

Chernoff bounds, and some applications

Lecturers: Diego Cifuentes and Michel Goemans

Note: These lecture notes are used in the recitation on guiding text. At some places during the notes, you are given a few versions for the text that could be included in the lecture notes. If you are reading these notes for the pre-recitation assignment, see the instructions there. If you are reading these notes to learn about the Chernoff bound: all text versions shown here are accurate.

1 Introduction

Location A

Version A1 This lecture note proves Chernoff bounds for a sum of independent Bernoulli trials. We start by proving the Chebychev inequality in Section 2. In Section 3 we introduce the Chernoff bounds, and in Section 4 we prove them.

Version A2 This lecture note presents Chernoff bounds, which bound the probability that a sum of independent random variables is a certain distance from its expected value. As a warmup, in Section 2 we prove the Chebychev inequality; this proof will be important later. Section 3 introduces the Chernoff bounds and illustrates their value. Section 4 proves the Chernoff bounds.

Version A3 In the lecture notes on probability, you saw Chebychev's inequality, which bounds the probability that a random variable is a certain distance from its expected value. In these lecture notes, you'll see that when Chebychev is used to bound a sum of random variables (e.g., a sum of die rolls, or an average of draws from a distribution), its application requires only that those variables be pairwise independent. If they are instead mutually independent (i.e., each variable is independent of any subset of the others), a natural question is whether we can use this information to obtain a tighter bound on the sum than that provided by Chebychev.

It turns out that in some cases we can do so, especially when the number of variables added is very large (e.g., when more than 1000 draws are taken from a binomial distribution). This lecture note presents such a stronger bound: the Chernoff bound in Theorem 4. A key step in its proof is using exponentiation to convert the sum of random variables to a product so mutual independence can be used.

The proof of the Chernoff bound uses strategies similar to those used in a simpler proof of Chebychev's inequality, so we begin with the simpler Chebychev proof in Section 2; this section also demonstrates that applying the Chebychev inequality to a sum of random variables requires only pairwise independence. In Section 3 we introduce the Chernoff bound, and illustrate that it is stronger than Chebychev's inequality for sums of many mutually independent variables. Section 4 is devoted to proving the Chernoff bound.

[End of Location A](#)

2 Chebychev's inequality

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 $X : S \rightarrow \mathbb{R}$ be a random variable with expectation $\mathbb{E}(X)$ and variance $\text{Var}(X)$. Then, for any $a > 0$,*

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

In the probability lecture notes, we proved 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 $X : S \rightarrow \mathbb{R}$ 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 $\sum_{s \in \bar{A}} p(s)X(s) \geq 0$. 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

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

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

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

Step 2 follows by Markov's inequality, and Step 3 uses the definition of variance. In Step 1, both sides of the inequality were squared so that the definition of variance could be used in Step 3. □

The proof of the stronger Chernoff bound will use a similar strategy, but instead of squaring both sides of the inequality before applying Markov's inequality we will exponentiate so that the sum of random variables can later be converted to a product.

Remark. Even though Markov's and Chebyshev's inequalities use information about only 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 and providing a specific value a for which Theorem 2 and Theorem 1 are tight, i.e. hold with equality.

Deviation of a sum of independent random variables

Location B

[Version B1](#) We now apply Chebychev's inequality to a sum of independent random variables.

[Version B2](#) *You may write this version in the March 20 recitation.*

Theorem 3. Let X_1, X_2, \dots, X_n be independent random variables with $\mathbb{E}(X_i) = \mu_i$ and $\text{Var}(X_i) = \sigma_i^2$. 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 $\text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i)$ for independent variables. \square

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 $\text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i)$. 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 3 is tight when the variables $\{X_i\}$ are only 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) Show that Chebyshev's Inequality is asymptotically tight for the random variable $Z = \sum_{S \subseteq [d]} Y_S$.

[Location C](#)

[Version C1](#) In the next section we introduce Chernoff bounds, which are named after Herman Chernoff, Emeritus Professor of Applied Mathematics here at MIT!

[Version C2](#) *You may write this version in the March 20 recitation*

3 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 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 the X_i are mutually 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 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 4 can be upper bounded by $e^{-\mu\delta^2/3}$. By combining them, we can obtain the following simple and useful bound:

Corollary 5. With X and X_1, \dots, 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: 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

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

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: $\mathbb{P}(|S_n/n - 2/3| \geq 1/6 \mid p \text{ is prime}) \leq \frac{8}{n}$.) Similarly if p is not prime then

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

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 5, using $\mathbb{E}(S_n) = 2n/3$,

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

Taking $\delta = 1/4$ we obtain

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

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.

4 Proof of Theorem 4

The proof of the Chernoff bound is more complex than the proofs of other tail bounds we've seen so far. 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. This lets us exploit the mutual independence of the X_i 's.

Sentence D1: To prove the Chernoff bound, we begin by moving the sum $X = \sum_{i=1}^n X_i$ into the exponent so we can later change addition to multiplication and use the mutual independence of the X_i .

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

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

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

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

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

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

Sentence D2: We can now bound $\mathbb{E}(e^{sX_i})$ for each X_i individually by using that the variables X_i are either 1 or 0, with probability p or $1 - p$, respectively.

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

Sentence D3: Using Equation 6, we obtain

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

Thus, Inequality 4 becomes

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

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

Sentence D4: 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:

$$\mathbb{P}(X \geq (1 + \delta)\mu) \leq \frac{e^{(e^s - 1)\mu}}{e^{s(1 + \delta)\mu}} = e^{(e^s - 1)\mu - s(1 + \delta)\mu} \quad \text{for any } s \geq 0.$$

Sentence D5: We choose $s = \ln(1 + \delta)$. Thus,

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

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

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

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 uses Equation (5) taking $s = \ln(1 - \delta)$, and applies the following inequality for the logarithm of $(1 - \delta)$ in the range $0 < \delta < 1$:

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

Details are left as an exercise.

The moment generating function

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

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

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

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

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.

MIT OpenCourseWare
<https://ocw.mit.edu>

18.200 Principles of Discrete Applied Mathematics
Spring 2024

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.