Tail Bounds: Chebyshev, WLLN and Chernoff

Lecturer: Ankur Moitra

In this section, we will develop tools for bounding the probability that a random variable deviates from its expectation. These bounds will progress from ones that use only relatively weak assumptions, to ones that use stronger assumptions but give much tighter control.

1 Chebyshev and Markov

Our first tail bound is known as Chebyshev's inequality. Suppose you know the mean and variance of a random variable f. Is there some way that you can put a bound on the probability that the random variable is a long way away from its mean? This is exactly what Chebyshev's inequality does.

Theorem 1 (Chebyshev's Inequality)**.** *Let $X : S \to \mathbb{R}$ be a random variable with expectation $\mathbb{E}(X)$ and variance $Var(X)$. Then, for any $a > 0$,*

$$\mathbb{P}\Big(\big|X - \mathbb{E}(X)\big| \geq a\Big) \leq \frac{\mathrm{Var}(X)}{a^2}.$$

We could prove Chebychev's inequality from first principles, but we can also derive it easily from Markov's inequality which applies only to non-negative random variables and gives us a bound depending on the expectation of the random variable.

Theorem 2 (Markov's Inequality)**.** *Let $X : S \to \mathbb{R}$ be a non-negative random variable. Then, for any $a > 0$,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

The intuition behind the proof is simple: If the probability that $X \geq a$ were too large, then it would contribute too much to the expectation and since the random variable is nonnegative its contribution cannot be cancelled out.

Proof. Let A denote the event $\{X \geq a\}$. Then

$$\mathbb{E}(X) = \sum_{s \in S} p(s)X(s) = \sum_{x \in A} p(s)X(s) + \sum_{s \in \bar{A}} p(s)X(s).$$

As X is non-negative, we have $\sum_{s \in \neg A} p(s)X(s) \geq 0$. Hence,

$$\mathbb{E}(X) \geq \sum_{s \in A} p(s)X(s) \geq a \sum_{s \in A} p(s) = a \cdot \mathbb{P}(A).$$

□

Chebyshev's inequality can now be derived from Markov's inequality. The intuition is along similar lines: If the probability of $\big|X - \mathbb{E}(X)\big| \geq a$ were too large, then it would contribute too much to the variance of X.

Proof of Chebyshev's inequality. Apply Markov's Inequality to the nonnegative random variable $(X-\mathbb{E}(X))^2$. Notice that

$$\mathbb{P}\Big[\big|X-\mathbb{E}(X)\big|\geq a\Big]=\mathbb{P}\Big[\big(X-\mathbb{E}(X)\big)^2\geq a^2\Big] \tag{1}$$

$$\leq \frac{\mathbb{E}\Big[\big(X-\mathbb{E}(X)\big)^2\Big]}{a^2} \tag{2}$$

$$=\frac{\mathrm{Var}(X)}{a^2}. \tag{3}$$

Step 1 holds because the events $\big|X-\mathbb{E}(X)\big|\geq a$ and $\big(X-\mathbb{E}(X)\big)^2\geq a^2$ are the same and hence they have the same probability. Step 2 follows by Markov's inequality, applied to the nonnegative random variable $Y=\big|X-\mathbb{E}(X)\big|$. And finally Step 3 uses the definition of variance. □

Remark. Markov's and Chebyshev's inequalities give us bounds on the probability of a random variable being far from its expectation. It is natural to wonder if they are *tight*, meaning that there are random variables that meet the conditions of the respective theorems for which the inequalities are actually equalities and hence cannot be improved. This turns out to be the case, and and as an exercise you can verify this by constructing non-trivial (i.e. non-constant) random variables and providing a specific value a for which Theorem 2 and Theorem 1 hold with equality.

2 Weak law of large numbers

Using Chebyshev's inequality, we can now show the so-called *weak law of large numbers*. This law says that if we have a random variable f (say the value resulting from the roll of a die) and take many independent copies of it, the average value of all these copies will be very close to the expected value of f. More formally, the weak law of large numbers is the following.

Theorem 3 (Weak law of large numbers)**.** *Fix $\epsilon>0$. Let $f_1,\cdots,f_n$ be n independent copies of a random variable f. Let*

$$g_n=\frac{1}{n}(f_1+f_2+\cdots f_n).$$

Then

$$\mathbb{P}\left[|g_n-\mathbb{E}(f)|\geq\epsilon\right]\to 0$$

as $n\to\infty$.

In plain English, the probability that g_n deviates from the expected value of f by at least ϵ becomes arbitrarily small as n grows arbitrarily large.

The weak law of large numbers can be proved by using Chebyshev's inequality applied to g_n. For this, we need to know $\mathbb{E}(g_n)$ and $\mathrm{Var}(g_n)$. By linearity of expectations, we have

$$\mathbb{E}(g_n)=\mathbb{E}\left[\frac{1}{n}(f_1+\cdots+f_n)\right]=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(f_i)=\frac{n}{n}\mathbb{E}(f)=\mathbb{E}(f).$$

For the variance, we get

$$\begin{aligned}\mathrm{Var}[g_n] &= \mathrm{Var}\left[\frac{1}{n}(f_1+\cdots+f_n)\right] \\ &= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} f_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[f_i] \\ &= \frac{1}{n}\mathrm{Var}[f],\end{aligned}$$

the third equality being true since the f_i's are *independent*. Thus, as n tends to infinity, $\mathbb{E}(g_n)$ remains constant while $\mathrm{Var}[g_n]$ tends to 0. For example, we saw that the roll of a fair die gives a variance of $\frac{35}{12}$. If we were to roll the die 1000 times and average all 1000 values, we would get a random variable whose expected value is still 3.5 but whose variance is much smaller, it is only $\frac{35}{12{,}000}$.

Now that we know the expected value and variance of g_n, we can simply use Chebyshev's inequality on g_n to get:

$$\mathbb{P}[|g_n - \mathbb{E}(f)| \geq \epsilon] \leq \frac{\mathrm{Var}[g_n]}{\epsilon^2} = \frac{\mathrm{Var}[f]}{n\epsilon^2},$$

and indeed this probability tends to 0 as n tends to infinity. This proves the weak law of large numbers.

3 Sums of Independent Random Variables

Although Markov's and Chebyshev's inequalities are tight for general random variables, it's natural to ask whether they are tight for special random variables. Suppose our random variable $X = \sum_{i=1}^{n} X_i$ is a sum of independent random variables. In this case, we can show:

Theorem 4. *Let $X_1, X_2, \ldots, X_n$ be independent random variables with $\mathbb{E}(X_i) = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2$. Then, for any $a > 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\mu_i\right| \geq a\right) \leq \frac{\sum_{i=1}^{n}\sigma_i^2}{a^2}.$$

Proof. This follows from Chebyshev's Inequality applied to $\sum_{i=1}^{n} X_i$ and the fact that $\mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n}\mathrm{Var}(X_i)$ for independent variables. □

In particular, for identically distributed random variables with expectation μ and variance σ^2, we obtain

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$$

for any $\epsilon > 0$. We have derived this when discussing the Weak Law of Large Numbers.

Can this result be improved or is it tight? At a first glance, you may suspect that this is tight, as we have made use of all our assumptions. In particular, we exploited the independence of the variables $\{X_i\}$ to get $\mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n}\mathrm{Var}(X_i)$. However, the proof of this fact about variances *uses only the pairwise independence* of the variables $\{X_i\}$.

Pairwise independence requires that every *pair* of variables be independent of each other, but *mutual* independence is stronger: it requires that every *subset* of the variables be independent.

Definition 1. The random variables X_i with $i \leq n$ are *pairwise independent* if, for all couples $i \neq j \in [n]$ and all x, y,

$$\mathbb{P}(X_i = x \wedge X_j = y) = \mathbb{P}(X_i = x) \cdot \mathbb{P}(X_j = y).$$

The variables X_i are *jointly* or *mutually independent* if, for all subsets $S \subseteq [n]$,

$$\mathbb{P}\left(\bigwedge_{i \in S} (X_i = x_i)\right) = \prod_{i \in S} \mathbb{P}(X_i = x_i).$$

Indeed, it is possible to show that Theorem 4 is tight when the variables $\{X_i\}$ are only pairwise independent.

Hard Exercise Let $X_1, \ldots, X_d$ be independent random variables that take value 1 or -1, each with probability 1/2. For each $S \subseteq [d]$, define the random variable $Y_S = \prod_{i \in S} X_i$. i) Show that the variables $\{Y_S\}$ are pairwise independent. ii) Show that Chebyshev's Inequality is asymptotically tight for the random variable $Z = \sum_{S \subseteq [d]} Y_S$.

We've seen that Chebyshev's inequality is tight for a sum of pairwise independent variables, but the question remains of whether a tighter bound than Chebyshev can be given for the sum of random variables that are mutually independent. In the next section, we will see that we can indeed obtain *stronger bounds* under this stronger assumption. These bounds are known as Chernoff bounds, after Herman Chernoff, Emeritus Professor of Applied Mathematics here at MIT!

4 Chernoff Bound

There are many different forms of Chernoff bounds, each tuned to slightly different assumptions. We will prove the statement of the bound for the simple case of a sum of independent Bernoulli trials, i.e. the case in which each random variable takes only the values 0 or 1. For example, this corresponds to the case of tossing unfair coins, each with its own probability of heads, and counting the total number of heads.

Theorem 5 (Chernoff Bounds)**.** *Let $X = \sum_{i=1}^{n} X_i$, where $X_i = 1$ with probability p_i and $X_i = 0$ with probability $1 - p_i$, and the X_i are mutually independent. Let $\mu = \mathbb{E}(X) = \sum_{i=1}^{n} p_i$. Then*

(i) **Upper Tail:** $\mathbb{P}\Big(X \geq (1+\delta)\mu\Big) \leq e^{-\frac{\delta^2}{2+\delta}\mu}$ *for all $\delta > 0$;*

(ii) **Lower Tail:** $\mathbb{P}\Big(X \leq (1-\delta)\mu\Big) \leq e^{-\mu\delta^2/2}$ *for all $0 < \delta < 1$;*

Notice that the lower and upper tail take slightly different forms. Curiously, this is necessary and boils down to the use of different approximations of the logarithmic function. More general versions of this bound exist, where this asymmetry is not present, but they are more complicated, as they involve the entropy of the distribution at the exponent.

For $\delta \in (0, 1)$, both lower and upper tails in Theorem 5 can be upper bounded by $e^{-\mu\delta^2/3}$. By combining them, we can obtain the following simple and useful bound:

Corollary 6. With X and $X_1, \ldots, X_n$ as before, and $\mu = \mathbb{E}(X)$,

$$\mathbb{P}\Big(|X - \mu| \geq \delta\mu\Big) \leq 2e^{-\mu\delta^2/3} \quad \text{for all } 0 < \delta < 1.$$

Example application: boosting success probability

Before proceeding with the proof of the Chernoff bound, let's see how it can be much stronger than Chebyshev's inequality. Consider the following situation: Suppose you have a randomized algorithm $\mathcal{A}$ for testing whether a number p is prime or not, and it outputs the correct answer with probability 2/3 (where the probability is over the random choices made by the algorithm). Note that the algorithm can make a mistake in either direction; say that the number is prime when it is not (i.e. it is composite) or vice versa. What

if you want to boost the success of the algorithm, by repeating it many times using mutually independent trials and then taking the majority vote of its answers?

Assume that we repeat the algorithm n times and we let S_n denote the number of times that the algorithm outputs "prime". We would output "prime" if $S_n/n \geq 1/2$. Then Chebyshev's inequality tells us that if p really is prime then

$$\mathbb{P}\Big(\big|S_n/n - 2/3\big| \geq \epsilon\Big) \leq \frac{2}{9n\epsilon^2}.$$

So for example, $\mathbb{P}(|S_n/n - 2/3| \geq 1/6) \leq \frac{8}{n}$. (To be formal, we should have written the conditioning on our assumption: $\mathbb{P}\left(|S_n/n - 2/3| \geq 1/6 \;\middle|\; p \text{ is prime}\right) \leq \frac{8}{n}$.) Similarly if p is not prime then

$$\mathbb{P}\Big(\big|S_n/n - 1/3\big| \geq 1/6\Big) \leq \frac{8}{n}.$$

And hence if we want the failure probability (in either case) of our algorithm to be minuscule – say 10^{-10} – then we need to repeat the algorithm approximately $8 \cdot 10^{10}$ times.

But in fact the Chernoff bound tells us that we can repeat it many fewer times. From Corollary 6, using $\mathbb{E}(S_n) = 2n/3$,

$$\mathbb{P}\Big(\big|S_n - 2n/3\big| \geq \delta(2n/3)\Big) \leq 2e^{-2n\delta^2/9}.$$

Taking $\delta = 1/4$ we obtain

$$\mathbb{P}\Big(\big|S_n/n - 2/3\big| \geq 1/6\Big) \leq 2e^{-n/72}.$$

This is a *massive* improvement over the Chebyshev bound because to get the probability to be 10^{-10} we only need to repeat it less than 10^4 times instead.

5 Proof of Theorem 5

The proof of the Chernoff bound is more complex than the proofs of other tail bounds we've seen so far. But at a conceptual level it follows a similar strategy as in the proof of Chebyshev's inequality but with the important difference that instead of squaring both sides of the inequality to get a new random variable that is nonnegative to which we can apply Markov's inequality, we will exponentiate the sum of random variables instead. When we take the expectation, we will be able to exploit mutual independence of the X_i's.

For any $s > 0$ and $a \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{P}(X \geq a) &= \mathbb{P}\left(e^{sX} \geq e^{sa}\right) \\
&\leq \frac{\mathbb{E}\left(e^{sX}\right)}{e^{sa}} \qquad \text{by Markov's inequality.}
\end{aligned} \tag{4}$$

So we have some upper bound on $\mathbb{P}(X > a)$ in terms of $\mathbb{E}\left(e^{sX}\right)$. Similarly, for any $s > 0$, we have

$$\begin{aligned}
\mathbb{P}(X \leq a) &= \mathbb{P}\Big(e^{-sX} \geq e^{-sa}\Big) \\
&\leq \frac{\mathbb{E}\left(e^{-sX}\right)}{e^{-sa}}.
\end{aligned} \tag{5}$$

Since X is a sum of random variables $X_1, \ldots, X_n$, then

$$\begin{aligned}
\mathbb{E}\left(e^{sX}\right) &= \mathbb{E}\left(e^{s\sum_{i=1}^n X_i}\right) \\
&= \mathbb{E}\left(\prod_{i=1}^n e^{sX_i}\right). \\
&= \prod_{i=1}^n \mathbb{E}\left(e^{sX_i}\right) \qquad \text{by mutual independence.}
\end{aligned} \tag{6}$$

We can now bound $\mathbb{E}\left(e^{sX_i}\right)$ for each X_i individually by using that the variables X_i are either 1 or 0, with probability p or $1-p$, respectively.

$$\begin{aligned}\mathbb{E}\left(e^{sX_i}\right) &= p_i \cdot e^s + (1-p_i)\cdot 1 \qquad \text{by definition of expectation}\\ &= 1 + p_i(e^s-1)\\ &\le e^{p_i(e^s-1)} \qquad \text{using } 1+y \le e^y \text{ with } y = p_i(e^s-1).\end{aligned}$$

We've weakened the bound here in order to create a nice exponent that enables us to eliminate the p_i by using $\sum_{i=1}^n p_i = \mathbb{E}(X) = \mu$. Indeed, using Equation 6, we obtain

$$\mathbb{E}\left(e^{sX}\right) = \prod_{i=1}^n \mathbb{E}\left(e^{sX_i}\right) \le \prod_{i=1}^n e^{p_i(e^s-1)} = e^{(e^s-1)\sum_{i=1}^n p_i} = e^{(e^s-1)\mu}. \tag{7}$$

Thus, Inequality 4 becomes

$$\mathbb{P}(X \ge a) \le \frac{e^{(e^s-1)\mu}}{e^{sa}},$$

for any $s > 0$ and $a \in \mathbb{R}$.

The purpose of the Chernoff bound is to bound the deviation of X from its expected value, μ; so for the upper tail we use $a = (1+\delta)\mu$, where δ indicates the deviation:

$$\mathbb{P}\Big(X \ge (1+\delta)\mu\Big) \le \frac{e^{(e^s-1)\mu}}{e^{s(1+\delta)\mu}} = e^{(e^s-1)\mu - s(1+\delta)\mu} \qquad \text{for any } s \ge 0.$$

We choose $s = \ln(1+\delta)$ since that's where the bound is as small as possible, as can be seen by taking the derivative of the exponent as a function of s. Thus,

$$\mathbb{P}\Big(X \ge (1+\delta)\mu\Big) \le \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu.$$

The final manipulations simplify the bound. Taking the natural logarithm of the right-hand side yields

$$\mu\Big(\delta - (1+\delta)\ln(1+\delta)\Big).$$

Using the following inequality for $x > 0$ (left as an exercise):

$$\ln(1+x) \ge \frac{x}{1+x/2},$$

we obtain

$$\mu\Big(\delta - (1+\delta)\ln(1+\delta)\Big) \le -\frac{\delta^2}{2+\delta}\mu.$$

Hence, we have the desired bound for the upper tail:

$$\mathbb{P}\Big(X \ge (1+\delta)\mu\Big) \le \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu \le e^{-\frac{\delta^2}{2+\delta}\mu}.$$

The proof of the lower tail is entirely analogous. It uses Equation (5) taking $s = \ln(1-\delta)$, and applies the following inequality for the logarithm of $(1-\delta)$ in the range $0 < \delta < 1$:

$$\ln(1-\delta) \ge -\delta + \frac{\delta^2}{2}.$$

Details are left as an exercise.

The moment generating function

The expected value $\mathbb{E}\left(e^{sX}\right)$ was a key player in the proof of the Chernoff bound. The function

$$M_X(s) = \mathbb{E}\left(e^{sX}\right) \qquad \text{(defined from } \mathbb{R} \text{ to } \mathbb{R}\text{)},$$

is known as the *moment generating function* of the random variable X. The reason for the name is related to the Taylor expansion of e^{sX}; assuming it converges, we have

$$M_X(s) = \mathbb{E}\Big(1 + sX + \tfrac{1}{2}s^2X^2 + \tfrac{1}{3!}s^3X^3 + \cdots\Big) = \sum_{i=0}^{\infty} \frac{1}{i!} s^i \mathbb{E}\left(X^i\right).$$

The terms $\mathbb{E}\left(X^i\right)$ are called "moments" and encode important information about the distribution; notice that the first moment ($i = 1$) is just the expectation, and the second moment is closely related to the variance. So the moment generating function encodes information of all of these moments in some way.

18.200 Principles of Discrete Applied Mathematics

Spring 2024