9.3 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 9.8 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_n$ 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*

$$\textit{Upper tail}: \qquad \Pr\left[S_n \geq \mu + \lambda\right] \leq \exp\left\{-\frac{\lambda^2}{2\mu + \lambda}\right\}. \tag{9.8}$$

$$\textit{Lower tail}: \qquad \Pr\left[S_n \leq \mu - \lambda\right] \leq \exp\left\{-\frac{\lambda^2}{3\mu}\right\}. \tag{9.9}$$

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 9.9 (Double-Sided Concentration Bound)**.** *For $X_1, \dots, X_n$ 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 $\beta \in [0, 1]$,*

$$\Pr\left[S_n \notin \mu(1 \pm \beta)\right] \leq 2\, e^{-\beta^2 \mu/3}. \tag{9.10}$$

9.3.1 The Proof

Proof of Theorem 9.8. We only prove (9.8); the proof for (9.9) 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,

$$\begin{aligned}
\Pr[S_n \geq \mu + \lambda)] &= \Pr[e^{tS_n} \geq e^{t(\mu+\lambda)}] \\
&\leq \frac{\mathbb{E}[e^{tS_n}]}{e^{t(\mu+\lambda)}} \quad \text{(using Markov's inequality)} \\
&= \frac{\prod_i \mathbb{E}[e^{tX_i}]}{e^{t(\mu+\lambda)}} \quad \text{(using independence)}
\end{aligned} \tag{9.11}$$

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

$$\Pr[S_n \geq \mu + \lambda] \leq \frac{\prod_i \mathbb{E}[e^{tX_i}]}{e^{t(\mu+\lambda)}} \tag{9.12}$$

$$\leq \frac{\prod_i \exp(\mu_i(e^t - 1))}{e^{t(\mu+\lambda)}} \tag{9.13}$$

$$\leq \frac{\exp(\mu(e^t - 1))}{e^{t(\mu+\lambda)}} \qquad (\text{since } \mu = \sum_i \mu_i)$$

$$= \exp(\mu(e^t - 1) - t(\mu + \lambda)). \tag{9.14}$$

Since this calculation holds for all positive t, and we want the tightest upper bound, we should minimize the expression (9.14). Setting the derivative w.r.t. t to zero gives $t = \ln(1 + \lambda/\mu)$ which is non-negative for $\lambda \geq \mu$.

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.

$$\Pr[S_n \geq \mu + \lambda] \leq \frac{e^{\lambda}}{(1 + \lambda/\mu)^{\mu+\lambda}}. \tag{9.15}$$

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

$$\Pr[S_n \geq \mu + \lambda] \leq \left(\frac{e^{\beta}}{(1+\beta)^{1+\beta}}\right)^{\mu}, \tag{9.16}$$

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

$$\frac{\beta}{1 + \beta/2} \leq \ln(1 + \beta) \tag{9.17}$$

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

$$(9.16) \overset{(9.17)}{\leq} \exp\left\{\frac{-\beta^2\mu}{2+\beta}\right\} = \exp\left\{\frac{-\lambda^2}{2\mu + \lambda}\right\},$$

where the last expression follows by algebraic manipulation. This proves the upper tail bound (9.8); 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 (9.12) to (9.13) goes through again. This completes the proof of Theorem 9.8. □

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 9.8 are impossible in general if we do not assume some kind of boundedness and independence.

9.3.2 *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 (9.11). Now define the log-MGF of the underlying r.v. X to be

$$\psi(t) := \mathbb{E}[e^{tX}]. \tag{9.18}$$

The expression (9.11) 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

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

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

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.

Theorem 9.10 (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*

$$\Pr[S_n \geq \mu + \lambda] \leq \exp\left(-n \cdot \psi^*\left(\frac{\mu+\lambda}{n}\right)\right). \tag{9.19}$$

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

$$\begin{aligned}\mathbb{E}[e^{tX)}] &= \frac{1}{\sqrt{2\pi}\sigma}\int_{x\in\mathbb{R}} e^{tx}e^{-\frac{x^2}{2\sigma^2}}dx \\ &= e^{t^2\sigma^2/2}\cdot\frac{1}{\sqrt{2\pi}\sigma}\int_{x\in\mathbb{R}} e^{-\frac{(x-t\sigma^2)^2}{2\sigma^2}}dx = e^{t^2\sigma^2/2}.\end{aligned} \tag{9.20}$$

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 (9.19) for the sum of n normal $N(0, \sigma^2)$ variables says:

$$\Pr[S_n \geq \lambda] \leq e^{-\frac{\lambda^2}{2n\sigma^2}}. \tag{9.21}$$

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

$$\Pr[G \geq \lambda] \leq e^{-\frac{\lambda^2}{2\sigma^2}}. \tag{9.22}$$

In fact, you may have noticed that for Gaussians, the two statements (9.21) and (9.22) 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.

9.3.3 *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

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

(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 Bin$(n, 1/n)$, the expected load is 1. Now (9.8) 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 (9.15) 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 (9.8) and (9.9) 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 9.8