18.310 lecture notes February 21, 2015

Chernoff bounds, and some applications

Lecturer: Michel Goemans

1 Preliminaries

Before we venture into Chernoff bound, let us recall Chebyshev's inequality which gives a simple bound on the probability that a random variable deviates from its expected value by a certain amount.

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 $a \in \mathbb{R}$:*

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

We gave a proof from first principles, but we can also derive it easily from Markov's inequality which only applies 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 $a > 0$,*

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

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 requires the variance of the random variable but can be derived from Markov's inequality.

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

$$\mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = \mathrm{Var}(X).$$

□

Even though Markov's and Chebyshev's Inequality only use information about the expectation and the variance of the random variable under consideration, they are essentially tight for a general random variable.

Exercise. Verify this by constructing non-trivial (i.e. non-constant) random variables for which Theorem 2 and Theorem 1 are tight, i.e. hold with equality.

2 Deviation of a sum on independent random variables

As we are not able to improve Markov's Inequality and Chebyshev's Inequality in general, it is worth to consider whether we can say something stronger for a more restricted, yet interesting, class of random variables. This idea brings us to consider the case of a random variable that is the sum of a number of independent random variables.

This scenario is particularly important and ubiquitous in statistical applications. Examples of such random variables are the number of heads in a sequence of coin tosses, or the average support obtained by a political candidate in a poll.

Can Markov's and Chebyshev's Inequality be improved for this particular kind of random variable? Before confronting this question, let us check what Chebyshev's Inequality (the stronger of the two) gives us for a sum of independent random variables.

Theorem 3. *Let $X_1, X_2, \dots, X_n$ be independent random variables with $\mathbb{E}(X_i) = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2$. Then, for any $a > 0$:*

$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$

$\sigma_{ij} = 0$

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

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 dervied 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 $\mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$. Notice, however, that this last step actually *only uses the pairwise independence* of the variables $\{X_i\}$, i.e. the fact that, for all couples $i \neq j \in [n]$ and all $x, y \in \mathbb{R}$:

$$\mathbb{P}(X_i = x \wedge X_j = y) = \mathbb{P}(X_i = x) \cdot \mathbb{P}(X_j = y). \tag{1}$$

Indeed, it is possible to show that Theorem 3 is tight when all the variables $\{X_i\}$ are just guaranteed to be pairwise independent.

Hard Exercise Let $X_1, \dots, X_d$ 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) Let $Z = \sum_{S \subseteq D} Y_S$. Show that Chebyshev's Inequality is asymptotically tight for Z.

We are now ready to tackle the case of a sum of independent random variables. Recall that we are now using the following strong version of independence (also known as joint or mutual independence), which guarantees the same property of Equation 1 for any subset $S \subseteq [n]$ of random variables:

$$\forall S \subseteq [n],\ \mathbb{P}(\bigwedge_{i \in S} X_i = x_i) = \prod_{i \in S} \mathbb{P}(X_i = x_i).$$

In this case, the proof of Theorem 3 is too weak as it does not rely on the joint independence. In the next section, we will see that we can indeed obtain *stronger bounds* under this stronger assumpiton. These bounds are known as Chernoff bounds, after Herman Chernoff, Emeritus Professor of Applied Mathematics here at MIT!

3 Chernoff Bound

There are many different forms of Chernoff bounds, each tuned to slightly different assumptions. We will start with 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 only takes 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 4 (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 all* X_i *are independent. Let* $\mu = \mathbb{E}(X) = \sum_{i=1}^n p_i$*. Then*

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

(ii) **Lower Tail:** $\mathbb{P}(X \leq (1-\delta)\mu) \leq e^{-\mu\delta^2/2}$ *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 approximation of the logarithmic function. There exist more general versions of this bound, where this asymmetry is not present, but they are more complicated, as the involve the entropy of the distribution at the exponent.

For $\delta \in (0,1)$, we can combine the lower and upper tails in Theorem 4 to obtain the following simple and useful bound:

Corollary 5. With X and $X_1, \ldots, X_n$ as before, and $\mu = \mathbb{E}(X)$,

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

Example application: coin tossing

Suppose we have a fair coin. Repeatedly toss the coin, and let S_n be the number of heads from the first n tosses. Then the weak law of large numbers tells us that $\mathbb{P}(|S_n/n - 1/2| \geq \epsilon) \to 0$ as $n \to \infty$. But what can we say about this probability for some fixed n? If we go back to the proof of the weak law that we gave in terms of Chebyshev's inequality, we find that it tells us that

$$\mathbb{P}(|S_n/n - 1/2| \geq \epsilon) \leq \frac{1}{4n\epsilon^2}.$$

So for example, $\mathbb{P}(|S_n/n - 1/2| \geq 1/4) \leq \frac{4}{n}$.

But we can apply Chernoff instead of Chebyshev; what do we get then? From Corollary 5, using $\mathbb{E}(S_n) = n/2$,

$$\mathbb{P}(|S_n - n/2| \geq \delta(n/2)) \leq 2e^{-n\delta^2/6}.$$

Taking $\delta = 1/2$ we obtain $\mathbb{P}(|S_n/n - 1/2| \geq 1/4) \leq 2e^{-n/24}$. This is a *massive* improvement over the Chebyshev bound! Let's try this now with a much smaller δ: let $\delta = \sqrt{6 \ln n/n}$. Then we obtain

$$\mathbb{P}(|S_n/n - 1/2| \geq \tfrac{1}{2}\sqrt{6 \ln n/n}) \leq 2e^{-\ln n} = 2\frac{1}{n}.$$

If instead we take δ just twice as large, $\delta = 2\sqrt{6 \ln n/n}$,

$$\mathbb{P}(|S_n/n - 1/2| \geq \sqrt{6 \ln n/n}) \leq 2e^{-4 \ln n} = 2\frac{1}{n^4}.$$

3.1 Proof idea and moment generating function

For completeness, we give a proof of Theorem 4. Let X be any random variable, and $a \in \mathbb{R}$. We will make use of the same idea which we used to prove Chebyshev's inequality from Markov's inequality. For any $s > 0$,

$$\begin{aligned}\mathbb{P}(X \geq a) &= \mathbb{P}(e^{sX} \geq e^{sa}) \\ &\leq \frac{\mathbb{E}(e^{sX})}{e^{sa}} \qquad \text{by Markov's inequality.} \qquad (2)\end{aligned}$$

(Recall that to obtain Chebyshev, we squared both sides in the first step, here we exponentiate.) So we have some upper bound on $\mathbb{P}(X > a)$ in terms of $\mathbb{E}(e^{sX})$. Similarly, for any $s > 0$, we have

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

The key player in this reasoning is the *moment generating function* M_X of the random variable X, which is a function from $\mathbb{R}$ to $\mathbb{R}$ defined by

$$M_X(s) = \mathbb{E}\left(e^{sX}\right).$$

The reason for the name is related to the Taylor expansion of e^{sX}; assuming it converges, we have

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

The terms $\mathbb{E}(X^i)$ 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.

Moment generating functions behave wonderfully with respect to addition of independent random variables:

Lemma 1. *If $X = \sum_{i=1}^n X_i$ where $X_1, X_2, \ldots, X_n$ are independent random variables, then*

$$M_X(s) = \prod_{i=1}^n M_{X_i}(s).$$

Proof.

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

□

This lemma allows us to prove a Chernoff bound by bounding the moment generating function of each X_i individually.

3.2 Proof of Theorem 4

Before proceeding to prove the theorem, we compute the form of the moment generating function for a single Bernoulli trial. Our goal is to then combine this expression with Lemma 1 in the proof of Theorem 4.

Lemma 2. *Let Y be a random variable that takes value 1 with probability p and value 0 with probability $1-p$. Then, for all $s \in \mathbb{R}$:*

$$M_Y(s) = \mathbb{E}(e^{sY}) \leq e^{p(e^s-1)}.$$

Proof. We have:

$$\begin{aligned} M_Y(s) &= \mathbb{E}(e^{sY}) \\ &= p \cdot e^s + (1-p) \cdot 1 \qquad \text{by definition of expectation} \\ &= 1 + p(e^s - 1) \\ &\leq e^{p(e^s-1)} \qquad \text{using } 1 + y \leq e^y \text{ with } y = p(e^s - 1). \end{aligned}$$

□

We are now ready to prove Theorem 4 by combining Lemma 1 and 2.

Proof of Theorem 4. Applying Lemma 1 and Lemma 2, we obtain

$$M_X(s) \leq \prod_{i=1}^{n} e^{p_i(e^s-1)} = e^{(e^s-1)\sum_{i=1}^n p_i} \leq e^{(e^s-1)\mu}, \tag{3}$$

using that $\sum_{i=1}^n p_i = \mathbb{E}(X) = \mu$.

For the proof of the upper tail, we can now apply the strategy described in Equation 2, with $a = (1+\delta)\mu$ and $s = \ln(1+\delta)$.

$$\begin{aligned} \mathbb{P}(X \geq (1+\delta)\mu) &\leq e^{-s(1+\delta)\mu} e^{(e^s-1)\mu)} \\ &= \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu. \end{aligned}$$

Our choice of s is motivated as follows: we are trying to make our upper bound for the tail probability to be as small as possible. To do this, we can minimize our expression for the upper bound as a function of s. Taking the derivative of the exponent shows that this minimum is achieved exactly at $s = \log(1+\delta)$.

Taking the natural logarithm of the right-hand side yields

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

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

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

we obtain

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

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

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

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

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

Details are left as an exercise. □

4 Other versions of Chernoff Bound

Chernoff bound can be applied to more general settings than that of Bernoulli variables. In particular, the following version of the bound applies to bounded random variables, regardless of their distribution!

Theorem 6. *Let $X_1, X_2, \ldots, X_n$ be random variables such that $a \leq X_i \leq b$ for all i. Let $X = \sum_{i=1}^n X_i$ and set $\mu = \mathbb{E}(X)$. Then, for all $\delta > 0$:*

(i) **Upper Tail:** $\mathbb{P}(X \geq (1+\delta)\mu) \leq e^{-\frac{2\delta^2\mu^2}{n(b-a)^2}}$;

(ii) **Lower Tail:** $\mathbb{P}(X \leq (1-\delta)\mu) \leq e^{-\frac{\delta^2\mu^2}{n(b-a)^2}}$.