10

Concentration of Measure

Consider the following questions:

1. You distribute n tasks among n machines, by sending each task to a machine uniformly and independently at random: while any machine has unit expected load, what is the maximum load (i.e., the maximum number of tasks assigned to any machine)?

2. You want to estimate the bias p of a coin by repeatedly flipping it and then taking the sample mean. How many samples suffice to be within $\pm\varepsilon$ of the correct answer p with confidence $1-\delta$?

3. How many unit vectors can you choose in $\mathbb{R}^n$ that are almost orthonormal? I.e., they must satisfy $|\langle v_i, v_j \rangle| \leq \varepsilon$ for all $i \neq j$?

4. A n-dimensional hyercube has $N = 2^n$ nodes. Each node $i \in [N]$ contains a packet p_i, which is destined for node π_i, where π is a permutation. The routing happens in rounds. At each round, each packet traverses at most one edge, and each edge can transmit at most one packet. Find a routing policy where each packet reaches its destination in $O(n)$ rounds, regardless of the permutation π.

All these questions can be answered by the same basic tool, which goes by the name of *Chernoff bounds* or *concentration inequalities* or *tail inequalities* or *concentration of measure*, or tens of other names. The basic question is simple: *if we have a real-valued function $f(X_1, X_2, \ldots, X_m)$ of several independent random variables X_i, such that it is "not too sensitive to each coordinate", how often does it deviate far from its mean?* To make it more concrete, consider this—

> Given n independent random variables $X_1, \ldots, X_n$, each bounded in the interval $[0,1]$, let $S_n = \sum_{i=1}^n X_i$. What is
>
> $$\Pr\left[S_n \notin (1 \pm \varepsilon)\, \mathbb{E}S_n\right]?$$

This question will turn out to have relations to convex geometry, to online learning, to many other areas. But of greatest interest to us, this question will solve many problems in algorithm analysis, including the above four. Let us see some basic results, and then give the answers to the four questions.

10.1 Asymptotic Analysis

We will be concerned with *non-asymptotic analysis*, i.e., the qualitative behavior of sums (and other Lipschitz functions) of finite number of (bounded) independent random variables. Before we begin that, a few words about the asymptotic analysis, which concerns the convergence of averages of infinite sequences of random variables.

Given a sequence of random variables $\{X_n\}$ and another random variable Y, the following two notions of convergence can be defined.

Definition 10.1 (Convergence in Probability)**.** $\{X_n\}$ converges in probability to Y if for every $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}(|X_n - Y| > \varepsilon) = 0 \tag{10.1}$$

This is denoted by $X_n \xrightarrow{\text{p}} Y$.

Definition 10.2 (Convergence in Distribution)**.** Let $F_X(.)$ denote the CDF of a random variable X. $\{X_n\}$ converges in distribution to Y if

$$\lim_{n\to\infty} F_{X_n}(t) = F_Y(t) \tag{10.2}$$

for all points t where the distribution function F_Y is continuous. This is denoted by $X_n \xrightarrow{\text{d}} Y$.

There are many results known here, and we only mention the two well-known results below. The *weak law of large numbers* states that the average of independent and identically distributed (i.i.d.) random variables converges in probability to their mean.

Theorem 10.3 (Weak law of large numbers)**.** *Let S_n denote the sum of n i.i.d. random variables, each with mean μ and variance $\sigma^2 < \infty$, then*

$$S_n/n \xrightarrow{p} \mu. \tag{10.3}$$

The *central limit theorem* tells us about the distribution of the sum of a large collection of i.i.d. random variables. Let $N(0,1)$ denote the standard normal variable with mean 0 and variance 1, whose probability density function is $f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$.

Theorem 10.4 (Central limit theorem)**.** *Let S_n denote the sum of n i.i.d. random variables, each with mean μ and variance $\sigma^2 < \infty$, then*

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} N(0,1). \tag{10.4}$$

There are many powerful asymptotic results in the literature; see need to give references here.

10.2 Non-Asymptotic Convergence Bounds

Our focus will be on the behavior of finite sequences of random variables. The central question here will be: what is the chance of deviating far from the mean? Given an r.v. X with mean μ, and some deviation $\lambda > 0$, the quantity

$$\Pr[X \geq \mu + \lambda]$$

is called the *upper tail*, and the analogous quantity

$$\Pr[X \leq \mu - \lambda]$$

is the *lower tail*. We are interested in bounding these tails for various values of λ.

10.2.1 Markov's inequality

Most of our results will stem from the most basic of all results: *Markov's inequality*. This inequality qualitatively generalizes that idea that a random variable cannot always be above its mean, and gives a bound on the upper tail.

Theorem 10.5 (Markov's Inequality). *Let X be a non negative random variable and $\lambda > 0$, then*

$$\mathbb{P}(X \geq \lambda) \leq \frac{\mathbb{E}(X)}{\lambda} \tag{10.5}$$

With this in hand, we can start substituting various non-negative functions of random variables X to deduce interesting bounds. For instance, the next inequality looks at both the mean $\mu := \mathbb{E}X$ and the variance $\sigma^2 := \mathbb{E}[(X-\mu)^2]$ of a random variable, and bounds both the upper and lower tails.

10.2.2 Chebychev's Inequality

Theorem 10.6 (Chebychev's inequality). *For any random variable X with mean μ and variance σ^2, we have*

$$\Pr[|X - \mu| \geq \lambda] \leq \frac{\sigma^2}{\lambda^2}.$$

Proof. Using Markov's inequality on the non-negative r.v. $Y = (X - \mu)^2$, we get

$$\Pr[Y \geq \lambda^2] \leq \frac{\mathbb{E}[Y]}{\lambda^2}.$$

The proof follows from $\Pr[Y \geq \lambda^2] = \Pr[|X - \mu| \geq \lambda]$. □

10.2.3 Examples: The First Bounds

Example 1 (Coin Flips): Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli random variables with $\Pr[X_i = 0] = 1 - p$ and $\Pr[X_i = 1] = p$. (Im other words, these are the outcomes of independently flipping n coins, each with *bias* p.) Let $S_n := \sum_i^n X_i$ be the *number of heads*. Then S_n is distributed as a binomial random variable $\text{Bin}(n, p)$, with

$$\mathbb{E}[S_n] = np \qquad \text{and} \qquad \text{Var}[S_n] = np(1-p).$$

Recall that linearity of expectations for r.v.s X, Y means $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$. For *independent* we have $\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y]$.

Applying Markov's inequality for the upper tail gives

$$\Pr[S_n - pn \geq \beta n] \leq \frac{pn}{pn + \beta n} = \frac{1}{1 + (\beta/p)}.$$

So, for $p = 1/2$, this is $\frac{1}{1+2\beta} \approx 1 - O(\beta)$ for small values of $\beta > 0$. However, Chebychev's inequality gives a much tighter bound:

$$\Pr[|S_n - pn| \geq \beta n] \leq \frac{np(1-p)}{\beta^2 n^2} < \frac{p}{\beta^2 n}.$$

In particular, this already says that the sample mean S_n/n lies in the interval $p \pm \beta$ with probability at least $1 - \frac{p}{\beta^2 n}$. Equivalently, to get confidence $1 - \delta$, we just need to set $\delta \geq \frac{p}{\beta^2 n}$, i.e., take $n \geq \frac{p}{\beta^2 \delta}$. (We will see a better bound soon.)

Concretely, to get within an additive 1% error of the correct bias p with probability 99.9%, set $\beta = 0.01$ and $\delta = 0.001$, so taking $n \geq 10^7 \cdot p$ samples suffices.

Example 2 (Balls and Bins): Throw n balls uniformly at random and independently into n bins. Then for a fixed bin i, let L_i denote the number of balls in it. Observe that L_i is distributed as a $\text{Bin}(n, 1/n)$ random variable. Markov's inequality gives a bound on the probability that L_i deviates from its mean 1 by $\lambda \gg 1$ as

$$\Pr\left[L_i \geq 1 + \lambda\right] \leq \frac{1}{1+\lambda} \approx \frac{1}{\lambda}.$$

However, Chebychev's inequality gives a much tighter bound as

$$\Pr\left[|L_i - 1| \geq \lambda\right] \leq \frac{(1 - 1/n)}{\lambda^2} \approx \frac{1}{\lambda^2}.$$

So setting $\lambda = 2\sqrt{n}$ says that the probability of any fixed bin having more than $2\sqrt{n} + 1$ balls is at most $\frac{(1-1/n)}{4n}$. Now a union bound over all bins i means that, with probability at least $n \cdot \frac{(1-1/n)}{4n} \leq 1/4$, the load on every bin is at most $1 + 2\sqrt{n}$.

Doing this argument with Markov's inequality would give a trivial upper bound of $1 + 2n$ on the load. This is useless, since there are at most n balls, so the load can never be more than n.

Example 3 (Random Walk): Suppose we start at the origin and at each step move a unit distance either left or right uniformly randomly and independently. We can then ask about the behaviour of the final position after n steps. Each step (X_i) can be modelled as a *Rademacher* random variable with the following distribution.

A random sign is also called a *Rademacher random variable*, the name Bernoulli being already taken for a random bit in $\{0, 1\}$.

$$X_i = \begin{cases} 1 & \text{w.p. } \frac{1}{2} \\ -1 & \text{w.p. } \frac{1}{2} \end{cases}$$

The position after n steps is given by $S_n = \sum_{i=1}^{n} X_i$, with mean and variance being $\mu = 0$ and $\sigma^2 = n$ respectively. Applying Chebyshev's inequality on S_n with deviation $\lambda = t\sigma = t\sqrt{n}$, we get

$$\Pr\left[S_n > t\sqrt{n}\right] \leq \frac{1}{t^2}. \qquad (10.6)$$

We will soon see how to get a tighter tail bound.

10.2.4 Higher-Order Moment Inequalities

All the bounds in the examples above can be improved by using higher-order moments of the random variables. The idea is to use the same recipe as in Chebychev's inequality.

Theorem 10.7 ($2k^{th}$-Order Moment inequalities)**.** *Let $k \in \mathbb{Z}_{\geq 0}$. For any random variable X having mean μ, and finite moments upto order $2k$, we have*

$$\Pr[|X - \mu| \geq \lambda] \leq \frac{\mathbb{E}((X-\mu)^{2k})}{\lambda^{2k}}.$$

Proof. The proof is exactly the same: using Markov's inequality on the non-negative r.v. $Y := (X - \mu)^{2k}$,

$$\Pr[|X - \mu| \geq \lambda] = \Pr[Y \geq \lambda^{2k}] \leq \frac{\mathbb{E}[Y]}{\lambda^{2k}}. \qquad \square$$

We can get stronger tail bounds for large values of k, however it becomes increasingly tedious to compute $E((X - \mu)^{2k})$ for the random variables of interest.

Example 3 (Random Walk, continued): If we consider the fourth moment of S_n:

$$\begin{aligned}
\mathbb{E}\left[(S_n)^4\right] &= \mathbb{E}\Big[\sum_{i=1}^{n} X_i\Big] \\
&= \mathbb{E}\Big[\sum_i X_i^4 + 4\sum_{i<j} X_i^3 X_j + 6\sum_{i<j} X_i^2 X_j^2 + 12\sum_{i<j<k} X_i^2 X_j X_k + 24\sum_{i<j<k<l} X_i X_j X_k X_l\Big] \\
&= n + 6\binom{n}{2},
\end{aligned}$$

where we crucially used that the r.v.s are independent and mean-zero, hence terms like $X_i^3 X_j$, $X_i^2 X_j X_k$, and $X_i X_j X_k X_l$ all have mean zero. Now substituting this expectation in the fourth-order moment inequality, we get a stronger tail bound for $\lambda = t\sigma = t\sqrt{n}$.

$$\Pr\left[|S_n| \geq t\sqrt{n}\right] \leq \frac{\mathbb{E}\left[(S_n)^4\right]}{t^4 n^2} = \frac{n + 6\binom{n}{2}}{t^4 n^2} = \frac{\Theta(1)}{t^4}. \qquad (10.7)$$

Compare this with the bound in (10.6).

10.2.5 Digression: The Right Answer for Random Walks

We can actually explicitly computing $\Pr(S_n = k)$ for sums of Rademacher random variables. Indeed, we just need to choose the positions for +1 steps, which means

$$\frac{\Pr[S_n = 2\lambda]}{\Pr[S_n = 0]} = \frac{\binom{n}{\frac{n}{2}+\lambda}}{\binom{n}{\frac{n}{2}}}.$$

For large n, we can use Stirling's formula $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$:

$$\frac{\Pr[S_n = 2\lambda]}{\Pr[S_n = 0]} \approx \frac{(\frac{n}{2})^{n/2}(\frac{n}{2})^{n/2}}{(\frac{n}{2}+\lambda)^{(n/2+\lambda)}(\frac{n}{2}-\lambda)^{(n/2-\lambda)}} = \frac{1}{(1+\frac{2\lambda}{n})^{\frac{n}{2}+\lambda}(1-\frac{2\lambda}{n})^{\frac{n}{2}-\lambda}}.$$

If $\lambda \ll n$, then we can approximate $1 + k\frac{\lambda}{n}$ by $e^{k\frac{\lambda}{n}}$:

$$\frac{\Pr[S_n = 2\lambda]}{\Pr[S_n = 0]} \approx e^{\frac{-2\lambda}{n}(\frac{n}{2}+\lambda)} e^{\frac{2\lambda}{n}(\frac{n}{2}-\lambda)} = e^{\frac{-4\lambda^2}{n}}.$$

Finally, substituting $\lambda = t\sigma = t\sqrt{n}$, we get

$$\Pr[S_n = 2\lambda] \approx \Pr[S_n = 0] \cdot e^{-4t^2}.$$

This shows that most of the probability mass lies in the region $|S_n| \leq O(\sqrt{n})$, and drops off exponentially as we go further. And indeed, this is the bound we will derive next—we will get slightly weaker constants, but we will avoid these tedious approximations.

10.3 Chernoff bounds, and Hoeffding's inequality

The main bound of this section is a bit of a mouthful, but as Ryan O'Donnell says in his notes, you should memorize it "like a poem". I find it lies in a sweet spot: it is not difficult to remember, and still is very broadly applicable:

The provenance of these bounds is again quite complicated. There's Herman Chernoff's paper, which derives the corresponding inequality for i.i.d. Bernoulli random variables. Wassily Hoeffding gives the generalization for independent random variables all taking values in some bounded interval $[a, b]$. Moreover, Chernoff attributes his result to another Herman, namely Herman Rubin. Then there's Harald Cramér (of the Cramér-Rao fame, not of Cramer's rule). And there's the bound by Sergei Bernstein, many years earlier, which is at least as strong. . .

Theorem 10.8 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be n independent random variables taking values in $[0, 1]$. Let $S_n := \sum_{i=1}^n X_i$, with mean $\mu := \mathbb{E}[S_n] = \sum_i \mathbb{E}[X_i]$. Then for any $\lambda \geq 0$ we have*

$$\textit{Upper tail}: \qquad \Pr\left[S_n \geq \mu + \lambda\right] \leq \exp\left\{-\frac{\lambda^2}{2\mu+\lambda}\right\}. \qquad (10.8)$$

$$\textit{Lower tail}: \qquad \Pr\left[S_n \leq \mu - \lambda\right] \leq \exp\left\{-\frac{\lambda^2}{3\mu}\right\}. \qquad (10.9)$$

Before we prove the bound, let's give a simpler version that suffices for many settings; here we assume the deviation λ is smaller than the mean, and hence can be written as $\beta\mu$ for $\beta \in [0, 1]$.

Corollary 10.9 (Double-Sided Concentration Bound). *For $X_1, \ldots, X_n$ independent r.v.s taking values in $[0,1]$, Let $S_n := \sum_{i=1}^n X_i$ have mean $\mu := \mathbb{E}[S_n]$. Then for any $\beta \in [0,1]$,*

$$\Pr\left[S_n \notin \mu(1 \pm \beta)\right] \leq 2\, e^{-\beta^2 \mu/3}. \tag{10.10}$$

10.3.1 *The Proof*

Proof of Theorem 10.8. We only prove (10.8); the proof for (10.9) is similar. The idea is to use Markov's inequality not on the square or the fourth power, but on a function which is fast-growing enough so that we get tighter bounds, and "not too fast" so that we can control the errors. So we consider the *Laplace transform*, i.e., the function

$$x \mapsto e^{tx}$$

for some value $t > 0$ to be chosen carefully. Since this map is monotone,

$$\begin{aligned}
\Pr[S_n \geq \mu + \lambda)] &= \Pr[e^{tS_n} \geq e^{t(\mu+\lambda)}] \\
&\leq \frac{\mathbb{E}[e^{tS_n}]}{e^{t(\mu+\lambda)}} \quad \text{(using Markov's inequality)} \\
&= \frac{\prod_i \mathbb{E}[e^{tX_i}]}{e^{t(\mu+\lambda)}} \quad \text{(using independence)} \qquad (10.11)
\end{aligned}$$

Bernoulli random variables: Assume that all the $X_i \in \{0,1\}$; we will remove this assumption later. Let the mean be $\mu_i = \mathbb{E}[X_i]$, so the *moment generating function* can be explicitly computed as

$$\mathbb{E}[e^{tX_i}] = 1 + \mu_i(e^t - 1) \leq \exp(\mu_i(e^t - 1)).$$

Substituting, we get

$$\begin{aligned}
\Pr[S_n \geq \mu + \lambda] &\leq \frac{\prod_i \mathbb{E}[e^{tX_i}]}{e^{t(\mu+\lambda)}} && (10.12)\\
&\leq \frac{\prod_i \exp(\mu_i(e^t - 1))}{e^{t(\mu+\lambda)}} && (10.13)\\
&\leq \frac{\exp(\mu(e^t - 1))}{e^{t(\mu+\lambda)}} \qquad \text{(since } \mu = \sum_i \mu_i) \\
&= \exp(\mu(e^t - 1) - t(\mu + \lambda)). && (10.14)
\end{aligned}$$

Since this calculation holds for all positive t, and we want the tightest upper bound, we should minimize the expression (10.14). Setting the derivative w.r.t. t to zero gives $t = \ln(1 + \lambda/\mu)$ which is non-negative for $\lambda \geq \mu$.

$$\Pr[S_n \geq \mu + \lambda] \leq \frac{e^{\lambda}}{(1 + \lambda/\mu)^{\mu+\lambda}}. \tag{10.15}$$

This bound on the upper tail is also one to be kept in mind; it often is useful when we are interested in large deviations where $\lambda \gg \mu$. One such example will be the load-balancing application with jobs and machines.

If we define $\beta := \lambda/\mu$ as the deviation in multiples of the mean, this quantity is

$$\Pr[S_n \geq \mu + \lambda] \leq \left(\frac{e^{\beta}}{(1+\beta)^{1+\beta}} \right)^{\mu}, \tag{10.16}$$

which is an expression that may be easy to deal with/memorize. And we can simplify even further: since

$$\frac{\beta}{1 + \beta/2} \leq \ln(1+\beta) \tag{10.17}$$

for all $\beta \geq 0$, so we get

$$(10.16) \overset{(10.17)}{\leq} \exp\left\{ \frac{-\beta^2 \mu}{2+\beta} \right\} = \exp\left\{ \frac{-\lambda^2}{2\mu + \lambda} \right\},$$

where the last expression follows by algebraic manipulation. This proves the upper tail bound (10.8); a similar proof gives us the lower tail as well.

Removing the assumption that $X_i \in \{0, 1\}$: If the r.v.s are not Bernoullis, then we define new Bernoulli r.v.s $Y_i \sim \text{Bernoulli}(\mu_i)$, which take value 0 with probability $1 - \mu_i$, and value 1 with probability μ_i, so that $\mathbb{E}[X_i] = \mathbb{E}[Y_i]$. Note that $f(x) = e^{tx}$ is convex for every value of $t \geq 0$; hence the function $\ell(x) = (1-x) \cdot f(0) + x \cdot f(1)$ satisfies $f(x) \leq \ell(x)$ for all $x \in [0, 1]$. Hence $\mathbb{E}[f(X_i)] \leq \mathbb{E}[\ell(X_i)]$; moreover $\ell(x)$ is a linear function so $\mathbb{E}[\ell(X_i)] = \ell(\mathbb{E}[X_i]) = \mathbb{E}[\ell(Y_i)]$, since X_i and Y_i have the same mean. Finally, $\ell(y) = f(y)$ for $y \in \{0, 1\}$. Putting all this together,

$$\mathbb{E}[e^{tX_i}] \leq \mathbb{E}[e^{tY_i}] = 1 + \mu_i(e^t - 1) \leq \exp(\mu_i(e^t - 1)),$$

so the step from (10.12) to (10.13) goes through again. This completes the proof of Theorem 10.8. ☐

Since the proof has a few steps, let's take stock of what we did:

i. Apply Markov's inequality on the function e^{tX},
ii. Use independence and linearity of expectations to break into e^{tX_i},
iii. Reduce to the Bernoulli case $X_i \in \{0, 1\}$,
iv. Compute the MGF (moment generating function) $\mathbb{E}[e^{tX_i}]$,
v. Choose t to minimize the resulting bound, and
vi. Use convexity to argue that Bernoullis are the "worst case".

You can get tail bounds for other functions of random variables by varying this template around; e.g., we will see an application for sums of independent normal (a.k.a. Gaussian) random variables in the next chapter.

Do make sure you see why the bounds of Theorem 10.8 are impossible in general if we do not assume some kind of boundedness and independence.

10.3.2 *The Generic Chernoff Bound*

Let's consider the case where the r.v.s X_i are identically distributed. Suppose we start off the same, and get to (10.11). Now define the log-MGF of the underlying r.v. X to be

$$\psi(t) := \mathbb{E}[e^{tX}]. \tag{10.18}$$

The expression (10.11) can be then written as

$$\exp\big(n\,\psi(t) - t(\mu+\lambda)\big) = \exp(-n(t(\mu+\lambda)/n - \psi(t))).$$

The tightest upper bound is obtained when the expression $t\lambda/n - \psi(t)$ is the largest. The *Legendre-Fenchel dual* of the function $\psi(t)$ is defined as

$$\psi^*(\lambda) := \sup_{t\geq 0}\, \{t\lambda - \psi(t)\},$$

This is also called the convex conjugate. Since it is the max of a collection of linear functions, one for each t, the dual function ψ^* is always convex, even if the original function ψ is not.

Exercise: if $\psi_1(t) \geq \psi_2(t)$ for all $t \geq 0$, then $\psi_1^*(\lambda) \leq \psi_2^*(\lambda)$ for all λ.

so we get the following concise statement, which we call the *generic Chernoff bound*:

Theorem 10.10 (Generic Chernoff Bound)**.** *Suppose S_n is the sum of n i.i.d. random variables, each having log-MGF $\psi(t)$. Let $\mu := \mathbb{E}[S_n]$. Then*

$$\Pr[S_n \geq \mu + \lambda] \leq \exp\left(-n\cdot\psi^*\left(\frac{\mu+\lambda}{n}\right)\right). \tag{10.19}$$

For the rest of the proof of the Chernoff bound, we can just focus on computing the dual $\psi^*(\lambda)$ of the log-MGF $\psi(t)$. Let's see some examples:

1. The first example is when $X \sim N(0,\sigma^2)$, then

$$\begin{aligned}\mathbb{E}[e^{tX)}] &= \frac{1}{\sqrt{2\pi}\sigma}\int_{x\in\mathbb{R}} e^{tx}e^{-\frac{x^2}{2\sigma^2}}dx \\ &= e^{t^2\sigma^2/2}\cdot\frac{1}{\sqrt{2\pi}\sigma}\int_{x\in\mathbb{R}} e^{-\frac{(x-t\sigma^2)^2}{2\sigma^2}}dx = e^{t^2\sigma^2/2}.\end{aligned} \tag{10.20}$$

   Hence, for $X \sim N(0,\sigma^2)$ r.v.s, we have

$$\psi(t) = \frac{t^2\sigma^2}{2} \qquad \text{and} \qquad \psi^*(\lambda) = \frac{\lambda^2}{2\sigma^2},$$

   the latter by basic calculus. Now the generic Chernoff bound (10.19) for the sum of n normal $N(0,\sigma^2)$ variables says:

$$\Pr[S_n \geq \lambda] \leq e^{-\frac{\lambda^2}{2n\sigma^2}}. \tag{10.21}$$

   This is even interesting when $n = 1$, in which case we get that for a $N(0,\sigma^2)$ random variable G,

$$\Pr[G \geq \lambda] \leq e^{-\frac{\lambda^2}{2\sigma^2}}. \tag{10.22}$$

In fact, you may have noticed that for Gaussians, the two statements (10.21) and (10.22) are equivalent, using the fact that the sum of n independent $N(0,\sigma^2)$ r.v.s is itself a $N(0,n\sigma^2)$ r.v..

2. How about a Rademacher $\{-1, +1\}$-valued r.v. X? The MGF is

$$\mathbb{E}[e^{tX}] = \frac{e^t + e^{-t}}{2} = \cosh t = 1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \cdots \leq e^{t^2/2},$$

so

$$\psi(t) = \frac{t^2}{2} \qquad \text{and} \qquad \psi^*(\lambda) = \frac{\lambda^2}{2}.$$

Note that

$$\psi_{\text{Rademacher}}(t) \leq \psi_{N(0,1)}(t) \implies \psi^*_{\text{Rademacher}}(\lambda) \geq \psi^*_{N(0,1)}(\lambda).$$

This means the upper tail bound for a single Rademacher is at least as strong as that for the standard normal.

3. And what about a centered Bernoulli with bias p? The log-MGF is

$$\psi(t) := \log \mathbb{E}[e^{tX}] = \log((1-p) + pe^t),$$

and a little calculus shows that the dual is

$$\psi^*(\lambda) = \lambda \log \frac{\lambda}{p} + (1-\lambda) \log \frac{1-\lambda}{1-p}.$$

Interestingly this function has a name: it is *Kullback-Leibler divergence* $D_{KL}(\lambda \| p)$ between two Bernoulli distributions, one with bias λ and the other with bias p. In summary, if we write $\mu + \lambda = qn$ for some $q > p$, we have

$$\Pr[S_n \geq qn] \leq e^{-nD_{KL}(q\|p)}.$$

The KL divergence $D_{KL}(q\|p)$, also called the *relative entropy*, is a distance measure between two distributions. It is not symmetric, so be careful with the order of the arguments! We will see more of it when we discuss online learning and mirror descent.

We can also extend the generic Chernoff bound to sums of non-identical distributions using the AM-GM inequality: details here.

10.3.3 The Examples Again: New and Improved Bounds

Example 1 (Coin Flips): Since each r.v. is a Bernoulli(p), the sum $S_n = \sum_i X_i$ has mean $\mu = np$, and hence

$$\Pr\left[|S_n - np| \geq \beta n\right] \leq \exp\left(-\frac{\beta^2 n}{2p+\beta}\right) \leq \exp\left(-\frac{\beta^2 n}{2}\right).$$

(For the second inequality, we use that the interesting settings have $p + \beta \leq 1$.) Hence, if $n \geq \frac{2\ln(1/\delta)}{\beta^2}$, the empirical average S_n/n is within an additve β of the bias p with probability at least $1-\delta$. This has an exponentially better dependence on $1/\delta$ than the bound we obtained from Chebychev's inequality.

This is asymptotically the correct answer: consider the problem where we have n coins, $n-1$ of them having bias $1/2$, and one having bias $1/2 + 2\beta$. We want to find the higher-bias coin. One way is to estimate the bias of each coin to within β with confidence $1 - \frac{1}{2n}$, using the procedure above—which takes $O(\log n/\varepsilon^2)$ flips per coin—and then take a union bound. It turns out any algorithm needs $\frac{\Omega(n \log n)}{\varepsilon^2}$ flips, so this the bound we have is tight. .

Example 2 (Load Balancing): Since the load L_i on any bin i behaves like $\mathrm{Bin}(n, 1/n)$, the expected load is 1. Now (10.8) says:

$$\Pr[L_i \geq 1 + \lambda] \leq \exp\left(- \frac{\lambda^2}{2+\lambda} \right).$$

If we set $\lambda = \Theta(\log n)$, the probability of the load L_i being larger than $1 + \lambda$ is at most $1/n^2$. Now taking a union bound over all bins, the probability that any bin receives at least $1 + \lambda$ balls is at most $\frac{1}{n}$. I.e., the maximum load is $O(\log n)$ balls with high probability.

In fact, the correct answer is that the maximum load is $(1 + o(1))\frac{\ln n}{\ln \ln n}$ with high probability. For example, the proofs in cite show this. Getting this precise bound requires a bit more work, but we can get an asymptotically correct bound by using (10.15) instead, with a setting of $\lambda = \frac{C \ln n}{\ln \ln n}$ with a large constant C.

Moreover, this shows that the asymmetry in the bounds (10.8) and (10.9) is essential. A first reaction would have been to believe our proof to be weak, and to hope for a better proof to get

$$\Pr[S_n \geq (1+\beta)\mu] \leq \exp(-\beta^2 \mu / c)$$

for some constant $c > 0$, for all values of β. This is not possible, however, because it would imply a max-load of $\Theta(\sqrt{\log n})$ with high probability.

The situation where $\lambda \leq \mu$ is often called the *Gaussian regime*, since the bound on the upper tail behaves like $\exp(-\lambda^2/\mu) = \exp(-\beta^2 \mu)$, with $\beta = \lambda/\mu$. In other cases, the upper tail bound behaves like $\exp(-\lambda)$, and is said to be the *Poisson regime*.

Example 3 (Random Walk): In this case, the variables are $[-1, 1]$ valued, and hence we cannot apply the bounds from Theorem 10.8 directly. But define $Y_i = \frac{1+X_i}{2}$ to get Bernoulli(1/2) variables, and define $T_n = \sum_{i=1}^n Y_i$. Since $T_n = S_n/2 + n/2$,

$$\begin{aligned}
\Pr\left[|S_n| \geq t\sqrt{n}\right] &= \Pr\left[|T_n - n/2| \geq (t/2)\sqrt{n}\right] \\
&\leq 2 \exp\left\{ - \frac{(t^2/n) \cdot (n/2)}{2 + \sqrt{t/n}} \right\} && \text{using (10.8)} \\
&\leq 2 \exp(-t^2/6).
\end{aligned}$$

In general, if X_i takes values in $[a, b]$, we can define $Y_i := \frac{X_i - a}{b-a}$ and then use Theorem 10.8.

Recall from §10.2.5 that the tail bound of $\approx \exp(-t^2/O(1))$ is indeed in the right ballpark.

10.4 Other concentration bounds

Many of the extensions address the various assumptions of Theorem 10.8: that the variables are bounded, that they are independent, and that the function S_n is the *sum* of these r.v.s. Add details and refs to this section.

But before we move on, let us give the bound that Sergei Bernstein gave in the 1920s: it uses knowledge about the variance of the random variable to get a potentially sharper bound than Theorem 10.8.

Theorem 10.11 (Bernstein's inequality)**.** *Consider n independent random variables $X_1, \ldots, X_n$ with $|X_i - \mathbb{E}[X_i]| \leq 1$ for each i. Let $S_n := \sum_i X_i$ have mean μ and variance σ^2. Then for any $\lambda \geq 0$ we have*

$$\Pr[|S_n - \mu| \geq \lambda] \leq 2 \exp\left(-\frac{\lambda^2}{2\sigma^2 + 2\lambda/3} \right).$$

10.4.1 *Mildly Correlated Variables*

The only place we used independence in the proof of Theorem 10.8 was in (10.11). So if we have some set of r.v.s where this inequality holds even without independence, the proof can proceed unchanged. Indeed, one such case is when the r.v.s are *negatively correlated*. Loosely speaking, this means that if some variables are "high" then it makes more likely for the other variables to be "low". Formally, $X_1, \ldots, X_n$ are *negatively associated* if for all disjoint sets A, B and for all monotone increasing functions f, g, we have

$$\mathbb{E}[f(X_i : i \in A) \cdot g(X_j : j \in B)] \leq \mathbb{E}[f(X_i : i \in A)] \cdot \mathbb{E}[g(X_j : j \in B)].$$

We can use this in the step (10.11), since the function e^{tx} is monotone increasing for $t > 0$.

Negative association arises in many settings: say we want to choose a subset S of k items out of a universe of size n, and let $X_i = \mathbf{1}_{i \in S}$ be the indicator for whether the i^{th} item is selected. The variables $X_1, \ldots, X_n$ are clearly not independent, but they are negatively associated.

10.4.2 *Martingales*

A different and powerful set of results can be obtained when we stop considering random variables are not independent, but allow variables X_j to take on values that depend on the past choices $X_1, X_2, \ldots, X_{j-1}$ but in a controlled way. One powerful formalization is the notion of a *martingale*. A *martingale difference sequence* is a sequence of r.v.s $Y_1, Y_2, \ldots, Y_n$, such that $\mathbb{E}[Y_i \mid Y_1, \ldots, Y_{i-1}] = 0$ for each i. (This is true for mean-zero independent r.v.s, but may be true in other settings too.)

Theorem 10.12 (Hoeffding-Azuma inequality)**.** *Let $Y_1, Y_2, \ldots, Y_n$ be a martingale difference sequence with $|Y_i| \leq c_i$ for each i, for constants c_i.*

Then for any $t \geq 0$,

$$\Pr\left[|\sum_{i=1}^{n} Y_i| \geq \lambda\right] \leq 2\exp\left(-\frac{\lambda^2}{2\sum_{i=1}^{n} c_i^2}\right).$$

For instance, applying the Azuma-Hoeffding bounds to the random walk in Example 3, where each Y_i is a Rademacher r.v. gives $\Pr[|S_n| \geq t\sqrt{n}] \leq 2e^{-t^2/8}$, which is very similar to the bounds we derived above. But we can also consider, e.g., a "bounded" random walk that starts at the origin, say, and stops whenever it reaches either $-\ell$ or $+r$. In this case, the step size $Y_i = 0$ with unit probability if $\sum_{j=1}^{i-1} Y_j \in \{-\ell, r\}$, else it is $\{\pm 1\}$ independently and uniformly at random.

10.4.3 *Going Beyond Sums of Random Variables*

The Azuma-Hoeffding inequality can be used to bound functions of $X_1, \ldots, X_n$ other than their sum—and there are many other bounds for more general classes of functions. In all these cases we want any single variable to affect the function only in a limited way—i.e., the function should be Lipschitz. One popular packaging was given by Colin McDiarmid:

Theorem 10.13 (McDiarmid's inequality)**.** *Consider n independent r.v.s $X_1, \ldots, X_n$, with X_i taking values in a set A_i for each i, and a function $f : \prod A_i \to \mathbb{R}$ satisfying $|f(x) - f(x')| \leq c_i$ whenever x and x' differ only in the i^{th} coordinate. Let $\mu := \mathbb{E}[f(X_1, \ldots, X_n)]$ be the expected value of the random variable $f(\overline{X})$. Then for any non-negative β,*

$$\textit{Upper tail}: \qquad \Pr[f(X) \geq \mu(1+\beta)] \leq \exp\left(-\frac{2\mu^2\beta^2}{\sum_i c_i^2}\right)$$

$$\textit{Lower tail}: \qquad \Pr[f(X) \leq \mu(1-\beta)] \leq \exp\left(-\frac{2\mu^2\beta^2}{\sum_i c_i^2}\right)$$

This inequality does not assume very much about the function, except it being c_i-Lipschitz in the i^{th} coordinate; hence we can also use this to the truncated random walk example above, or for many other applications.

10.4.4 *Moment Bounds vs. Chernoff-style Bounds*

One may ask how moment bounds relate to Chernoff-Hoeffding bounds: Philips and Nelson showed that Chernoff-style bounds obtained using the approach of bounding the moment-generating function are never stronger than moment bounds:

T.K. Philips and R. Nelson (1995)

Theorem 10.14. *Consider n independent random variables $X_1, \ldots, X_n$, each with mean 0. Let $S_n = \sum X_i$. Then*

$$\Pr[S_n \geq \lambda] \leq \min_{k \geq 0} \frac{\mathbb{E}[X^k]}{\lambda^k} \leq \inf_{t \geq 0} \frac{\mathbb{E}[e^{tX}]}{e^{t\lambda}}.$$

10.4.5 Matrix-Valued Random Variables

Finally, an important line of research considers concentration for vector-valued and matrix valued functions of independent (and mildly dependent) r.v.s. One object that we will see in a homework, and also in later applications, is the matrix-valued case: here the notation $A \succeq 0$ means the matrix is positive-semidefinite (i.e., all its eigenvalues are non-negative), and $A \succeq B$ means $A - B \succeq 0$. See, e.g., the lecture notes by Joel Tropp!

Theorem 10.15 (Matrix Chernoff bounds)**.** *Consider n independent symmetric matrices $X_1, \ldots, X_n$ of dimension d. Moreover, $I \succeq X_i \succeq 0$ for each i, i.e., the eigenvalues of each matrix are between 0 and 1. If $\mu_{max} := \lambda_{max}(\sum \mathbb{E}[X_i])$ is the largest eigenvalue of their expected sum, then*

$$\Pr\left[\lambda_{max}\left(\sum X_i\right) \geq \mu_{max} + \gamma\right] \leq d \;\exp\left(-\frac{\gamma^2}{2\mu_{max} + \gamma}\right).$$

As an example, if we are throwing n balls into n bins, then we can let matrix X_i have a single 1 at position (j, j) if the i^{th} ball falls into bin j, and zeros elsewhere. Now the sum of these matrices has the loads of the bins on the diagonal, and the maximum eigenvalue is precisely the highest load. This bound therefore gives that the probability of a bin with load $1 + \gamma$ is at most $n \cdot e^{\gamma^2/(2+\gamma)}$—again implying a maximum load of $O(\log n)$ with high probability.

But we can use this for a lot more than just diagonal matrices (which can be reasoned about using the scalar-valued Chernoff bounds, plus the naïve union bound). Indeed, we can sample edges of a graph at random, and then talk about the eigenvalues of the resulting adjacency matrix (or more interestingly, of the resulting Laplacian matrix) using these bounds. We will discuss this in a later chapter.

10.5 Application #1: Oblivious Routing on the Hypercube

Now we return to fourth application mentioned at the beginning of the chapter. (The first two applications have already been considered above, the third will be covered as a homework problem.)

The setting is the following: we are given the d-dimensional hypercube Q_d, with $n = 2^d$ vertices. We have $n = 2^d$ vertices, each labeled with a d-bit vector. Each vertex i has a single packet (which we also call packet i), destined for vertex $\pi(i)$, where π is a permutation on the nodes $[n]$.

Packets move in synchronous rounds. Each edge is bi-directed, and at most one packet can cross each directed edge in each round. Moreover, each packet can cross at most one edge per round. So if $uv \in E(Q_d)$, one packet can cross from u to v, and one from v to u, in a round. Each edge e has an associated waiting queue W_e; so each node has d queues, one for each edge leaving it. If several packets want to cross an edge e in the same round, only one can cross; the rest wait in the queue W_e and try again the next round. We assume the queues are allowed to grow to arbitrary size (though one can also show queue length bounds in the algorithm below). The goal is to get a simple routing scheme that delivers the packets in $O(d)$ rounds, no matter what permutation π needs to be routed.

One natural proposal is the *bit-fixing routing* scheme: each packet i looks at its current position u, finds the first bit position where u differs from $\pi(i)$, and flips the bit (which corresponds to traversing an edge out of u). For example:

$$0001010 \to 1001010 \to 1101010 \to 1100010 \to 1100011.$$

However, this proposal can create "congestion hotspots" in the network, and therefore delay some packets by $2^{\Omega(d)}$. In fact, it turns out any deterministic *oblivious* strategy (that does not depend on the actual sources and destinations) must have a delay of $\Omega(\sqrt{2^d/d})$ rounds.

Suppose we choose a permutation π such that

$$\pi(\mathbf{w0}) = \mathbf{0w},$$

where $\mathbf{w}, \mathbf{0} \in \{0,1\}^{d/2}$. All these $2^{d/2}$ packets have to pass through the all-zeros node in the bit-fixing routing scheme; since this node can send out at most d packets at each timestep, need at least $2^{d/2}/d$ rounds.

10.5.1 *A Randomized Algorithm...*

Here's a lovely randomized strategy, due to Les Valiant, and to Valiant and Brebner. It requires no centralized control, and is optimal in the sense of requiring $O(d)$ rounds (with high probability) on any permutation π.

Valiant (1982)

> Each node i picks a randomized midpoint R_i independently and uniformly from $[n]$: it sends its packet to R_i. Then after $5d$ rounds have elapsed, the packets proceed to their final destinations $\pi(i)$. All routing is done using bit-fixing.

10.5.2 *... and its Analysis*

Theorem 10.16. *The random midpoint algorithm above succeeds in delivering the packets in at most* $10d$ *rounds, with probability at least* $1 - \frac{2}{n}$.

Proof. We only prove that all packets reach their midpoints by time $5d$, with high probability. The argument for the second phase is then identical. Let P_i be the bit-fixing path from i to the midpoint R_i, and define

$$S_i := \{j \neq i \mid \text{ path } P_j \text{ shares an edge with } P_i\}.$$

Claim 10.17. Any two paths P_i and P_j intersect in one contiguous segment.

Proof. (Exercise.) This is where using a consistent routing strategy like bit-fixing helps. □

Claim 10.18. Packet i reaches midpoint R_i by time at most $|P_i| + |S_i|$.

Proof. Consider the path $P_i = \langle e_1, e_2, \dots, e_\ell \rangle$ taken by packet i. If S_i were empty, clearly packet i would reach its destination in time $|P_i|$; we now show how to charge each timestep that packet i is delayed to a distinct packet in S_i. For that, we first define the notion of *lag*. For any edge $e_k \in P_i$, we say every packet in W_{e_k} at the beginning of timestep t has lag $t - k$. Note that all packets in the same queue at the same time have the same lag. Now:

Observe: the lags are defined for packets in S_i according to the numbering of edges in P_i, not the numbering of their own paths.

1. Each packet j in $S_i \cup \{i\}$ either reaches its destination on P_i or it leaves P_i (forever, by Claim 10.17) after traversing some last edge $e_k \in P_i$. Call this traversal of e_k the *final traversal* for packet j, and call its lag value just before this final traversal its *final lag*.

2. Suppose packet i traverses the last edge e_ℓ on its path and reaches its destination at timestep T. Since it has lag $T - \ell = T - |P_i|$ just before it traverses the edge, it reaches the destination at time $|P_i|$ plus its final lag. So it suffices to show that i's final lag is at most $|S_i|$.

3. The initial lag (at time $t = 1$) of this packet i is $(1 - 1) = 0$, since it belongs to queue W_{e_1} at the very beginning. The lag of this packet never decreases over time as it makes its way along the path. Indeed, if it is in W_{e_k} at the beginning of some timestep t, and it traverses the edge, it now belongs to $w_{e_{k+1}}$ at the start of timestep $t + 1$, and its new lag is $(t + 1) - (k + 1) = t - k$ and therefore unchanged.

4. Else suppose packet i's lag increases from some value L to $L + 1$ at some timestep. This is because $i \in W_{e_k}$ for some k at the beginning of time $t = L + k$, but some other packet $j \in S_i$ from queue W_{e_k} was sent across the edge e_k at this timestep. In this case, imagine packet i gives packet j a *token* numbered L. So there is a single token generated for each increase in i's lag, each with a different number.

5. We show (in the next bullet point) how to maintain the invariant that at the beginning of each time, any token numbered L still on the path P_i is carried by some packet in S_i with current lag L. This implies that when a packet in S_i makes its final traversal and it has some final lag L', it is either carrying a single token numbered L' at that time or no token at all. Since each token is carried by some packet, this means there can be at most $|S_i|$ tokens overall, and hence i's final lag value is at most $|S_i|$.

6. To ensure the invariant, note that when j got the token numbered L from i, packet j had lag value L. Now as long as j does not get delayed as it proceeds along the path, its lag remains L (and it keeps the token). If it does get delayed, say while waiting in queue $W_{e_{k'}}$ while some other packet j' (having the same lag value L, because they were sharing the same queue) traverses the edge $e_{k'}$, packet j gives its token numbered L to this j'. This maintains the invariant.

□

Finally, we bound the size of S_i by a concentration bound. Since R_i is chosen uniformly at random from $\{0,1\}^d$, the labels of i and R_i differ in $d/2$ bits in expectation. Hence P_i has expected length $d/2$. There are $d2^d = dn$ (directed) edges, and all $n = 2^d$ paths behave symmetrically, so the expected number of paths P_j using any edge e is $\frac{n \cdot d/2}{dn} = 1/2$.

Claim 10.19. $\Pr[|S_i| \geq 4d] \leq e^{-2d}$.

Proof. If X_{ij} is the indicator of the event that P_i and P_j intersect, then $|S_i| = \sum_{j \neq i} X_{ij}$, i.e., it is a sum of a collection of independent $\{0,1\}$-valued random variables. Now conditioned on any choice of P_i (which is of length at most d), the expected number of paths using each edge in it is at most $1/2$, so the conditional expectation of S_i is at most $d/2$. Since this holds for any choice of P_i, the unconditional expectation $\mu = \mathbb{E}[S_i]$ is also at most $d/2$.

Now apply the Chernoff bound to S_i with $\lambda = 4d - \mu$ and $\mu \leq d/2$ to get

$$\Pr[|S_i| \geq 4d] \leq \exp\left\{-\frac{(4d-\mu)^2}{2\mu + (4d-\mu)}\right\} \leq e^{-2d}.$$

Note that we could apply the bound even though the variables X_{ij} were not i.i.d., and moreover we did not need estimates for $\mathbb{E}[X_{ij}]$, just an upper bound for their expected sum. □

Now applying a union bound over all $n = 2^d$ packets i means that all n packets reach their midpoints within $d + 4d$ steps with probability $1 - 2^d \cdot e^{-2d} \geq 1 - e^{-d} \geq 1 - 1/n$. Similarly, the second phase has a probability at most $1/n$ of failing to complete in $5d$ steps, completing the proof. □

A different strategy would be to let each packet pick a random permutation and fix the bits according to that permutation. Sadly, this approach gives delay $2^{\Omega(d)}$. This is true even if each node picks its permutation independently. One bad example appears in Valiant's original paper (see Section 5 "The Necessity for Phase A") and shows that you can fix a permutation that "gangs up" on some node, even if the bit-fixing order is random.

10.6 Application #2: Graph Sparsification

10.7 Application #3: The Power of Two Choices