationary distribution of $K$ is simply the uniform distribution on $\mathbf{X}$. To define the Metropolis-Hastings algorithm we assume we have:

1. An unnormalized probability distribution specified by a function $\kappa : \mathbf{X} \to [0, 1]$.

2. A Markov chain $K$ that is reversible with respect to $\kappa$.
3. Algorithms for sampling state transitions of $K$ and for computing the function $\kappa$.

The Markov chain $K$ is called the *proposal distribution* for the Metropolis-Hastings procedure. As stated earlier, in many applications $\kappa(x) = 1$ for all $x$ (i.e., the normalization of $\kappa$ is the uniform distribution on $\mathbf{X}$) and the reversibility condition $\kappa_x K_{xy} = \kappa_y K_{yx}$ simply states that the Markov transition matrix $K$ is a symmetric matrix.

Now for $x \neq y$ define

$$P_{xy} = K_{xy} \cdot \kappa_x \cdot \frac{\min\{w_x, w_y\}}{w_x}, \tag{6.1}$$

and define $P_{xx} = 1 - \sum_{y \neq x} P_{xy}$. Note that

$$\sum_{y \neq x} P_{xy} = \kappa_x \cdot \left( \sum_{y \neq x} K_{xy} \frac{\min\{w_x, w_y\}}{w_x} \right) \leq \kappa_x \cdot \left( \sum_{y \neq x} K_{xy} \right) \leq \kappa_x \leq 1,$$

so $P_{xx} \geq 0$. Thus, $P$ is indeed a Markov transition matrix.

**Lemma 6.13** The Markov chain $P$ defined by Equation (6.1) is reversible with respect to $w$.

*Proof.* Consider any $x, y \in \mathbf{X}$. If $x = y$ then the equation $w_x P_{xy} = w_y P_{yx}$ holds trivially. Otherwise,

$$w_x P_{xy} = K_{xy} \cdot \kappa_x \cdot \min\{w_x, w_y\}$$
$$w_y P_{yx} = K_{yx} \cdot \kappa_y \cdot \min\{w_y, w_x\}.$$

The lemma follows because $\min\{w_x, w_y\} = \min\{w_y, w_x\}$ and because our assumption that $K$ is reversible with respect to $\kappa$ implies $K_{xy}\kappa_x = K_{yx}\kappa_y$. ■

An algorithm to simulate state transitions of the Markov chain $P$ can be described as follows. Suppose the current state of the Markov chain is $x \in \mathbf{X}$.

1. Using the sampling oracle for Markov chain $K$, sample "proposed state" $y \in \mathbf{X}$ with probability $K_{xy}$.
2. Compute $w_x, w_y$, and $\kappa_x$.
3. With probability $\frac{\min w_x, w_y}{w_x} \cdot \kappa_x$, transition to state $y$.
4. Otherwise, remain at state $x$.

▪ **Example 6.14 — Glauber dynamics for sampling $q$-colorings.** To illustrate the Metropolis-Hastings procedure, we show how to use it to define a simple Markov chain whose unique stationary distribution is the uniform distribution over proper $q$-colorings of a graph $G = (V, E)$. For two labelings $x, y : V \to [q]$ define their Hamming distance as

$$d(x, y) = \#\{v \in V \mid x(v) \neq y(v)\}.$$

Assume that $q$ is large enough that the graph whose vertices are proper $q$-colorings of $G$, and whose edges are pairs of colorings whose Hamming distance is 1, constitutes a

non-empty connected graph. (If this graph is not connected, the Markov chain defined here will be reducible and it will have multiple stationary distributions.)

We will take $\kappa(x) = 1$ for all $x \in \mathbf{X}$, and for our proposal distribution we will define $n = |V|$ and

$$K_{xy} = \begin{cases} \frac{1}{nq} & \text{if } d(x,y) = 1 \\ \frac{1}{q} & \text{if } x = y \\ 0 & \text{otherwise.} \end{cases}$$

A state transition of $K$ can be simulated by the following algorithm: starting from state $x$, sample vertex $v \in V$ and color $c \in [q]$ independently and uniformly at random, and let $y$ be the state obtained from $x$ by recoloring $v$ with color $c$ and leaving all other colors the same. From the definition of $K$ it follows easily that $K_{xy} = K_{yx}$, i.e. $K$ is reversible with respect to $\kappa$.

Recall that our goal is to design a Markov chain whose stationary distribution is the uniform distribution on proper $q$-colorings of $G$. In other words, we want to draw samples from the distribution given by the unnormalized density function $w$ such that $w(x) = 1$ when $x$ is a proper coloring and $w(x) = 0$ otherwise. To simulate a state transition of the Markov chain $P$ defined by the Metropolis-Hastings procedure we do the following steps, starting from state $x$. Assume that $x$ is a proper coloring.

1. *Sample "proposed state" $y \in \mathbf{X}$ with probability $K_{xy}$.*
   In other words, sample vertex $v \in V$ and color $c \in [q]$ independently and uniformly at random, and let $y$ be the state obtained from $x$ by recoloring $v$ with color $c$ and leaving all other colors the same.
2. *Compute $w_x, w_y$, and $\kappa_x$.*
   By assumption, $x$ is a proper coloring, so $w_x = \kappa_x = 1$. Recall from above that $w_y = 1$ if and only if $y$ is a proper coloring. Since $x$ is a proper coloring and $y$ is obtained from $x$ by recoloring $v$, we only need to check whether every edge incident to $v$ remains properly colored. In other words, to execute this step we merely need to test whether vertex $v$ has any neighbor whose color is already $c$. If so, $w_y = 0$; otherwise, $w_y = 1$.
3. *With probability $\frac{\min\{w_x, w_y\}}{w_x} \cdot \kappa_x$, transition to state $y$.*
   The probability in question is 1 if the color of every neighbor of $v$ is different from $c$, and 0 otherwise.
4. *Otherwise, remain at state $x$.*

Hence, the Metropolis-Hastings Algorithm in this case corresponds to the following very simple procedure. The starting state of the Markov chain is any proper coloring of $G$. To simulate one state transition, we sample a uniformly random vertex $v$ and uniformly random color $c$, and we change the color of $v$ to $c$ if and only if the color of every neighbor of $v$ is different from $c$. This Markov chain on the set of proper colorings of $G$ is called *Glauber dynamics*. ∎

## 6.3 Mixing time

The ability to efficiently simulate state transitions of a Markov chain whose stationary distribution is $\pi$ doesn't necessarily imply the ability to efficiently draw samples from $\pi$,

or from a distribution close to $\pi$. The reason is that the Markov chain might be *slowly mixing*: for small — or even moderately large — values of $t$, the state distribution after $t$ steps, $\pi_t$, might be quite far from the eventual stationary distribution, $\pi$. Distance between distributions is often measured using the *total variation distance* (also known as statistical distance):

$$\|\pi - \pi'\|_{TV} = \max_{S \subset \mathbf{X}} \{|\pi(S) - \pi'(S)|\} = \tfrac{1}{2}\|\pi - \pi'\|_1.$$

(The second equation can be confirmed by observing that the maximum of $|\pi(S) - \pi'(S)|$ is attained when $S = \{x \mid \pi(x) > \pi'(x)\}$.)

**Definition 6.15** For any $\varepsilon > 0$ and any irreducible Markov chain $P$, the *$\varepsilon$-mixing time* $\tau_P(\varepsilon)$ is defined to be the smallest $t_0$ such that for all initial state distributions $\pi_0$ and all $t \geq t_0$, the time-$t$ state distribution $\pi_t = \pi_0 P^t$ satisfies $\|\pi_t - \pi\|_{TV} \leq \varepsilon$, where $\pi$ denotes the stationary distribution of $P$.

Theorem 6.10 shows that when $P$ is irreducible and aperiodic, the mixing time $\tau_P(\varepsilon)$ depends logarithmically on $1/\varepsilon$ as $\varepsilon \to 0$. On the other hand, since we are primarily interested in Markov chains whose state space $|\mathbf{X}|$ is exponentially large (i.e., exponential in the size of the problem description) it is usually very important to understand how $\tau_P(\varepsilon)$ depends on $|\mathbf{X}|$.

**Definition 6.16** A Markov chain $P$ is called *rapidly mixing* if its mixing time $\tau_P(\varepsilon)$ is bounded above by a polynomial function of $\log|\mathbf{X}/\varepsilon|$.

Determining which Markov chains are rapidly mixing and which ones aren't is a very active research area. In the following section we will present a very useful technique for proving rapid mixing of Markov chains.

## 6.4 Coupling

This section presents a method for bounding the mixing time of a Markov chain by "coupling" two parallel executions of the Markov chain that start from different states but converge toward occupying the same state as time progresses.

**Definition 6.17** If $\pi, \pi'$ are two probability distributions on a sample set $\mathbf{X}$, a *coupling* of $\pi$ and $\pi'$ is a probability measure $\nu$ on ordered pairs $(x, x') \in \mathbf{X} \times \mathbf{X}$ such that the marginal distribution of $x$ is $\pi$ and the marginal distribution of $x'$ is $\pi'$. In other words, for every set $S \subseteq \mathbf{X}$,

$$\nu(S \times \mathbf{X}) = \pi(S), \qquad \nu(\mathbf{X} \times S) = \pi'(S).$$

The total variation distance has an important characterization in terms of coupling.

**Lemma 6.18** $\|\pi - \pi'\|_{TV} = \inf\{\nu(x \neq x') \mid \nu \text{ a coupling of } \pi, \pi'\}$.

*Proof.* Let $\Delta = \{(x,x) \mid x \in \mathbf{X}\} \subseteq \mathbf{X} \times \mathbf{X}$, and let $\Delta^c$ denote the complementary set,

$$\Delta^c = \{(x,x') \mid x \neq x'\} \subset \mathbf{X} \times \mathbf{X}.$$

The probability denoted by $\nu(x \neq x')$ in the lemma statement can also (more accurately) be written as $\nu(\Delta^c)$. If $\nu$ is a coupling of $\pi$ and $\pi'$, then for every set $S \subseteq \mathbf{X}$,

$$\pi(S) - \pi'(S) = \nu(S \times \mathbf{X}) - \nu(\mathbf{X} \times S) \leq \nu(S \times (\mathbf{X} \setminus S)) \leq \nu(\Delta^c).$$

Since the inequality holds for every $S \subseteq \mathbf{X}$ and every coupling $\nu$, it follows that

$$\sup_{S \subseteq \mathbf{X}} \|\pi(S) - \pi'(S)\| \leq \inf\{\nu(\Delta^c) \mid \nu \text{ a coupling of } \pi, \pi'\}.$$

The left side is $\|\pi - \pi'\|_{TV}$, so we have proven an inequality between the two sides of the equation asserted by the lemma. To prove the opposite inequality, we directly construct a coupling $\nu$ such that $\|\pi - \pi'\|_{TV} = \nu(\Delta^c)$. For this purpose, let $\delta = \|\pi - \pi'\|_{TV}$. If $\delta = 0$ then $\pi = \pi'$ and the coupling can simply be defined by setting $\nu(x,x) = \pi(x) = \pi'(x)$ for all $x \in \mathbf{X}$ and $\nu(x,x') = 0$ for $x \neq x'$. If $\delta > 0$ then for each $x \in \mathbf{X}$ let

$$\begin{aligned}\delta(x) &= (\pi(x) - \pi'(x))^+ = \max\{\pi(x) - \pi'(x), 0\} \\ \delta'(x) &= (\pi'(x) - \pi(x))^+ = \max\{\pi'(x) - \pi(x), 0\}\end{aligned}$$

and define

$$\nu(x,x') = \begin{cases} \min\{\pi(x), \pi'(x)\} & \text{if } x = x' \\ \delta^{-1} \cdot \delta(x) \cdot \delta'(x') & \text{if } x \neq x'. \end{cases}$$

If $S = \{x \mid \pi(x) > \pi'(x)\}$ then $\pi(S) - \pi'(S) = \|\pi - \pi'\|_{TV} = \delta$. This justifies the following identities.

$$\sum_{x \in \mathbf{X}} \delta(x) = \sum_{x : \pi(x) > \pi'(x)} (\pi(x) - \pi'(x)) = \pi(S) - \pi'(S) = \delta \tag{6.2}$$

$$\sum_{x' \in \mathbf{X}} \delta'(x') = \sum_{x : \pi'(x) \geq \pi(x)} (\pi'(x) - \pi(x)) = \pi'(\mathbf{X} \setminus S) - \pi(\mathbf{X} \setminus S) = \delta. \tag{6.3}$$

Using these identities we can see that $\nu$ is a coupling of $\pi$ and $\pi'$.

$$\sum_{x' \in \mathbf{X}} \nu(x,x') = \min\{\pi(x), \pi'(x)\} + \sum_{x' \neq x} \delta^{-1} \cdot \delta(x) \cdot \delta'(x') = \min\{\pi(x), \pi'(x)\} + \delta^{-1} \cdot \delta(x) \cdot \sum_{x' \neq x} \delta'(x').$$

If $\pi(x) \leq \pi'(x)$ then $\min\{\pi(x), \pi'(x)\} = \pi(x)$ and $\delta(x) = 0$, so the right side equals $\pi(x)$ as required by the definition of a coupling. If $\pi(x) > \pi'(x)$ then $\delta'(x) = 0$, so the right side is equal to $\min\{\pi(x), \pi'(x)\} + \delta^{-1} \cdot \delta(x) \cdot \sum_{x' \in \mathbf{X}} \delta'(x')$. According to equation (6.2) the sum equals $\delta$, so the entire right side is equal to $\min\{\pi(x), \pi'(x)\} + \delta(x)$, which equals $\pi(x)$. Thus, in either case, $\sum_{x' \in \mathbf{X}} \nu(x,x') = \pi(x)$ as required by the definition of coupling. The proof that $\sum_{x \in \mathbf{X}} \nu(x,x') = \pi'(x')$ follows similarly. Finally, to prove that $\nu(\Delta^c) = \delta$, we calculate

$$\nu(\Delta) = \sum_{x \in \mathbf{X}} \min\{\pi(x), \pi'(x)\} = \sum_{x \in \mathbf{X}} (\pi(x) - \delta(x)) = \sum_{x \in \mathbf{X}} \pi(x) - \sum_{x \in \mathbf{X}} \delta(x) = 1 - \delta$$

and subtract both sides of this equation from 1. ■

A special case of coupling two probability distributions occurs when both of the probability distributions are Markov chains with the same transition matrix.

**Definition 6.19** A *Markov coupling* with transition matrix $P$ and initial state distributions $\pi_0, \pi_0'$ is a probability distribution over sequences of pairs $\{(X_t, X_t') \mid t = 0, 1, \ldots\}$ such that:

1. The distributions of the random sequences $X_0, X_1, X_2, \ldots$ and $X_0', X_1', X_2', \ldots$ are both Markov chains with transition matrix $P$.

2. The distribution of $X_0$ is $\pi_0$, and the distribution of $X'_0$ is $\pi'_0$.

Although each of the random state sequences $X_0, X_1, \ldots$ and $X'_0, X'_1, \ldots$ in a Markov coupling must evolve according to the transition matrix $P$, they may use shared randomness to evolve in a correlated way. In particular, by constructing Markov couplings in which $X_t$ and $X'_t$ tend to become more similar over time, we can bound mixing times of Markov chains.

**Lemma 6.20 — Markov Coupling Lemma.** Let $P$ be a Markov transition matrix with stationary distribution $\pi$. For any $t_0 \in \mathbb{N}$ and $\varepsilon > 0$, the mixing time bound $\tau_P(\varepsilon) \leq t_0$ is implied by the following sufficient condition: every initial state distribution $\pi_0$ has a Markov coupling with transition matrix $P$ and initial state distributions $\pi_0, \pi$, satisfying $\Pr(X_t \neq X'_t) \leq \varepsilon$ for all $t \geq t_0$.

*Proof.* Let $\pi = \pi'_0$ be the stationary distribution of $P$. Since $X'_0$ is distributed according to $\pi$ and $\pi$ is stationary for $P$, the distribution of $X'_t$ must be equal to $\pi$ for every $t > 0$ as well. Letting $\pi_t$ denote the distribution of $X_t$, we find that the joint distribution of the pair $(X_t, X'_t)$ is a coupling of $\pi_t$ with $\pi$. Lemma 6.18 now implies that $\|\pi_t - \pi\|_{TV} \leq \varepsilon$ for all $t \geq t_0$, hence $\tau_P(\varepsilon) \leq t_0$. ■

▪ **Example 6.21 — Lazy random walk on the hypercube.** As a first example of a Markov coupling, let us analyze the following Markov chain with state space $probspc = \{0,1\}^n$, called "lazy random walk on the hypercube." Given state $X_t \in \{0,1\}^n$, the following state $X_{t+1}$ is sampled by setting $X_{t+1} = X_t$ with probability $\frac{1}{2}$, and otherwise choosing one of the $n$ bits of $X_t$ uniformly at random and flipping that bit to obtain $X_{t+1}$. The state transition dynamics can equivalently be described by a process that inverts the order of the two random decisions, i.e. which bit to flip and whether or not to be "lazy" and remain at $X_t$.

1. Sample a coordinate $i_t \in [n]$ uniformly at random.
2. Sample a uniformly random bit $b_t \in \{0,1\}$.
3. Let $X_{t+1}$ equal $X_t$ with the $i_t^{\text{th}}$ bit set to $b_t$.

To analyze the mixing time of the lazy random walk on the hypercube, we will use a Markov coupling. Specifically, consider two Markov chains $X_0, X_1, \ldots$ and $X'_0, X'_1, \ldots$ that start from (potentially) different initial states but evolve according to the sampling rule described above, at each time step $t$ using the same random index $i_t$ and random bit $b_t$ to define the transitions $X_t \to X_{t+1}$ and $X'_t \to X'_{t+1}$. An easy inductive argument establishes that for all $t \geq 0$, the strings $X_t$ and $X'_t$ match on the set of coordinates indexed by $I_t = \{i_s \mid s < t\}$. This is true for $t = 0$ since $I_0$ is the empty set. For $t > 0$ and $i \in I_t$, either we have $i = i_t$, in which case that $i^{\text{th}}$ coordinates of $X_t$ and $X'_t$ are both equal to $b_t$, or $i \in I_{t-1}$. In the latter case, by the inductive hypothesis the $i^{\text{th}}$ coordinates of $X_{t-1}$ and $X'_{t-1}$ are equal, and then there is no way for them to become unequal since our Markov coupling doesn't allow any coordinates of $X_t$ and $X'_t$ to become unequal if they were equal in the preceding time step.

Having established that $X_t$ and $X'_t$ match in every coordinate indexed by $I_t$, we can conclude that $X_t = X'_t$ whenever $I_t = [n]$. Analyzing the $\varepsilon$-mixing time of the lazy random walk on the hypercube thus boils down to analyzing the question, "What is the distribution of the earliest time $\tau$ such that $I_\tau = [n]$?" Since the sequence $i_0, i_1, \ldots$ is a sequence of independent uniform draws from the set $[n]$, this question is just a restatement of the

coupon collector problem! Our analysis of the coupon collector problem implies that the answer is

$$\tau_P(\varepsilon) = n \ln n + O\left(n/\sqrt{\varepsilon}\right).$$

■

### 6.4.1 Analyzing Card Shuffling via Coupling

One of the most famous applications of mixing time analysis is to *card shuffling*: how many times must one shuffle a deck of cards, if one wants the resulting permutation of the cards to be close to uniformly distributed? This is a question about the mixing time of a Markov chain whose states are permutations of the cards, and whose transition probabilities $P_{\sigma\tau}$ represent the probability that a deck of cards initially ordered according to the permutation $\sigma$ becomes ordered according to $\tau$ after one shuffle.

Of course, to analyze the mixing time of the card-shuffling Markov chain, we need to specify a mathematical model of the act of shuffling. The most popular such model is the Gilbert-Shannon-Reeds model. The *GSR shuffle* of a deck of $n$ cards is the random permutation obtained by the following procedure, which resembles the physical process of a "riffle shuffle."

1. First, one divides the deck into two halves, a "left half" and a "right half", consisting of the first $m$ cards and the last $n-m$ cards respectively, where $m$ is a random sample from the binomial distribution $\mathrm{B}\left(n, \frac{1}{2}\right)$, i.e. the probability of sampling $m$ cards in the left half is $\binom{n}{m} / 2^n$.
2. Next, the left and right halves of the deck are randomly interleaved to form the shuffled deck. The shuffled deck is assembled from the bottom up, by iteratively selecting the last remaining card from the left or right half, choosing between them with probability proportional to the number of cards remaining in each half. In other words, if $\ell$ cards remain in the left half and $r$ cards remain in the right half, the probability that the next card placed into the shuffled deck is from the left half is $\ell/(\ell+r)$ and the probability that it is from the right half is $r/(\ell+r)$.

The following equivalent description of the process is somewhat simpler to state, bears less resemblance to the physical act of shuffling a deck of cards by riffling two halves together.

1. A random binary string $x \in \{0,1\}^n$ is sampled.
2. Let $I_0$ denote the set of $i \in [n]$ such that $x_i = 0$, and let $I_1$ denote the set of $i \in [n]$ such that $x_i = 1$. Let $m$ denote the number of elements in $I_0$.
3. The first $m$ cards of the deck are matched, in an order-preserving manner, to the positions in the permuted deck identified by the index set $I_0$.
4. The last $n-m$ cards of the deck are matched, in an order-preserving manner, to the positions in the permuted deck identified by the index set $I_1$.

Bounding the $\varepsilon$-mixing time of the Markov chain defined by the GSR shuffle is equivalent to finding a value of $t$ such that the composition of $t$ random GSR shuffles is $\varepsilon$-close to a uniformly random permutation in total variation distance. Equivalently, we are looking for a $t$ such that the *inverse* of the composition of $t$ random GSR shuffles is $\varepsilon$-close to uniformly random. Restating the problem in this way is convenient because the second

description of the GSR shuffle above admits a beautifully simple description of a procedure for sampling the *inverse* permutation.

1. Each card selects an element of $\{0,1\}$ independently and uniformly and random.
2. The cards that selected 0 are moved to the front of the deck, preserving their order.
3. The cards that selected 1 are moved to the end of the deck, preserving their order.

We can couple two executions of the inverse-GSR Markov chain by having the cards choose identical random bits in both branches of the coupling. Starting from initial permutations $X_0$ and $X_0'$ respectively, the states reached after $t$ transitions are permutations defined by the following rule for sorting cards.

1. Each card $i$ selects a length-$t$ string $b_t(i) \in \{0,1\}^t$ independently and uniformly at random.
2. Cards are sorted in lexicographically increasing order of the strings $b_t(i)$, breaking ties according to the card's position in $X_0$ or $X_0'$, depending whether we are defining the permutation $X_t$ or $X_t'$.

The probability that $X_t \neq X_t'$ is bounded above by the probability that there exists a pair of cards $i \neq j$ such that $b_t(i) = b_t(j)$. Thus, to bound the mixing time of the GSR shuffle, we are led to the following question: *if $n$ elements of $\{0,1\}^t$ are sampled independently and uniformly at random, how large must $t$ be so that the probability of two elements being equal is less than $\varepsilon$?* We learned how to resolve this type of question when we learned about the birthday paradox. The expected number of collisions is $\binom{n}{2}/2^t$, so if $t$ is large enough that $\binom{n}{2}/2^t < \varepsilon$ then the collision probability will be less than $\varepsilon$. We may conclude that the $\varepsilon$-mixing time of the GSR shuffle satisfies

$$\tau_P(\varepsilon) \leq \log_2\left(\frac{\binom{n}{2}}{\varepsilon}\right) < 2\log_2(n) + \log_2\left(\tfrac{1}{\varepsilon}\right).$$

Bayer and Diaconis famously worked out a tight analysis of the mixing time of the GSR shuffle, showing that $\tau_P(\varepsilon) = \frac{3}{2}\log_2(n) + \Theta(1)$ where the $\Theta(1)$ term depends on $\varepsilon$.

### 6.4.2 Analyzing Glauber Dynamics via Coupling

Recall the Glauber dynamics for sampling a uniformly random $q$-coloring of an undirected graph $G$. This is the Markov chain whose states are proper colorings of $G$, and whose transition dynamics are described by the following sampling process: in state $x : V(G) \to [q]$, sample a uniformly random vertex $v$ and color $c$, and let $y : V(G) \to [q]$ be the function defined by setting

$$y(u) = \begin{cases} c & \text{if } u = v \\ x(u) & \text{if } u \neq v. \end{cases}$$

If $y$ is a proper coloring then transition from $x$ to $y$, otherwise remain in state $x$.

In this section we will prove Glauber dynamics mixes rapidly when $q > 4\Delta$, where $\Delta$ is the maximum degree of a vertex of $G$. There is a long-standing conjecture that Glauber dynamics mixes rapidly whenever $q > \Delta + 1$. At present, however, the best known result in this direction asserts that Glauber dynamics mixed rapidly whenever $\frac{q-1}{\Delta} > \alpha$, where $\alpha \approx 1.763\ldots$ is the solution to the equation $e^{1/x} = x$.

To analyze Glauber dynamics we will use the Markov Coupling Lemma. The construction of the Markov coupling is very simple to describe. Starting from states $X_0$ and $X_0'$ sampled from some arbitrary initial distribution $\pi_0$ and from the stationary distribution, respectively, we repeatedly update the pair of states by choosing the same vertex $v$ and color $c$ in both Markov chains. To bound the probability of the event $X_t \neq X_t'$, we will analyze the Hamming distance

$$d(X_t, X_t') = \#\{v \mid X_t(v) \neq X_t'(v)\}.$$

How does the Hamming distance change when both sides of the coupling undergo a Markov transition corresponding to choosing vertex $v$ and color $c$?

1. If $X_t(v) \neq X_t'(v)$, and $X_{t+1}(v) = X_{t+1}'(v) = c$, then the Hamming distance decreases by 1. Let us call this event a *color merge*.
2. If $X_t(v) = X_t'(v)$ but $X_{t+1}(v) \neq X_{t+1}'(v)$, then the Hamming distance increases by 1. We will call this event a *color split*. A color split occurs when $v$ is recolored with color $c$ on one side of the coupling, but on the other side the recoloring doesn't take place because a neighbor of $v$ is already colored with $c$.
3. In all other cases, the Hamming distance is unchanged.

Let $d_t = d(X_t, X_t')$. To estimate the probability of a color merge, observe that the probability of sampling a vertex $v$ such that $X_t(v) \neq X_t'(v)$ is $d_t/n$, and when such a vertex $v$ is sampled, a color merge takes place unless we sample a color $c$ which is among the colors of $v$'s neighbors in $X_t$ or $X_t'$. Since $v$ has $\Delta$ or fewer neighbors, there are at least $q - 2\Delta$ colors that are not used by $v$'s neighbors in either $X_t$ or $X_t'$. Hence, the probability of a color merge is at least:

$$\Pr(\text{color merge}) \geq \frac{d_t}{n} \cdot \frac{q - 2\Delta}{q}.$$

Now let's estimate the probability of a color split. In order for such an event to take place, $v$ must have a neighbor $w$ such that $X_t(w) = c$ and $X_t'(w) \neq c$ or $X_t(w) \neq c$ and $X_t'(w) = c$. When this happens, we will say that the color split is *blamed on* the directed edge $(v, w)$. Every color split can be blamed on at least one directed edge, possibly more than one. Now, in order for a directed edge $(v, w)$ to be blamed for a color split, $w$ must be among the $d_t$ vertices whose colors in $X_t$ and $X_t'$ differ, $c$ must be one of the two elements of the set $\{X_t(w), X_t'(w)\}$, and $v$ must be one of the (at most) $\Delta$ neighbors of $w$, so

$$\Pr(\text{color split}) \leq \mathbb{E}[\text{number of blamed edges}] \leq d_t \cdot \frac{2}{q} \cdot \frac{\Delta}{n} = \frac{d_t}{n} \cdot \frac{2\Delta}{q}.$$

Combining these two bounds, we find that

$$\begin{aligned}
\mathbb{E}[d_{t+1} \mid d_t] &= d_t - \Pr(\text{color merge}) + \Pr(\text{color split}) \\
&\leq d_t - \frac{d_t}{n} \cdot \frac{q - 2\Delta}{q} + \frac{d_t}{n} \cdot \frac{2\Delta}{q} \\
&= \left(1 - \frac{q - 4\Delta}{qn}\right) \cdot d_t.
\end{aligned}$$

By induction on $t$,

$$\mathbb{E}[d_t] \leq \left(1 - \frac{q - 4\Delta}{qn}\right)^t \cdot d_0 < \exp\left(-\frac{q - 4\Delta}{qn} \cdot t\right) \cdot n. \tag{6.4}$$

When $t \geq \frac{q}{q-4\Delta} \cdot n \ln(n/\varepsilon)$, the right side of (6.4) is less than or equal to $\varepsilon$. Hence, by Lemma 6.20, the $\varepsilon$-mixing time of Glauber dynamics is bounded above by $\frac{q}{q-4\Delta} \cdot n \ln(n/\varepsilon)$.

# 7. Probability in Vector Spaces

This chapter introduces some important and commonly used probability distributions, especially the Gaussian distribution which is ubiquitous in statistics, data science, and all of the natural and social sciences. We begin by briefly reviewing some material from probability theory. In doing so, we adopt an unorthodox approach that emphasizes random variables and the operations one can perform on them, rather than the traditional approach of starting with sample spaces, events, and probabilities.

## 7.1 Review of Random Variables

Our presentation of probability will focus on *random variables*. A random variable $X$ taking values in a set $T$ can be thought of as a variable whose value definitely belongs to $T$, but the value is undetermined until $X$ is randomly sampled. If $\phi(x)$ is a Boolean predicate on $T$ (i.e., a mapping from $T$ to $\{\text{TRUE}, \text{FALSE}\}$) then there is a number $\Pr(\phi(X))$ in $[0,1]$ called the *probability of the event* $\phi(X)$.

Technically, $\Pr(\phi(X))$ is only defined when $\phi$ is "measurable." We will not give the definition of measurable here, but we will say that when $T$ is a vector space and $\phi$ is any predicate that can defined using continuous functions, equations, and inequalities, $\phi$ is measurable. For example, if $X$ is a real-valued random variable, the predicate $\phi(x) = (x \geq 0)$ is measurable and its probability, written as $\Pr(X \geq 0)$, is a well-defined number between 0 and 1. Any Boolean predicate that could be defined in an ordinary programming language is measurable. Henceforth when we use the word "predicate" we always mean "measurable predicate."

In addition to being well-defined and non-negative, probabilities must satisfy the following properties:

1. **normalization:** $\Pr(X \in T) = 1$.

2. **finite additivity:** If $\phi_0$ and $\phi_1$ are *mutually exclusive*, meaning no $x \in T$ satisfies $\phi_0(x)$ and $\phi_1(x)$, then

$$\Pr(\phi_0(X) \vee \phi_1(X)) = \Pr(\phi_0(X)) + \Pr(\phi_1(X)).$$

3. **monotone convergence:** If $\phi_1, \phi_2, \ldots$ is a countable sequence of predicates,

$$\Pr(\exists n \in \mathbb{N}\ \phi_n(X)) = \lim_{N \to \infty} \Pr(\exists n \leq N\ \phi_n(X)).$$

Two random variables $X$ and $Y$, taking values in $T$, are said to *have the same distribution*, or to be *identically distributed*, if the equation $\Pr(\phi(X)) = \Pr(\phi(Y))$ holds for every predicate $\phi$. We will denote the relation "$X$ and $Y$ are identically distributed" by the notation $X \sim Y$. This is an equivalence relation on the set of $T$-valued random variables, and its equivalence classes are called *probability distributions on* $T$. We will sometimes use calligraphic font to refer to probability distributions, and we will abuse notation and write $X \sim \mathcal{X}$ when $X$ is a random variable and $\mathcal{X}$ is a probability distribution, to denote that $\mathcal{X}$ is the distribution of $X$, i.e. that $X$ belongs to the equivalence class $\mathcal{X}$. (The notation $X \in \mathcal{X}$ already expresses this relationship, since an equivalence class is by definition a set. However it's not customary to think of probability distributions as sets, and it's more customary to write $X \sim \mathcal{X}$ when the distribution of $X$ is $\mathcal{X}$.)

If $X$ is a random variable and $G$ is a function, then one can construct another random variable $Y = G(X)$. The distribution of $Y$ is defined by the property that for every predicate $\phi$, $\Pr(\phi(Y)) = \Pr(\phi(G(X)))$. (Once again, there is a technicality that $F$ must be what is called a "measurable function". The set of measurable functions includes any function on a vector space that can be defined using continuous functions and if-then statements whose conditional is a measurable predicate is a measurable. Any function that can be written in an ordinary programming language is measurable. Henceforth, when we use the word "function" we implicitly mean "measurable function.")

For a random variable $X$ taking values in $T$, we say that $X$ is *supported* in a subset $S \subseteq T$ if $\Pr(X \in S) = 1$.

### 7.1.1 Finitely supported random variables

Given a finite set $S \subseteq T$ and a function $p : S \to [0,1]$ satisfying $\sum_{s \in S} p(s) = 1$, we can construct a $T$-valued random variable $X$ such that $\Pr(X = s) = p(s)$ for all $s \in S$. Such an $X$ is called a *finitely-supported random variable*, and its *support set* is the set $\{s \in S \mid p(s) > 0\}$. The distribution of a finitely-supported random variable is uniquely determined by its support set and by the probabilities of each element of the support set.

### 7.1.2 Independence

Two random variables $X, Y$ are *independent* if they satisfy the equation

$$\Pr(\phi(X) \wedge \psi(Y)) = \Pr(\phi(X)) \cdot \Pr(\psi(Y))$$

for every two predicates $\phi, \psi$. More generally, a (possibly infinite) set of random variables $\{X_i \mid i \in \mathcal{I}\}$ is *mutually independent* if the following equation holds whenever $\phi_1, \ldots, \phi_n$ is a finite sequence of predicates and $i(1), \ldots, i(n)$ is a finite sequence of distinct indices in $\mathcal{I}$:

$$\Pr(\phi_1(X_{i(1)}) \wedge \phi_2(X_{i(2)}) \wedge \cdots \wedge \phi_n(X_{i(n)})) = \prod_{k=1}^{n} \Pr(\phi_k(X_{i(k)})).$$

If $X$ and $Y$ are two random variables, then one can always construct a pair of independent random variables $(X', Y')$ having the same distributions as $X$ and $Y$, respectively. More generally, for any (possibly infinite) index set $\mathcal{I}$, if we are given a probability distribution $\mathcal{X}_i$ for each $i \in \mathcal{I}$, then one can construct an $\mathcal{I}$-indexed family $\{X_i \mid i \in \mathcal{I}\}$ of mutually independent random variables, such that $X_i \sim \mathcal{X}_i$ for all $i \in \mathcal{I}$.

### 7.1.3 Real-valued random variables

If $X$ is a random variable taking values in the real numbers, its *cumulative distribution function* $F_X$ (known as the CDF, for short) is the function

$$F_X(\theta) = \Pr(X \leq \theta).$$

It is a theorem that if two $\mathbb{R}$-valued random variables have the same CDF then they are identically distributed.

**Lemma 7.1** If $X$ is a real-valued random variable then its CDF, $F_X$, is a non-decreasing function that satisfies

$$\lim_{\theta \to \infty} F_X(\theta) = 1, \qquad \lim_{\theta \to -\infty} F_X(\theta) = 0.$$

*Proof.* If $\theta_0 < \theta_1$ then

$$F_X(\theta_1) = \Pr(X \leq \theta_0) + \Pr(\theta_0 < X \leq \theta_1) \geq \Pr(X \leq \theta_0) = F_X(\theta_0),$$

so $F_X$ is non-increasing. Since $F_X(\theta)$ is bounded below by 0 and above by 1 for all $\theta$, and $F_X$ is non-increasing, it follows that $\lim_{\theta \to \infty} F_X(\theta)$ and $\lim_{\theta \to -\infty} F_X(\theta)$ exist. By monotone convergence,

$$\lim_{\theta \to \infty} F_X(\theta) = \lim_{n \to \infty} F_X(n) = \Pr(\exists n \in \mathbb{N}\, X \leq n) = 1,$$

since for every real number is less than some natural number. Similarly,

$$\lim_{\theta \to -\infty} F_X(\theta) = 1 - \lim_{\theta \to -\infty} 1 - F_X(\theta) = 1 - \lim_{n \to \infty} 1 - F_X(-n) = \Pr(\exists n \in \mathbb{N}\, X > -n) = 1,$$

since every real number is greater than $-n$ for some $n \in \mathbb{N}$. ■

An important distribution on $\mathbb{R}$ is the *uniform distribution* on $[0,1]$. This is the distribution whose CDF is

$$F_{\text{unif}}(\theta) = \begin{cases} 0 & \text{if } \theta \leq 0 \\ \theta & \text{if } 0 < \theta < 1 \\ 1 & \text{if } \theta \geq 1. \end{cases}$$

Equivalently, a random variable $X$ supported in $[0,1]$ is uniformly distributed if and only if the binary digits of $X$ (after the decimal point) are mutually independent and each of them is 0 or 1 with equal probability.

**Lemma 7.2** If $X$ is a real-valued random variable whose CDF, $F_X$, is continuous, then the random variable $Y = F_X(X)$ is uniformly distributed in $[0,1]$.

*Proof.* Consider any $\theta \in (0,1)$. Since $F_X(\theta)$ converges to 0 and 1 as $\theta$ tends to $-\infty$ and $\infty$, respectively, and $F_X$ is continuous, the intermediate value theorem guarantees that the set $F^{-1}(\{\theta\})$ is non-empty. Let $t$ denote the maximum element of $F^{-1}(\{\theta\})$. (It is a non-empty, closed, bounded subset of $\mathbb{R}$, so it has a maximum element.) Then, $X \leq t$ if and only if $F_X(X) \leq \theta$. Hence,

$$\Pr(Y \leq \theta) = \Pr(F_X(X) \leq \theta) = \Pr(X \leq t) = F_X(t) = \theta.$$

Since this equation holds for all $\theta \in (0,1)$, $Y$ is uniformly distributed. ■

**Corollary 7.3** If $X$ is a random variable whose CDF, $F_X$, is continuous and strictly increasing, and $Y$ is uniformly distributed in $[0,1]$, then $X$ and $F_X^{-1}(Y)$ are identically distributed.

Corollary 7.3 gives a useful recipe for drawing random samples from a distribution with specified CDF, $F$: one draws a uniformly random sample from $[0,1]$ and applies the function $F^{-1}$.

▪ **Example 7.4** A random variable $X$ is *exponentially distributed with rate $r$* if it satisfies

$$\Pr(X > \theta) = e^{-r\theta}.$$

Equivalently, $X$ is exponentially distributed with rate $r$ if its CDF is $F_X(\theta) = 1 - e^{-r\theta}$. Using Corollary 7.3 we can see that one way to sample an exponentially distributed random variable with rate $r$ is to sample a uniformly random number $Y \in [0,1]$ and apply the transformation $X = \frac{1}{r}\ln(\frac{1}{1-y})$. ▪

### 7.1.4 Probability density

If $V$ is a finite-dimensional vector space and $f : V \to [0,\infty)$ is a function satisfying $\int_V f(\mathrm{x})\,d\mathrm{x} = 1$ then one can construct a random variable $X$ whose distribution satisfies $\Pr(X \in S) = \int_S f(\mathrm{x})\,d\mathrm{x}$ for every (measurable) subset $S \subset V$. We say that $f$ is the probability density function of $X$. In the special case when $V = \mathbb{R}$, if $X$ has probability density function $f$ then its CDF is $F_X(\theta) = \int_{-\infty}^{\theta} f(x)\,dx$. Conversely, if the CDF of a real-valued random variable is differentiable, then the derivative of the CDF is a probability density function for that random variable.

If $X$ and $Y$ are independent random variables taking values in vector spaces $V$ and $W$, respectively, and $X$ and $Y$ have density functions $f, g$, respectively, then the random variable $(X,Y)$, which takes values in $V \times W$, has density function $h$ defined by

$$h(\mathrm{x},\mathrm{y}) = f(\mathrm{x})g(\mathrm{y}).$$

### 7.1.5 Expected value

If $X$ is a random variable taking values in $[0,\infty]$ its expected value (also known as its expectation) is defined by the formula

$$\mathbb{E}[X] = \int_{\theta=0}^{\infty} \Pr(X > \theta)\,d\theta = \int_{\theta=0}^{\infty} (1 - F_X(\theta))\,d\theta.$$

The standard definition of expected value represents it as the weighted average of the possible values of $X$, weighted by their respective probabilities. That definition turns out to be equivalent to the formula above; the following two lemmas state and prove the equivalence, first for the case when $X$ has finite support and then for the case when $X$ has a probability density function.

**Lemma 7.5** If $X$ is a finitely-support random variable with support set $S \subset [0,\infty]$ then $\mathbb{E}[X] = \sum_{s\in S} s \cdot \Pr(X = s)$.

*Proof.* Enumerate the elements of $S$ in increasing order as $s_1 \leq s_2 \leq \ldots \leq s_n$ and let $p_i = \Pr(x = s_i)$. For notational convenience let $s_0 = 0$. Then we have

$$\sum_{i=1}^{n} s_i p_i = \sum_{i=1}^{n}\sum_{j=1}^{i}(s_j - s_{j-1})p_i = \sum_{j=1}^{n}\sum_{i=j}^{n}(s_j - s_{j-1})p_i = \sum_{j=1}^{n}(s_j - s_{j-1})\Pr(X > s_{j-1}) \quad (7.1)$$

In addition we have

$$\int_0^\infty \Pr(X > \theta)\,d\theta = \sum_{j=1}^{n}\int_{s_{j-1}}^{s_j}\Pr(X > \theta)\,d\theta = \sum_{j=1}^{n}(s_j - s_{j-1})\Pr(X > s_{j-1}). \quad (7.2)$$

The right sides of Equations (7.1) and (7.2) are identical. The left sides are, respectively, equal to $\sum_{s\in S} s \cdot \Pr(X = s)$ and $\mathbb{E}[X]$, which completes the proof of the lemma. ■

**Lemma 7.6** If $X$ is a $[0,\infty)$-valued random variable that has a probability density function $f_X$, then

$$\mathbb{E}[X] = \int_0^\infty \theta\, f_X(\theta)\,d\theta.$$

*Proof.* The probability density satisfies $f_X(\theta) = \frac{d}{d\theta}F_X(\theta)$. Using integration by parts we find that

$$\int_0^\infty \theta\, f_X(\theta)\,d\theta == \int_{\theta=0}^\infty (1 - F_X(\theta))\,d\theta + \left(\lim_{\theta\to\infty}\theta\cdot(1-F_X(\theta))\right) = \mathbb{E}[X] + \lim_{\theta\to\infty}\theta\cdot(1-F_X(\theta)). \quad (7.3)$$

The proof divides now into two cases. If the limit on the right side of Equations (7.3) is zero, then we are done. Otherwise, there is some $\varepsilon > 0$ such that the set

$$\Theta_\varepsilon = \{\theta \mid \theta\cdot(1 - F_X(\theta) > \varepsilon\}$$

is unbounded. In this case we claim that both the left and right sides of Equation (7.3) are infinite. Define an infinite sequence of positive numbers $\theta_1, \theta_2, \ldots$ recursively, by choosing $\theta_1$ to be any element of $\Theta_\varepsilon$ and choosing $\theta_{n+1}$ to be any element of $\Theta_\varepsilon$ that exceeds $2\theta_n$. Define $\theta_0 = 0$ for notational convenience. Then for any $\theta \in [\theta_{n-1}, \theta_n]$ we have $1 - F_X(\theta) \geq 1 - F_X(\theta_n)$, so

$$\begin{aligned}\int_0^\infty (1 - F_X(\theta))\,d\theta = \sum_{n=1}^{\infty}\int_{\theta_{n-1}}^{\theta_n}(1 - F_X(\theta))\,d\theta &\geq \sum_{n=1}^{\infty}\int_{\theta_{n-1}}^{\theta_n}(1 - F_X(\theta_n))\,d\theta \\ &= \sum_{n=1}^{\infty}(\theta_n - \theta_{n-1})(1 - F_X(\theta_n)) > \sum_{n=1}^{\infty}\frac{\theta_n}{2}(1 - F_x(\theta_n)).\end{aligned}$$

The sum on the right side is infinite because each summand is greater than $\frac{\varepsilon}{2}$. Hence, the right side of Equation (7.3) is infinite, as claimed. As for the left side of (7.3),

$$\begin{aligned}\int_0^\infty \theta f_X(\theta)\,d\theta = \sum_{n=1}^{\infty}\int_{\theta_{n-1}}^{\theta_n} \theta f_X(\theta)\,d\theta &\geq \sum_{n=1}^{\infty}\int_{\theta_{n-1}}^{\theta_n} \theta_{n-1} f_X(\theta)\,d\theta \\ &= \sum_{n=1}^{\infty} \theta_{n-1}(F_X(\theta_n) - F_X(\theta_{n-1})) \\ &= \sum_{n=1}^{\infty} \theta_{n-1}\left[(1-F_X(\theta_{n-1})) - (1-F_X(\theta_n))\right] \\ &= \sum_{n=1}^{\infty} (\theta_n - \theta_{n-1})(1-F_X(\theta_n)) > \sum_{n=1}^{\infty} \frac{\theta_n}{2}(1-F_X(\theta_n))\end{aligned}$$

Again, the sum on the last line is infinite because each summand is at least $\varepsilon/2$. ■

For a random variable $X$ that takes both positive and negative values in $\mathbb{R}$, define $X^+ = \max\{0, X\}$ and $X^- = \min\{0, X\}$. Both $X^+$ and $-X^-$ are non-negative random variables. If at least one of them has finite expectation, then $\mathbb{E}[X]$ is defined by the equation

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[-X^-].$$

If $\mathbb{E}[X^+] = \mathbb{E}[-X^-] = \infty$ then the expectation of $X$ is undefined.

An important property of the expectation operator is *linearity of expectation*: for real-valued random variables $X, Y$, we have

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

whenever the terms on the left and right sides are well-defined. Linearity of expectation also holds for countable sums: if $X_1, X_2, \ldots$ is an infinite sequence of random variables such that either

1. $\sum_{n=1}^{\infty} |\mathbb{E}[X_i]| < \infty$, or
2. each variable $X_n$ is supported on $[0, \infty]$,

then

$$\mathbb{E}\left[\sum_{n=1}^{\infty} X_n\right] = \sum_{n=1}^{\infty} \mathbb{E}[X_n].$$

For a random variable $X$ taking values in $\mathbb{R}^n$, one can define the expectation $\mathbb{E}[X]$ coordinatewise. In other words, the $i^{\text{th}}$ coordinate of $\mathbb{E}[X]$ is the expectation of the $i^{\text{th}}$ coordinate of $X$. Using linearity of expectation for scalar-valued random variables, one can prove that the expectations of vector-valued random variables satisfy the following version of linearity of expectation: for any random variables $X, Y$ taking values in $\mathbb{R}^n$ and any $n \times n$ matrices $A$ and $B$,

$$\mathbb{E}[AX + BY] = A\,\mathbb{E}[X] + B\,\mathbb{E}[Y].$$

If $X$ is a random variable taking values in a finite-dimensional vector space $V$, its expectation is defined by choosing a based vector space structure $\beta : \mathbb{R}^n \to V$, and defining $\mathbb{E}[X] = \beta(\mathbb{E}[\beta^{-1}(X)])$. Using linearity of expectation, one can verify that the vector $\mathbb{E}[X]$ defined by this equation does not depend on the choice of based vector space structure.

We present the following lemma about expectations of products of independent random variables without proof.

**Lemma 7.7** If $X, Y$ are independent random variables and $f, g$ are real-valued functions, then

$$\mathbb{E}[f(X)\,g(Y)] = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)].$$

Another useful fact about expected values is Markov's Inequality, which bounds the probability that a non-negative random variable exceeds its expected value by a specified factor.

**Lemma 7.8 — Markov's Inequality.** If $X$ is a random variable taking values in $[0,\infty)$ and $\mathbb{E}[X] < \infty$, then for all $\theta > 0$,

$$\Pr(X \geq \theta) \leq \frac{\mathbb{E}[X]}{\theta}.$$

*Proof.* The function $G(t) = \Pr(X \geq t)$ is non-negative and non-increasing in $t$, so

$$\mathbb{E}[X] = \int_0^\infty \Pr(X \geq t)\,dt \geq \int_0^\theta \Pr(X \geq t)\,dt \geq \int_0^\theta \Pr(X \geq \theta)\,dt = \theta \cdot \Pr(X \geq \theta).$$

Dividing both sides by $\theta$ we obtain Markov's Inequality. ■

### 7.1.6 Variance and covariance

If $X$ is a real-valued random variable whose expectation is well-defined and finite, the variance of $X$ is defined by

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

An important property of the variance is that when one sums up a sequence of independent random variables, the variance of their sum equals the sum of their variances.

**Lemma 7.9** If $X_1, X_2, \ldots, X_n$ are independent real-valued random variables, each with finite variance, then

$$\mathrm{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

*Proof.* We will prove the $n = 2$ case of the lemma, i.e. that the relation $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ holds when $X$ and $Y$ are independent. The full lemma then follows easily by induction on $n$, using $X = X_n$ and $Y = X_1 + \cdots + X_{n-1}$.

Let $\bar{x} = \mathbb{E}[X]$ and $\bar{y} = \mathbb{E}[Y]$. Using the definition of variance, along with linearity of expectation, we find that

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbb{E}\left[(X - \bar{x} + Y - \bar{y})^2\right] \\
&= \mathbb{E}\left[(X-\bar{x})^2\right] + 2\mathbb{E}\left[(X-\bar{x})(Y-\bar{y})\right] + \mathbb{E}\left[(Y-\bar{y})^2\right] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathbb{E}\left[(X-\bar{x})(Y-\bar{y})\right].
\end{aligned}$$

Since $X$ and $Y$ are assumed to be independent we can apply Lemma 7.7 to conclude that

$$\mathbb{E}\left[(X-\bar{x})(Y-\bar{y})\right] = \mathbb{E}[X-\bar{x}] \cdot \mathbb{E}[Y-\bar{y}] = 0$$

which concludes the proof. ■

The covariance of two real-valued random variables $X$ and $Y$ is defined by

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right].$$

If $X$ and $Y$ are independent one can check, using linearity of expectation, that their covariance is zero.

For a vector-valued random variable $X$ taking values in $\mathbb{R}^n$, the covariance matrix $\mathrm{Cov}(X)$ is the $n \times n$ matrix whose $(i,j)$ entry is $\mathrm{Cov}(X_i, X_j)$. Equivalently, $\mathrm{Cov}(X)$ can be defined using the formula

$$\mathrm{Cov}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])\,(X - \mathbb{E}[X])^\top\right].$$

## 7.2 Gaussian distributions

The *normal distribution* on $\mathbb{R}$ is the probability distribution with density function $f(x) = \frac{1}{Z} e^{-\frac{1}{2}x^2}$, where the normalizing factor $\frac{1}{Z}$ is chosen to ensure that $\int_{-\infty}^{\infty} f(x)\,dx = 1$, as required for a probability density function. The normal distribution (and its multi-dimensional generalization, the Gaussian distribution) is the most important distribution in continuous probability theory. One reason for its importance is the Central Limit Theorem, which says that (under mild conditions) the distribution of the average of $n$ identically distributed random variables converges to a normal distribution, when suitably shifted and rescaled.

**Theorem 7.10 — Central Limit Theorem.** Let $X_1, X_2, \ldots$ be an infinite sequence of identically distributed real-valued random variables, each with finite expectation $\mu$ and finite variance $\sigma^2$. Then as $n \to \infty$,

$$\frac{\sqrt{n}}{\sigma}\left(\frac{X_1 + \cdots + X_n}{n} - \mu\right) \xrightarrow{d} \mathcal{N}(0,1).$$

The relation $\xrightarrow{d}$ in the theorem statement is called "convergence in distribution." It means that if $F_n$ denotes the CDF of the random variable on the left side and $F$ denotes the CDF of the random variable on the right side, then $F_n(\theta) \to F(\theta)$ as $n \to \infty$, uniformly in $\theta$. In other words, for every $\varepsilon > 0$ there is some $n_0 < \infty$ such that for all $n > n_0$ and all $\theta \in \mathbb{R}$, $|F_n(\theta) - F(\theta)| < \varepsilon$.

Unfortunately there is no closed-form expression for the CDF of the normal distribution. This raises the question of how to sample normally-distributed random variables. Fortunately there is a clever trick that allows drawing two independent normally-distributed random variables at once. This is based on the observation that if $X$ and $Y$ are independent, normally-distributed random variables, then the probability density function of the pair $(X,Y)$ is

$$f(x,y) = \left(\frac{1}{Z} e^{-\frac{1}{2}x^2}\right)\left(\frac{1}{Z} e^{-\frac{1}{2}y^2}\right) = \frac{1}{Z^2} e^{-\frac{1}{2}\left(x^2+y^2\right)}.$$

Now, represent the pair $(X,Y)$ in polar coordinates as $(R,\Theta)$ where $R$ and $\Theta$ are random variables satisfying $X = R\cos(\Theta)$, $Y = R\sin(\Theta)$. We have

$$\frac{1}{Z^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(x^2+y^2\right)}\,dx\,dy = \frac{1}{Z^2}\int_0^{2\pi}\int_0^{\infty} e^{-\frac{1}{2}r^2} r\,dr\,d\theta \tag{7.4}$$

The extra factor of $r$ in the integrand is attributable to the change-of-variables formula for integrals in polar coordinates, $dxdy = rdrd\theta$. It makes a huge difference because $re^{-\frac{1}{2}r^2}$ is the derivative of $1-e^{-\frac{1}{2}r^2}$. Hence, we can perform the substitution $u = \frac{1}{2}r^2$ and rewrite the integral as

$$\frac{1}{Z^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)}dxdy = \frac{1}{Z^2}\int_0^{2\pi}\int_0^{\infty} e^{-u}dud\theta. \tag{7.5}$$

This integral formula has a few consequences.

1. It's easy to evaluate the right side and find that it equals $\frac{2\pi}{Z^2}$. Since the left side must be equal to 1 (integrating a random variable's probability density over its support set always yields 1) we may conclude that $Z = \sqrt{2\pi}$. Therefore,

   *The normal distribution* $\mathcal{N}(0,1)$ *has density* $f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$.

2. From the right side of Equation (7.5) we can deduce that $R$ and $\Theta$ are independent random variables, $\Theta$ is uniformly distributed in $[0,2\pi)$, and $U = \frac{1}{2}R^2$ is exponentially distributed with rate 1. Therefore, one can use the following procedure to draw samples from $\mathcal{N}(0,1)$.

   a. *Sample* $\Theta$ *uniformly at random from* $[0,2\pi)$.
   b. *Sample* $Z$ *uniformly at random from* $[0,1]$.
   c. *Let* $U = \ln(\frac{1}{1-Z})$.
   d. *Let* $R = \sqrt{2U}$.
   e. *Let* $X = R\cos(\Theta)$.

3. An exponentially distributed random variable with rate 1 has expected value 1, so $\frac{1}{2}\mathbb{E}[R^2] = 1$. Since $R^2 = X^2+Y^2$ and $X,Y$ are identically distributed random variables with $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, we have $\mathrm{Var}(X) = \mathbb{E}[X^2] = \frac{1}{2}\mathbb{E}[X^2+Y^2] = \frac{1}{2}\mathbb{E}[R^2] = 1$. Therefore,

   *A random variable with distribution* $\mathcal{N}(0,1)$ *has variance 1.*

If $X$ is a random sample from $\mathcal{N}(0,1)$, then the random variable $Y = \sigma X + \mu$ has expectation $\mu$ and variance $\sigma^2$, because

$$\begin{aligned}\mathbb{E}[Y] &= \sigma\mathbb{E}[X] + \mu = \mu\\ \mathrm{Var}[Y] &= \mathbb{E}[(Y-\mu)^2] = \mathbb{E}[(\sigma X)^2] = \sigma^2\mathbb{E}[X^2] = \sigma^2.\end{aligned}$$

The distribution of $Y = \sigma X + \mu$ is denoted by $\mathcal{N}(\mu,\sigma^2)$ and is called the *Gaussian distribution with mean* $\mu$ *and variance* $\sigma^2$.

### 7.2.1 Moments and cumulants of the normal distribution

Recall the moment generating function and cumulant generating function of a distribution.

$$\begin{aligned}M_X(t) &= \mathbb{E}[e^{tX}]\\ K_X(t) &= \ln M_X(t).\end{aligned}$$

When $X$ is drawn from the $\mathcal{N}(0,1)$ distribution, we can evaluate $M_X(t)$ and $K_x(t)$ by direct calculation. Let $f(x) = \frac{1}{Z}e^{-x^2/2}$ denote the probability density function of $\mathcal{N}(0,1)$, where $Z = \sqrt{2\pi}$. We have

$$\begin{aligned} M_X(t) = \mathbb{E}[e^{tX}] &= \int_{-\infty}^{\infty} e^{tx} f(x)\,dx \\ &= \frac{1}{Z}\int_{-\infty}^{\infty} e^{tx-x^2/2}\,dx \\ &= \frac{1}{Z}\int_{-\infty}^{\infty} e^{t^2/2}\cdot e^{-(x-t)^2/2}\,dx \\ &= e^{t^2/2}\cdot\frac{1}{Z}\int_{-\infty}^{\infty} e^{-u^2/2}\,du = e^{t^2/2}, \end{aligned}$$

where the last line was derived using the substitution $u = x - t$. Taking the natural logarithm of both sides, we find that

$$K_X(t) = \tfrac{1}{2}t^2.$$

From the Taylor series for $M_X(t)$ and $K_X(t)$ we deduce that the moments and cumulants of the normal distribution are as follows.

$$m_n(X) = \begin{cases} 0 & \text{if } n \text{ is odd} \\ \frac{(2k)!}{2^k\cdot k!} & \text{if } n = 2k \text{ is even} \end{cases}$$

$$\kappa_n(X) = \begin{cases} 0 & \text{if } n \neq 2 \\ 1 & \text{if } n = 2 \end{cases}$$

The fact that $\kappa_2$ is the only non-zero cumulant of the normal distribution explains why the central limit theorem holds. The following argument is not a rigorous proof but offers some insight into the justification for the central limit theorem.

Suppose $X_1, X_2, \ldots$ are independent, identically distributed random variables with finite mean and variance, whose distribution is not necessarily Gaussian. Our goal is to show that the distribution of the random variable $X^{(n)} = \frac{\sqrt{n}}{\sigma}\left(\frac{X_1+\cdots+X_n}{n} - \mu\right)$ converges to $\mathcal{N}(0,1)$, where $\mu$ and $\sigma^2$ denote the mean and variance of each $X_i$. Let $Y_i = \frac{X_i-\mu}{\sigma\sqrt{n}}$ and note that $X^{(n)} = Y_1 + \cdots + Y_n$.

To gain some quantative understanding of the distribution of $X^{(n)}$, and why it is similar to $\mathcal{N}(0,1)$, it helps to look at the cumulant generating function $K_{X^{(n)}}(t)$. The following facts will be extremely helpful.

1. If $A$ and $B$ are independent random variables, then $K_{A+B}(t) = K_A(t) + K_B(t)$. This follows by taking the logarithm of both sides of the equation

$$M_{A+B}(t) = \mathbb{E}[e^{At+Bt}] = \mathbb{E}[e^{At}]\cdot\mathbb{E}[e^{Bt}] = M_A(t)\cdot M_B(t).$$

2. If $A$ is a random variable and $c$ is a constant, then $K_{A+c}(t) = K_A(t) + ct$. This is a special case of the preceding observation, where $B$ is taken to be a random variable whose value is deterministically equal to $c$.

3. If $A$ is a random variable with mean $\mu$ and variance $\sigma^2$ then

$$K_A(t) = \mu t + \frac{1}{2}\sigma^2 t^2 + \sum_{m=3}^{\infty} \frac{\kappa_m(A)}{m!} t^m.$$

This is a restatement of the fact that the first two cumulants of $A$ are its mean and variance.

4. If $A$ is a random variable and $\alpha$ is a scalar, then $K_{\alpha A}(t) = K_A(\alpha t)$. This follows by taking logarithms of both sides of the equation

$$M_{\alpha A}(t) = \mathbb{E}[e^{\alpha A t}] = \mathbb{E}[e^{\alpha t A}] = M_A(\alpha t).$$

Combining these facts, we find that

$$\begin{aligned}
\forall i \quad K_{X_i}(t) &= \mu t + \frac{1}{2}\sigma^2 t^2 + \sum_{m=3}^{\infty} \frac{\kappa_m(X_i)}{m!} t^m \\
\forall i \quad K_{X_i-\mu}(t) &= \frac{1}{2}\sigma^2 t^2 + \sum_{m=3}^{\infty} \frac{\kappa_m(X_i)}{m!} t^m \\
\forall i \quad K_{Y_i}(t) = K_{(X_i-\mu)/(\sigma\sqrt{n})}(t) &= \frac{1}{2}\frac{\sigma^2 t^2}{\sigma^2 n} + \sum_{m=3}^{\infty} \frac{\kappa_m(X_i)}{m!} \frac{t^m}{\sigma^m n^{m/2}} \\
K_{X^{(n)}}(t) = \sum_{i=1}^{n} K_{Y_i}(t) &= \frac{1}{2}t^2 + \sum_{m=3}^{\infty} \frac{\kappa_m(X_i)}{m!} \frac{t^m}{\sigma^m n^{(m/2)-1}}.
\end{aligned}$$

As $n \to \infty$, every term of the infinite sum on the right side of the last line converges to zero. So, if the series converges absolutely for all $t$ (which is a sufficient condition for interchanging the limit with the summation) we could conclude that for all $t$,

$$\lim_{n\to\infty} K_{X^{(n)}}(t) = \frac{1}{2}t^2 = K_{\mathcal{N}(0,1)}(t).$$

Furthermore, if the pointwise convergence $K_{X^{(n)}}(t) \to K_{\mathcal{N}(0,1)}(t)$ were sufficient to conclude that $X^{(n)} \xrightarrow{d} \mathcal{N}(0,1)$, this argument would establish the central limit theorem.

Both of the unresolved technicalities in the above argument can be dealt with by working with a different generating function, the *characteristic function*, which is simply the moment generating function evaluated along the pure imaginary number line instead of the real number line:

$$\varphi_X(t) = M_X(it) = \mathbb{E}[e^{itX}].$$

The relation $\lim_{n\to\infty} \varphi_{X^{(n)}}(t) = \varphi_{\mathcal{N}(0,1)}(t)$ can be proven using a line of reasoning similar to the argument sketched above for cumulant generating functions. Then, the proof of the central limit theorem finishes up by applying a powerful theorem called Lévy's Continuity Theorem, which says that if $\lim_{n\to\infty} \varphi_{X^{(n)}}(t) = \varphi_X(t)$ for all $t \in \mathbb{R}$, then $X^{(n)} \xrightarrow{d} X$.

### 7.2.2 Multivariate Gaussian distributions

For vector-valued random variables taking values in $\mathbb{R}^n$, the counterpart of the normal distribution is the *multivariate normal distribution* $\mathcal{N}(0, \mathbb{1})$, which is the distribution of a

random vector whose coordinates are independent random samples from $\mathcal{N}(0,1)$. In other words, the density of $\mathcal{N}(0,\mathbb{1})$ is the function

$$f(\mathrm{x}) = \left(\tfrac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}\left(x_1^2+x_2^2+\cdots+x_n^2\right)} = \left(\tfrac{1}{2\pi}\right)^{d/2} e^{-\frac{1}{2}\langle \mathrm{x},\mathrm{x}\rangle}. \tag{7.6}$$

If $X \sim \mathcal{N}(0,\mathbb{1})$, then its expectation and covariance matrix are $\mathbb{E}[X] = 0$ and $\mathrm{Cov}(X) = \mathbb{1}$, respectively.

If $X \sim \mathcal{N}(0,\mathbb{1})$ then the distribution of $X$ has two key properties that are evident from Equation (7.6).

1. The $n$ coordinates of $X$ are independent random variables.
2. The distribution of $X$ is rotation-invariant. In other words, for any orthogonal matrix $Q$, the random variable $QX$ has the same distribution as $X$.

A surprising number of identities regarding normally distributed random variables can be derived from these observations.

**Lemma 7.11** If $X_1,\ldots,X_n$ are independent random variables, each distributed according to $\mathcal{N}(0,1)$, then $\frac{1}{\sqrt{n}}(X_1+\cdots+X_n)$ also has the distribution $\mathcal{N}(0,1)$. More generally, for any coefficients $a_1,\ldots,a_n \in \mathbb{R}$, not all equal to zero, the random variable $Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$ has the distribution $\mathcal{N}(0, a_1^2+\cdots+a_n^2)$.

*Proof.* Let $\sigma = \sqrt{a_1^2+\cdots+a_n^2}$, and observe that the vector $\mathrm{a} = \frac{1}{\sigma}(a_1,a_2,\ldots,a_n)$ satisfies $\|a\|_2 = 1$. Hence, there exists an orthogonal matrix $Q$ whose first row is a. The random vector $X = (X_1,\ldots,X_n)$ has the distribution $\mathcal{N}(0,\mathbb{1})$, so $QX \sim \mathcal{N}(0,\mathbb{1})$ as well. The first coordinate of the vector $QX$ is $Y/\sigma$, hence $Y/\sigma \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,\sigma^2)$. ■

If $X$ is a $\mathbb{R}^n$-valued random variable with distribution $\mathcal{N}(0,\mathbb{1})$, $B$ is an invertible $n \times n$ matrix, and $\mu$ is any vector in $\mathbb{R}^n$, then the distribution of $Y = BX + \mu$ is called a *multivariate Gaussian distribution.* The expectation of $Y$ is $\mu$ and its covariance is

$$\mathrm{Cov}(Y) = \mathbb{E}\left[(Y-\mu)(Y-\mu)^\top\right] = \mathbb{E}\left[(BX)(BX)^\top\right] = B\mathbb{E}[XX^\top]B^\top = BB^\top,$$

since $\mathbb{E}[XX^\top] = \mathrm{Cov}(X) = \mathbb{1}$. The distribution of $Y$ is denoted by $\mathcal{N}(\mu, BB^\top)$. The density of $Y$ can be calculated as follows. Let $T$ denote the function $T(\mathrm{x}) = B\mathrm{x}+\mu$. Its inverse is the function $T^{-1}(\mathrm{y}) = B^{-1}(\mathrm{y}-\mu)$. A small ball $\mathcal{B}$ of volume $\varepsilon > 0$ centered at y is mapped by $T^{-1}$ to a small ellipsoid $\mathcal{E}$ of volume $|\det(B^{-1})| \cdot \varepsilon$ centered at $\mathrm{x} = T^{-1}(\mathrm{y})$. We have

$$\Pr(Y \in \mathcal{B}) = \Pr(X \in \mathcal{E}) = \left[|\det(B^{-1})|\left(\tfrac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}\langle \mathrm{x},\mathrm{x}\rangle} + o(1)\right] \cdot \varepsilon,$$

where $o(1)$ denotes an error term that converges to zero as $\varepsilon \to 0$. Thus, the density of $Y$ at y is $|\det(B^{-1})|\left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}\langle \mathrm{x},\mathrm{x}\rangle}$. Now, recalling that $\mathrm{x} = T^{-1}(\mathrm{y}) = B^{-1}(\mathrm{y}-\mu)$, we have

$$\langle \mathrm{x},\mathrm{x}\rangle = \left\langle B^{-1}(\mathrm{y}-\mu), B^{-1}(\mathrm{y}-\mu)\right\rangle = \left\langle \mathrm{y}-\mu, (B^{-1})^\top B^{-1}(\mathrm{y}-\mu)\right\rangle = \left\langle \mathrm{y}-\mu, (BB^\top)^{-1}(\mathrm{y}-\mu)\right\rangle.$$

The right side depends only on $BB^\top$, not on $B$. Thus, if two multivariate Gaussian random variables have the same mean $\mu$ and the same covariance matrix $\Sigma = BB^\top$, then they are identically distributed and their density is

$$f(\mathrm{y}) = \det(\Sigma)^{-1/2}\left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}\left\langle \mathrm{y}-\mu, \Sigma^{-1}(\mathrm{y}-\mu)\right\rangle}.$$

**Lemma 7.12** If $A$ is a $d \times n$ matrix of rank $d$ and $X \sim \mathcal{N}(0, \mathbb{1})$ is a $\mathbb{R}^n$-valued multivariate normal random variable, then $Y = AX$ is a $\mathbb{R}^d$-valued Gaussian random variable with distribution $\mathcal{N}(0, AA^\top)$.

*Proof.* Using the singular value decomposition, write $A$ as $A = USV^\top$ where $S$ is a $d \times n$ matrix whose diagonal entries, $S_{ii}$, are equal to the singular values of $A$ and whose off-diagonal entries, $S_{ij}$ $(i \neq j)$, are all equal to zero. We can factor $S$ as $S = D[\mathbb{1}\ 0]$, where $D$ is a $d \times d$ diagonal matrix with the singular values of $A$ on the diagonal, and $[\mathbb{1}\ 0]$ is a $d \times n$ matrix formed by juxtaposing the $d \times d$ identity matrix with a $d \times (n-d)$ block of zeros. Then

$$Y = UD[\mathbb{1}\ 0]V^\top X.$$

Let $W = [\mathbb{1}\ 0]V^\top X$. Since the distribution of $X$ is rotation-invariant and $V^\top$ is a rotation matrix, the distribution of $W$ is the same as the distribution of $[\mathbb{1}\ 0]X$, i.e. the first $d$ coordinates of $X$. In other words, $W \sim \mathcal{N}(0, \mathbb{1})$, where $\mathbb{1}$ now refers to the $d \times d$ identity matrix rather than $n \times n$ identity. The matrix $B = UD$ is invertible, and we have derived above that when $W \sim \mathcal{N}(0, \mathbb{1})$ and $Y = BW$ for an invertible matrix $B$, then $Y \sim \mathcal{N}(0, BB^\top)$. To finish up, note that

$$BB^\top = UD^2U^\top = USS^\top U^\top = AA^\top,$$

so $Y \sim \mathcal{N}(0, AA^\top)$ as claimed. ■

We remark that Lemma 7.11 corresponds to the special case of Lemma 7.12 where $A$ has only one row.

## 7.3 Matrix Concentration Inequalities

Analogous to the Chernoff and Hoeffding bounds, there are important probabilistic inequalities asserting that a sum of independent random matrices is unlikely to be far from its expected value. These *matrix concentration inequalities* have powerful applications to the analysis of randomized algorithms and to understanding the performance of algorithms applied to random high-dimensional datasets.

In this section we will present two important matrix concentration inequalities, one for sums of independent *bounded* matrices and one for singular values of *Gaussian* matrices. In later sections of these notes we'll see applications of both.

### 7.3.1 The Matrix Chernoff Bound

The matrix Chernoff bound is a generalization of the Chernoff bound for scalar-valued random variables. The Chernoff bound assumes the random variables are independent and that each of them obeys some inequalities, namely $0 \leq X_i \leq 1$ for all $i$, and it concludes that the sum $X = X_1 + \cdots + X_N$ obeys some inequalities with high probability, namely $(1-\varepsilon)\mathbb{E}[X] \leq X \leq (1+\varepsilon)\mathbb{E}[X]$. Similarly, the matrix Chernoff bound assumes that some symmetric-matrix-valued random variables $X_i$ are independent and that they obey some inequalities, and it concludes that their sum obeys some inequalities with high probability. The inequalities are expressed in terms of the *Loewner order* on symmetric matrices, defined as follows.

**Definition 7.13** The *Loewner order* is the partial ordering defined on $n \times n$ symmetric matrices, for any finite $n$, by the relation

$$A \succeq B \quad \Leftrightarrow \quad \forall \mathrm{x} \in \mathbb{R}^n \ \langle \mathrm{x}, A\mathrm{x} \rangle \geq \langle \mathrm{x}, B\mathrm{x} \rangle \quad \Leftrightarrow \quad A - B \text{ is positive semi-definite.}$$

**Theorem 7.14** Suppose $X_1, X_2, \ldots, X_N$ are independent random symmetric matrices satisfying $0 \preceq X_i \preceq \mathbb{1}$ for all $i \in [N]$. Furthermore, suppose their sum $X = X_1 + X_2 + \cdots + X_N$ satisfies $a\mathbb{1} \preceq \mathbb{E}[X] \preceq b\mathbb{1}$. Then for $0 \leq \varepsilon < 1$,

$$\Pr\left(X \not\succeq (1-\varepsilon)a\mathbb{1}\right) \ \leq \ ne^{-\varepsilon^2 a/2}$$

and

$$\Pr\left(X \not\preceq (1+\varepsilon)b\mathbb{1}\right) \ \leq \ ne^{-\varepsilon^2 b/3}.$$

A proof can be found in Joel Tropp's monograph "An Introduction to Matrix Concentration Inequalities" [**Tropp**]. The proof is conceptually similar to the proof of the Chernoff bound for scalar-valued random variables, making use of a matrix-valued version of the moment generating function. However, to justify all the steps of the proof one needs inequalities from matrix analysis that are beyond the scope of this course.

While omitting the proof of the Chernoff bound, we offer here some interpretations.

1. When $n = 1$ we recover the usual Chernoff bound by setting $a = b = \mathbb{E}[X]$. This is because a $1 \times 1$ matrix is automatically symmetric, and the Loewner order on $1 \times 1$ matrices is just the standard ordering of the real numbers.

2. If we allow $n > 1$ but we restrict $X_1, \ldots, X_N$ to be diagonal matrices, then the Loewner order is just the entrywise partial ordering, where $A \succeq B$ if and only if each diagonal entry of $A$ is greater than or equal to the corresponding diagonal entry of $B$. We can apply the Chernoff bound to the $i^{\text{th}}$ diagonal entry to deduce

$$\Pr\left(X_{ii} \leq (1-\varepsilon)\mathbb{E}[X_{ii}]\right) \leq e^{-\varepsilon^2 a/2} \qquad \text{and} \qquad \Pr\left(X_{ii} \geq (1+\varepsilon)\mathbb{E}[X_{ii}]\right) \leq e^{-\varepsilon^2 b/3}.$$

   The matrix Chernoff bound (for the special case of diagonal matrices) follows by taking the union bound over all $n$ diagonal entries. Thus, in some sense, the matrix Chernoff bound generalizes the Chernoff-bound-plus-union-bound combination.

3. Let $\succeq_{\text{ew}}$ denote the entry-wise ordering on matrices, where $A \succeq_{\text{ew}} B$ if and only if $A_{ij} \geq B_{ij}$ for all $(i, j) \in [n]^2$. Generalizing the reasoning above about the case of diagonal matrices, we can deduce the following simple matrix concentration inequality. If $X_1, \ldots, X_N$ are independent random matrices satisfying $0 \preceq_{\text{ew}} X_i \preceq_{\text{ew}} 11^\top$, and their sum $X = X_1 + \cdots + X_N$ satisfies $a11^\top \preceq_{\text{ew}} \mathbb{E}[X] \preceq_{\text{ew}} b11^\top$, then

$$\Pr\left(X \not\succeq_{\text{ew}} (1-\varepsilon)\mathbb{E}[X]\right) \leq n^2 e^{-\varepsilon^2 a/2} \qquad \text{and} \qquad \Pr\left(X \not\preceq_{\text{ew}} (1+\varepsilon)\mathbb{E}[X]\right) \leq n^2 e^{-\varepsilon^2 b/3}.$$

   Compared to the matrix Chernoff bound there are two differences. The minor difference is the factor $n^2$ rather than $n$ on the right side. The major difference is the use of the entrywise ordering $\succeq_{\text{ew}}$ rather than the Loewner ordering $\succeq$. Inequalities in the Loewner ordering tend to be stronger (hence more useful) because the relation $A \succeq B$ asserts infinitely many inequalities $\langle \mathrm{x}, A\mathrm{x} \rangle \geq \langle \mathrm{x}, B\mathrm{x} \rangle$, one for each vector $\mathrm{x} \in \mathbb{R}^n$, whereas the relation $A \succeq_{\text{ew}} B$ asserts only $n^2$ inequalities $A_{ij} \geq B_{ij}$.

### 7.3.2 Singular values of Gaussian random matrices

Matrices of independent Gaussian random variables are frequently used in randomized algorithms for linear algebraic problems. The following tail bound concerning their singular values is highly useful for analyzing such algorithms.

**Theorem 7.15** Suppose $d \leq n$. If $A$ is an $m \times n$ matrix with independent entries each drawn from the distribution $\mathcal{N}(0, \frac{1}{n})$, then for all $\varepsilon > 0$, with probability at least $1 - 2e^{-\varepsilon^2 n/2}$ the singular values of $A$ satisfy

$$1 + \varepsilon + \sqrt{m/n} \geq \sigma_1(A) \geq \cdots \geq \sigma_m(A) \geq 1 - \varepsilon - \sqrt{m/n}.$$

Again, the proof of the theorem is beyond the scope of these notes; see [**DavidsonSzarek**] for a thorough treatment. Here, we limit ourselves to making a few remarks sketching how to prove a weaker version of the result.

- The conclusion of the theorem asserts that all the eigenvalues of $AA^\top$ are between $(1-\varepsilon-\sqrt{m/n})^2$ and $(1+\varepsilon+\sqrt{m/n})^2$. Note that, when reinterpreted as a statement about eigenvalues of $AA^\top$, Theorem 7.15 almost follows from the matrix Chernoff bound. If $a_i$ denote the $i^{\text{th}}$ column of $A$, then the identity $AA^\top = \sum_{i=1}^n a_i a_i^\top$ expresses $AA^\top$ as a sum of independent random matrices, each of them symmetric and positive semidefinite. Since the matrix entries $a_{ij}$ are independent $\mathcal{N}(0, \frac{1}{n})$ random variables, the expectation of $a_i a_i^\top$ is $\frac{1}{n}\mathbb{1}$. So, $AA^\top$ is a sum of $n$ i.i.d. random symmetric positive semidefinite matrices, and $\mathbb{E}[AA^\top] = \mathbb{1}$. If each summand $a_i a_i^\top$ were bounded above by $\mathbb{1}$ in the Loewner ordering, we could apply the matrix Chernoff bound to conclude that $(1-\varepsilon)\mathbb{1} \preceq AA^\top \preceq (1+\varepsilon)\mathbb{1}$ with high probability.
- Unfortunately, it is not the case that $a_i a_i^\top \preceq \mathbb{1}$ with probability 1. Instead, we have the inequality $a_i a_i^\top \preceq \|a_i\|_2^2 \mathbb{1}$, which follows from the Cauchy-Schwartz inequality:

$$\forall \mathrm{x} \in \mathbb{R}^n \quad \left\langle \mathrm{x}, a_i a_i^\top \mathrm{x} \right\rangle = \left( \langle a_i, \mathrm{x} \rangle \right)^2 \leq \|a_i\|_2^2 \|\mathrm{x}\|_2^2 = \|a_i\|_2^2 \langle \mathrm{x}, \mathbb{1}\mathrm{x} \rangle .$$

- Observe that

$$\mathbb{E}\left[\|a_i\|_2^2\right] = \sum_{i=1}^n \mathbb{E}[a_{ij}^2] = n \cdot \left(\frac{1}{n}\right) = 1.$$

  Recall from Problem Set 4 that $\Pr\left(\|a_i\|_2^2 > 1+\delta\right)$ is exponentially small in $n$.
- So, at the cost of an exponentially small failure probability, we can condition on the event that $\|a_i\|_2^2 \leq 1+\delta$ for all $i$. Since the $n$ columns of the matrix $A$ are independent random vectors and the conditioning event for column $i$ depends only on the random vector $a_i$, the vectors $a_i$ remain conditionally independent when we condition on the event $\mathcal{E} = \left\{\forall i \ \|a_i\|_2^2 \leq 1+\delta\right\}$.
- Conditional on event $\mathcal{E}$, all of the random matrices $\frac{1}{1+\delta} a_i a_i^\top$ are independent and are sandwiched between 0 and $\mathbb{1}$ in the Loewner order, so we can apply the matrix Chernoff bound to their sum, $AA^T$, to conclude that the conditional probabilities (given $\mathcal{E}$) of the events $AA^T \preceq (1-\delta)\mathbb{E}[AA^T \,|\, \mathcal{E}]$ and $AA^T \succeq (1+\delta)\mathbb{E}[AA^T \,|\, \mathcal{E}]$ are exponentially small in $n$.

- Since $\mathcal{E}$ is such a low-probability event, and the distribution of each random column vector $a_i$ is so light-tailed, one can show that

$$(1-o(1))\mathbb{1} \preceq \mathbb{E}[AA^T \mid \mathcal{E}] \preceq \mathbb{E}[AA^T] = \mathbb{1}$$

  where the $o(1)$ term tends to zero (exponentially fast) as $n \to \infty$.
- A version of Theorem 7.15 with a weaker exponential tail bound (e.g., a worse constant in the exponent) can be derived by combining the steps of reasoning above.

## 7.4 Algorithms Based on Random Projections

In the design of algorithms and data structures for linear-algebraic problems, random matrices play a role analogous to random hash functions in the design of discrete algorithms. A hash function maps a large set of keys to a much smaller set of buckets. Although collisions (two keys mapping to the same bucket) are inevitable, they are rare enough that any small set of keys are probably hashed to distinct buckets. Similarly, a random rectangular matrix (with more columns than rows) defines a linear transformation from a high-dimensional vector space to a lower-dimensional space. Again, collisions are inevitable but if the random matrix is sampled from a suitable distribution — for example, if the matrix entries are independent Gaussians — then any specific low-dimensional subspace of the domain has a high probability of being mapped *nearly isometrically* to a subspace of the range, meaning that the mapping nearly preserves the distance between every pair of vectors. The following definition and lemma make this observation precise.

**Lemma 7.16** Let $A \in \mathbb{R}^{m\times n}$ be a random matrix with independent entries each drawn from the Gaussian distribution $\mathcal{N}(0, \frac{1}{m})$. Consider any $\varepsilon > 0$ and any linear subspace $W \subseteq \mathbb{R}^n$ of dimension $d \leq \varepsilon^2 m$. With probability at least $1 - 2e^{-\varepsilon^2 m/2}$, left-multiplication by $A$ preserves the length of every vector in $W$ to within a factor between $1-2\varepsilon$ and $1+2\varepsilon$. In other words, $A$ satisfies the following property with probability at least $1-2e^{-\varepsilon^2 n/2}$:

$$\forall \mathrm{w} \in W \quad (1-2\varepsilon)\|\mathrm{w}\|_2 \leq \|A\mathrm{w}\|_2 \leq (1+2\varepsilon)\|\mathrm{w}\|_2. \tag{7.7}$$

*Proof.* Let $B$ denote an $n \times d$ matrix whose first $d$ columns constitute an orthonormal basis of $W$. We have

$$\max\{\|A\mathrm{w}\|_2 : \mathrm{w} \in W, \|\mathrm{w}\|_2 = 1\} = \max\{\|AB\mathrm{x}\|_2 : \mathrm{x} \in \mathbb{R}^d, \|\mathrm{x}\|_2 = 1\} = \sigma_1(AB)$$
$$\min\{\|A\mathrm{w}\|_2 : \mathrm{w} \in W, \|\mathrm{w}\|_2 = 1\} = \min\{\|AB\mathrm{x}\|_2 : \mathrm{x} \in \mathbb{R}^d, \|\mathrm{x}\|_2 = 1\} = \sigma_d(AB)$$

Hence, Inequality (7.7) is equivalent to the assertion that $1+2\varepsilon \geq \sigma_1(AB) \geq \sigma_d(AB) \geq 1-2\varepsilon$. To prove these inequalities, we will show that the matrix $AB$ has independent Gaussian entries and then apply Theorem 7.15.

Let $Q$ be an $n \times n$ orthogonal matrix whose first $d$ columns constitute the matrix $B$. If $\mathrm{a}_1, \mathrm{a}_2, \ldots, \mathrm{a}_m$ are the rows of $A$, then they are independent random vectors sampled from the distribution $\mathcal{N}(0, \frac{1}{n}\mathbb{1})$, which is a rotation-invariant distribution. Since $Q$ is an orthogonal matrix, it follows that the row vectors $\mathrm{a}_1 Q, \mathrm{a}_2 Q, \ldots, \mathrm{a}_m Q$ are also independent random samples from $\mathcal{N}(0, \frac{1}{n}\mathbb{1})$, i.e. the matrices $A$ and $AQ$ are identically distributed. Since $AB$ consists of the first $d$ columns of $AQ$, it follows that $AB$ is a $m \times d$ matrix of

independent entries drawn from $\mathcal{N}(0, \frac{1}{m})$. By assumption, $d \leq \varepsilon^2 m$ so $\sqrt{d}m \leq \varepsilon$. Applying Theorem 7.15 to the matrix $(AB)^\top$ allows us to conclude that with probability at least $1 - 2e^{-\varepsilon^2 m/2}$, $1 + 2\varepsilon \geq \sigma_1(AB) \geq \sigma_d(AB) \geq 1 - 2\varepsilon$, as desired. ■

### 7.4.1 Dimensionality Reduction and the Johnson-Lindenstrauss Lemma

Suppose you have a dataset consisting of vectors $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_N$ in $\mathbb{R}^n$. For example, this could be a collection of photos, each represented as a vector. The representation of a photo could be a vector of raw pixel values or, more likely, the output of an image processing algorithm. We are primarily interested in the case when $N$ and $n$ are both quite large. We will be implicitly assuming that the encoding of data as vectors has the property that similarity of data items translates to proximity, in the $L_2$ norm, between their corresponding vectors.

We are interested in projecting the data into a lower dimension, $m$, such that all distances between pairs of points are approximately preserved. This greatly reduces the computational cost of working with the data (e.g., searching for points near a specified query point) and the communication cost of sending information about the data points over a network.

In this section we will analyze a very simple dimensionality reduction algorithm due to Johnson and Lindenstrauss. The idea is simply to project the data from $\mathbb{R}^n$ to $\mathbb{R}^m$ using a linear transformation represented by a matrix with independent, identically distributed Gaussian entries. For now we will leave the dimension of the target space, $m$, as an indeterminate. Later we will see that for the purpose of preserving distances up to multiplicative error $\varepsilon$, it is appropriate to set $m = O\left(\frac{\log N}{\varepsilon^2}\right)$.

**Lemma 7.17 — Johnson-Lindenstrauss Lemma.** For any $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_N \in \mathbb{R}^n$ and any $0 < \varepsilon, \delta < 1$, if $m > 16\ln(N/\delta)/\varepsilon^2$ and $A$ is a $m \times n$ random matrix with independent entries drawn from the distribution $\mathcal{N}(0, \frac{1}{m})$, then with probability at least $1 - \delta^2$ the inequality

$$(1-\varepsilon)\|\mathrm{x}_i - \mathrm{x}_j\|_2 \leq \|A\mathrm{x}_i - A\mathrm{x}_j\|_2 \leq (1+\varepsilon)\|\mathrm{x}_i - \mathrm{x}_j\|_2$$

holds for all $1 \leq i, j \leq N$.

*Proof.* Assume without loss of generality that all of the vectors $\mathrm{x}_1, \ldots, \mathrm{x}_N$ are distinct. For any given pair $\mathrm{x}_i, \mathrm{x}_j$, the vector $\mathrm{y} = \mathrm{x}_i - \mathrm{x}_j$ spans a one-dimensional subspace $W \subseteq \mathbb{R}^n$. We may apply Lemma 7.16, equating the parameter $\varepsilon$ in the statement of that lemma with $\varepsilon/2$ here. The dimension of $W$ is less than $(\varepsilon/2)^2 m$ so the hypotheses of the lemma are satisfied. Consequently, the probability that $A$ distorts the length of y by a factor lying outside the interval $[1-\varepsilon, 1+\varepsilon]$ is less than

$$2e^{-\varepsilon^2 m/8} = 2e^{-2\ln(N/\delta)} = \frac{2\delta^2}{N^2} < \frac{\delta^2}{\binom{N}{2}}.$$

Taking the union bound, i.e. summing over all pairs $\mathrm{x}_i, \mathrm{x}_j$, establishes the lemma's conclusion. ■

### 7.4.2 Sparse Recovery

We have seen that a random projection from $\mathbb{R}^n$ to $\mathbb{R}^{O(\log(n)/\varepsilon^2)}$ approximately preserves the distance between every two elements of a finite set of $n$ vectors. In this section we will see that it also approximately preserves the distance between every two *sparse* vectors, i.e. those with few non-zero components. Putting this fact to use, we will show how to efficiently recover a sparse vector x given the vector $A$x, where $A$ is a Gaussian random matrix.

**Definition 7.18** A vector $\mathrm{x} \in \mathbb{R}^n$ is $s$-sparse if at least $n-s$ coordinates of x are equal to zero. A matrix $A$ satisfies the *$s$-restricted isometry property with constant $\varepsilon_s$* if the inequalities

$$(1-\varepsilon_s)\|\mathrm{x}\|_2^2 \leq \|A\mathrm{x}\|_2^2 \leq (1+\varepsilon_s)\|\mathrm{x}\|_2^2 \tag{7.8}$$

are satisfied for every $s$-sparse vector x.

**Proposition 7.19** For every $s \geq 1$ and $0 < \varepsilon_s, \delta < 1$, if $n \geq s$ and $m > \frac{50}{\varepsilon_s^2}\left[s\ln(n) + \ln(2/\delta)\right]$, then a matrix $A \in \mathbb{R}^{m\times n}$ with independent random entries sampled from $\mathcal{N}(0, \frac{1}{m})$ satisfies the $s$-restricted isometry property with constant $\varepsilon_s$, with probability at least $1-\delta$.

*Proof.* The inequalities $\left(1-\frac{2\varepsilon_s}{5}\right)^2 > 1-\varepsilon_s$ and $\left(1+\frac{2\varepsilon_s}{5}\right)^2 < 1+\varepsilon_s$ are satisfied whenever $0 < \varepsilon_s < 1$. Hence, if we let $\varepsilon = \varepsilon_s/5$, then the inequalities $(1-2\varepsilon)\|\mathrm{w}\|_2 \leq \|A\mathrm{w}\|_2 \leq (1+2\varepsilon)\|\mathrm{w}\|_2$ for a vector w are sufficient to imply $(1-\varepsilon_s)\|\mathrm{w}\|_2^2 \leq \|A\mathrm{w}\|_2^2 \leq (1+\varepsilon_s)\|\mathrm{w}\|_2^2$.

For any subset $J \subseteq [n]$ let $W_J$ be the $s$-dimensional subspace of $\mathbb{R}^n$ consisting of vectors x satisfying $\mathrm{x}_i = 0$ for all $i \notin J$. The $s$-sparse vectors in $\mathbb{R}^n$ are precisely the union of the subspaces $W_J$ as $J$ ranges over the $s$-element subsets of $[n]$. By assumption, $s \leq \varepsilon^2 m$, so Lemma 7.16 ensures that for any fixed $J$, the probability that Inequality (7.8) is violated by some $\mathrm{x} \in W_J$ is at most $2e^{-\varepsilon^2 m/2} = 2e^{-\varepsilon_s^2 m/50}$. Taking the union bound over $s$-element sets $J$, we find that $A$ satisfies the $s$-restricted isometry property with constant $\varepsilon_s$, with probability at least $1 - 2\binom{n}{s}e^{-\varepsilon_s^2 m/50}$. This probability is greater than $1-\delta$ provided that $m > \frac{50}{\varepsilon_s^2}\left[s\ln(n) + \ln(2/\delta)\right]$. ∎

The main application of matrices with the restricted isometry property is to solve an inverse problem called *sparse recovery* where the aim is to identify a sparse vector $\mathrm{x} \in \mathbb{R}^n$ given the value of $\mathrm{b} = A\mathrm{x} \in \mathbb{R}^m$. When $m < n$ this is an underdetermined linear system, meaning there are infinitely many vectors y solving the equation $A\mathrm{y} = \mathrm{b}$. The set of all such solutions forms a $(n-m)$-dimensional affine subspace of $\mathbb{R}^n$, but we will see that there is a unique $s$-sparse solution provided that $A$ satisfies the $3s$-restricted isometry property with $\varepsilon < \frac{1}{3}$. Furthermore, we'll see that there is an efficient algorithm to find the sparse vector x satisfying $A\mathrm{x} = \mathrm{b}$.

**Definition 7.20** A vector $\mathrm{z} \in \mathbb{R}^n$ is *mostly $s$-sparse* if there is an index set $J \subseteq [n]$ with $|J| \leq s$ such that

$$\sum_{i\in J}|z_i| \geq \sum_{i\notin J}|z_i|.$$

By definition, a matrix with the $s$-restricted isometry property approximately preserves

the 2-norm of every $s$-sparse vector. Our next lemma shows that the length of every mostly $s$-sparse vector is also approximately preserved, albeit with a worse approximation factor, if we make the stronger assumptions that the matrix satisfies the $(3s)$-restricted isometry property and that $\varepsilon < \frac{1}{3}$. (The upper bound on $\varepsilon$ is used to ensure that the constant factor on the right side of Inequality (7.9) below is strictly positive.)

**Lemma 7.21** Suppose $A$ is a matrix that satisfies the $(3s)$-restricted isometry property with constant $\varepsilon > 0$. If z is mostly $s$-sparse then

$$\|Az\|_2 \geq \frac{1}{2}\left(\sqrt{1-\varepsilon} - \sqrt{\tfrac{1+\varepsilon}{2}}\right)\|z\|_2. \tag{7.9}$$

*Proof.* Without loss of generality assume that the coordinates of z are ordered such that $|z_1| \geq |z_2| \geq \cdots \geq |z_n|$. Also assume without loss of generality that $n = (2m+1)s$ for some positive integer $m$. (Otherwise, pad the vector z with zeros and increase the number of columns of $A$ from $n$ to $(2m+1)s$, while continuing to satisfy the restricted isometry property.)

Break the coordinate range $[n] = [(2m+1)s]$ into $m+1$ blocks $J_0, J_1, \ldots, J_m$ such that $J_0$ consists of the first $s$ coordinates, $J_1$ consists of the next $2s$ coordinates, $J_2$ consists of the next $2s$ coordinates after that, and so on. In other words,

$$J_\ell = \{i \mid i > 0,\ (2\ell-1)s < i \leq (2\ell+1)s\}.$$

Let $z_\ell$ be a vector obtained from z by preserving the coordinates in block $J_\ell$ and setting all other coordinates to zero. In other words,

$$(z_\ell)_i = \begin{cases} z_i & \text{if } (2\ell-1)s < i \leq (2\ell+1)s \\ 0 & \text{otherwise.} \end{cases}$$

Since z is mostly $s$-sparse, and we are assuming the coordinates are sorted so that $|z_1| \geq |z_2| \geq \cdots \geq |z_n|$, we have

$$\|z_0\|_1 \geq \|z_1 + z_2 + \cdots + z_m\|_1. \tag{7.10}$$

Another useful observation stemming from the way coordinates are ordered is that $2s \cdot \|z_{i+1}\|_\infty \leq \|z_i\|_1$, because the absolute value of *every* coordinate of $z_{i+1}$ is less than or equal to the absolute value of *every* coordinate of $z_i$. Combining this observation with the inequality $\|z_{i+1}\|_2 \leq \sqrt{2s}\|z_{i+1}\|_\infty$, we obtain

$$\|z_{i+1}\|_2 \leq \frac{1}{\sqrt{2s}}\|z_i\|_1.$$

Now, we can bound $\|Az\|_2$ from below as follows.

$$\begin{aligned}
\|Az\|_2 &= \|A(z_0+z_1)+Az_2+Az_3+\cdots+Az_m\|_2 \\
&\geq \|A(z_0+z_1)\|_2 - \big(\|Az_2\|_2+\|Az_3\|_2+\cdots+\|Az_m\|_2\big) \\
&\geq \sqrt{1-\varepsilon}\,\|z_0+z_1\|_2 - \sqrt{1+\varepsilon}\,\big(\|z_2\|_2+\|z_3\|_2+\cdots+\|z_m\|_2\big) \qquad (7.11)\\
&\geq \sqrt{1-\varepsilon}\|z_0\|_2 - \sqrt{\frac{1+\varepsilon}{2s}}\big(\|z_1\|_1+\|z_2\|_1+\cdots+\|z_{m-1}\|_1\big) \\
&= \sqrt{1-\varepsilon}\|z_0\|_2 - \sqrt{\frac{1+\varepsilon}{2s}}\|z_1+z_2+\cdots+z_{m-1}\|_1 \\
&\geq \sqrt{1-\varepsilon}\|z_0\|_2 - \sqrt{\frac{1+\varepsilon}{2s}}\|z_0\|_1. \qquad (7.12)
\end{aligned}$$

In line (7.11) we have used the inequalities $\sqrt{1-\varepsilon}\|z_0+z_1\|_2 \leq \|A(z_0+z_1)\|_2$ and $\sqrt{1+\varepsilon}\|z_i\|_2 \geq \|Az_i\|_2$, both of which are justified by the $(3s)$-restricted isometry property with constant $\varepsilon$.

Let $\sigma$ be a vector in $\{-1,0,1\}^n$ with the same sign pattern and sparsity pattern as $z_0$, meaning that

$$\sigma_i = \begin{cases} 1 & \text{if } z_{0i} > 0 \\ 0 & \text{if } z_{0i} = 0 \\ -1 & \text{if } z_{01} < 0. \end{cases}$$

Then $\langle \sigma, z_0 \rangle = \|z_0\|_1$, so the Cauchy-Schwartz inequality implies

$$\|z_0\|_1 \leq \|\sigma\|_2\,\|z_0\|_2 = \sqrt{s}\,\|z_0\|_2.$$

Substituting this bound into inequality (7.12) above, we find that

$$\|Az\|_2 \geq \left(\sqrt{1-\varepsilon} - \sqrt{\frac{1+\varepsilon}{2}}\right)\|z_0\|_2. \qquad (7.13)$$

To conclude the proof of the lemma we need to show $\|z_0\|_2 \geq \frac{1}{2}\|z\|_2$. Let $t = \frac{1}{s}\|z_0\|_1 = \frac{1}{s}\left(|z_1|+|z_2|+\cdots+|z_s|\right)$ and observe $t \geq |z_s|$. Every component of the vector $w = \frac{1}{t}\left(z_1+z_2+\cdots+z_m\right)$ belongs to the interval $[-1,1]$, because $|z_i| \leq |z_s| \leq t$ for $i > s$. Hence,

$$\begin{aligned}
\|w\|_2^2 &= \sum_{i=1}^n w_i^2 \leq \sum_{i=1}^n |w_i| = \|w\|_1 \\
\|z-z_0\|_2^2 &= t^2\|w\|_2^2 \leq t^2\|w\|_1 = t\|z_1+\cdots+z_m\|_1 \leq t\|z_0\|_1 = \frac{1}{s}\|z_0\|_1^2 \leq \|z_0\|_2^2.
\end{aligned} \qquad (7.14)$$

By the triangle inequality, $\|z\|_2 \leq \|z_0\|_2 + \|z-z_0\|_2$. Combined with Inequality (7.14), this implies $\|z\|_2 \leq 2\|z_0\|_2$ and completes the proof of the lemma. ■

We will use Lemma 7.21 to analyze the following algorithm for sparse recovery: of all the vectors $x$ that satisfy $Ax = b$, output one with minimum $L_1$ norm. The $L_1$ norm is a convex function, so the problem can be solved efficiently using a convex minimization algorithm, such as gradient descent.

**Proposition 7.22** Suppose $A$ is a matrix that satisfies the $(3s)$-restricted isometry property with constant $\varepsilon < \frac{1}{3}$, $\mathrm{x}_0$ is an $s$-sparse vector, and $\mathrm{b} = A\mathrm{x}_0$. Among the solutions of the equation $A\mathrm{x} = \mathrm{b}$, the vector $\mathrm{x}_0$ is the unique one with minimum $L_1$ norm.

*Proof.* Suppose $\mathrm{x}_1$ is a solution of minimum $L_1$ norm to the equation $A\mathrm{x} = \mathrm{b}$. We must prove that $\mathrm{x}_1 = \mathrm{x}_0$. Let $\mathrm{z} = \mathrm{x}_1 - \mathrm{x}_0$, and observe that $A\mathrm{z} = A\mathrm{x}_1 - A\mathrm{x}_0 = \mathrm{b} - \mathrm{b} = 0$. Let $J = \{i \mid x_{0i} \neq 0\}$ and observe $|J| \leq s$. We have

$$\begin{aligned}\|\mathrm{x}_1\|_1 &= \sum_{i=1}^{n} |x_{1i}| = \sum_{i=1}^{n} |x_{0i} + z_i| = \sum_{i\in J} |x_{0i} + z_i| + \sum_{i\notin J} |z_i| \\ &\geq \sum_{i\in J} |x_{0i}| - \sum_{i\in J} |z_i| + \sum_{i\notin J} |z_i| = \|\mathrm{x}_0\|_1 - \sum_{i\in J} |z_i| + \sum_{i\notin J} |z_i|.\end{aligned}$$

Since $\|\mathrm{x}_1\|_1 \leq \|\mathrm{x}_0\|_1$ by our choice of $\|\mathrm{x}_1\|$, it follows that $\sum_{i\in J} |z_i| \geq \sum_{i\notin J} |z_i|$, i.e. z is mostly $s$-sparse. By Lemma 7.21,

$$0 = \|A\mathrm{z}\|_2 \geq \frac{1}{2}\left(\sqrt{1-\varepsilon} - \sqrt{\frac{1+\varepsilon}{2}}\right)\|\mathrm{z}\|_2.$$

Our assumption $\varepsilon < \frac{1}{3}$ implies $1 - \varepsilon > \frac{1+\varepsilon}{2}$, so the factor $\frac{1}{2}\left(\sqrt{1-\varepsilon} - \sqrt{\frac{1+\varepsilon}{2}}\right)$ on the right side is strictly positive. It follows that $\|\mathrm{z}\|_2 = 0$, so $0 = \mathrm{z} = \mathrm{x}_1 - \mathrm{x}_0$, as desired. ■