CS229 Supplemental Lecture notes
Hoeffding's inequality

John Duchi

1 Basic probability bounds

A basic question in probability, statistics, and machine learning is the following: given a random variable Z with expectation $\mathbb{E}[Z]$, how likely is Z to be close to its expectation? And more precisely, how close is it likely to be? With that in mind, these notes give a few tools for computing bounds of the form

$$\mathbb{P}(Z \geq \mathbb{E}[Z] + t) \quad \text{and} \quad \mathbb{P}(Z \leq \mathbb{E}[Z] - t) \tag{1}$$

for $t \geq 0$.

Our first bound is perhaps the most basic of all probability inequalities, and it is known as Markov's inequality. Given its basic-ness, it is perhaps unsurprising that its proof is essentially only one line.

Proposition 1 (Markov's inequality)**.** *Let $Z \geq 0$ be a non-negative random variable. Then for all $t \geq 0$,*

$$\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}[Z]}{t}.$$

Proof We note that $\mathbb{P}(Z \geq t) = \mathbb{E}[\mathbf{1}\{Z \geq t\}]$, and that if $Z \geq t$, then it must be the case that $Z/t \geq 1 \geq \mathbf{1}\{Z \geq t\}$, while if $Z < t$, then we still have $Z/t \geq 0 = \mathbf{1}\{Z \geq t\}$. Thus

$$\mathbb{P}(Z \geq t) = \mathbb{E}[\mathbf{1}\{Z \geq t\}] \leq \mathbb{E}\left[\frac{Z}{t}\right] = \frac{\mathbb{E}[Z]}{t},$$

as desired. $\square$

Essentially all other bounds on the probabilities (1) are variations on Markov's inequality. The first variation uses second moments—the variance—of a random variable rather than simply its mean, and is known as Chebyshev's inequality.

Proposition 2 (Chebyshev's inequality)**.** *Let Z be any random variable with* $\mathrm{Var}(Z) < \infty$*. Then*

$$\mathbb{P}(Z \geq \mathbb{E}[Z] + t \text{ or } Z \leq \mathbb{E}[Z] - t) \leq \frac{\mathrm{Var}(Z)}{t^2}$$

for $t \geq 0$*.*

Proof The result is an immediate consequence of Markov's inequality. We note that if $Z \geq \mathbb{E}[Z] + t$, then certainly we have $(Z - \mathbb{E}[Z])^2 \geq t^2$, and similarly if $Z \leq \mathbb{E}[Z] - t$ we have $(Z - \mathbb{E}[Z])^2 \geq t^2$. Thus

$$\begin{aligned}\mathbb{P}(Z \geq \mathbb{E}[Z] + t \text{ or } Z \leq \mathbb{E}[Z] - t) &= \mathbb{P}((Z - \mathbb{E}[Z])^2 \geq t^2) \\ &\stackrel{(i)}{\leq} \frac{\mathbb{E}[(Z - \mathbb{E}[Z])^2]}{t^2} = \frac{\mathrm{Var}(Z)}{t^2},\end{aligned}$$

where step (i) is Markov's inequality. □

A nice consequence of Chebyshev's inequality is that averages of random variables with finite variance converge to their mean. Let us give an example of this fact. Suppose that Z_i are i.i.d. and satisfy $\mathbb{E}[Z_i] = 0$. Then $\mathbb{E}[Z_i] = 0$, while if we define $\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$ then

$$\mathrm{Var}(\bar{Z}) = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Z_i\right)^2\right] = \frac{1}{n^2}\sum_{i,j \leq n} \mathbb{E}[Z_i Z_j] = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[Z_i^2] = \frac{\mathrm{Var}(Z_1)}{n}.$$

In particular, for any $t \geq 0$ we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i\right| \geq t\right) \leq \frac{\mathrm{Var}(Z_1)}{nt^2},$$

so that $\mathbb{P}(|\bar{Z}| \geq t) \to 0$ for any $t > 0$.

2 Moment generating functions

Often, we would like sharper—even exponential—bounds on the probability that a random variable Z exceeds its expectation by much. With that in mind, we need a stronger condition than finite variance, for which moment generating functions are natural candidates. (Conveniently, they also play nicely with sums, as we will see.) Recall that for a random variable Z, the *moment generating function* of Z is the function

$$M_Z(\lambda) := \mathbb{E}[\exp(\lambda Z)], \tag{2}$$

which may be infinite for some λ.

2.1 Chernoff bounds

Chernoff bounds use of moment generating functions in an essential way to give exponential deviation bounds.

Proposition 3 (Chernoff bounds)**.** *Let Z be any random variable. Then for any $t \geq 0$,*

$$\mathbb{P}(Z \geq \mathbb{E}[Z] + t) \leq \min_{\lambda \geq 0} \mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}]e^{-\lambda t} = \min_{\lambda \geq 0} M_{Z-\mathbb{E}[Z]}(\lambda)e^{-\lambda t}$$

and

$$\mathbb{P}(Z \leq \mathbb{E}[Z] - t) \leq \min_{\lambda \geq 0} \mathbb{E}[e^{\lambda(\mathbb{E}[Z] - Z)}]e^{-\lambda t} = \min_{\lambda \geq 0} M_{\mathbb{E}[Z]-Z}(\lambda)e^{-\lambda t}.$$

Proof We only prove the first inequality, as the second is completely identical. We use Markov's inequality. For any $\lambda > 0$, we have $Z \geq \mathbb{E}[Z] + t$ if and only if $e^{\lambda Z} \geq e^{\lambda \mathbb{E}[Z] + \lambda t}$, or $e^{\lambda(Z - \mathbb{E}[Z])} \geq e^{\lambda t}$. Thus, we have

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) = \mathbb{P}(e^{\lambda(Z - \mathbb{E}[Z])} \geq e^{\lambda t}) \stackrel{(i)}{\leq} \mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}]e^{-\lambda t},$$

where the inequality (i) follows from Markov's inequality. As our choice of $\lambda > 0$ did not matter, we can take the best one by minizing the right side of the bound. (And noting that certainly the bound holds at $\lambda = 0$.) $\square$

The important result is that Chernoff bounds "play nicely" with summations, which is a consequence of the moment generating function. Let us assume that Z_i are independent. Then we have that

$$M_{Z_1+\cdots+Z_n}(\lambda) = \prod_{i=1}^{n} M_{Z_i}(\lambda),$$

which we see because

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} Z_i\right)\right] = \mathbb{E}\left[\prod_{i=1}^{n} \exp(\lambda Z_i)\right] = \prod_{i=1}^{n} \mathbb{E}[\exp(\lambda Z_i)],$$

by of the independence of the Z_i. This means that when we calculate a Chernoff bound of a sum of i.i.d. variables, we need only calculate the moment generating function for *one* of them. Indeed, suppose that Z_i are i.i.d. and (for simplicity) mean zero. Then

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{n} Z_i \geq t\right) &\leq \frac{\prod_{i=1}^{n} \mathbb{E}\left[\exp(\lambda Z_i)\right]}{e^{\lambda t}} \\ &= (\mathbb{E}[e^{\lambda Z_1}])^n e^{-\lambda t},\end{aligned}$$

by the Chernoff bound.

2.2 Moment generating function examples

Now we give several examples of moment generating functions, which enable us to give a few nice deviation inequalities as a result. For all of our examples, we will have very convienent bounds of the form

$$\boxed{M_Z(\lambda) = \mathbb{E}[e^{\lambda Z}] \leq \exp\left(\frac{C^2\lambda^2}{2}\right) \quad \text{for all } \lambda \in \mathbb{R},}$$

for some $C \in \mathbb{R}$ (which depends on the distribution of Z); this form is *very* nice for applying Chernoff bounds.

We begin with the classical normal distribution, where $Z \sim \mathcal{N}(0, \sigma^2)$. Then we have

$$\mathbb{E}[\exp(\lambda Z)] = \exp\left(\frac{\lambda^2\sigma^2}{2}\right),$$

which one obtains via a calculation that we omit. (You should work this out if you are curious!)

A second example is known as a Rademacher random variable, or the random sign variable. Let $S = 1$ with probability $\frac{1}{2}$ and $S = -1$ with probability $\frac{1}{2}$. Then we claim that

$$\mathbb{E}[e^{\lambda S}] \le \exp\left(\frac{\lambda^2}{2}\right) \quad \text{for all } \lambda \in \mathbb{R}. \tag{3}$$

To see inequality (3), we use the Taylor expansion of the exponential function, that is, that $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$. Note that $\mathbb{E}[S^k] = 0$ whenever k is odd, while $\mathbb{E}[S^k] = 1$ whenever k is even. Then we have what

$$\begin{aligned}\mathbb{E}[e^{\lambda S}] &= \sum_{k=0}^{\infty} \frac{\lambda^k \mathbb{E}[S^k]}{k!} \\ &= \sum_{k=0,2,4,\ldots} \frac{\lambda^k}{k!} = \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!}.\end{aligned}$$

Finally, we use that $(2k)! \ge 2^k \cdot k!$ for all $k = 0, 1, 2, \ldots$, so that

$$\mathbb{E}[e^{\lambda S}] \le \sum_{k=0}^{\infty} \frac{(\lambda^2)^k}{2^k \cdot k!} = \sum_{k=0}^{\infty} \left(\frac{\lambda^2}{2}\right)^k \frac{1}{k!} = \exp\left(\frac{\lambda^2}{2}\right).$$

Let us apply inequality (3) in a Chernoff bound to see how large a sum of i.i.d. random signs is likely to be.

We have that if $Z = \sum_{i=1}^{n} S_i$, where $S_i \in \{\pm 1\}$ is a random sign, then $\mathbb{E}[Z] = 0$. By the Chernoff bound, it becomes immediately clear that

$$\mathbb{P}(Z \ge t) \le \mathbb{E}[e^{\lambda Z}] e^{-\lambda t} = \mathbb{E}[e^{\lambda S_1}]^n e^{-\lambda t} \le \exp\left(\frac{n\lambda^2}{2}\right) e^{-\lambda t}.$$

Applying the Chernoff bound technique, we may minimize this in $\lambda \ge 0$, which is equivalent to finding

$$\min_{\lambda \ge 0} \left\{ \frac{n\lambda^2}{2} - \lambda t \right\}.$$

Luckily, this is a convenient function to minimize: taking derivatives and setting to zero, we have $n\lambda - t = 0$, or $\lambda = t/n$, which gives

$$\mathbb{P}(Z \ge t) \le \exp\left(-\frac{t^2}{2n}\right).$$

In particular, taking $t = \sqrt{2n \log \frac{1}{\delta}}$, we have

$$\mathbb{P}\left(\sum_{i=1}^{n} S_i \geq \sqrt{2n \log \frac{1}{\delta}}\right) \leq \delta.$$

So $Z = \sum_{i=1}^{n} S_i = O(\sqrt{n})$ with extremely high probability—the sum of n independent random signs is essentially never larger than $O(\sqrt{n})$.

3 Hoeffding's lemma and Hoeffding's inequality

Hoeffding's inequality is a powerful technique—perhaps the most important inequality in learning theory—for bounding the probability that sums of bounded random variables are too large or too small. We will state the inequality, and then we will prove a weakened version of it based on our moment generating function calculations earlier.

Theorem 4 (Hoeffding's inequality)**.** *Let $Z_1, \ldots, Z_n$ be independent bounded random variables with $Z_i \in [a, b]$ for all i, where $-\infty < a \leq b < \infty$. Then*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

and

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \leq -t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

for all $t \geq 0$.

We prove Theorem 4 by using a combination of (1) Chernoff bounds and (2) a classic lemma known as Hoeffding's lemma, which we now state.

Lemma 5 (Hoeffding's lemma)**.** *Let Z be a bounded random variable with $Z \in [a, b]$. Then*

$$\mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right) \quad \text{for all } \lambda \in \mathbb{R}.$$

Proof We prove a slightly weaker version of this lemma with a factor of 2 instead of 8 using our random sign moment generating bound and an inequality known as *Jensen's inequality* (we will see this very important inequality later in our derivation of the EM algorithm). Jensen's inequality states the following: if $f : \mathbb{R} \to \mathbb{R}$ is a *convex* function, meaning that f is bowl-shaped, then

$$f(\mathbb{E}[Z]) \le \mathbb{E}[f(Z)].$$

The simplest way to remember this inequality is to think of $f(t) = t^2$, and note that if $\mathbb{E}[Z] = 0$ then $f(\mathbb{E}[Z]) = 0$, while we generally have $\mathbb{E}[Z^2] > 0$. In any case, $f(t) = \exp(t)$ and $f(t) = \exp(-t)$ are convex functions.

We use a clever technique in probability theory known as *symmetrization* to give our result (you are not expected to know this, but it is a very common technique in probability theory, machine learning, and statistics, so it is good to have seen). First, let Z' be an independent copy of Z with the same distribution, so that $Z' \in [a, b]$ and $\mathbb{E}[Z'] = \mathbb{E}[Z]$, but Z and Z' are independent. Then

$$\mathbb{E}_Z[\exp(\lambda(Z - \mathbb{E}_Z[Z]))] = \mathbb{E}_Z[\exp(\lambda(Z - \mathbb{E}_{Z'}[Z']))] \overset{(i)}{\le} \mathbb{E}_Z[\mathbb{E}_{Z'} \exp(\lambda(Z - Z'))],$$ f(z') = exp(\lambda (z-expectation_z'[z']))

where $\mathbb{E}_Z$ and $\mathbb{E}_{Z'}$ indicate expectations taken with respect to Z and Z'. Here, step (i) uses Jensen's inequality applied to $f(x) = e^{-x}$. Now, we have

$$\mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))] \le \mathbb{E}\left[\exp\left(\lambda(Z - Z')\right)\right].$$ Think it is missing something here, missing a E perhaps?

Now, we note a curious fact: the difference $Z - Z'$ is symmetric about zero, so that if $S \in \{-1, 1\}$ is a random sign variable, then $S(Z - Z')$ has exactly the same distribution as $Z - Z'$. So we have

$$\begin{aligned}
\mathbb{E}_{Z,Z'}[\exp(\lambda(Z - Z'))] &= \mathbb{E}_{Z,Z',S}[\exp(\lambda S(Z - Z'))] \\
&= \mathbb{E}_{Z,Z'}\left[\mathbb{E}_S\left[\exp(\lambda S(Z - Z')) \mid Z, Z'\right]\right].
\end{aligned}$$

Now we use inequality (3) on the moment generating function of the random sign, which gives that

$$\mathbb{E}_S\left[\exp(\lambda S(Z - Z')) \mid Z, Z'\right] \le \exp\left(\frac{\lambda^2(Z - Z')^2}{2}\right).$$

But of course, by assumption we have $|Z - Z'| \le (b - a)$, so $(Z - Z')^2 \le (b - a)^2$. This gives

$$\mathbb{E}_{Z,Z'}[\exp(\lambda(Z - Z'))] \le \exp\left(\frac{\lambda^2(b - a)^2}{2}\right).$$

This is the result (except with a factor of 2 instead of 8). $\square$

Now we use Hoeffding's lemma to prove Theorem 4, giving only the upper tail (i.e. the probability that $\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq t$) as the lower tail has a similar proof. We use the Chernoff bound technique, which immediately tells us that

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq t\right) &= \mathbb{P}\left(\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq nt\right) \\
&\leq \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i])\right)\right]e^{-\lambda nt} \\
&= \left(\prod_{i=1}^{n}\mathbb{E}[e^{\lambda(Z_i - \mathbb{E}[Z_i])}]\right)e^{-\lambda nt} \stackrel{(i)}{\leq} \left(\prod_{i=1}^{n}e^{\frac{\lambda^2(b-a)^2}{8}}\right)e^{-\lambda nt}
\end{aligned}$$

where inequality (i) is Hoeffding's Lemma (Lemma 5). Rewriting this slightly and minimzing over $\lambda \geq 0$, we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq t\right) \leq \min_{\lambda \geq 0}\exp\left(\frac{n\lambda^2(b-a)^2}{8} - \lambda nt\right) = \exp\left(-\frac{2nt^2}{(b-a)^2}\right),$$

as desired.