These notes introduce 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.

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 $\{\textsc{true}, \textsc{false}\}$) 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 ϕ is "measurable." We will not give the definition of measurable here, but we will say that when T is a vector space and ϕ is any predicate that can defined using continuous functions, equations, and inequalities, ϕ 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 ϕ_0 and ϕ_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 ϕ. 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 ϕ, $\Pr(\phi(Y)) = \Pr(\phi(F(X)))$. (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 $\Pr(X \in S) = 1$.

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.

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 ϕ, ψ. 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 $i \in \mathcal{I}$.

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 1.1. *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 θ, 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 1.2. *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 θ tends to $-\infty$ and ∞, 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 1.3. *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 1.3 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 1.1. 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 1.3 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 $X = \frac{1}{r} \ln(\frac{1}{1-y})$.

1.4 Probability density

If V is a finite-dimensional vector space and $f : V \to [0, \infty)$ is a function satisfying $\int_V f(\mathbf{x})\, d\mathbf{x} = 1$ then one can construct a random variable X whose distribution satisfies $\Pr(X \in S) = \int_S f(\mathbf{x})\, d\mathbf{x}$ 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

$$h(\mathbf{x}, \mathbf{y}) = f(\mathbf{x})g(\mathbf{y}).$$

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 1.4. *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

$$\sum_{i=1}^{n} s_i p_i = \sum_{i=1}^{n} \sum_{j=1}^{i} (s_j - s_{j-1}) p_i = \sum_{j=1}^{n} \sum_{i=j}^{n} (s_j - s_{j-1}) p_i = \sum_{j=1}^{n} (s_j - s_{j-1}) \Pr(X > s_{j-1}) \tag{1}$$

In addition we have

$$\int_0^\infty \Pr(X > \theta)\, d\theta = \sum_{j=1}^{n} \int_{s_{j-1}}^{s_j} \Pr(X > \theta)\, d\theta = \sum_{j=1}^{n} (s_j - s_{j-1}) \Pr(X > s_{j-1}). \tag{2}$$

The right sides of Equations (1) and (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 1.5. *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

$$\int_0^\infty \theta\, f_X(\theta)\, d\theta == \int_{\theta=0}^\infty (1 - F_X(\theta))\, d\theta + \left(\lim_{\theta \to \infty} \theta \cdot (1 - F_X(\theta)) \right) = \mathbb{E}[X] + \lim_{\theta \to \infty} \theta \cdot (1 - F_X(\theta)). \tag{3}$$

The proof divides now into two cases. If the limit on the right side of Equations (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 (3) are infinite. Define an infinite sequence of positive numbers $\theta_1, \theta_2, \ldots$ recursively, by choosing θ_1 to be any element of Θ_ε and choosing θ_{n+1} to be any element of Θ_ε 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 (3) is infinite, as claimed. As for the left side of (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^{th} coordinate of $\mathbb{E}[X]$ is the expectation of the i^{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 1.6. *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 1.7 (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 θ we obtain Markov's Inequality. □

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 1.8. *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 1.6 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

$$\mathrm{Cov}(X) = \mathbb{E}\left[(X-\mathbb{E}[X])\,(X-\mathbb{E}[X])^\top\right].$$

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 2.1 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be an infinite sequence of identically distributed real-valued random variables, each with finite expectation μ and finite variance σ^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 θ. 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

$$f(x, y) = \left(\frac{1}{Z}e^{-\frac{1}{2}x^2}\right)\left(\frac{1}{Z}e^{-\frac{1}{2}y^2}\right) = \frac{1}{Z^2}e^{-\frac{1}{2}(x^2+y^2)}.$$

Now, represent the pair (X, Y) in polar coordinates as (R, Θ) where R and Θ are random variables satisfying $X = R\cos(\Theta),\ Y = R\sin(\Theta)$. We have

$$\frac{1}{Z^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)}dx\,dy = \frac{1}{Z^2}\int_0^{2\pi}\int_0^{\infty} e^{-\frac{1}{2}r^2} r\,dr\,d\theta \tag{4}$$

The extra factor of r in the integrand is attributable to the change-of-variables formula for integrals in polar coordinates, $dx\,dy = r\,dr\,d\theta$. 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

$$\frac{1}{Z^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)}dx\,dy = \frac{1}{Z^2}\int_0^{2\pi}\int_0^{\infty} e^{-u}du\,d\theta. \tag{5}$$

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 disibutrion $\mathcal{N}(0, 1)$ *has density* $f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$.

2. From the right side of Equation (5) we can deduce that R and Θ are independent random variables, Θ 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* Θ *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 μ and variance σ^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 μ and variance σ^2.*

2.1 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* $\mathcal{N}(\mathbf{0}, \mathbb{1})$, 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 $\mathcal{N}(\mathbf{0}, \mathbb{1})$ is the function

$$f(\mathbf{x}) = \left(\tfrac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}\left(x_1^2 + x_2^2 + \cdots + x_n^2\right)} = \left(\tfrac{1}{2\pi}\right)^{d/2} e^{-\frac{1}{2}\langle \mathbf{x}, \mathbf{x}\rangle}. \tag{6}$$

If $X \sim \mathcal{N}(\mathbf{0}, \mathbb{1})$, then its expectation and covariance matrix are $\mathbb{E}[X] = \mathbf{0}$ and $\mathrm{Cov}(X) = \mathbb{1}$, respectively.

If $X \sim \mathcal{N}(\mathbf{0}, \mathbb{1})$ then the distribution of X has two key properties that are evident from Equation (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 2.2. *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 $\mathbf{a} = \frac{1}{\sigma}(a_1, a_2, \ldots, a_n)$ satisfies $\|a\|_2 = 1$. Hence, there exists an orthogonal matrix Q whose first row is $\mathbf{a}$. The random vector $X = (X_1, \ldots, X_n)$ has the distribution $\mathcal{N}(\mathbf{0}, \mathbb{1})$, so $QX \sim \mathcal{N}(\mathbf{0}, \mathbb{1})$ as well. The first coordinate of the vector QX is Y/σ, 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 $\mathcal{N}(\mathbf{0}, \mathbb{1})$, B is an invertible $n \times n$ matrix, and $\boldsymbol{\mu}$ is any vector in $\mathbb{R}^n$, then the distribution of $Y = BX + \boldsymbol{\mu}$ is called a *multivariate Gaussian distribution.* The expectation of Y is $\boldsymbol{\mu}$ and its covariance is

$$\mathrm{Cov}(Y) = \mathbb{E}\left[(Y-\boldsymbol{\mu})(Y-\boldsymbol{\mu})^\top\right] = \mathbb{E}\left[(BX)(BX)^\top\right] = B\mathbb{E}[XX^\top]B^\top = B\,B^\top,$$

since $\mathbb{E}[XX^\top] = \mathrm{Cov}(X) = \mathbb{1}$. The distribution of Y is denoted by $\mathcal{N}(\boldsymbol{\mu}, BB^\top)$. The density of Y can be calculated as follows. Let T denote the function $T(\mathbf{x}) = B\mathbf{x} + \boldsymbol{\mu}$. Its inverse is the function $T^{-1}(\mathbf{y}) = B^{-1}(\mathbf{y} - \boldsymbol{\mu})$. A small ball $\mathcal{B}$ of volume $\varepsilon > 0$ centered at $\mathbf{y}$ is mapped by T^{-1} to a small ellipsoid $\mathcal{E}$ of volume $|\det(B^{-1})| \cdot \varepsilon$ centered at $\mathbf{x} = T^{-1}(\mathbf{y})$. We have

$$\Pr(Y \in \mathcal{B}) = \Pr(X \in \mathcal{E}) = \left[|\det(B^{-1})| \left(\tfrac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}\langle \mathbf{x}, \mathbf{x}\rangle} + o(1) \right] \cdot \varepsilon,$$

where $o(1)$ denotes an error term that converges to zero as $\varepsilon \to 0$. Thus, the density of Y at $\mathbf{y}$ is $|\det(B^{-1})| \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}\langle \mathbf{x}, \mathbf{x}\rangle}$. Now, recalling that $\mathbf{x} = T^{-1}(\mathbf{y}) = B^{-1}(\mathbf{y} - \boldsymbol{\mu})$, we have

$$\langle \mathbf{x}, \mathbf{x}\rangle = \left\langle B^{-1}(\mathbf{y} - \boldsymbol{\mu}), B^{-1}(\mathbf{y} - \boldsymbol{\mu}) \right\rangle = \left\langle \mathbf{y} - \boldsymbol{\mu}, (B^{-1})^\top B^{-1}(\mathbf{y} - \boldsymbol{\mu}) \right\rangle = \left\langle \mathbf{y} - \boldsymbol{\mu}, (BB^\top)^{-1}(\mathbf{y} - \boldsymbol{\mu}) \right\rangle.$$

The right side depends only on $BB^\top$, not on B. Thus, if two multivariate Gaussian random variables have the same mean μ and the same covariance matrix $\Sigma = BB^\top$, then they are identically distributed and their density is

$$f(\mathbf{y}) = \det(\Sigma)^{-1/2} \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}\langle \mathbf{y} - \boldsymbol{\mu}, \Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu})\rangle}.$$

Lemma 2.3. *If A is a $d \times n$ matrix of rank d and $X \sim \mathcal{N}(\mathbf{0}, \mathbb{1})$ is a $\mathbb{R}^n$-valued multivariate normal random variable, then $Y = AX$ is a $\mathbb{R}^d$-valued Gaussian random variable with distribution $\mathcal{N}(\mathbf{0}, AA^\top)$.*

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 $S = D\,[\mathbb{1}\ \mathbf{0}]$, where D is a $d \times d$ diagonal matrix with the singular values of A on the diagonal, and $[\mathbb{1}\ \mathbf{0}]$ 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

$$Y = UD\,[\mathbb{1}\ \mathbf{0}]\,V^\top X.$$

Let $W = [\mathbb{1}\ \mathbf{0}]\,V^\top X$. 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 $[\mathbb{1}\ \mathbf{0}]\,X$, i.e. the first d coordinates of X. In other words, $W \sim \mathcal{N}(\mathbf{0}, \mathbb{1})$, 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 $W \sim \mathcal{N}(\mathbf{0}, \mathbb{1})$ and $Y = BW$ for an invertible matrix B, then $Y \sim \mathcal{N}(\mathbf{0}, BB^\top)$. To finish up, note that

$$BB^\top = UD^2U^\top = USS^\top U^\top = AA^\top,$$

so $Y \sim \mathcal{N}(\mathbf{0}, AA^\top)$ as claimed. □

We remark that Lemma 2.2 corresponds to the special case of Lemma 2.3 where A has only one row.

3 The Chernoff and Hoeffding Bounds

When one averages n independent random variables, each with bounded mean and variance, the average is unlikely to differ from its expected value by much more than $\frac{1}{\sqrt{n}}$. In Section 2 we encountered one important theorem that gives a precise and quantitative interpretation to this intuition, namely the Central Limit Theorem. However, the Central Limit Theorem is an asymptotic statement that holds as $n \to \infty$. In the analysis of randomized algorithms, and in the average-case analysis of algorithms applied to random datasets, it is often important to make use of non-asymptotic theorems that bound the probability that an average of n independent random variables will be far from its expected value, for fixed n rather than $n \to \infty$. This section presents two such theorems, the Chernoff and Hoeffding Bounds, along with a few applications.

3.1 The cumulant generating function of a random variable

For a real-valued random variable X, the *cumulant generating function* $K_X(t)$ is defined by the formula

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

The importance of this function lies in the fact that it behaves additively with respect to summation of independent random variables.

Lemma 3.1. *If X, Y are independent real-valued random variables then*

$$K_{X+Y}(t) = K_X(t) + K_Y(t)$$

for all t.

Proof. We have

$$K_{X+Y}(t) = \ln\left(\mathbb{E}\left[e^{t(X+Y)}\right]\right) = \ln\left(\mathbb{E}\left[e^{tX}e^{tY}\right]\right) = \ln\left(\mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right]\right) = \ln\left(\mathbb{E}\left[e^{tX}\right]\right) + \ln\left(\mathbb{E}\left[e^{tY}\right]\right),$$

where the equation $\mathbb{E}\left[e^{tX}e^{tY}\right] = \mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right]$ follows from the independence of X and Y. □

Another useful property of the cumulant generating function is its behavior under scaling.

Lemma 3.2. *For any real-valued random variable X and any $\lambda, t \in \mathbb{R}$,*

$$K_{\lambda X}(t) = K_X(\lambda t).$$

Proof.

$$K_{\lambda X}(t) = \ln\left(\mathbb{E}\left[e^{t\lambda X}\right]\right) = K_X(\lambda t).$$

□

When one expands the cumulant generating function K_X as a power series in t near $t = 0$, the coefficients of the power series recover many useful parameters of the distribution of X.

$$\begin{aligned}
K_X(t) &= \ln\left(\mathbb{E}\left[1 + tX + \frac{1}{2}t^2X^2 + \frac{1}{6}t^3X^3 + \cdots\right]\right) \\
&= \ln\left(1 + \mathbb{E}[X]t + \frac{1}{2}\mathbb{E}[X^2]t^2 + \frac{1}{6}\mathbb{E}[X^3]t^3 + \cdots\right) \\
&= \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}\left(\mathbb{E}[X]t + \frac{1}{2}\mathbb{E}[X^2]t^2 + \frac{1}{6}\mathbb{E}[X^3]t^3 + \cdots\right)^k \\
&= \left(\mathbb{E}[X]t + \frac{1}{2}\mathbb{E}[X^2]t^2 + \frac{1}{6}\mathbb{E}[X^3]t^3 + \cdots\right) - \left(\frac{1}{2}\mathbb{E}[X]^2t^2 + \frac{1}{2}\mathbb{E}[X]\mathbb{E}[X^2]t^3 + \cdots\right) \\
&\quad + \left(\frac{1}{3}\mathbb{E}[X]^3t^3 + \cdots\right) + O(t^4) \\
&= \mathbb{E}[X]t + \frac{1}{2}\left(\mathbb{E}[X^2] - \mathbb{E}[X]^2\right)t^2 + \frac{1}{6}\left(\mathbb{E}[X^3] - 3\mathbb{E}[X]\mathbb{E}[X^2] + 2\mathbb{E}[X]^3\right) + O(t^4).
\end{aligned}$$

The expectation and variance of X occur as the coefficients of t and $\frac{1}{2}t^2$, respectively. The coefficient of $\frac{1}{n!}t^n$ is called the n^{th} cumulant of X and is denoted by $\kappa_n(X)$. The identity $\kappa_n(X + Y) = \kappa_n(X) + \kappa_n(Y)$, valid for all $n \in \mathbb{N}$ and all independent random variables X, Y, follows from Lemma 3.1. This generalizes the familiar facts that expectation and variance behave additively when applied to sums of independent random variables.

The cumulant generating function forms the centerpiece of one proof of the Central Limit Theorem. To begin with, let us calculate the cumulant generating function of X when $X \sim \mathcal{N}(0, 1)$.

$$\begin{aligned}
\mathbb{E}\left[e^{tX}\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx}e^{-\frac{1}{2}x^2}\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2-2tx)}\,dx = \frac{1}{\sqrt{2\pi}}e^{\frac{1}{2}t^2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-t)^2}\,dx = e^{\frac{1}{2}t^2} \\
K_X(t) &= \ln\left(\mathbb{E}\left[e^{tX}\right]\right) = \frac{1}{2}t^2
\end{aligned}$$

Now consider an infinite sequence of independent, identically distributed random variables $Y_1, Y_2, \ldots$ with $\mathbb{E}[Y_i] = 0$ and $\mathrm{Var}(Y_i) = 1$ for all i. Denote the cumulant generating function of each Y_i by $K_Y(t)$. Since $\mathbb{E}[Y_i] = 0$ and $\mathrm{Var}(Y_i) = 1$, we have $K_Y(t) = \frac{1}{2}t^2 + O(t^3)$. The Central Limit Theorem is equivalent to the assertion that $Z_n = \frac{1}{\sqrt{n}}(Y_1 + \cdots Y_n) \xrightarrow{d} \mathcal{N}(0, 1)$ as $n \to \infty$. Using Lemmas 3.1 and 3.2, we have

$$K_{Z_n}(t) = K_{Y_1+\cdots+Y_n}\left(\frac{t}{\sqrt{n}}\right) = nK_Y\left(\frac{t}{\sqrt{n}}\right) = n\left(\frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + O\left(\frac{t}{\sqrt{n}}\right)^3)\right) = \frac{1}{2}t^2 + O\left(\frac{t^3}{\sqrt{n}}\right),$$

so $K_{Z_n}(t) \to \frac{1}{2}t^2$ as $n \to \infty$. Recall that $\frac{1}{2}t^2$ is the cumulant generating function of the distribution $\mathcal{N}(0, 1)$. The remainder of the proof needs to show that the convergence of the cumulant generating function of Z_n to that of $\mathcal{N}(0, 1)$ implies the convergence of the distribution of Z_n to $\mathcal{N}(0, 1)$. This is the more difficult part of the proof, and we will omit discussion of it here.

The cumulant generating function furnishes a useful way to bound the probability of a random variable deviating significantly from its expected value.

Lemma 3.3. *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. □

3.2 The Chernoff Bound

Let $X_1, X_2, \ldots, X_n$ 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 $X_1 + \cdots + X_n$ differs from its expectation by a factor lying outside the interval $[1 - \varepsilon, 1 + \varepsilon]$. We will assume throughout this section that $0 < \varepsilon < 1$.

Lemma 3.4. *For any random variable X taking values in $[0, 1]$, the cumulant generating function K_X satisfies*

$$K_X(t) \leq (e^t - 1)\mathbb{E}[X]$$

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

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

The lemma follows upon taking the natural logarithm of both sides and using the inequality $\ln(1 + z) \leq z$, which is valid for all $z > 0$. □

Theorem 3.5 (Chernoff bound). *If $X_1, X_2, \ldots, X_n$ are independent random variables taking values in $[0, 1]$ and $X = X_1 + \cdots + X_n$, then*

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

Proof. Using Lemmas 3.1 and 3.4, together with $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i]$, we find that $K_X(t) \leq (e^t - 1)\mathbb{E}[X]$ for all $t \in \mathbb{R}$. Now, from Lemma 3.3 we have

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

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 Lemma 3.3 yields

$$\Pr(X \leq (1-\varepsilon)\mathbb{E}[X]) \;<\; e^{(e^{-t}-1+(1-\varepsilon)t)\mathbb{E}[X]}.$$

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

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

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 3.5 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 Chernoff bound concerns that probability that the ratio $X/\mathbb{E}[X]$ is far from 1, i.e. it pertains to multiplicative deviations of X from its expected value. In Section 3.3 we present an exponential tail bound for additive deviations, i.e. a bound on the probability that $|X - \mathbb{E}[X]|$ is far from zero.

4. The probability of large deviation is exponentially small as a function of $\mathbb{E}[X]$, not as a function of n. Even if n is very large, it's possible that the distribution of X is not very concentrated around its expected value. For example, if $X_1, \ldots, X_{n-1}$ are deterministically equal to 0, and X_n 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.

5. In the exponential function on the right side of the Chernoff bound, the dependence on ε 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.*"

3.3 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 using an application of Lemma 3.3, and the key ingredient will be a lemma that furnishes an upper bound on the cumulant generating function of a random variable.

Lemma 3.6 (Hoeffding's Lemma)**.** *If X is a random variable supported on an interval $[a, b]$, with expected value μ, 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, $K_X(t) = \frac{1}{2} K_X''(u) t^2$ 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, the distribution of Y is the unique distribution satisfying the property that for every function g, $\mathbb{E}[g(Y)] = \mathbb{E}[e^{uX} g(X)]/\mathbb{E}[e^{uX}]$. Using the function $g(z) = e^{tz}$, we find that $\mathbb{E}[e^{tY}] = \mathbb{E}[e^{(u+t)X}]/\mathbb{E}[e^{uX}]$. Taking the logarithm of both sides, we obtain $K_Y(t) = K_X(u + t) - K_X(u)$, as desired.

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 3.7. *(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 3.3,

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

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

3.4 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.

3.4.1 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 ε:

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

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 δ, 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 ε/M because we are trying to estimate a quantity belonging to an interval of length M, and we tolerate additive error up to ε. On the other hand, the number of samples depends on logarithmically on the inverse of the "confidence parameter," δ, which governs the maximum failure probability that is deemed tolerable.

3.4.2 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 $L : \mathcal{H} \times T \to [0, 1]$. 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 3.8. *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 ϕ, 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 ϕ, 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 $\left|\overline{L}(h) - \frac{1}{N}\sum_{i=1}^{N} L(h_j, Z_i)\right| \geq \frac{\varepsilon}{2}$. 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 $\overline{L}(h)$, 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 3.9 below) 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. □

In the proof we used the following simple fact about the probability of a disjunction of finitely many predicates, called the Union Bound.

Lemma 3.9 (Union Bound)**.** *If X is a random variable and $\psi_1(X), \ldots, \psi_m(X)$ are Boolean predicates, then*

$$\Pr\left(\vee_{j=1}^{m}\psi_j(X)\right) \leq \sum_{j=1}^{m} \Pr(\psi_j(X)).$$

Proof. Our plan is to use the property that probabilities are finitely additive. However, we can't use finite additivity directly because the predicates $\psi_1, \ldots, \psi_m$ are not (necessarily) mutually exclusive. To work this obstacle, we construct mutually exclusive predicates $\phi_1, \phi_2, \ldots, \phi_m$ as follows:

$$\phi_j(X) = \psi_j(X) \wedge \neg\left(\vee_{k=1}^{j-1}\psi_k(X)\right).$$

By induction on j, the relation

$$\vee_{k=1}^{j} \phi_k(X) = \vee_{k=1}^{j}\psi_k(X) \tag{7}$$

holds. Consequently, for all $j > i$ we have

$$\phi_j(X) \Rightarrow \neg\left(\vee_{k=1}^{j-1}\psi_k(X)\right) \Rightarrow \neg\left(\vee_{k=1}^{j-1}\phi_k(X)\right) \Rightarrow \neg\phi_i(X)$$

which shows that $\phi_1(X), \ldots, \phi_m(X)$ are mutually exclusive and justifies

$$\Pr\left(\wedge_{j=1}^{m}\phi_j(X)\right) = \sum_{j=1}^{m} \Pr(\phi_j(X)). \tag{8}$$

If $\upsilon_j(X) = \psi_j(X) \wedge \overline{\phi_j}(X)$ then $\phi_j(X)$ and $\upsilon_j(X)$ are mutually exclusive and $\psi_j(X) = \phi_j(X) \vee \upsilon_j(X)$ so

$$\Pr(\psi_j(X)) = \Pr(\phi_j(X)) + \Pr(\upsilon_j(X)) \geq \Pr(\phi_j(X)). \tag{9}$$

Applying Equation (7) with $j = m$, and using Equations (**??**), we obtain

$$\Pr\left(\vee_{j=1}^{m}\psi_j(X)\right) = \Pr\left(\vee_{j=1}^{m}\phi_j(X)\right) = \sum_{j=1}^{m}\phi_j(X) \leq \sum_{j=1}^{m}\psi_j(X),$$

which completes the proof of the Union Bound. □

3.5 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 Π

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.

$$\forall x \quad \Pr(A(x,r) \neq \Pi(x)) \leq \frac{1}{3}. \tag{10}$$

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 (10) 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 3.10. *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 (10), 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 3.10 has the following consequence for complexity theory. A decision problem Π 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 Π 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 3.11. *If* Π *is a decision problem in* BPP *then* Π *belongs to* P/poly.

Proof. Let A be a randomized polynomial-time algorithm for Π that satisfies property (10). 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 3.10, 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 Π. □

A very natural and worthy goal is to eliminate the non-uniformity in Theorem 3.11 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.

4 Tail bounds for matrices

A recurring theme in the analysis of algorithms is that one is applying an algorithm to n random samples from a data distribution, and one wants to ensure that, with high probability, the output is the same (or similar) to what we would obtain if we could run the algorithm on the entire distribution. We already saw one example of this theme in Section 3.4.2, where the algorithm in question was simply picking out the hypothesis with the least average loss on a data set, from among a finite set of hypotheses. In this section we will analyze a more complicated example — using the top singular vector of a set of data samples to estimate the top singular vector of a covariance matrix — that requires using the exponential tail bounds for random scalars from Section 3 to derive an exponential tail bound for random matrices.

It takes quite a few steps to reach the main result in this section, so we begin by presenting a road map.

1. In Lemma 4.1 we construct a finite set of vectors on the (Euclidean) unit sphere in $\mathbb{R}^d$, such that every other vector on the sphere is close to at least one element of the set.

2. We use this finite set of vectors in Lemma 4.2 to prove an "enhanced union bound" that non-trivially bounds the probability of the union of infinitely many events, when each event refers to a particular inner product deviating significantly from its expected value.

3. In **??** we apply the enhanced union bound to derive an exponential tail bound for sums of random matrices, asserting that with high probability, when the random sum is multiplied by *any* unit vector, the product is close to its expected value.

4. If Section 4.1 we apply the matrix tail bound to analyze the distribution of the top singular vector of a set of random samples from a distribution.

Lemma 4.1. *Let $\mathbb{S}^{d-1}$ denote the unit sphere in $\mathbb{R}^d$, i.e. the set of all $\mathbf{x}$ such that $\|x\|_2 = 1$. For every positive integer d and every $0 < \gamma < \frac{1}{2}$ there is a set $C(d,\gamma) \subset \mathbb{S}^{d-1}$ with at most $(2e/\gamma)^d$ elements, such that for all $\mathbf{x} \in \mathbb{S}^{d-1}$ there exists $\mathbf{w} \in C(d,\gamma)$ with $\langle \mathbf{w}, \mathbf{x}\rangle > 1-\gamma$. Also, for all $\mathbf{x} \in \mathbb{S}^{d-1}$, the vector $(1-\gamma)\mathbf{x}$ is a convex combination of elements of $C(d,\gamma)$.*

Proof. Let $k = \lceil \frac{d}{2\gamma} \rceil$. For every d-tuple of non-negative integers $I = (i_1, i_2, \ldots, i_d)$ such that $i_1 + \cdots + i_d = k$, and every d-tuple of signs $S = (s_1, \ldots, s_d) \in \{\pm 1\}^d$, let

$$\mathbf{w}_{I,S} = k^{-1/2} \begin{bmatrix} s_1\sqrt{i_1} \\ s_2\sqrt{i_2} \\ \vdots \\ s_d\sqrt{i_d} \end{bmatrix}.$$

Let $C(d,\gamma)$ denote the set of all such vectors $\mathbf{w}_{I,S}$. By construction, the squared 2-norm of any such vector is $k^{-1}\sum_{j=1}^d s_j^2 i_j = k^{-1}(i_1 + \cdots + i_d) = 1$, so $C(d,\gamma) \subset \mathbb{S}^{d-1}$. For any $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{S}^{d-1}$, define d-tuples $I = (i_1, \ldots, i_d)$ and $S = (s_1, \ldots, s_d)$ by letting s_j be the sign of x_j, letting $i_j = \lfloor k x_j^2 \rfloor$ for $j < d$, and letting $i_d = k - (i_1 + \cdots + i_{d-1})$. We aim to prove that $\langle \mathbf{w}_{I,S}, \mathbf{x}\rangle > 1-\gamma$. By construction, for all j, $i_j > kx_j^2 - 1$, hence

$$\langle \mathbf{w}_{I,S}, \mathbf{x}\rangle = k^{-1/2}\sum_{j=1}^d s_j\sqrt{i_j}x_j = \sum_{j=1}^d \sqrt{\frac{i_j}{k}}|x_j| > \sum_{j=1}^d \left(\max\{x_j^2 - \frac{1}{k}, 0\}\right)^{1/2}|x_j|.$$

Using the identity

$$\left(|x_j| - \frac{1}{|x_j|k}\right)^2 = \left(x_j^2 - \frac{1}{k}\right)\left(1 - \frac{1}{x_j^2 k}\right),$$

we see that the right side is less than or equal to $x_j^2 - \frac{1}{k}$ when $x_j^2 - \frac{1}{k} \geq 0$, hence

$$\left(\max\{x_j^2 - \frac{1}{k}, 0\}\right)^{1/2} \geq \max\{|x_j| - \frac{1}{|x_j|k}, 0\}.$$

Let J denote the set of indices $j \in \{1, 2, \ldots, d\}$ such that $x_j^2 \geq \frac{1}{k}$. We find that

$$\langle \mathbf{w}_{I,S}, \mathbf{x}\rangle > \sum_{j=1}^d \left(\max\{x_j^2 - \frac{1}{k}, 0\}\right)^{1/2}|x_j| \geq \sum_{j\in J}\left(|x_j| - \frac{1}{|x_j|k}\right)|x_j| \geq \sum_{j\in J} x_j^2 - \frac{d}{k} = 1 - \sum_{j\notin J} x_j^2 - \frac{d}{k} > 1 - \frac{2d}{k},$$

because the inequality $x_j^2 < \frac{1}{k}$ for all $j \notin J$ implies $\sum_{j\notin J} x_j^2 < \frac{d}{k}$. By our choice of k, $\frac{2d}{k} \leq \gamma$, hence $\langle \mathbf{w}_{I,S}, \mathbf{x}\rangle > 1-\gamma$ as desired.

Next we reason about the cardinality of $C(d,\gamma)$. The number of d-tuples of non-negative integers that sum up to k is given by the binomial coefficient $\binom{d+k-1}{d}$, which is less than $\binom{2k}{d}$ since $k > d$ by

our assumption $\gamma < \frac{1}{2}$. Now using the identity[1] $\binom{m}{\ell} \le \left(\frac{em}{\ell}\right)^{\ell}$, we find that

$$|C(d,\gamma)| \le 2^d \binom{2k}{d} \le (4ek/d)^d = (2e/\gamma)^d.$$

Finally, we must prove that for every $\mathbf{x} \in \mathbb{S}^{d-1}$, $(1-\gamma)\mathbf{x}$ is a convex combination of elements of $C(d,\gamma)$. Let K denote the set of all convex combinations of elements of $C(d,\gamma)$, and assume, by way of contradiction, that $(1-\gamma)\mathbf{x} \notin K$. Since K is a closed, convex subset of $\mathbb{R}^d$, the intersection of all halfspaces containing K is equal to K. Hence, for some non-zero $\mathbf{a} \in \mathbb{R}^d$ and some scalar $b \in \mathbb{R}$, the halfspace $\mathcal{H} = \{\mathbf{z} \mid \langle \mathbf{a}, \mathbf{z} \rangle \le b\}$ contains every point of K but does not contain $(1-\gamma)\mathbf{x}$. Scaling $\mathbf{a}$ and b by the same positive scale factor, if necessary, we can assume $\|\mathbf{a}\|_2 = 1$. Then, we know from earlier in the proof that there must exist some $\mathbf{w} \in C(d,\gamma)$ such that $\langle \mathbf{a}, \mathbf{w} \rangle > 1-\gamma$. As $\mathbf{w} \in K$, this implies $1-\gamma < b$. Now the Cauchy-Schwartz inequality implies

$$\langle \mathbf{a}, (1-\gamma)\mathbf{x} \rangle \le (1-\gamma)\|\mathbf{a}\|_2 \|\mathbf{x}\|_2 = 1-\gamma < b,$$

contradicting our assumption that $(1-\gamma)\mathbf{x} \notin \mathcal{H}$. □

As explained earlier, Lemma 4.1 can be used to prove an "enhanced union bound" that allows us to show that the probability of a union of infinitely many events (one for each vector in $\mathbb{R}^d$) is small, if we can show that the probability of each individual one of the events is small. The enhanced union bound will be stated in terms of a multiplicative notion of approximation, called relative error, defined as follows.

Definition 4.1. If $y, \hat{y} \in \mathbb{R}$, and $y \neq 0$, the *relative error* of $\hat{y}$ approximating y is

$$\eta(\hat{y}, y) = \left| \frac{\hat{y}}{y} - 1 \right|.$$

If $y = 0$, then $\eta(\hat{y}, y) = 0$ if $\hat{y} = 0$ and otherwise $\eta(\hat{y}, y)$ is infinite.

In other words, $\eta(\hat{y}, y) < \varepsilon$ when $\hat{y}$ is between $(1-\varepsilon)y$ and $(1+\varepsilon)y$. As an example, the Chernoff bound shows that when $X_1, \ldots, X_n$ are independent random variables taking values in $[0,1]$ and $X = X_1 + \cdots + X_n$, then for any $\varepsilon > 0$ the relation $\eta(X, x) \le \varepsilon$ holds with probability at least $1 - e^{-\frac{1}{3}\varepsilon^2 \mathbb{E}[X]}$.

Lemma 4.2 (Enhanced Union Bound)**.** *Suppose P is a random symmetric positive semidefinite matrix in $\mathbb{R}^d$ and $Q = \mathbb{E}[P]$. Then for any $0 < \beta < 1$ and $0 < \gamma < 1$,*

$$\Pr\left(\sup_{\mathbf{x} \in \mathbb{R}^d} \eta(\langle \mathbf{x}, P\mathbf{x} \rangle, \langle \mathbf{x}, Q\mathbf{x} \rangle) \ge \beta \right) \le \left(\frac{2e}{\gamma}\right)^{2d} \sup_{\mathbf{x} \in \mathbb{R}^d} \Pr\left(\eta(\langle \mathbf{x}, P\mathbf{x} \rangle, \langle \mathbf{x}, Q\mathbf{x} \rangle) \ge \tfrac{1}{2}(1-\gamma)^2\beta \right). \tag{11}$$

[1] The inequality $\binom{m}{\ell} \le \left(\frac{em}{\ell}\right)^{\ell}$ follows because $\int_1^{\ell} \ln(x)\,dx \le \sum_{j=2}^{\ell} \ln(j)$ implies $(\ell/e)^{\ell} \le \ell!$, which in turn implies $\binom{m}{\ell} \le \frac{m^{\ell}}{\ell!} \le \left(\frac{em}{\ell}\right)^{\ell}$.

Before proving the lemma, we pause to interpret Inequality (11). For any fixed $\mathbf{x} \in \mathbb{R}^d$, the quantity $\langle \mathbf{x}, P\mathbf{x} \rangle$ is a random scalar, and $\langle \mathbf{x}, Q\mathbf{x} \rangle$ is its expected value. The relation $\eta(\langle \mathbf{x}, P\mathbf{x} \rangle, \langle \mathbf{x}, Q\mathbf{x} \rangle) \geq \beta$ means that $\langle \mathbf{x}, P\mathbf{x} \rangle$ deviates from its expected value by a factor at least $1 + \beta$, or at most $1 - \beta$. This is exactly the type of event that is shown to be unlikely by the Chernoff Bound. In fact, in the applications we will be looking at, P will be a sum of independent random matrices, so for any fixed $\mathbf{x} \in \mathbb{R}^d$, the random variable $\langle \mathbf{x}, P\mathbf{x} \rangle$ is a sum of independent, scalar-valued random variables, and the probability of $\eta(\langle \mathbf{x}, P\mathbf{x} \rangle, \langle \mathbf{x}, Q\mathbf{x} \rangle) \geq \beta$ will be analyzed using Chernoff bounds and their generalizations.

Now, let's think about the two sides of Inequality (11). They differ in three ways.

1. On the left, the supremum is inside the probability. On the right, it is outside. In other words, to say that the right side is small means that *for every fixed vector* $\mathbf{x}$*, the probability that* $\langle \mathbf{x}, P\mathbf{x} \rangle$ *approximates* $\langle \mathbf{x}, Q\mathbf{x} \rangle$ *is high*, whereas saying that the left side is small means that *with high probability, for all* $\mathbf{x}$ *the quantity* $\langle \mathbf{x}, P\mathbf{x} \rangle$ *approximates* $\langle \mathbf{x}, Q\mathbf{x} \rangle$. The two statements sound very similar but are actually quite different. For example, if D is a random date on the calendar, then *for every fixed human being, the probability that their birthday differs from* D *is quite high.* However, the probability that *every human being's birthday differs from* D is zero: no matter what date we randomly sample, there are quite a few human beings with that birthday. So, to prove that the right side being small implies that the left side is small, we will actually need to use some facts about vectors in $\mathbb{R}^d$, it doesn't just follow automatically from the laws of logic and probability.

2. On the right, the relative error is $(1 - \gamma)^2\beta$ rather than β. To make the numbers more concrete, if we want to conclude that with high probability, for all $\mathbf{x}$ the quantity $\langle \mathbf{x}, P\mathbf{x} \rangle$ is a 20% relative error approximation to $\langle \mathbf{x}, Q\mathbf{x} \rangle$, we will need to assume that for every $\mathbf{x}$, with high probability $\langle \mathbf{x}, P\mathbf{x} \rangle$ is a 5% relative error approximation to $\langle \mathbf{x}, Q\mathbf{x} \rangle$. (In this example, we used $\beta = 0.2$ and $\gamma = 0.5$.)

3. On the right side, there is an extra factor of $(2e/\gamma)^{2d}$. As you might have guessed, this comes from taking a union bound over pairs of elements of $C(d, \gamma)$. In order for the enhanced union bound to be meaningful, the right side needs to be less than 1. Therefore, the bound is only meaningful when $\Pr(\eta(\langle \mathbf{x}, P\mathbf{x} \rangle, \langle \mathbf{x}, Q\mathbf{x} \rangle) \geq (1 - \gamma)^2\beta)$ is exponentially small in d. Fortunately, exponential tail bounds such as the Chernoff bound are strong enough to enable such a conclusion.

Proof of Lemma 4.2. We first prove the lemma in the case when $Q = \mathbb{1}$. Later we will explain how the general case follows.

Let

$$C^{(2)}(d, \gamma) = C(d, \gamma) \cup \{\mathbf{w} \pm \mathbf{w}' \mid \mathbf{w}, \mathbf{w}' \in C(d, \gamma)\}.$$

Observe that if $C(d, \gamma)$ has N elements then $C^{(2)}(d, \gamma)$ has N^2 elements: the N elements of $C(d, \gamma)$ and the $2\binom{N}{2}$ elements of the form $\mathbf{w} \pm \mathbf{w}'$. Assuming $Q = \mathbb{1}$ we will show the implication

$$\exists \mathbf{x} \in \mathbb{R}^d \; \eta(\langle \mathbf{x}, P\mathbf{x} \rangle, \langle \mathbf{x}, Q\mathbf{x} \rangle) \geq \beta) \implies \exists \mathbf{w} \in C^{(2)}(d, \gamma) \; \eta(\langle \mathbf{w}, P\mathbf{w} \rangle, \langle \mathbf{w}, Q\mathbf{w} \rangle) \geq \tfrac{1}{2}(1 - \gamma)^2\beta). \quad (12)$$

The probability of the event on the left side equals the left side of Inequality (11), whereas the probability of the event on the right side is bounded above by the right side of Inequality (11) (by the Union Bound). Hence, Inequality (11) will be established once we show that (12) is valid.

If $\mathbf{x} \in \mathbb{R}^d$ satisfies $\eta(\langle \mathbf{x}, P\mathbf{x}\rangle, \langle \mathbf{x}, Q\mathbf{x}\rangle) \geq \beta)$ then $\mathbf{x} \neq \mathbf{0}$. Since the quantity $\eta(\langle \mathbf{x}, P\mathbf{x}\rangle, \langle \mathbf{x}, Q\mathbf{x}\rangle)$ is invariant to scaling $\mathbf{x}$, we may assume without loss of generality that $\|\mathbf{x}\|_2 = 1 - \gamma$. Then, $\mathbf{x}$ is a convex combination of elements of $C(d, \gamma)$. Say $\mathbf{x} = \sum_{i=1}^m a_i \mathbf{w}_i$ with $a_i \geq 0$, $\sum_{i=1}^m a_i = 1$, and $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\} \subseteq C(d, \gamma)$.

Since we are assuming $Q = \mathbb{1}$ for now, the inequality $\eta(\langle \mathbf{x}, P\mathbf{x}\rangle, \langle \mathbf{x}, Q\mathbf{x}\rangle) \geq \beta)$ can be rewritten as $|\langle \mathbf{x}, (P - \mathbb{1})\mathbf{x}\rangle - \|\mathbf{x}\|_2^2| \geq \beta\|\mathbf{x}\|_2^2 = (1-\gamma)^2\beta$. Using $\mathbf{x} = \sum_{i=1}^m a_i \mathbf{w}_i$ we rewrite this as

$$\sum_{i=1}^m \sum_{j=1}^m a_i a_j \left\langle \mathbf{w}_i, (P - \mathbb{1})\mathbf{w}_j \right\rangle \geq (1-\gamma)^2\beta. \tag{13}$$

The coefficients $a_i a_j$ are non-negative, and $\sum_{i=1}^m \sum_{j=1}^m a_i a_j = 1$, so the left side of Inequality (13) is a weighted average of the inner products $\left\langle \mathbf{w}_i, (P - \mathbb{1})\mathbf{w}_j \right\rangle$. Consequently at least one of those inner products is great than or equal to $(1-\gamma)^2\beta$. Now, we use the so-called *polarization identity*[2]

$$\langle \mathbf{y}, A\mathbf{z}\rangle = \frac{1}{4}\left[\langle \mathbf{y} + \mathbf{z}, A(\mathbf{y} + \mathbf{z})\rangle - \langle \mathbf{y} - \mathbf{z}, A(\mathbf{y} - \mathbf{z})\rangle\right]$$

with $\mathbf{y} = \mathbf{w}_i$, $\mathbf{z} = \mathbf{w}_j$, $A = P - \mathbb{1}$, to conclude that

$$\begin{aligned} 4(1-\gamma)^2\beta &\leq \left\langle \mathbf{w}_i + \mathbf{w}_j, (P - \mathbb{1})(\mathbf{w}_i + \mathbf{w}_j)\right\rangle - \left\langle \mathbf{w}_i - \mathbf{w}_j, (P - \mathbb{1})(\mathbf{w}_i - \mathbf{w}_j)\right\rangle \\ &\leq \left|\left\langle \mathbf{w}_i + \mathbf{w}_j, (P - \mathbb{1})(\mathbf{w}_i + \mathbf{w}_j)\right\rangle\right| + \left|\left\langle \mathbf{w}_i - \mathbf{w}_j, (P - \mathbb{1})(\mathbf{w}_i - \mathbf{w}_j)\right\rangle\right|, \end{aligned} \tag{14}$$

hence there exists $\mathbf{w} \in C^{(2)}(d, \gamma)$ such that $|\langle \mathbf{w}, (P - \mathbb{1})\mathbf{w}\rangle| > 2(1-\gamma)^2\beta \geq \frac{1}{2}(1-\gamma)^2\beta\langle \mathbf{w}, \mathbf{w}\rangle$. This completes the proof of the implication (12), which finishes the proof of the enhanced union bound in the case $Q = \mathbb{1}$.

For the general case, since Q is a symmetric, positive definite matrix it has an orthonormal basis of eigenvectors $\mathbf{b}_1, \ldots, \mathbf{b}_d$ with corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d = 0$. Let d' denote the number of positive eigenvalues, and let B be the $d' \times d$ matrix whose rows are the vectors $\sqrt{\lambda_i}\mathbf{b}_i^\top$. Observe that $Q = B^\top B$, or equivalently $Q = \sum_{i=1}^d \lambda_i \mathbf{b}_i \mathbf{b}_i^\top$. To confirm this equation, right-multiply both sides by the vector $\mathbf{b}_i$ and verify that the two sides are equal for $i = 1, 2, \ldots, d$.

Now let A denote the $d' \times d$ matrix whose rows are the vectors $\sqrt{\frac{1}{\lambda_i}}\mathbf{b}_i^\top$, and observe that $AB^\top$ is the $d' \times d'$ identity matrix. (This follows from the orthonormality of the basis $\{\mathbf{b}_1, \ldots, \mathbf{b}_d\}$.) Hence, if $P' = APA^\top$, then

$$\mathbb{E}[P'] = AQA^\top = (AB^\top)(BA^\top) = \mathbb{1}_{d'}.$$

Let V denote the nullspace of Q and let W denote the orthogonal complement of V. In other words, $V = \{\mathbf{v} \mid Q\mathbf{v} = \mathbf{0}\}$ and $W = \{\mathbf{w} \mid \langle \mathbf{w}, \mathbf{v}\rangle = 0 \quad \forall \mathbf{v} \in V\}$. For any vector $\mathbf{v} \in V$, we have

[2] The polarization identity is valid in any inner product space, when A is self-adjoint with respect to the inner product. $P - \mathbb{1}$ is self-adjoint with respect to the standard inner product on $\mathbb{R}^d$ because it is a symmetric matrix.

$\Pr(\langle \mathbf{v}, P\mathbf{v}\rangle \geq 0) = 1$, because P is supported on the set of positive semidefinite matrices, and $\mathbb{E}[\langle \mathbf{v}, P\mathbf{v}\rangle] = \langle \mathbf{v}, Q\mathbf{v}\rangle = 0$. A random variable that is supported on the non-negative real numbers and has expected value 0 must take the value 0 with probability 1, hence $\Pr(\langle \mathbf{v}, P\mathbf{v}\rangle = 0) = 1$. For a vector $\mathbf{v}$ and positive semidefinite matrix P, $\langle \mathbf{v}, P\mathbf{v}\rangle = 0$ implies $P\mathbf{v} = \mathbf{0}$, hence we have shown that for all $\mathbf{v} \in V$, $\Pr(P\mathbf{v} = \mathbf{0}) = 1$. Now if $\mathbf{x}$ is any vector in $\mathbb{R}^d$, we can write $\mathbf{x} = \mathbf{v} + \mathbf{w}$ where $\mathbf{v} \in V$ and $\mathbf{w} = W$. Then

$$\begin{aligned}
\langle \mathbf{x}, P\mathbf{x}\rangle &= \langle \mathbf{v}, P\mathbf{v}\rangle + 2\langle \mathbf{v}, P\mathbf{w}\rangle + \langle \mathbf{w}, P\mathbf{w}\rangle = \langle \mathbf{w}, P\mathbf{w}\rangle \\
\langle \mathbf{x}, Q\mathbf{x}\rangle &= \langle \mathbf{v}, Q\mathbf{v}\rangle + 2\langle \mathbf{v}, Q\mathbf{w}\rangle + \langle \mathbf{w}, Q\mathbf{w}\rangle = \langle \mathbf{w}, Q\mathbf{w}\rangle \\
\eta(\langle \mathbf{x}, P\mathbf{x}\rangle, \langle \mathbf{x}, Q\mathbf{x}\rangle) &= \eta(\langle \mathbf{w}, P\mathbf{w}\rangle, \langle \mathbf{w}, Q\mathbf{w}\rangle) \qquad (15)
\end{aligned}$$

The linear transformation $\mathbf{w} \mapsto B\mathbf{w}$ is an isomorphism from W to $\mathbb{R}^{d'}$ with inverse isomorphism $\mathbf{y} \mapsto A^\top \mathbf{y}$. Using these isomorphisms we find that

$$\begin{aligned}
&\Pr\left(\sup_{\mathbf{x}\in\mathbb{R}^d} \eta(\langle \mathbf{x}, P\mathbf{x}\rangle, \langle \mathbf{x}, Q\mathbf{x}\rangle) \geq \beta\right) \\
&= \Pr\left(\sup_{\mathbf{w}\in W} \eta(\langle \mathbf{w}, P\mathbf{w}\rangle, \langle \mathbf{w}, Q\mathbf{w}\rangle) \geq \beta\right) && \text{by Equation (15)} \\
&= \Pr\left(\sup_{\mathbf{y}\in\mathbb{R}^{d'}} \eta(\langle A^\top\mathbf{y}, PA^\top\mathbf{y}\rangle, \langle A^\top\mathbf{y}, QA^\top\mathbf{y}\rangle) \geq \beta\right) && \text{using isomorphism } \mathbf{y} \mapsto A^\top\mathbf{y} \\
&= \Pr\left(\sup_{\mathbf{y}\in\mathbb{R}^{d'}} \eta(\langle \mathbf{y}, P'\mathbf{y}\rangle, \langle \mathbf{y}, \mathbf{y}\rangle) \geq \beta\right) && \text{using } P' = APA^\top \text{ and } \mathbb{1}_{d'} = AQA^\top \\
&\leq \left(\frac{2e}{\gamma}\right)^{2d} \sup_{\mathbf{z}\in\mathbb{R}^{d'}} \Pr\left(\eta(\langle \mathbf{z}, P'\mathbf{z}\rangle, \langle \mathbf{z}, \mathbf{z}\rangle) \geq \tfrac{1}{2}(1-\gamma)^2\beta\right) && \text{from the } Q = \mathbb{1} \text{ case of the lemma} \\
&= \left(\frac{2e}{\gamma}\right)^{2d} \sup_{\mathbf{w}\in W} \Pr\left(\eta(\langle \mathbf{x}, P\mathbf{x}\rangle, \langle \mathbf{x}, Q\mathbf{x}\rangle) \geq \tfrac{1}{2}(1-\gamma)^2\beta\right) && \text{using isomorphism } \mathbf{z} \mapsto B\mathbf{z} \\
&= \left(\frac{2e}{\gamma}\right)^{2d} \sup_{\mathbf{x}\in\mathbb{R}^d} \Pr\left(\eta(\langle \mathbf{x}, P\mathbf{x}\rangle, \langle \mathbf{x}, Q\mathbf{x}\rangle) \geq \tfrac{1}{2}(1-\gamma)^2\beta\right) && \text{by Equation (15).}
\end{aligned}$$

□

Proposition 4.3. *Let P be a random symmetric positive semidefinite matrix in $\mathbb{R}^{d\times d}$, and let $Q = \mathbb{E}[P]$. Suppose there exist constants α_0, κ, n such that for all α in the interval $(0, \alpha_0)$ and all vectors $\mathbf{x} \in \mathbb{R}^d$, the random scalar quantity $\langle \mathbf{x}, P\mathbf{x}\rangle$ satisfies*

$$\Pr(\eta(\langle \mathbf{x}, P\mathbf{x}\rangle, \langle \mathbf{x}, Q\mathbf{x}\rangle) \geq \alpha) \leq 2e^{-\kappa\alpha^2 n}.$$

Then for all $\beta \in (0,\ 2\alpha_0)$,

$$\Pr\left(\sup_{\mathbf{x}\in\mathbb{R}^d} \eta(\langle \mathbf{x}, P\mathbf{x}\rangle, \langle \mathbf{x}, Q\mathbf{x}\rangle) \geq \beta\right) < 2\exp\left(7d - \frac{\kappa}{9}\beta^2 n\right).$$

Proof. Apply Lemma 4.2 with $\gamma = 2e^{-5/2}$, and let $\alpha = \frac{1}{2}(1-\gamma)^2\beta$. By our choice of γ and our assumption $\beta < 2\alpha_0$, we have $\alpha < \alpha_0$. Hence,

$$\sup_{\mathbf{x}\in\mathbb{R}^d} \Pr(\eta(\langle \mathbf{x}, P\mathbf{x}\rangle, \langle \mathbf{x}, Q\mathbf{x}\rangle) \geq \alpha) \leq e^{-\kappa\alpha^2 n} = 2e^{-\frac{\kappa}{4}(1-\gamma)^4\beta^2 n} < 2e^{-\frac{\kappa}{9}\beta^2 n}.$$

Applying Lemma 4.2,

$$\Pr\left(\sup_{\mathbf{x}\in\mathbb{R}^d} \eta(\langle \mathbf{x}, P\mathbf{x}\rangle)) \geq \beta\right) < 2\left(\frac{2e}{\gamma}\right)^{2d} e^{-\frac{\kappa}{9}\beta^2 n} = 2e^{7d}e^{-\frac{\kappa}{9}\beta^2 n},$$

which completes the proof. □

Remark 4.1. In applications of Proposition 4.3 it is often the case that P is a sum of independent random matrices, and that the parameter n is equal to, or is a function of, the number of independent summands. In these applications the parameter κ, on the other hand, is typically a universal constant that doesn't scale with n. This explains the (somewhat strange and arbitrary) decision to distinguish the constant κ from the parameter n in the statement and proof of Proposition 4.3, even though κ and n never appear separately but only in the form of the product κn. In other words, the decision to represent the number κn as a product of a constant κ and a parameter n is purely for mnemonic purposes, to make it easier to work out the implications of Proposition 4.3 in settings where there is a natural "problem size" parameter, n, representing the number of data points, vertices of a graph, or some such quantity.

4.1 Analyzing the SVD on random samples

As a first application of Proposition 4.3, we analyze the sample complexity of estimating the top singular vector of a Gaussian covariance matrix, given independent random samples from the distribution.

Suppose $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^d$ are independent random samples from the distribution $\mathcal{N}(\mathbf{0}, BB^\top)$. Let $A^\top$ denote a $n \times d$ matrix whose n rows are the vectors $\mathbf{a}_1^\top, \ldots, \mathbf{a}_n^\top$, and let $\hat{v}$ denote the top right singular vector of $A^\top$. We wish to understand how large n must be, as a function of d, ε, δ, so that

$$\Pr\left(\langle \hat{\mathbf{v}}, \mathbf{v}_1\rangle \geq 1 - \varepsilon\right) \geq 1 - \delta,$$

where $\mathbf{v}_1$ is the top right singular vector of $B^\top$.

To see how this problem relates to matrix tail bounds such as Proposition 4.3, observe that

$$\mathbb{E}[AA^\top] = \sum_{i=1}^{n} \mathbb{E}\left[\mathbf{a}_i.\mathbf{a}_i^\top\right] = nBB^\top,$$

so if we define $P_i = \mathbf{a}_i\mathbf{a}_i^\top$, $P = P_1 + \cdots + P_n = AA^\top$, and $Q = \mathbb{E}[P] = nBB^\top$, we see that Proposition 4.3 asserts that for n large enough, with high probability $AA^\top$ is a good approximation to $nBB^\top$, in the sense that for all $\mathbf{x} \in \mathbb{R}^d$, the inner product $\langle \mathbf{x}, AA^\top\mathbf{x}\rangle = \|A^\top\mathbf{x}\|_2^2$ approximates the inner product $\langle \mathbf{x}, nBB^\top\mathbf{x}\rangle = n\|B^\top\mathbf{x}\|_2^2$ with small relative error. Since $\hat{\mathbf{v}}$ is the unit vector that

maximizes $\|A^\top \mathbf{x}\|_2^2$ and $\mathbf{v}_1$ is the unit vector that maximizes $n\|B^\top \mathbf{x}\|_2^2$, if $\|A^\top \mathbf{x}\|_2^2 \approx n\|B^\top \mathbf{x}\|_2^2$ for all $\mathbf{x}$, then it should be possible to prove that the $\hat{\mathbf{v}}$ is a good approximation to $\mathbf{v}_1$.

Before we can apply Proposition 4.3 we need to be able to show that for all $\mathbf{x} \in \mathbb{R}^d$, the random variable $\langle \mathbf{x}, P\mathbf{x} \rangle$ approximates its expected value with small relative error. This would be easy to do, using the Chernoff bound, if the variable $\langle \mathbf{x}, P\mathbf{x} \rangle$ were a sum of independent $[0,1]$-valued random variables, but it's not. Instead

$$\langle \mathbf{x}, P\mathbf{x} \rangle = \sum_{i=1}^{n} \langle \mathbf{x}, P_i \mathbf{x} \rangle = \sum_{i=1}^{n} \langle \mathbf{x}, \mathbf{a}_i \mathbf{a}_i^\top \mathbf{x} \rangle = \sum_{i=1}^{n} \langle \mathbf{x}, \mathbf{a}_i \rangle^2 ,$$

and each summand $\langle \mathbf{x}, \mathbf{a}_i \rangle$ a Gaussian random variable with distribution $\mathcal{N}(0, \|B^\top \mathbf{x}\|_2^2)$. Thus, we need a tail bound for sums of squares of Gaussian random variables that is qualitatively similar to the Chernoff bound. The following lemma serves that purpose.

Lemma 4.4. *If $X_1, X_2, \ldots, X_n$ are independent, identically distributed Gaussian random variables, each with expected value zero, and $Y = X_1^2 + \cdots + X_n^2$, then for $0 < \alpha < 1$ we have*

$$\Pr\left(\eta(Y, \mathbb{E}Y) > \alpha\right) < 2e^{-\frac{1}{8}\alpha^2 n}.$$

Proof. For any $\lambda > 0$, the distribution of $\eta(Y, \mathbb{E}Y)$ is uneaffected if we replace Y with λY. Hence, we may assume without loss of generality that each X_i has variance 1. Letting $Y_i = X_i^2$, the cumulant generating function of Y_i is

$$K(t) = \ln \mathbb{E}\left[e^{tY_i}\right] = \ln \mathbb{E}\left[e^{tX_i^2}\right]. \tag{16}$$

To compute the expected value on the right side we can directly evaluate the integral.

$$\begin{aligned}
\mathbb{E}\left[e^{tX_i^2}\right] &= \sqrt{\frac{1}{2\pi}} \int_{-\infty}^{\infty} e^{tx^2 - \frac{1}{2}x^2}\, dx \\
&= \sqrt{\frac{1}{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(1-2t)x^2}\, dx \\
&= \sqrt{\frac{1}{1-2t}} \cdot \sqrt{\frac{1}{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2}\, du \\
&= \sqrt{\frac{1}{1-2t}},
\end{aligned}$$

for $t < \frac{1}{2}$. (Here, we used the substitution $u = \sqrt{1-2t}x$ to evaluate the integral, and the assumption $t < \frac{1}{2}$ was necessary in order for the factor $\sqrt{1-2t}$ to be a positive real number.) Substituting this formula for the expected value into Equation (16), we obtain the formula

$$K(t) = \frac{1}{2} \ln\left(\frac{1}{1-2t}\right), \tag{17}$$

which is valid for $t < \frac{1}{2}$. Now, since Y is the sum of n independent random variables each with cumulant generating function $K(t) = \frac{1}{2}\ln\left(\frac{1}{1-2t}\right)$, the cumulant generating function of Y is $K_Y(t) =$

$\frac{n}{2}\ln\left(\frac{1}{1-2t}\right)$. Using Lemma 3.3, we have

$$\Pr(Y \geq (1+\alpha)\mathbb{E}[Y]) \leq \exp\left(\frac{n}{2}\ln\left(\frac{1}{1-2t}\right) - t(1+\alpha)\mathbb{E}[Y]\right) = \exp\left(\frac{n}{2}\ln\left(\frac{1}{1-2t}\right) - (1+\alpha)tn\right). \quad (18)$$

The exponent on the right is minimized when $t = \frac{\alpha}{2(1+\alpha)}$ and $\frac{1}{1-2t} = 1+\alpha$. Then

$$\Pr(Y \geq (1+\alpha)\mathbb{E}[Y]) \leq \exp\left(\frac{n}{2}(\ln(1+\alpha) - \alpha)\right) < \exp\left(-\frac{1}{8}\alpha^2 n\right), \quad (19)$$

where we have used the inequality $\alpha - \ln(1+\alpha) > \frac{1}{4}\alpha^2$, which is valid for $0 < \alpha < 1$. (The validity can be confirmed by verifying that the left and right sides are equal when $\alpha = 0$, and the derivative of the left side exceeds the derivative of the right side when $0 < \alpha < 1$.)

A second application of Lemma 3.3 with $t = \frac{\alpha}{2(1-\alpha)}$ and $\frac{1}{1+2t} = 1-\alpha$ implies

$$\Pr(Y \leq (1-\alpha)\mathbb{E}[Y]) \leq \exp\left(\frac{n}{2}\ln\left(\frac{1}{1+2t}\right) + (1-\alpha)tn\right) \leq \exp\left(\frac{n}{2}(\alpha + \ln(1-\alpha))\right) < \exp\left(-\frac{1}{4}\alpha^2 n\right),$$

where we have used the inequality $-\alpha - \ln(1-\alpha) > \frac{1}{2}\alpha^2$, which is valid for $0 < \alpha < 1$. (Again, the validity can be confirmed by verifying that the left and right sides are equal when $\alpha = 0$, and the derivative of the left side exceeds the derivative of the right side when $0 < \alpha < 1$.) □

Corollary 4.5. *If $X_1, X_2, \ldots, X_n$ are independent, identically distributed Gaussian random variables, each with expected value zero, and $Y = X_1^2 + \cdots + X_n^2$, then for $0 < \varepsilon < 1$ we have*

$$\Pr\left(Y \leq (1-\varepsilon)^2\mathbb{E}Y\right) \leq e^{-\frac{1}{2}\varepsilon^2 n}, \qquad \Pr\left(Y \geq (1+\varepsilon)^2\mathbb{E}Y\right) \leq e^{-\frac{1}{2}\varepsilon^2 n}.$$

Proof. Using Inequality (19) from the proof of Lemma 4.4, with $\alpha = 2\varepsilon + \varepsilon^2$, we find that

$$\Pr\left(Y \geq (1+\varepsilon)^2\mathbb{E}Y\right) = \Pr\left(Y \geq (1+\alpha)\mathbb{E}Y\right) \leq \exp\left(\frac{n}{2}(\ln(1+\alpha) - \alpha)\right).$$

The inequality $\ln(1+\varepsilon) \leq \varepsilon$ along with the equation $1+\alpha = (1+\varepsilon)^2$ imply

$$\ln(1+\alpha) - \alpha = 2\ln(1+\varepsilon) - 2\varepsilon - \varepsilon^2 = 2(\ln(1+\varepsilon) - \varepsilon) - \varepsilon^2 \leq -\varepsilon^2,$$

so $\Pr(Y \geq (1+\varepsilon)^2\mathbb{E}Y) \leq \exp(-\frac{1}{2}\varepsilon^2 n)$ as claimed.

Using Inequality (20) from the proof of Lemma 4.4, with $\alpha = 2\varepsilon - \varepsilon^2$, we find that

$$\Pr\left(Y \geq (1-\varepsilon)^2\mathbb{E}Y\right) = \Pr\left(Y \geq (1-\alpha)\mathbb{E}Y\right) \leq \exp\left(\frac{n}{2}(\alpha + \ln(1-\alpha))\right).$$

The inequality $\ln(1-\varepsilon) \leq -\varepsilon$ along with the equation $1-\alpha = (1-\varepsilon)^2$ imply

$$\alpha + \ln(1-\alpha) = 2\varepsilon - \varepsilon^2 + 2\ln(1-\varepsilon) = 2(\varepsilon + \ln(1-\varepsilon)) - \varepsilon^2 \leq -\varepsilon^2,$$

so $\Pr(Y \leq (1-\varepsilon)^2\mathbb{E}Y) \leq \exp(-\frac{1}{2}\varepsilon^2 n)$ as claimed. □

We now have all the ingredients in place to analyze the sample complexity of estimating the top singular vector of a Gaussian covariance matrix using SVD.

Proposition 4.6. *For a $d \times d$ matrix B with singular values $\sigma_1 > \sigma_2 \geq \sigma_3 \geq \sigma_d$, suppose that A is a $d \times n$ matrix whose columns are independent samples from $\mathcal{N}(\mathbf{0}, BB^\top)$. Let $\mathbf{v}_1$ denote the top right singular vector of $B^\top$, and let $\hat{\mathbf{v}}$ denote the top right singular vector of $A^\top$. If $0 < \varepsilon < 0.3$, $\delta > 0$, and $n > \frac{100}{\varepsilon^2}\left(1 - \frac{\sigma_2^2}{\sigma_1^2}\right)^{-2}(7d + \ln(2/\delta))$ then*

$$\Pr\left(|\langle \hat{\mathbf{v}}, \mathbf{v}_1 \rangle| \geq 1 - \varepsilon\right) \geq 1 - \delta.$$

Proof. Let $\mathbf{a}_1, \ldots, \mathbf{a}_n$ denote the columns of the matrix A. For any $\mathbf{x} \in \mathbb{R}^d$ we have

$$\left\langle \mathbf{x}, AA^\top \mathbf{x} \right\rangle = \sum_{i=1}^{n} \left\langle \mathbf{x}, \mathbf{a}_i \right\rangle^2,$$

and each of the inner products $\langle \mathbf{x}, \mathbf{a}_i \rangle$ is independently sampled from a Gaussian distribution with mean 0 and variance $\|B^\top \mathbf{x}\|_2^2$. According to Lemma 4.4, for any $0 < \alpha < 1$,

$$\Pr\left(\eta(\left\langle \mathbf{x}, AA^\top \mathbf{x} \right\rangle, \left\langle \mathbf{x}, nBB^\top \mathbf{x} \right\rangle) \geq \alpha\right) < 2e^{-\frac{1}{8}\alpha^2 n}.$$

Now applying Proposition 4.3 with $\kappa = \frac{1}{8}$, we find that for any $0 < \beta < 1$,

$$\Pr\left(\sup_{\mathbf{x} \in \mathbb{R}^d} \eta(\left\langle \mathbf{x}, AA^\top \mathbf{x} \right\rangle, \left\langle \mathbf{x}, nBB^\top \mathbf{x} \right\rangle) \geq \beta\right) < 2e^{7d - \frac{\beta^2}{72}n}. \tag{20}$$

Let $\beta = \left(1 - \sigma_2^2/\sigma_1^2\right)\left(\varepsilon - \frac{1}{2}\varepsilon^2\right)$. Since $\varepsilon < 0.3$ we have

$$\begin{aligned}
\beta^2 &= \left(1 - \frac{\sigma_2^2}{\sigma_1^2}\right)^2 \varepsilon^2 \left(1 - \frac{1}{2}\varepsilon\right)^2 > \left(1 - \frac{\sigma_2^2}{\sigma_1^2}\right)^2 \varepsilon^2 \cdot (0.85)^2 \\
\beta^2 n &> 100 \cdot (0.85)^2 \cdot (7d + \ln(1/\delta)) > 72(7d + \ln(2/\delta)) \\
7d - \frac{\beta^2}{72} n &< -\ln(2/\delta) = \ln(\delta/2)
\end{aligned}$$

hence the right side of Inequality (20) is less than δ.

We have shown that with probability at least $1 - \delta$, it holds that

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \eta(\left\langle \mathbf{x}, AA^\top \mathbf{x} \right\rangle, \left\langle \mathbf{x}, nBB^\top \mathbf{x} \right\rangle) < \beta. \tag{21}$$

To conclude the proof we need to show that this property implies $|\langle \hat{\mathbf{v}}, \mathbf{v}_1 \rangle| \geq 1 - \varepsilon$. By the definition of the SVD, $\hat{\mathbf{v}}$ is the unit vector that maximizes $\left\langle \mathbf{v}, AA^\top \mathbf{v} \right\rangle$. In particular,

$$\left\langle \mathbf{v}_1, AA^\top \mathbf{v}_1 \right\rangle \leq \left\langle \mathbf{v}, AA^\top \mathbf{v} \right\rangle. \tag{22}$$

Now let $\hat{v} = a\mathbf{v}_1 + b\mathbf{w}$, where $\mathbf{w}$ is a unit vector orthogonal to $\mathbf{v}_1$, and $a^2 + b^2 = 1$. Recall from the definition of singular values that for all unit vectors $\mathbf{w}$ orthogonal to $\mathbf{v}_1$, $\|B^\top \mathbf{w}\|_2 \leq \sigma_2$, which implies $\langle \mathbf{w}, BB^\top \mathbf{w} \rangle \leq \sigma_2^2$. Applying Inequality (21), we find

$$\langle \mathbf{v}_1, AA^\top \mathbf{v}_1 \rangle \geq (1-\beta) \langle \mathbf{v}_1, nBB^\top \mathbf{v}_1 \rangle = (1-\beta) n\sigma_1^2 \tag{23}$$

$$\begin{aligned} \langle \hat{\mathbf{v}}, AA^\top \hat{\mathbf{v}} \rangle &\leq (1+\beta) \left(\langle a\mathbf{v}_1 + b\mathbf{w}, nBB^\top (a\mathbf{v}_1 + b\mathbf{w}) \rangle \right) \\ &= (1+\beta) n (a^2\sigma_1^2 + b^2\sigma_2^2) \\ &= (1+\beta) n\sigma_1^2 \left(1 - b^2 \left(1 - \frac{\sigma_2^2}{\sigma_1^2} \right) \right). \end{aligned} \tag{24}$$

Combining Inequalities (21), (23), (24), we find that

$$\begin{aligned} (1-\beta) n\sigma_1^2 &\leq (1+\beta) n\sigma_1^2 \left(1 - b^2 \left(1 - \frac{\sigma_2^2}{\sigma_1^2} \right) \right) \\ \frac{1-\beta}{1+\beta} &\leq 1 - b^2 \left(1 - \frac{\sigma_2^2}{\sigma_1^2} \right) \\ \frac{2\beta}{1+\beta} &\geq b^2 \left(1 - \frac{\sigma_2^2}{\sigma_1^2} \right) \\ 2\beta \left(1 - \frac{\sigma_2^2}{\sigma_1^2} \right)^{-1} &\geq b^2 (1+\beta) \geq b^2 \\ 2\varepsilon - \varepsilon^2 &\geq b^2 \\ 1 - (2\varepsilon - \varepsilon^2) &\leq 1 - b^2 = a^2. \end{aligned}$$

Since the left side of the last line is $(1-\varepsilon)^2$, we have shown that $(1-\varepsilon)^2 \leq a^2$, or $|a| \geq 1 - \varepsilon$. Recalling that $\hat{\mathbf{v}} = a\mathbf{v}_1 + b\mathbf{w}$ with $\langle \mathbf{w}, \mathbf{v}_1 \rangle = 0$, we have

$$|\langle \hat{\mathbf{v}}, \mathbf{v}_1 \rangle| = |a \langle \mathbf{v}_1, \mathbf{v}_1 \rangle + b \langle \mathbf{w}, \mathbf{v}_1 \rangle| = |a| \geq 1 - \varepsilon,$$

as desired. □

4.2 The Ahlswede-Winter Inequality

In this section we present a stronger exponential tail bound for random matrices, in which the probability of violating the bound has only polynomial dependence on the dimension, rather than the exponential dependence in Proposition 4.3. Unfortunately the proof of the inequality is beyond the scope of these notes.

Theorem 4.7 (Ahlswede-Winter Inequality). *Suppose $P_1, X_2, \ldots, P_m$ are mutually independent random, symmetric, positive semidefinite $d \times d$ matrices, let $P = P_1 + \cdots + P_m$, and let $Q = \mathbb{E}[P]$. If $r > 0$ is a scalar such that for all i and all $\mathbf{x} \in \mathbb{R}^d$, $\langle \mathbf{x}, P_i \mathbf{x} \rangle \leq \frac{1}{r} \langle \mathbf{x}, Q\mathbf{x} \rangle$ with probability 1, then for all $\beta \in (0,1)$,*

$$\Pr\left(\sup_{\mathbf{x} \in \mathbb{R}^d} \eta(\langle \mathbf{x}, P\mathbf{x} \rangle, \langle \mathbf{x}, Q\mathbf{x} \rangle) \geq \beta \right) \leq 2d \cdot e^{-\frac{1}{4}\beta^2 r}. \tag{25}$$

To compare the Ahlswede-Winter Inequality with Proposition 4.3, it is useful to note that the assumption $\langle \mathbf{x}, P_i \mathbf{x} \rangle \leq \frac{1}{r} \langle \mathbf{x}, Q\mathbf{x} \rangle$ implies, using the Chernoff Bound with the $[0,1]$-valued random variables $Y_i = r\frac{\langle \mathbf{x}, P_i \mathbf{x} \rangle}{\langle \mathbf{x}, Q\mathbf{x} \rangle}$, that for all $\mathbf{x} \in \mathbb{R}^d$ and all $0 < \alpha < 1$,

$$\Pr\left(\eta(\langle \mathbf{x}, P\mathbf{x} \rangle, \langle \mathbf{x}, Q\mathbf{x} \rangle) \geq \alpha\right) \leq 2e^{-\frac{1}{3}\alpha^2 r} = 2e^{-\kappa\alpha^2 n}$$

where $\kappa = r/(3n)$. Hence, we can conclude using Proposition 4.3 that

$$\Pr\left(\sup_{\mathbf{x} \in \mathbb{R}^d} \eta(\langle \mathbf{x}, P\mathbf{x} \rangle, \langle \mathbf{x}, Q\mathbf{x} \rangle) \geq \beta\right) \leq 2e^{7d} \cdot e^{-\frac{\kappa}{9}\beta^2 n} = 2e^{7d} \cdot e^{-\frac{1}{27}\beta^2 r}$$

The Ahlswede-Winter Inequality improves upon this bound in two respects: the constant $\frac{1}{27}$ is improved to $\frac{1}{4}$, but much more importantly the factor e^{7d} is improved to d. Recall that Proposition 4.3 is proven by taking the union bound over the exponentially number of vectors that constitute the set $C(d, 2e^{-5/2})$, and the factor $2e^{7d}$ on the right side of the inequality reflects the exponential number of "bad events" in this application of the union bound. Replacing this factor of $2e^{7d}$ with $2d$ on the right side of the Ahlswede-Winter Inequality means that the inequality is as efficient, quantitatively, as a union bound over only $2d$ "bad events." Conceptually, it is as if we only need to worry about $\eta(\langle \mathbf{x}, P\mathbf{x} \rangle, \langle \mathbf{x}, Q\mathbf{x} \rangle)$ being large as $\mathbf{x}$ ranges over a basis of $\mathbb{R}^d$. In actuality, the proof of the Ahlswede-Winter Inequality doesn't use the union bound at all, so this discussion is only meant to give some intuition about how to evaluate the power of the inequality in quantitative terms.

4.3 The Stochastic Block Model

As a second application of matrix tail bounds, we investigate the "community detection"problem in which one is given the adjacency matrix of an undirected graph (e.g., a social network) and one wants to find a labeling of the vertices such that edges occur more frequently between vertices that share a label than between differently-labeled vertices.

A frequently used model of community structure in networks is the *stochastic block model*, which defines a probability distribution over graphs with a given vertex set, V. In this model, each vertex u of the graph has a label σ_u and the probability that the graph contains an edge between u and v, conditional on the labeling $\boldsymbol{\sigma}$, is

$$p_{uv} = \begin{cases} p_{\text{in}} & \text{if } \sigma_u = \sigma_v \\ p_{\text{out}} & = \text{if } \sigma_u \neq \sigma_v. \end{cases}$$

Here, p_{in} and p_{out} are two parameters of the model satisfying $0 \leq p_{\text{out}} < p_{\text{in}} \leq 1$. The labeling $\boldsymbol{\sigma}$ may be modeled as random or non-random. In our model we will treat $\boldsymbol{\sigma}$ as a fixed, non-random labeling. However, the input to the community detection problem consists only of the vertices and edges of the graph; the labels of the vertices are not revealed in the input but must instead be (approximately) inferred from the given data.

In this section we will make two simplifying assumptions.

1. There are only two possible labels for a vertex: the labeling $\boldsymbol{\sigma}$ will be represented as a vector in $\{\pm 1\}^n$. Note that this means two vertices u, v have the same label or different labels according to whether $\sigma_u \sigma_v = 1$ or $\sigma_u \sigma_v = -1$.

2. The graph has an even number of vertices, n, and the labeling divides them into two groups of equal size: $\sum_{u \in V} \sigma_u = 0$.

Under these assumptions, we will be interested in the question: for what values of $n, p_{\text{in}}, p_{\text{out}}$ can we calculate a labeling $\hat{\boldsymbol{\sigma}}$ that is significantly more correlated with $\boldsymbol{\sigma}$ than a random labeling? One way of formalizing this objective is to insist that $|\langle \hat{\boldsymbol{\sigma}}, \boldsymbol{\sigma} \rangle| \geq (1 - 2\varepsilon)n$ for some specified $\varepsilon > 0$ When one finds a labeling that meets this objective, it means that either $\hat{\boldsymbol{\sigma}}$ or its negation, $-\hat{\boldsymbol{\sigma}}$, labels at least $1 - \varepsilon$ fraction of the vertices correctly.

We can begin to see how this subject relates to random matrices when we think about the adjacency matrix of the graph G. In the vector space $\mathbb{R}^V$ of real-valued functions on the vertices of G, there is a standard basis containing, for each vertex v, a vector $\mathbf{e}_v$ that represents the function taking the value 1 at v and 0 everywhere else. An edge $e = \{u, v\}$ in the graph is represented in the adjacency matrix by setting two entries to the value 1: one in row u and column v, another in row v and column u. In other words, the adjacency matrix of a graph G with vertex set V and edge set E is

$$A_G = \sum_{\{u,v\} \in E} \mathbf{e}_u \mathbf{e}_v^\top + \mathbf{e}_v \mathbf{e}_u^\top.$$

When edges of G are sampled independently at random (as in the stochastic block model, when the labeling $\boldsymbol{\sigma}$ is fixed) the individual terms of this sum become independent random matrices. However, we are not quite ready to apply Proposition 4.3 to reason about the random matrix A_G, since the summands $\mathbf{e}_u \mathbf{e}_v^\top + \mathbf{e}_v \mathbf{e}_u^\top$ are not positive definite matrices. To circumvent this difficulty, we will make use of a different sum of random matrices, where the terms are symmetric and positive semidefinite by construction.

$$L_G = \sum_{\{u,v\} \in E} (\mathbf{e}_u - \mathbf{e}_v)(\mathbf{e}_u - \mathbf{e}_v)^\top \tag{26}$$

The matrix L_G is called the *Laplacian matrix* of G. Its digaonal entries are the degrees of the vertices of G, because the matrix $\mathbf{e}_u \mathbf{e}_u^\top$ appears once in the sum for every edge $\{u, v\}$. Its off-diagonal entries match those of $-A_G$, the negation of the adjacency matrix. For the stochstic block model with labeling $\boldsymbol{\sigma}$ we can calculate $\mathbb{E}[L_G]$ quite easily. Let $p = \frac{1}{2}(p_{\text{in}} + p_{\text{out}})$ and $q = \frac{1}{2}(p_{\text{in}} - p_{\text{out}})$, so that $p_{\text{in}} = p + q$ while $p_{\text{out}} = p - q$. We find that

$$\mathbb{E}[L_G]_{u,v} = \begin{cases} \frac{n}{2}(p_{\text{in}} + p_{\text{out}}) - p_{\text{in}} & \text{if } u = v \\ -p_{\text{in}} & \text{if } u \neq v,\ \sigma_u = \sigma_v \\ -p_{\text{out}} & \text{if } \sigma_u \neq \sigma_v \end{cases} = pn\delta_{u,v} - p - q\sigma_u\sigma_v. \tag{27}$$

This formula for the entries of the expected Laplacian, $E[L_G]$, can be summarized more succinctly by the equation

$$\mathbb{E}[L_G] = pn\mathbb{1} - p\mathbf{1}\mathbf{1}^\top - q\boldsymbol{\sigma}\boldsymbol{\sigma}^\top. \tag{28}$$

From Equation (28) we can deduce[3] the eigenvectors and eigenvalues of the expected Laplacian matrix $\mathbb{E}[L_G]$. The smallest eigenvalue is 0, with eigenvector $\mathbf{1}$. The next-smallest eigenvalue is $(p-q)n$, with eigenvector $\boldsymbol{\sigma}$. The remaining eigenvalue is pn, with multiplicity $n-2$. The eigenspace corresponding to this eigenvalue is the $(n-2)$-dimensional space of vectors orthogonal to both $\mathbf{1}$ and $\boldsymbol{\sigma}$.

Now let's consider the random matrix L_G. The smallest eigenvalue is again 0, with eigenvector $\mathbf{1}$, because in the sum defining the Laplacian L_G, each of the summands $(\mathbf{e}_u - \mathbf{e}_v)(\mathbf{e}_u - \mathbf{e}_v)^\top$ has $\mathbf{1}$ in its nullspace. When $\mathbf{x}$ ranges over unit vectors orthogonal to $\mathbf{1}$, the expected value of the inner product $\langle \mathbf{x}, L_G\mathbf{x}\rangle$ is minimized when $\mathbf{x}$ is parallel to $\boldsymbol{\sigma}$, which for unit vectors $\mathbf{x}$ occurs when $\mathbf{x} = \pm\sqrt{1/n}\cdot\boldsymbol{\sigma}$. Hence, a natural plan for identifying the community structure in the network is to compute the eigenvector $\mathbf{y}$ of L_G corresponding to its second-smallest eigenvalue. This vector will be orthogonal to $\mathbf{1}$, and if the eigenvectors of L_G are close to those of $\mathbb{E}[L_G]$ then $\mathbf{y}$ should be close to $\boldsymbol{\sigma}$.

Algorithm 2 Stochastic block model algorithm

1: Compute the Laplacian matrix L_G and its eigenvectors.
2: Let $\mathbf{y}$ denote the eigenvector corresponding to the second-smallest eigenvalue of L_G.
3: Output the labeling $\hat{\boldsymbol{\sigma}}$ defined by setting $\hat{\sigma}_v = +1$ if $y_v \geq 0$, $\hat{\sigma}_v = -1$ if $y_v < 0$.

To analyze the correctness of the algorithm, the following lemma concerning symmetric matrices and their eigenspaces will be useful.

Lemma 4.8. *Suppose A is a symmetric matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$ and corresponding eigenspaces $V_1, V_2, \ldots, V_k$. The spaces $V_1, \ldots, V_k$ are pairwise orthogonal (meaning each vector in V_i is orthogonal to each vector in V_j when $1 \leq i < j \leq k$) and every $\mathbf{x} \in \mathbb{R}^n$ can be written uniquely in the form $\mathbf{x} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_k\mathbf{x}_k$, where $\mathbf{x}_i$ is an element of V_i satisfying $\|\mathbf{x}_i\|_2 = 1$, for $i = 1, \ldots, k$. Furthermore, $\|_\|\mathbf{x}_2^2 = a_1^2 + a_2^2 + \cdots + a_k^2$, and*

$$\langle \mathbf{x}, A\mathbf{x}\rangle = \lambda_1 a_1^2 + \lambda_2 a_2^2 + \cdots + \lambda_k a_k^2.$$

[3] The derivation of the eigenspaces and eigenvalues of $\mathbb{E}[L_G]$ is accomplished by repeatedly applying the following principle: *if A and B are two matrices and $\mathbf{v}$ is an eigenvector of both A and B, with eigenvalues λ_A and λ_B respectively, then $\mathbf{v}$ is also an eigenvector of $A + B$ with eigenvalue $\lambda_A + \lambda_B$.*

Applying this principle to $\mathbb{E}[L_G] = pn\mathbb{1} - p\mathbf{1}\mathbf{1}^\top - q\boldsymbol{\sigma}\boldsymbol{\sigma}^\top$, we can reason about the eigenvalues as follows. First, the matrix $A = pn\mathbb{1}$ is a scalar multiple of the identity matrix, so *every* vector is an eigenvector of A with eigenvalue pn. Next, the matrix $B = -p\mathbf{1}\mathbf{1}^\top$ is a symmetric rank-one matrix, so it has two eigenspaces: a one-dimensional eigenspace generated by $\mathbf{1}$, with eigenvalue $-pn$, and an $(n-1)$-dimensional eigenspace consisting of all the vectors orthogonal to $\mathbf{1}$, with eigenvalue 0. Finally, the matrix $C = -q\boldsymbol{\sigma}\boldsymbol{\sigma}^\top$ is also a symmetric rank-one matrix, so it also has two eigenspaces: a one-dimensional eigenspace generated by $\boldsymbol{\sigma}$, with eigenvalue $-qn$, and an $(n-1)$-dimensional eigenspace consisting of all the vectors orthogonal to $\boldsymbol{\sigma}$, with eigenvalue 0. Putting all this information together, we conclude that the matrix $\mathbb{E}[L_G] = A + B + C$ has a one-dimensional eigenspace generated by $\mathbf{1}$, with eigenvalue $pn - pn = 0$, another one-dimensional eigenspace generated by $\boldsymbol{\sigma}$, with eigenvalue $pn - qn = (p-q)n$, and finally an $(n-2)$-dimensional eigenspace consisting of the vectors orthogonal to both $\mathbf{1}$ and $\boldsymbol{\sigma}$, with eigenvalue pn. Note that in order for us to derive the eigenspaces and eigenvalues using this method, it was convenient that $\mathbf{1}$ was an eigenvector of C and that $\boldsymbol{\sigma}$ was an eigenvector of B; both of these convenient facts were true because of our assumption that the two communities are of equal size, i.e. $\langle \mathbf{1}, \boldsymbol{\sigma}\rangle = 0$. Without that assumption, the calculation of the eigenspaces and eigenvectors of $\mathbb{E}[L_G]$ would have been more involved.

Proof. The lemma is clearly true when A is a diagonal matrix. For the general case, we use the fact that every symmetric matrix can be written in the form $A = QDQ^\top$ where Q is orthogonal and D is diagonal. Letting $W_1, \ldots, W_k$ denote the eigenspaces of D, we find that A has eigenspaces $V_i = QW_i$ for $i = 1, \ldots, k$, with the same eigenvalues. These are pairwise orthogonal since left-multiplication by Q preserves orthogonality. Furthermore, if $\mathbf{x} = Q\mathbf{y}$ then $\langle \mathbf{x}, A\mathbf{x} \rangle = \langle Q\mathbf{y}, QD\mathbf{y} \rangle = \langle \mathbf{y}, D\mathbf{y} \rangle$ since Q preserves inner products. Hence, the formula for $\langle \mathbf{x}, A\mathbf{x} \rangle$ follows from the corresponding formula for $\langle \mathbf{y}, D\mathbf{y} \rangle$. □

Proposition 4.9. *If $n \in \mathbb{N}$ satisfies the inequality $n > \frac{32p^2}{\varepsilon^2 q^2 (p-q)} \ln(2n/\delta)$ then with probability at least $1 - \delta$, Algorithm 2 outputs a labeling $\hat{\boldsymbol{\sigma}}$ such that either $\hat{\boldsymbol{\sigma}}$ or $-\hat{\boldsymbol{\sigma}}$ labels at least $(1 - \varepsilon)n$ vertices correctly.*

Proof. The proof will be an application of Theorem 4.7, which means that we need to write L_G as a sum of independent random symmetric positive semidefinite matrices. Let W denote the set of unordered pairs of vertices $\{u, v\}$, and for any $\{u, v\} \in W$ let Y_{uv} equal 1 if $\{u, v\} \in E$, 0 if not. Then we can express L_G as

$$L_G = \sum_{\{u,v\} \in W} Y_{uv} (\mathbf{e}_u - \mathbf{e}_v)(\mathbf{e}_u - \mathbf{e}_v)^\top = \sum_{\{u,v\} \in W} P_{uv},$$

where the matrices $P_{uv} = Y_{uv}(\mathbf{e}_u - \mathbf{e}_v)(\mathbf{e}_u - \mathbf{e}_v)^\top$ are independent, random, symmetric, positive semidefinite matrices. We wish to apply the Ahlswede-Winter Inequality, with $P = \sum_{\{u,v\} \in W} P_{uv}$ and $Q = \mathbb{E}[P] = L_G$. To do so, we need to find a value of r such that for all $\mathbf{x} \in \mathbb{R}^d$ and all $\{u, v\} \in W$, $\langle \mathbf{x}, P_{uv}\mathbf{x} \rangle \leq \frac{1}{r} \langle \mathbf{x}, Q\mathbf{x} \rangle$. We will decompose $\mathbf{x}$ as a sum of eigenvectors of Q and apply Lemma 4.8. Let $\bar{\mathbf{1}} = \sqrt{1/n} \cdot \mathbf{1}$ and $\bar{\boldsymbol{\sigma}} = \sqrt{1/n} \cdot \boldsymbol{\sigma}$ denote the unit-length eigenvectors of Q corresponding to the eigenvalues 0 and $(p - q)n$, respectively. If we write $\mathbf{x} = a\bar{\mathbf{1}} + b\bar{\boldsymbol{\sigma}} + c\mathbf{w}$, where $\mathbf{w}$ is orthogonal to $\mathbf{1}$ and $\boldsymbol{\sigma}$ and $\|\mathbf{w}\|_2 = 1$, then from our calculation of the eigenvalues of Q we know that

$$\langle \mathbf{x}, Q\mathbf{x} \rangle = b^2(p - q)n + c^2 pn \geq (b^2 + c^2)(p - q)n.$$

Meanwhile,

$$\begin{aligned}
\langle \mathbf{x}, P_{uv}\mathbf{x} \rangle &= Y_{uv} \langle \mathbf{x}, (\mathbf{e}_u - \mathbf{e}_v)(\mathbf{e}_u - \mathbf{e}_v)^\top \mathbf{x} \rangle \\
&= Y_{uv} \langle \mathbf{x}, \mathbf{e}_u - \mathbf{e}_v \rangle^2 \\
&= Y_{uv} \left\langle a\bar{\mathbf{1}} + b\bar{\boldsymbol{\sigma}} + c\mathbf{w}, \mathbf{e}_u - \mathbf{e}_v \right\rangle^2 \\
&= Y_{uv} \langle b\bar{\boldsymbol{\sigma}} + c\mathbf{w}, \mathbf{e}_u - \mathbf{e}_v \rangle^2 \\
&\leq Y_{uv} \|b\bar{\boldsymbol{\sigma}} + c\mathbf{w}\|_2^2 \|\mathbf{e}_u - \mathbf{e}_v\|_2^2 = 2Y_{uv}(b^2 + c^2) \leq 2(b^2 + c^2).
\end{aligned}$$

Hence, if we set $r = \frac{(p-q)n}{2}$ then $\langle \mathbf{w}, P_{uv}\mathbf{w} \rangle \leq \frac{1}{r} \langle \mathbf{w}, Q\mathbf{w} \rangle$ is guaranteed to hold. Applying the Ahlswede-Winter Inequality, we conclude that for any $\beta > 0$,

$$\Pr\left(\sup_{\mathbf{x} \in \mathbb{R}^n} \eta(\langle \mathbf{x}, P\mathbf{x} \rangle, \langle \mathbf{x}, Q\mathbf{x} \rangle) \geq \beta \right) \leq 2n \cdot e^{-\frac{p-q}{8}\beta^2 n}.$$

Let $\beta = \frac{\varepsilon q/2}{p - q + \varepsilon q/2}$ and observe that

$$\frac{\beta}{1 - \beta} = \frac{\varepsilon q}{2(p - q)}, \tag{29}$$

an equation which will justify this choice of β when we use it later on. Suppose that

$$\sup_{\mathbf{x}\in\mathbb{R}^n} \eta(\langle \mathbf{x}, P\mathbf{x}\rangle, \langle \mathbf{x}, Q\mathbf{x}\rangle) \leq \beta. \tag{30}$$

We wish to prove that under this assumption, the relation $|\langle \mathbf{y}, \bar{\boldsymbol{\sigma}}\rangle|^2 \geq 1-\varepsilon$ holds. The assumption (30) holds with probability at least $1 - 2n \cdot e^{-(p-q)\beta^2 n/8}$, which is greater than $1-\delta$ when $n > \frac{32p^2}{\varepsilon^2 q^2 (p-q)} \ln(2n/\delta)$.

Since $\mathbf{y}$ is orthogonal to $\mathbf{1}$, it may be written as a sum of eigenvectors of Q belonging to the other two eigenspaces: $\mathbf{y} = a\bar{\boldsymbol{\sigma}} + b\mathbf{w}$, where $\|\mathbf{w}\|_2 = 1$ and $\langle \mathbf{1}, \mathbf{w}\rangle = \langle \boldsymbol{\sigma}, \mathbf{w}\rangle = 0$. Assuming without loss of generality that $\|\mathbf{y}\|_2^2 = 1$, we have $a^2 + b^2 = 1$. Now, by assumption (30), we have

$$\begin{aligned}
\langle \bar{\boldsymbol{\sigma}}, L_G \bar{\boldsymbol{\sigma}}\rangle &\leq (1+\beta)\langle \bar{\boldsymbol{\sigma}}, Q\bar{\boldsymbol{\sigma}}\rangle = (1+\beta)(p-q)n \\
\langle \mathbf{y}, L_G \mathbf{y}\rangle &\geq (1-\beta)\langle \mathbf{y}, Q\mathbf{y}\rangle = (1-\beta)\left[a^2\langle \boldsymbol{\sigma}, Q\boldsymbol{\sigma}\rangle + b^2\langle \mathbf{w}, Q\mathbf{w}\rangle\right] \\
&= (1-\beta)\left[a^2(p-q)n + b^2 pn\right] = (1-\beta)\left[(p-q)n + b^2 qn\right].
\end{aligned}$$

Since $\mathbf{y}$ was chosen to minimize $\langle \mathbf{y}, L_G\mathbf{y}\rangle$ among all vectors orthogonal to $\mathbf{1}$, and $\bar{\boldsymbol{\sigma}}$ is orthogonal to $\mathbf{1}$, it must be the case that $\langle \mathbf{y}, L_G\mathbf{y}\rangle \leq \langle \bar{\boldsymbol{\sigma}}, L_G\bar{\boldsymbol{\sigma}}\rangle$. Hence,

$$\begin{aligned}
(1-\beta)\left[(p-q)n + b^2 qn\right] &\leq (1+\beta)(p-q)n \\
\frac{(p-q)n + b^2 qn}{(p-q)n} &\leq \frac{1+\beta}{1-\beta} \\
1 + \frac{b^2 q}{p-q} \leq 1 + \frac{2\beta}{1-\beta} &= 1 + \frac{\varepsilon q}{p-q}
\end{aligned}$$

where we have used Equation (29) on the last line. It follows that $b^2 \leq \varepsilon$ and $a^2 \geq 1-\varepsilon$, hence

$$|\langle \mathbf{y}, \bar{\boldsymbol{\sigma}}\rangle| = |a| \geq \sqrt{1-\varepsilon}.$$

Our final task is to prove that $\pm\hat{\boldsymbol{\sigma}}$ labels at least $1-\varepsilon$ fraction of vertices correctly. Assume without loss of generality that $\langle \mathbf{y}, \bar{\boldsymbol{\sigma}}\rangle \geq \sqrt{1-\varepsilon}$; otherwise, replace $\mathbf{y}$ with $-\mathbf{y}$ and $\hat{\sigma}$ with $-\hat{\sigma}$. Now let $\mathbf{z}$ denote the vector defined by

$$z_u = \begin{cases} \sqrt{1/n}\cdot\sigma_u & \text{if } y_u\sigma_u > 0 \\ 0 & \text{if } y_u\sigma_u \leq 0. \end{cases}$$

Observe that $y_u z_u \geq y_u \bar{\sigma}_u$ for all u, so $\langle \mathbf{y}, \mathbf{z}\rangle \geq \langle \mathbf{y}, \bar{\boldsymbol{\sigma}}\rangle \geq \sqrt{1-\varepsilon}$. By the Cauchy-Schwartz Inequality, $\|\mathbf{y}\|_2\|\mathbf{z}\|_2 \geq \langle \mathbf{y}, \mathbf{z}\rangle$. Since $\|\mathbf{y}\|_2 = 1$, we find that $\|\mathbf{z}\|_2^2 \geq \langle \mathbf{y}, \mathbf{z}\rangle^2 \geq 1-\varepsilon$. Recalling how $\mathbf{z}$ was defined, this means there are at least $(1-\varepsilon)n$ vertices u such that $y_u\sigma_u > 0$. Every such u will be correctly labeled by Algorithm 2. □

Remark 4.2. Using the inequality $\sqrt{2n/\delta} > \ln(2n/\delta)$, we can deduce that the inequality $n > \frac{32p^2}{\varepsilon^2 q^2(p-q)}\ln(2n/\delta)$ is satisfied whenever $n > \frac{2048p^4}{\varepsilon^4 q^4 (p-q)^2 \delta}$. However, the latter bound is typically far greater than the true sample complexity of community detection, since the inequality $\sqrt{2n/\delta} > \ln(2n/\delta)$ has a great deal of slack when n is large and δ is small.

5 Randomized Algorithms

In Sections 3 and 4 we saw some applications of probability theory to average-case analysis of algorithms: the study of how algorithms perform on typical samples from a probability distribution. Three examples were the analysis of empirical risk minimization for hypothesis selection, singular value decomposition on Gaussian random samples, and community detection in the stochastic block model.

The other area where probability theory meets analysis of algorithms is *randomized algorithms*, in which the input data is not assumed to be random, but the algorithm uses internal randomness (metaphorically, it tosses coins) and the goal is to show that even for worst-case inputs, the algorithm satisfies some correctness (or approximate correctness) property, with high probability over the outcomes of its own coin tosses.

5.1 Dimensionality Reduction

Suppose you have a dataset consisting of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ in $\mathbb{R}^d$. 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 d 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, k, 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 $\mathbb{R}^d$ to $\mathbb{R}^k$ 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, k, as an indeterminate. Later we will see that for the purpose of preserving distances up to multiplicative error ε, it is appropriate to set $k = O\left(\frac{\log n}{\varepsilon^2}\right)$.

Lemma 5.1 (Johnson-Lindenstrauss Lemma)**.** *For any $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ and any $0 < \varepsilon, \delta < 1$, if $k > 4\ln(n/\delta)/\varepsilon^2$ and R is a $k \times d$ random matrix with independent entries drawn from the distribution $\mathcal{N}(0, \frac{1}{k})$, then with probability at least $1 - \delta$ the inequality*

$$(1-\varepsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2 \leq \|R\mathbf{x}_i - R\mathbf{x}_j\|_2 \leq (1+\varepsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2$$

holds for all $1 \leq i, j \leq n$.

Proof. Consider any vector $\mathbf{y} \in \mathbb{R}^d$. (Later we will be setting $\mathbf{y} = \mathbf{x}_i - \mathbf{x}_j$.) Our first task will be to compute the expected value of $\|R\mathbf{y}\|_2^2$. Denote the rows of R by $\mathbf{r}_1, \ldots, \mathbf{r}_k$. Then the components of the vector $R\mathbf{y}$ are $\langle \mathbf{r}_1, \mathbf{y} \rangle, \langle \mathbf{r}_2, \mathbf{y} \rangle, \ldots \langle \mathbf{r}_k, \mathbf{y} \rangle$. Letting $Y = \|R\mathbf{y}\|_2^2$ we see that $Y = \sum_{i=1}^k X_i^2$

where the random variables $X_i = \langle \mathbf{r}_i, \mathbf{y} \rangle$ are Gaussian with distribution $\mathcal{N}(0, \frac{1}{k}\|\mathbf{y}\|_2^2)$. This implies that $\mathbb{E}[X_i^2] = \frac{1}{k}\|\mathbf{y}\|_2^2$ for all i and that $\mathbb{E}[Y] = \|\mathbf{y}\|_2^2$. Furthermore, using steps from the proof of Lemma 4.4, for $0 < \varepsilon < 1$ we have

$$\Pr(Y \geq (1+\varepsilon)^2\|\mathbf{y}\|_2^2) \leq e^{-\frac{1}{2}\varepsilon^2 k}, \qquad \Pr(Y \leq (1-\varepsilon)^2\|\mathbf{y}\|_2^2) \leq e^{-\frac{1}{2}\varepsilon^2 k} \tag{31}$$

As $\mathbf{y}$ ranges over all $\binom{n}{2}$ vectors of the form $\mathbf{x}_i - \mathbf{x}_j$, we can use the Union Bound together with inequality (31) to verify that

$$\Pr\left(\forall i, j\ (1-\varepsilon)^2\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \leq \|R\mathbf{x}_i - R\mathbf{x}_j\|_2^2 \leq (1+\varepsilon)^2\|\mathbf{x}_i - \mathbf{x}_j\|_2^2\right) \geq 1 - 2\binom{n}{2}e^{-\frac{1}{2}\varepsilon^2 k}. \tag{32}$$

Since $2\binom{n}{2} < n^2$, the probability on the right side will be greater than $1-\delta$ if $e^{-\frac{1}{2}\varepsilon^2 k} < \delta/n^2$, which happens when $k > 4\ln(n/\delta)/\varepsilon^2$. □

5.2 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 $\mathbf{x}$ given the vector $R\mathbf{x}$, where R is a Gaussian random matrix.

Lemma 5.2. *For any $s \in \mathbb{N}$ and $0 < \varepsilon, \delta < 1$, if $k > 72\varepsilon^{-2}(\log(2/\delta) + 7s)$ and R is a $k \times s$ matrix with independent random entries drawn from $\mathcal{N}(0, \frac{1}{k})$, then with probability at least $1-\delta$ every $\mathbf{x} \in \mathbb{R}^s$ satisfies*

$$(1-\varepsilon)\|\mathbf{x}\|_2^2 \leq \|R\mathbf{x}\|_2^2 \leq (1+\varepsilon)\|\mathbf{x}\|_2^2. \tag{33}$$

Proof. Let $\mathbf{r}_1^\top, \ldots, \mathbf{r}_k^\top$ denote the rows of R. For each i, the random vector $\mathbf{r}_i$ is a sample from $\mathcal{N}(\mathbf{0}, \frac{1}{k}\mathbb{1})$, so $\mathbb{E}[\mathbf{r}_i\mathbf{r}_i^\top] = \frac{1}{k}\mathbb{1}$. Consequently the matrix $P = R^\top R = \sum_{i=1}^k \mathbf{r}_i\mathbf{r}_i^\top$ is a sum of independent random symmetric positive definite matrices, and

$$\mathbb{E}[P] = \sum_{i=1}^k \mathbb{E}[\mathbf{r}_i\mathbf{r}_i^\top] = k \cdot \frac{1}{k}\mathbb{1} = \mathbb{1}.$$

Applying Proposition 4.3 we find that

$$\Pr\left(\sup_{\mathbf{x}\in\mathbb{R}^s} \eta(\langle \mathbf{x}, R^\top R\mathbf{x}\rangle, \langle \mathbf{x}, \mathbf{x}\rangle) \geq \varepsilon\right) \leq 2\exp\left(7s - \tfrac{\varepsilon^2}{72}k\right) \leq \delta,$$

by our assumption that $\frac{\varepsilon^2}{72}k - 7s > \ln(2/\delta)$. Using the relations $\langle \mathbf{x}, R^\top R\mathbf{x}\rangle = \|R\mathbf{x}\|_2^2$ and $\langle \mathbf{x}, \mathbf{x}\rangle = \|\mathbf{x}\|_2^2$, we find that the relation $\eta(\langle \mathbf{x}, R^\top R\mathbf{x}\rangle, \langle \mathbf{x}, \mathbf{x}\rangle) \leq \varepsilon$ is equivalent to inequality (33) in the statement of the lemma. Therefore, that inequality holds with probability at least $1-\delta$ as claimed. □

Definition 5.1. A vector $\mathbf{x} \in \mathbb{R}^n$ is s-sparse if at least $n - s$ coordinates of $\mathbf{x}$ are equal to zero. A matrix R satisfies the *s-restricted isometry property with constant ε_s* if the inequalities

$$(1 - \varepsilon_s)\|\mathbf{x}\|_2^2 \leq \|R\mathbf{x}\|_2^2 \leq (1 + \varepsilon_s)\|\mathbf{x}\|_2^2 \tag{34}$$

are satisfied for every s-sparse vector $\mathbf{x}$.

Proposition 5.3. *For every $s \geq 3$ and $0 < \varepsilon, \delta < 1$, if $n \geq s$ and $k > 72\varepsilon^{-2}(s\ln(n) + \ln(2/\delta))$, then a $k \times n$ matrix R with independent random entries sampled from $\mathcal{N}(0, \frac{1}{k})$ satisfies the s-restricted isometry property with constant ε, with probability at least $1 - \delta$.*

Proof. For any subset $J \subseteq [n]$ with $|J| = s$, let R_J be the submatrix of R obtained by selecting the subset of columns indexed by J. Our plan is to use Lemma 5.2 to prove that with high probability R preserves the length of each vector $\mathbf{x}$ that satisfies $\mathbf{x}_i = 0$ for all $i \notin J$, then use the Union Bound over all choices of J to deduce that R satisfies the restricted isometry property with probability at least $1 - \delta$. Let m denote the number of subsets $J \subseteq [n]$ with $|J| = s$, and note that

$$m = \binom{n}{s} < \left(\tfrac{en}{s}\right)^s < n^s,$$

since the identity $\binom{n}{s} < \left(\frac{en}{s}\right)^s$ holds for all n, s and our assumption $s \geq 3$ implies $\frac{e}{s} < 1$. Note that $\ln(2m/\delta) < s\ln(n) + \ln(2/\delta)$. According to Lemma 5.2, with probability at least $1 - \delta/m$, the matrix R_J obtained by selecting the subset of columns of R indexed by J satisfies

$$\forall \mathbf{y} \in \mathbb{R}^s \quad (1 - \varepsilon)\|\mathbf{y}\|_2^2 \leq \|R_J\mathbf{y}\|_2^2 \leq (1 + \varepsilon)\|\mathbf{y}\|_2^2,$$

which implies that (34) is satisfied by every $\mathbf{x} \in \mathbb{R}^n$ with sparsity pattern J. Taking the union bound over all J, we find that with probability at least $1 - \delta$, R satisfies the s-restricted isometry property with constant ε. □

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 $\mathbf{x} \in \mathbb{R}^n$ given the value of $\mathbf{b} = R\mathbf{x} \in \mathbb{R}^k$. When $k < n$ this is an underdetermined linear system, meaning there are infinitely many vectors $\mathbf{y}$ solving the equation $R\mathbf{y} = \mathbf{b}$. The set of all such solutions forms a $(n-k)$-dimensional affine subspace of $\mathbb{R}^n$, but we will see that there is a unique s-sparse solution provided that R 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 $\mathbf{x}$ satisfying $R\mathbf{x} = \mathbf{b}$.

Definition 5.2. A vector $\mathbf{z} \in \mathbb{R}^n$ 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 ε is used to ensure that the constant factor on the right side of Inequality (35) below is strictly positive.)

Lemma 5.4. *Suppose A is a matrix that satisfies the $(3s)$-restricted isometry property with constant $\varepsilon > 0$. If $\mathbf{z}$ is mostly s-sparse then*

$$\|A\mathbf{z}\|_2 \geq \frac{1}{2}\left(\sqrt{1-\varepsilon} - \sqrt{\tfrac{1+\varepsilon}{2}}\right)\|\mathbf{z}\|_2. \tag{35}$$

Proof. Without loss of generality assume that the coordinates of $\mathbf{z}$ 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 $\mathbf{z}$ with zeros and increas 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 $\mathbf{z}_\ell$ be a vector obtained from $\mathbf{z}$ by preserving the coordinates in block J_ℓ and setting all other coordinates to zero. In other words,

$$(\mathbf{z}_\ell)_i = \begin{cases} \mathbf{z}_i & \text{if } (2\ell-1)s < i \leq (2\ell+1)s \\ 0 & \text{otherwise.} \end{cases}$$

Since $\mathbf{z}$ 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

$$\|\mathbf{z}_0\|_1 \geq \|\mathbf{z}_1 + \mathbf{z}_2 + \cdots + \mathbf{z}_m\|_1. \tag{36}$$

Another useful observation stemming from the way coordinates are ordered is that $\|\mathbf{z}_{i+1}\|_\infty \leq 2s \cdot \|\mathbf{z}_i\|_1$, because the absolute value of *every* coordinate of $\mathbf{z}_{i+1}$ is less than or equal to the absolute value of *every* coordinate of $\mathbf{z}_i$. Combining this observation with the inequality $\|\mathbf{z}_{i+1}\|_2 \leq \sqrt{2s}\|\mathbf{z}_{i+1}\|_\infty$, we obtain

$$\|_\|\mathbf{z}_{i+12} \leq \frac{1}{\sqrt{2s}}\|\mathbf{z}_i\|_1.$$

Now, we can bound $\|A\mathbf{z}\|_2$ from below as follows.

$$\begin{aligned}
\|A\mathbf{z}\|_2 &= \|A(\mathbf{z}_0 + \mathbf{z}_1) + A\mathbf{z}_2 + A\mathbf{z}_3 + \cdots + A\mathbf{z}_m\|_2 \\
&\geq \|A(\mathbf{z}_0 + \mathbf{z}_1)\|_2 - (\|_\|A\mathbf{z}_{22} + \|A\mathbf{z}_3\| + \cdots + \|A\mathbf{z}_m\|) \\
&\geq \sqrt{1-\varepsilon}\|\mathbf{z}_0 + \mathbf{z}_1\|_2 - \sqrt{1+\varepsilon}(\|\mathbf{z}_2\|_2 + \|\mathbf{z}_3\|_2 + \cdots + \|\mathbf{z}_m\|_2) \qquad (37) \\
&\geq \sqrt{1-\varepsilon}\|\mathbf{z}_0\|_2 - \sqrt{\frac{1+\varepsilon}{2s}}(\|\mathbf{z}_1\|_1 + \|\mathbf{z}_2\|_1 + \cdots + \|\mathbf{z}_{m-1}\|_1) \\
&= \sqrt{1-\varepsilon}\|\mathbf{z}_0\|_2 - \sqrt{\frac{1+\varepsilon}{2s}}\|\mathbf{z}_1 + \mathbf{z}_2 + \cdots + \mathbf{z}_{m-1}\|_1 \\
&\geq \sqrt{1-\varepsilon}\|\mathbf{z}_0\|_2 - \sqrt{\frac{1+\varepsilon}{2s}}\|\mathbf{z}_0\|_1. \qquad (38)
\end{aligned}$$

In line (37) we have used the inequalities $\sqrt{1-\varepsilon}\|_\|\mathbf{z}_0 + \mathbf{z}_{12} \leq \|A(\mathbf{z}_0 + \mathbf{z}_1)\|_2$ and $\sqrt{1+\varepsilon}\|\mathbf{z}_i\|_2 \geq \|A\mathbf{z}_i\|_2$, both of which are justified by the $(3s)$-restricted isometry property with constant ε.

Let $\boldsymbol{\sigma}$ be a vector in $\{-1, 0, 1\}^n$ with the same sign pattern and sparsity pattern as $\mathbf{z}_0$, 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 $\langle \boldsymbol{\sigma}, \mathbf{z}_0 \rangle = \|\mathbf{z}_0\|_1$, so the Cauchy-Schwartz inequality implies

$$\|\mathbf{z}_0\|_1 \leq \|\boldsymbol{\sigma}\|_2 \, \|\mathbf{z}_0\|_2 = \sqrt{s} \, \|\mathbf{z}_0\|_2.$$

Substituting this bound into inequality (38) above, we find that

$$\|A\mathbf{z}\|_2 \geq \left(\sqrt{1-\varepsilon} - \sqrt{\frac{1+\varepsilon}{2}} \right) \|\mathbf{z}_0\|_2. \tag{39}$$

To conclude the proof of the lemma we need to show $\|\mathbf{z}_0\|_2 \geq \frac{1}{2}\|\mathbf{z}\|$. Let $t = \frac{1}{s}\|\mathbf{z}_0\|_1 = \frac{1}{s}\left(|z_1| + |z_2| + \cdots + |z_s|\right)$ and observe $t \geq |z_s|$. Every component of the vector $\mathbf{w} = \frac{1}{t}\left(\mathbf{z}_1 + \mathbf{z}_2 + \cdots + \mathbf{z}_m\right)$ belongs to the interval $[-1, 1]$, because $|z_i| \leq |z_s| \leq t$ for $i > s$. Hence,

$$\begin{aligned} \|\mathbf{w}\|_2^2 &= \sum_{i=1}^{n} \mathbf{w}_i^2 \leq \sum_{i=1}^{n} |\mathbf{w}_i| = \|\mathbf{w}\|_1 \\ \|\mathbf{z} - \mathbf{z}_0\|_2^2 &= t^2 \|\mathbf{w}\|_2^2 \leq t^2 \|\mathbf{w}\|_1 = t\|\mathbf{z}_1 + \cdots + \mathbf{z}_m\|_1 \leq t\|\mathbf{z}_0\|_1 = \frac{1}{s}\|\mathbf{z}_0\|_1^2 \leq \|\mathbf{z}_0\|_2^2. \end{aligned} \tag{40}$$

By the triangle inequality, $\|\mathbf{z}\|_2 \leq \|\mathbf{z}_0\|_2 + \|\mathbf{z} - \mathbf{z}_0\|_2$. Combined with Inequality (40), this implies $\|\mathbf{z}\|_2 \leq 2\|\mathbf{z}_0\|_2$ and completes the proof of the lemma. □

We will use Lemma 5.4 to analyze the following algorithm for sparse recovery: of all the vectors $\mathbf{x}$ that satisfy $R\mathbf{x} = \mathbf{b}$, 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 5.5. *Suppose R is a matrix that satisfies the $(3s)$-restricted isometry property with constant $\varepsilon < \frac{1}{3}$, $\mathbf{x}_0$ is an s-sparse vector, and $\mathbf{b} = R\mathbf{x}_0$. Among the solutions of the equation $R\mathbf{x} = \mathbf{b}$, the vector $\mathbf{x}_0$ is the unique one with minimum L_1 norm.*

Proof. Suppose $\mathbf{x}_1$ is a solution of minimum L_1 norm to the equation $R\mathbf{x} = \mathbf{b}$. We must prove that $\mathbf{x}_1 = \mathbf{x}_0$. Let $\mathbf{z} = \mathbf{x}_1 - \mathbf{x}_0$, and observe that $R\mathbf{z} = R\mathbf{x}_1 - R\mathbf{x}_0 = \mathbf{b} - \mathbf{b} = \mathbf{0}$. Let $J = \{i \mid x_{0i} \neq 0\}$ and observe $|J| \leq s$. We have

$$\begin{aligned} \|\mathbf{x}_1\|_1 &= \sum_{i=1}^{n} |x_{1i}| = \sum_{i=1}^{n} |x_{0i} + z_i| = \sum_{i \in J} |x_{0i} + z_i| + \sum_{i \notin J} |z_i| \\ &\geq \sum_{i \in J} |x_{0i}| - \sum_{i \in J} |z_i| + \sum_{i \notin J} |z_i| = \|\mathbf{x}_0\|_1 - \sum_{i \in J} |z_i| + \sum_{i \notin J} |z_i|. \end{aligned}$$

Since $\|\mathbf{x}_1\|_1 \leq \|\mathbf{x}_0\|_1$ by our choice of $\|\mathbf{x}_1\|$, it follows that $\sum_{i\in J} |z_i| \geq \sum_{i\notin J} |z_i|$, i.e. $\mathbf{z}$ is mostly s-sparse. By Lemma 5.4,

$$0 = \|R\mathbf{z}\|_2 \geq \frac{1}{2}\left(\sqrt{1-\varepsilon} - \sqrt{\frac{1+\varepsilon}{2}}\right)\|\mathbf{z}\|_2.$$

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 $\|\mathbf{z}\|_2 = 0$, so $\mathbf{0} = \mathbf{z} = \mathbf{x}_1 - \mathbf{x}_0$, as desired. □

5.3 Data Streaming and Sketching

The random projections analyzed in Section 5.1 and Section 5.2 bear some resemblance to the random hash functions used in data structures and load balancing. In this section we survey some of the applications of hash functions 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 bits, where s is bounded by a polynomial function of b and $\log(n)$. Hence it is infeasible to store each data item, which would require space $s \geq b \cdot n$, and it's also infeasible to store a count of how many times each token was seen in the data stream, which would require space $s \geq 2^b \log(n)$.

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^{th} frequency moment, $\sum_j f_j^p$, where f_j denotes the number of occurrences of the token j in the stream.

5.3.1 Finding frequent elements

To illustrate the model, we begin by presenting 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 5.6. *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 $b_1, \ldots, b_k$. □

5.3.2 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 ε 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. For the sake of building intuition, suppose that $h(j)$ were an independent, uniformly distributed element of $[M]$ for each $j \in [m]$. Storing the description of such a hash function would require $m \log(M)$ bits of space, exceeding the space requirement of our algorithm, but for now let's just see what could be done with such a function h. Later we'll worry about using a hash function whose description can be stored in much less space.

The key observation is that if there are d distinct tokens in the stream, then the random variable $Z = \min\{h(a_i) \mid 1 \le i \le n\}$ 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

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

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

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

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

For future reference, we remark that we only used two properties of the random hash function h that we used when proving the three properties above.

1. *For every token a_i, the hash value $h(a_i)$ is uniformly distributed in $[M]$.* This was used to establish that $\mathbb{E}[X_{ik}] = \frac{k}{M}$.
2. *For every two distinct tokens a_i, a_j, the hash values $h(a_i)$ and $h(a_j)$ are independent random variables.* This was used in the calculation of $\text{Var}[Y_k]$, when we applied the identity $\mathbb{E}[X_{ik}X_{jk}] = \mathbb{E}[X_{ik}] \cdot \mathbb{E}[X_{jk}] = \left(\frac{k}{M}\right)^2$.

A probability distribution over hash functions is called *2-universal* if it satisfies these two properties. Below, we see how to construct randomized hash functions that satisfy both of these properties but can be stored in exponentially less space than a purely random hash function. For now, we proceed with the design and analysis of algorithms for approximating the number of distinct elements in a data stream, using any 2-universal randomized hash function h.

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

$$\Pr\left(\tfrac{M}{Z} > 6d\right) = \Pr\left(\tfrac{M}{6d} > Z\right) = \Pr\left(Y_k \geq 1\right) \leq \mathbb{E}[Y_k] = \tfrac{dk}{M} \leq \tfrac{1}{6}. \tag{41}$$

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

$$\begin{aligned}\Pr\left(\tfrac{M}{Z} < \frac{d}{6}\right) = \Pr\left(\tfrac{6M}{d} \leq Z\right) &= \Pr(Y_\ell = 0) \leq \Pr(|Y_\ell - \mathbb{E}Y_\ell| \geq \mathbb{E}Y_\ell) \\ &\leq \frac{\mathrm{Var}(Y_\ell)}{(\mathbb{E}Y_\ell)^2} \\ &< \frac{\mathbb{E}Y_{ell}}{(\mathbb{E}Y_\ell)^2} = \frac{1}{\mathbb{E}Y_\ell} = \frac{M}{d\ell} \leq \frac{1}{6} + \frac{1}{6M-5}.\end{aligned} \tag{42}$$

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^{th} smallest of the values $\{h(a_i)\}_{i=1}^n$, for $t > 1$. 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 εt 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 δ. This argument doesn't directly lead to the conclusion that tM/Z_t approximates d within ε, but a slight variation on the argument, using Y_q and Y_r for $q = \lfloor \frac{tM}{(1+\varepsilon)d} \rfloor$ and $r = \lfloor \frac{tM}{(1-\varepsilon)d} \rfloor$, does the trick.

Algorithm 3 Algorithm for estimating distinct elements

1: Set $t = \lceil \frac{2(1+\varepsilon)}{\varepsilon^2 \delta} \rceil$.
2: Choose $M \geq m$ 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 the $t \log(M)$ plus the amount of space required to store h. Later we will see that the space required for storing h is $2\log(M)$ when M

is a prime number. Since there is always a prime number between m and $2m$, we can ensure $\log(M) \le \log(m) + 1$. Also $t \le \frac{2}{\varepsilon^2\delta} + 1$, so the space required by the algorithm is $s = O\left(\frac{\log m}{\varepsilon^2\delta}\right)$.

Proposition 5.7. *When Algorithm 3 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. Set $q = \lfloor \frac{tM}{(1+\varepsilon)d} \rfloor$ and $r = \lfloor \frac{tM}{(1-\varepsilon)d} \rfloor$. If the algorithm outputs an estimate greater than $(1+\varepsilon)d$ it means that $Z_t \le q$ so $Y_q \ge t$. If it outputs an estimate less than $(1-\varepsilon)d$ it means that $Z_t > r$ so $Y_r < t$. We have $\mathbb{E}[Y_q] = dq/M$ and $\mathrm{Var}[Y_q] < \mathbb{E}[Y_q]$, so by Chebyshev's Inequality,

$$\Pr(Y_q \ge t) \le \frac{dq/M}{(dq/M - t)^2} \le \frac{t/(1+\varepsilon)}{(dq/M - t)^2}. \tag{43}$$

Similarly,

$$\Pr(Y_r < t) < \frac{dr/M}{(dr/M - t)^2} \le \frac{t/(1-\varepsilon)}{(dr/M - t)^2}. \tag{44}$$

To estimate the quantities on the right sides of inequalities (43) and (44), we use

$$\begin{aligned}
\tfrac{dq}{M} &\le \tfrac{t}{1+\varepsilon} = t - \tfrac{\varepsilon t}{1+\varepsilon} \\
\left(\tfrac{dq}{M} - t\right)^2 &\ge \frac{\varepsilon^2 t^2}{(1+\varepsilon)^2} = \frac{\varepsilon^2 t}{1+\varepsilon} \cdot \frac{t}{1+\varepsilon} \\
\tfrac{dr}{M} &\ge \tfrac{t}{1-\varepsilon} - \tfrac{d}{M} \ge t + \tfrac{\varepsilon t}{1-\varepsilon} - 1 \\
\left(\tfrac{dr}{M} - t\right)^2 &\ge \left(\tfrac{\varepsilon t}{1-\varepsilon} - 1\right)^2 = \left(\tfrac{\varepsilon t - 1 + \varepsilon}{1-\varepsilon}\right)^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
\Pr(Y_q \ge t) &\le \frac{t/(1+\varepsilon)}{(dq/M - t)^2} \le \frac{1+\varepsilon}{\varepsilon^2 t} \\
\Pr(Y_r > t) &< \frac{t/(1-\varepsilon)}{(dr/M - t)^2} \le \frac{(1-\varepsilon)t}{(\varepsilon t - 1 + \varepsilon)^2}.
\end{aligned}$$

To estimate the right side of the second line, we use

$$\begin{aligned}
\varepsilon^2 t > 1 > 1 - \varepsilon^2 &= (1-\varepsilon)(1+\varepsilon) \\
\tfrac{\varepsilon}{1+\varepsilon} &> \tfrac{1-\varepsilon}{\varepsilon t} \\
\left(1 - \tfrac{\varepsilon}{1+\varepsilon}\right)^2 &< \left(1 - \tfrac{1-\varepsilon}{\varepsilon t}\right)^2 \\
\tfrac{1-\varepsilon}{1+\varepsilon} = 1 - \tfrac{2\varepsilon}{1+\varepsilon} &< \left(1 - \tfrac{1-\varepsilon}{\varepsilon t}\right)^2 \\
\tfrac{(1-\varepsilon)\varepsilon^2 t^2}{1+\varepsilon} &< (\varepsilon t - 1 + \varepsilon)^2 \\
\tfrac{(1-\varepsilon)t}{(\varepsilon t - 1 + \varepsilon)^2} &< \tfrac{1+\varepsilon}{\varepsilon^2 t}.
\end{aligned}$$

Hence, both $\Pr(Y_q \le t)$ and $\Pr(Y_r > t)$ are less than or equal to $\frac{1+\varepsilon}{\varepsilon^2 t}$. The choice of $t = \lceil \frac{2(1+\varepsilon)}{\varepsilon^2\delta} \rceil$ ensures that $\frac{1+\varepsilon}{\varepsilon^2 t} \le \frac{\delta}{2}$. Hence, the probability that the algorithm fails to output a number in the range $[(1-\varepsilon)d, (1+\varepsilon)d]$ is at most $\frac{\delta}{2} + \frac{\delta}{2} = \delta$, as desired. □

5.3.3 A 2-universal hash family based on modular arithmetic

In this section we present a simple example of a 2-universal family of hash functions $[m] \to [M]$ when M is a prime number greater than or equal to m. In that case, for any $a, b \in [M]$ define $h_{ab}(x) \equiv ax + b \pmod{M}$. Note that one can store a description of the entire hash function h_{ab} using $2\log(M)$ bits, simply by storing the coefficients a and b.

Lemma 5.8. *If $a, b \in [M]$ are sampled independently and uniformly at random, then the distribution of the random hash function $h = h_{ab}$ is 2-universal.*

Proof. We need to prove that if $x \neq y \in [M]$, then $h(x)$ and $h(y)$ are uniformly distributed in $[M]$ and they are independent. In other words, we must prove for each pair of elements $i, j \in [M]$,

$$\Pr(h(x) = i \text{ and } h(y) = j) = \tfrac{1}{M^2}.$$

To do so, we recall that M is prime and $x \neq y$. Since $x - y$ is a non-zero integer between $-(M-1)$ and $M-1$, it is not divisible by M hence there exists some integer z such that $z(x-y) \equiv 1 \pmod{M}$. The equations $h(x) = i$ and $h(y) = j$ imply $i - j = h(x) - h(y) \equiv a(x-y) \pmod{M}$, which implies $z(i-j) \equiv az(x-y) \equiv a \pmod{M}$. Hence, $a \equiv z(i-j) \pmod{M}$ and $b \equiv i - ax = i - z(i-j)x \pmod{M}$ constitute the unique solution (mod M) to $h(x) = i$ and $h(y) = j$. The probability that this exact pair a, b is sampled when we draw a and b independently and uniformly at random from $[M]$ is $\frac{1}{M^2}$, as desired. □

5.3.4 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 $s = O(\text{poly}(\log n, \log m))$ 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 4 Count-Min Sketch

Given positive integers $B, t \ldots$
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, ℓ.
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 4. 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 t. To the extent that the counter exceeds t, the difference is due to hash collisions — other elements of the stream that hash to the same bucket as x. For large M, 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 5.9. *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 $f_x \le \hat{f}_x$ with probability 1 and $\hat{f}_x \le f_x + \frac{2n}{B}$ 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,

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

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

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

and the lemma follows. □

Corollary 5.10. *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 εn. The space complexity of the sketch with these parameters is $O(\log(mn)\log(1/\delta)/\varepsilon)$.*

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 $\varepsilon \|\mathbf{f}\|_2$. Here, $\mathbf{f}$ denotes the "frequency vector" of the stream, an m-dimensional vector whose x^{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 $\mathbf{f}_1 = n$. Note that $\mathbf{f}_2 \leq \mathbf{f}_1$ for any vector $\mathbf{f}$, so the error bound of $\varepsilon \|\mathbf{f}\|_2$ is never worse than the εn error bound of the Count-Min Sketch. However, $\|\mathbf{f}\|_2$ can be much smaller than n; for example, when the tokens are uniformly distributed we have $\|\mathbf{f}\|_2 \approx \frac{n}{\min\{\sqrt{m}, \sqrt{n}\}}$.

Algorithm 5 Count Sketch

Given positive integers $B, t \ldots$
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, ℓ.
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_ℓ 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 5.11. *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 $|\hat{f}_x - f_x| \leq \sqrt{\frac{3}{B}} \|\mathbf{f}\|_2$ 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_ℓ 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

$$\mathbb{E}[Z_{y\ell}] = \mathbb{E}[g_\ell(x) g_\ell(y) X_{y\ell} f_y] = \mathbb{E}[g_\ell(x)] \cdot \mathbb{E}[g_\ell(y)] \cdot \mathbb{E}[X_{y\ell}] \cdot f_y = 0, \tag{45}$$

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_ℓ which is independent of g_ℓ, and the values $g_\ell(x), g_\ell(y)$ are independent of one another by the pairwise-independence property of g_ℓ.

Using linearity of expectation we have

$$\mathbb{E}[g_\ell(x) \cdot C[h_\ell(x), \ell]] = \sum_{y=1}^{m} \mathbb{E}[Z_{y\ell}] = f_x + \sum_{y \neq x} \mathbb{E}[Z_{y\ell}] = f_x. \tag{46}$$

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_ℓ 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_ℓ, which is independent of g_ℓ.) 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

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

By Chebyshev's Inequality,

$$\Pr\left(|g_\ell(x) \cdot C[h_\ell(x), \ell] - f_x| \geq \sqrt{\tfrac{3}{B}}\,\|\mathbf{f}\|_2\right) \leq \frac{\mathrm{Var}[g_\ell(x) \cdot C[h_\ell(x), \ell]]}{\frac{3}{B}\|\mathbf{f}\|_2^2} = \frac{1}{3}. \tag{47}$$

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 $\sqrt{3/B} \cdot \|\mathbf{f}\|_2$. 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 $f_x - \sqrt{3/B} \cdot \|\mathbf{f}\|_2$ or greater than $f_x + \sqrt{3/B} \cdot \|\mathbf{f}\|_2$, at least $t/2$ of the estimates must differ from f_x by more than $\sqrt{3/B} \cdot \|\mathbf{f}\|_2$. 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_ℓ be a random variable which equals 1 if $|g_\ell(x) \cdot C[h_\ell(x), \ell] - f_x| \geq \sqrt{3/B} \cdot \|\mathbf{f}\|_2$, otherwise $W_\ell = 0$. Inequality (47) 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 5.12. *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 frequcency by no more than $\varepsilon\|\mathbf{f}\|_2$. The space complexity of the sketch with these parameters is $O\left(\log(mn)\log(1/\delta)/\varepsilon^2\right)$.*