15-359: Probability and Computing
Fall 2009

Lecture 10: More Chernoff Bounds, Sampling, and the Chernoff + Union Bound method

1 Chernoff Bound 2

Last lecture we saw the following:

Chernoff Bound 1: Let $X \sim \text{Binomial}(n, 1/2)$. Then for any $0 \le t \le \sqrt{n}$,

$$\begin{aligned} \mathbf{Pr}\left[X \ge \frac{n}{2} + t\frac{\sqrt{n}}{2}\right] &\le e^{-t^2/2}, \\ \text{and also } \mathbf{Pr}\left[X \le \frac{n}{2} - t\frac{\sqrt{n}}{2}\right] &\le e^{-t^2/2}. \end{aligned}$$

This bound tells us that if X is the sum of many independent Bernoulli(1/2)'s, it's extremely unlikely that X will deviate even a little bit from its mean. Let's rephrase the above a little. Taking $t = \epsilon\sqrt{n}$ in Chernoff Bound 1, we get

$$\left.\begin{array}{l} \mathbf{Pr}[X \ge (1+\epsilon)(n/2)] \\ \mathbf{Pr}[X \le (1-\epsilon)(n/2)] \end{array}\right\} \le e^{-\epsilon^2 n/2}.$$

Okay, that tells us about deviations for Binomial$(n, 1/2)$. Turns out that similar bounds are true for Binomial(n, p). And similar bounds are true for, say,

$$U = U_1 + \cdots + U_n$$

when the U_i's are *independent* random variables that are

$$U_i = \begin{cases} 0 & \text{w.p. } 1/3, \\ 1/2 & \text{w.p. } 1/3, \\ 1 & \text{w.p. } 1/3. \end{cases}$$

Actually, there are a dozens of versions of Chernoff bounds, covering dozens of variations; in other texts, you might see any of them. But as promised, you will only be required to memorize one more :) This one covers 95% of cases pretty well.

Chernoff Bound 2:
Let $X_1, \ldots, X_n$ be *independent* random variables.
They need not have the same distribution.
Assume that $0 \le X_i \le 1$ always, for each i.
Let $X = X_1 + \cdots + X_n$.
Write $\mu = \mathbf{E}[X] = \mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n]$.

Then for any $\epsilon \geq 0$,

$$\begin{aligned} \mathbf{Pr}[X \geq (1+\epsilon)\mu] &\leq \exp\left(-\frac{\epsilon^2}{2+\epsilon}\mu\right) \\ \text{and, } \mathbf{Pr}[X \leq (1-\epsilon)\mu] &\leq \exp\left(-\frac{\epsilon^2}{2}\mu\right). \end{aligned}$$

(Remember the notation $\exp(x) = e^x$.)

1.1 What's up with the $\frac{\epsilon^2}{2+\epsilon}$?

Chernoff Bound 2 takes some memorizing, we admit it. But it's a truly indispensable tool, so it's very much worth it. Actually, the second statement, the probability of going below the expectation, isn't *so* bad to remember; it's clean-looking at least. The first statement is, admittedly, slightly weird. By symmetry you were probably hoping that we were going to write

$$\exp\left(-\frac{\epsilon^2}{2}\mu\right)$$

on the right side of the first statement, like in the second statement. Unfortunately we can't; turns out, that's simply not true. Notice that

$$\frac{\epsilon^2}{2} \text{ is slightly bigger than } \frac{\epsilon^2}{2+\epsilon},$$

and therefore

$$-\frac{\epsilon^2}{2}\mu \text{ is slightly more negative than } -\frac{\epsilon^2}{2+\epsilon}\mu$$

(since $\mu \geq 0$ since all the X_i's are nonnegative), and therefore

$$\exp\left(-\frac{\epsilon^2}{2}\mu\right) \text{ is slightly smaller than } \exp\left(-\frac{\epsilon^2}{2+\epsilon}\mu\right).$$

So the bound for $\mathbf{Pr}[X \geq (1+\epsilon)\mu]$ is slightly bigger than the bound for $\mathbf{Pr}[X \leq (1-\epsilon)\mu]$, but unfortunately, it cannot be improved to the smaller quantity.

Now, ϵ is a user-selected parameter, and *most* of the time you, the user, select ϵ to be pretty small, smaller than 1. If indeed $\epsilon \leq 1$, we have

$$\frac{\epsilon^2}{2+\epsilon} \text{ is bigger than } \frac{\epsilon^2}{3},$$

and so we could put this slightly simpler quantity into our bound. Indeed, sometimes people state a simplified Chernoff Bound 2 like this:

Chernoff Bound 2′: Suppose we are in the setting of Chernoff Bound 2. Then for all $0 \leq \epsilon \leq 1$,

$$\begin{aligned} \mathbf{Pr}[X \geq (1+\epsilon)\mu] &\leq \exp\left(-\frac{\epsilon^2}{3}\mu\right) \\ \text{and, } \mathbf{Pr}[X \leq (1-\epsilon)\mu] &\leq \exp\left(-\frac{\epsilon^2}{2}\mu\right). \end{aligned}$$

That's a little easier to remember. And indeed, for the event $X \leq (1-\epsilon)\mu$ you would never care to take $\epsilon > 1$ anyway! But for the event $X \geq (1+\epsilon)\mu$ sometimes you *would* care to take $\epsilon > 1$. And Chernoff Bound 2′ doesn't really tell you what to do in this case. Whereas Chernoff Bound 2 does; for example, taking $\epsilon = 8$, it tells you

$$\mathbf{Pr}[X \geq 9\mu] \leq \exp(-6.4\mu).$$

1.2 More tricks and observations

Sometimes you simply want to upper-bound the probability that X is *far* from its expectation. For this, one thing we can do is take Chernoff Bound 2, intentionally *weaken* the second bound to $\exp(-\frac{\epsilon^2}{2+\epsilon}\mu)$ as well, and then add, concluding:

Two-sided Chernoff Bound: In the setting of Chernoff Bound 2,

$$\mathbf{Pr}[|X-\mu| \geq \epsilon\mu] \leq 2\exp\left(-\frac{\epsilon^2}{2+\epsilon}\mu\right).$$

What if we don't have $0 \leq X_i \leq 1$? Another twist is that sometimes you don't have that your r.v.'s X_i satisfy $0 \leq X_i \leq 1$. Sometimes they might satisfy, say, $0 \leq X_i \leq 10$. Don't panic! In this case, the trick is to define $Y_i = X_i/10$, and $Y = Y_1 + \cdots + Y_n$. Then we *do* have that $0 \leq Y_i \leq 1$ always, the Y_i's are independent assuming the X_i's are, $\mu_Y = \mathbf{E}[Y] = \mathbf{E}[X]/10 = \mu_X/10$, and we can use the Chernoff Bound on the Y_i's;

$$\begin{aligned}\mathbf{Pr}[X \geq (1+\epsilon)\mu_X] = \mathbf{Pr}[X/10 \geq (1+\epsilon)\mu_X/10] &= \mathbf{Pr}[Y \geq (1+\epsilon)\mu_Y] \\ &\leq \exp\left(-\frac{\epsilon^2}{2+\epsilon}\mu_Y\right) = \exp\left(-\frac{\epsilon^2}{2+\epsilon}\frac{\mu_X}{10}\right).\end{aligned}$$

So you lose a factor of 10 inside the final exponential probability bound, but you still get something pretty good. This is a key trick to remember!

Finally, one more point: Given that we have this super-general Chernoff Bound 2, why bother remembering Chernoff Bound 1? The reason is, Chernoff 2 is weaker than Chernoff 1. If $X \sim \text{Binomial}(n, 1/2)$, Chernoff Bound 1 tells us

$$\mathbf{Pr}[X \leq n/2 - t\sqrt{n}/2] \leq \exp(-t^2/2). \tag{1}$$

However, suppose we used Chernoff Bound 2. We have

$$\text{“}X \leq n/2 - t\sqrt{n}/2\text{”} \qquad \Leftrightarrow \qquad X \leq (1 - t/\sqrt{n})(n/2),$$

and so Chernoff Bound 2 gives us

$$\mathbf{Pr}[X \leq n/2 - t\sqrt{n}/2] \leq \exp\left(-\frac{(t/\sqrt{n})^2}{2} \cdot \frac{n}{2}\right) = \exp(-t^2/4).$$

This is only the *square-root* of the bound (1).

1.3 The proof

We will not prove Chernoff Bound 2. However the proof is not much more than an elaboration on our proof of Chernoff Bound 1. Again, the idea is to let $\lambda > 0$ be a small "scale", pretty close to ϵ, actually. You then consider the random variable $(1+\lambda)^X$, or relatedly, $e^{\lambda X}$. Finally, you bound, e.g.,

$$\mathbf{Pr}[X \geq (1+\epsilon)\mu] = \mathbf{Pr}[e^{\lambda X} \geq e^{(1+\epsilon)\mu}] \leq \frac{\mathbf{E}[e^{\lambda X}]}{e^{(1+\epsilon)\mu}},$$

using Markov's Inequality, and you can compute $\mathbf{E}[e^{\lambda X}]$ as

$$\mathbf{E}[e^{\lambda X}] = \mathbf{E}[e^{\lambda X_1 + \cdots + \lambda X_n}] = \mathbf{E}[e^{\lambda X_1} \cdots e^{\lambda X_n}] = \mathbf{E}[e^{\lambda X_1}] \cdots \mathbf{E}[e^{\lambda X_n}],$$

using the fact that $X_1, \ldots, X_n$ are independent.

2 Sampling

The #1 use of the Chernoff Bound is probably in Sampling/Polling. Suppose you want to know what fraction of the population approves of the current president. What do you do?

Well, you do a poll. Roughly speaking, you call up n random people and ask them if they approve of the president. Then you take this empirical fraction of people and claim that's a good estimate of the true fraction of the entire population that approves of the president. But is it a good estimate? And how big should n be?

Actually, there are *two* sources of error in this process. There's the probability that you obtain a "good" estimate. And there's the extent to which your estimate is "good". Take a close look at all those poll results that come out these days and you'll find that they use phrases like,

"This poll is accurate to within $\pm 2\%$*,* 19 *times out of* 20*."*

What this means is that they did a poll, they published an estimate of the true fraction of people supporting the president, and they make the following claim about their estimate: There is a $1/20$ chance that their estimate is just completely way off. Otherwise, i.e., with probability $19/20 = 95\%$, their estimate is within $\pm 2\%$ of the truth.

This whole 95% thing is called the "confidence" of the estimate, and its presence is inevitable. There's just no way you can legitimately say, "My polling estimate is 100% guaranteed to be within $\pm 2\%$ of the truth." Because if you sample n people at random, you know, there's a chance they all happen to live in Massachusetts, say (albeit an unlikely, much-less-than-5% chance), in which case your approval rating estimate for a Democratic president is going to much higher than the overall country-wide truth.

To borrow a phrase from Learning Theory, these polling numbers are "Probably Approximately Correct" — i.e., probably (with chance at least 95% over the choice of people), the empirical average is approximately (within $\pm 2\%$, say) correct (vis-a-vis the true fraction of the population).

2.1 Analysis

How do pollsters, and how can we, make such statements?

Let the true fraction of the population that approves of the president be p, a number in the range $0 \leq p \leq 1$. This is the "correct answer" that we are trying to elicit.

Suppose we ask n uniformly randomly chosen people for their opinion, and let each person be chosen *independently*. We are choosing people "*with* replacement". (I.e., it's possible, albeit a very slim chance, that we may ask the same person more than once.) Let X_i be the indicator random variable that the ith person we ask approves of the president. Here is the key observation:

Fact: $X_i \sim \text{Bernoulli}(p)$, and $X_1, \ldots, X_n$ are independent.

Let $X = X_1 + \cdots + X_n$, and let $\overline{X} = X/n$. The empirical fraction $\overline{X}$ is the estimate we will publish, our guess at p.

Question: How large does n have to be so that we get good "accuracy" with high "confidence"? More precisely, suppose our pollster boss wants our estimate to have accuracy θ and confidence $1 - \delta$, meaning

$$\mathbf{Pr}\left[|\overline{X} - p| \leq \theta\right] \geq 1 - \delta.$$

How large do we have to make n?

Answer: Let's start by using the Two-sided Chernoff Bound on X. Since $X \sim \text{Binomial}(n, p)$, we have $\mathbf{E}[X] = np$. So for any $\epsilon \geq 0$, we have

$$\begin{aligned} \mathbf{Pr}[|X - pn| \geq \epsilon pn] &\leq 2\exp\left(-\frac{\epsilon^2}{2+\epsilon} \cdot pn\right) \\ \Leftrightarrow \mathbf{Pr}[|\overline{X} - p| \geq \epsilon p] &\leq 2\exp\left(-\frac{\epsilon^2}{2+\epsilon} \cdot pn\right). \end{aligned}$$

Here the two events inside the $\mathbf{Pr}[\cdot]$ are the same event; we just divided by n.

We want accuracy θ; i.e., we want $\overline{X}$ to be within θ of p with high probability. (In our original example, $\theta = 2\% = .02$.) We need to get $\theta = \epsilon p$, so we should take $\epsilon = \theta/p$.[1] Doing so, we get

$$\mathbf{Pr}[|\overline{X} - p| \geq \theta] \quad \leq \quad 2\exp\left(-\frac{\theta^2/p^2}{2+\theta/p} \cdot pn\right) \quad = \quad 2\exp\left(-\frac{\theta^2}{2p+\theta} \cdot n\right).$$

Okay, what about getting confidence $1 - \delta$? Let's look at that bound on the right. Having that n inside the $\exp(-\cdot)$ is great — it tells us the bigger n is, the less chance that our estimate is off by more than θ. As for the $\frac{\theta^2}{2p+\theta}$, well, the bigger that term is, the better. The bigger p is, the smaller that factor is, but the biggest p could be is 1. I.e.,

$$\frac{\theta^2}{2p+\theta} \geq \frac{\theta^2}{2+\theta},$$

and therefore we have

$$\mathbf{Pr}[|\overline{X} - p| \geq \theta] \quad \leq \quad 2\exp\left(-\frac{\theta^2}{2+\theta} \cdot n\right).$$

[1] Worried about $p = 0$? In that case, $\overline{X}$ will correctly be 0 100% of the time!

So if we want confidence $1-\delta$ in the estimate (think, e.g., $\delta = 1/20$), we would like the right-hand side in the above to be at most δ.

$$\begin{aligned} \delta \geq 2\exp\left(-\frac{\theta^2}{2+\theta}\cdot n\right) \quad &\Leftrightarrow \quad \exp\left(\frac{\theta^2}{2+\theta}n\right) \geq \frac{2}{\delta} \\ &\Leftrightarrow \quad \frac{\theta^2}{2+\theta}n \geq \ln\frac{2}{\delta} \\ &\Leftrightarrow \quad n \geq \frac{2+\theta}{\theta^2}\ln\frac{2}{\delta}. \end{aligned}$$

We have thus proved the following very important theorem. (NB: As is traditional, we've called the accuracy "ϵ" in the below, rather than "θ".)

Sampling Theorem: Suppose we use independent, uniformly random samples to estimate p, the fraction of a population with some property. If the number of samples n we use satisfies

$$n \geq \frac{2+\epsilon}{\epsilon^2}\ln\frac{2}{\delta},$$

then we can assert that our estimate $\overline{X}$ satisfies

$$\overline{X} \in [p-\epsilon, p+\epsilon] \text{ with probability at least } 1-\delta.$$

Some comments:

- That range $[p-\epsilon, p+\epsilon]$ is sometimes called the *confidence interval*.
- Due to the slightly complicated statement of the bound, sometimes people will just write the slightly worse bounds
$$n \geq \frac{3}{\epsilon^2}\ln\frac{2}{\delta},$$
or even
$$n \geq O\left(\frac{1}{\epsilon^2}\ln\frac{2}{\delta}\right).$$
- One beauty of the Sampling Theorem is that the number of samples n you need *does not depend on the size of the total population*. In other words, it doesn't matter how big the country is, the number of samples you need to get a certain accuracy and a certain confidence only depends on that accuracy and confidence.
- In the example we talked about earlier we were interested in accuracy $\epsilon = 2\%$ and confidence 95%, meaning $\delta = 1/20$. So the Sampling Theorem tells us we need at least
$$n \geq \frac{2+.02}{(.02)^2}\ln\frac{2}{1/20} = 5050\ln 40 \approx 18600.$$
Not so bad: you only need to call 18600 or so folks! Er, well, actually, you need to get 18600 folks to respond. And you need to make sure that the events "person responds" and "person approves of the president" are independent. (Hmm... maybe being a pollster is not as easy as it sounds...)
- As you can see from the form of the bound in the Sampling Theorem, the really costly thing is getting high accuracy: $1/\epsilon^2$ is a fairly high price to have to pay in the number of samples. On the other hand, getting really high confidence is really cheap: because of the ln, it hardly costs anything to get δ really tiny.

3 The Chernoff + Union Bound method

Just as Linearity of Expectation and Indicator random variables often come together in glorious marriage, so too do the Chernoff Bound and the Union Bound. You remember the humble Union Bound, right?

Union Bound: $\mathbf{Pr}[B_1 \cup B_2 \cup \cdots \cup B_n] \leq \sum_{i=1}^{n} \mathbf{Pr}[B_i]$.

The idea behind the Chernoff + Union Bound method is the following: The Chernoff Bound is extraordinarily strong, usually showing that the probability a certain "bad event" happens is extremely tiny. Thus, even if very many different bad events exist, if you bound each one's probability by something extremely tiny, you can afford to just add up the probabilities. I.e.,

$$\mathbf{Pr}[\text{anything bad at all}] = \mathbf{Pr}[\text{Bad}_1 \cup \cdots \cup \text{Bad}_{\text{Large}}] \leq \sum_{i=1}^{\text{Large}} \mathbf{Pr}[\text{Bad}_i]$$

(this is a high-level picture :) and if the Chernoff Bound implies $\mathbf{Pr}[\text{Bad}_i] \leq$ minuscule for each i, we get

$$\mathbf{Pr}[\text{anything bad at all}] \leq \sum_{i=1}^{\text{Large}} \text{minuscule} = (\text{Large}) \times (\text{minuscule}) = \text{small}.$$

(Told you it was high-level.)

Let's do an example to make this concrete.

3.1 Random load balancing

Suppose you are a content delivery network — say, YouTube. Suppose that in a typical five-minute time period, you get a million content requests, and each needs to be served from one of your, say, 1000 servers. How should you distribute the requests (let's call them 'jobs') across your servers to balance the load? You might consider a round-robin policy, or a policy wherein you send each job to the server with the lowest load. But each of these requires maintaining some state and/or statistics, which might cause slight delays. You might instead consider the following extremely simple and lightweight policy, which is surprisingly effective: assign each job to a random server.

Let's abstract things slightly. Suppose we have k servers and n jobs. Assume all n jobs arrive very quickly, we assign each to a random server (independently), and the jobs take a while to process. What we are interested in the *load* of the servers.

ASSUMPTION: n is much bigger than k.

E.g., our YouTube example had $n = 10^6$, $k = 10^3$.

Question: The average "load" — jobs per server — will of course be n/k. But how close to perfectly balanced will things be? In particular, is it true that the *maximum* load is not much bigger than n/k, with high probability?

Answer: Yes! Let's do the analysis.

Let X_i denote the number of jobs assigned to server i, for $1 \leq i \leq k$.

Question: What is the distribution of the random variable X_i?

Answer: If you think a little bit carefully, you see that X_i is a binomial random variable: $X_i \sim \text{Binomial}(n, 1/k)$. To see it, just imagine staring at the ith server. For each of n trials/jobs, there is a $1/k$ chance that that job gets thrown onto this ith server.

Don't be confused by notation, by the way — we used to use subscripted X's like X_i to denote Bernoulli random variables, and their sums were Binomial random variables denoted X. Here we have that each X_i itself is a Binomial random variable.

Question: Are $X_1, \ldots, X_k$ independent random variables?

Answer: No! Here is one non-rigorous but intuitive reason: we *know* that it will always be the case that

$$\sum_{i=1}^{k} X_i = n.$$

So in particular, if I tell you the value of $X_1, \ldots, X_{k-1}$, you know exactly what X_k is. Bear in mind that this reasoning does not *formally* prove that $X_1, \ldots, X_k$ are not independent. But it's not hard to do that either. Here's one way: if $X_1, \ldots, X_k$ *were* independent, we'd have

$$\mathbf{Pr}[X_1 = n \cap X_2 = n \cap \cdots \cap X_k = n] = \mathbf{Pr}[X_1 = n] \cdot \mathbf{Pr}[X_2 = n] \cdots \mathbf{Pr}[X_k = n].$$

Now the left-hand side above is 0, since there's no way each server can have all n jobs! But the right-hand side is *nonzero*; each $\mathbf{Pr}[X_i = n] = (1/k)^n$.

But you know what? It's cool. We're going to be using the Union Bound, and one beauty of the Union Bound is that (like Linearity of Expectation), it does not care whether the events it involves are independent or not.

The average load after doing random load balancing is clearly n/k, just because there are n total jobs and k servers. Further, we have

$$\mathbf{E}[X_i] = n/k$$

for each i (by the formula $\mathbf{E}[\text{Binomial}(n, p)] = np$, e.g.). So, as is intuitive, each server is expected to have n/k jobs. But we are interested in the *maximum* load among the k servers...

Let

$$M = \max(X_1, X_2, \ldots, X_k),$$

a random variable representing the maximum load. Our goal is to show a statement of the form

$$\mathbf{Pr}[M \geq n/k + \text{c}] \leq \text{small}$$

where c is not too large. Then we'll be able to say, "With high probability, the maximum load is at most $n/k + c$." We'll let you in on a secret: we're eventually going to be able to take

$$c = 3\sqrt{\ln k}\sqrt{n/k}.$$

You might say to yourself, "uh, is that good?" Yeah, it's pretty good, actually. The $3\sqrt{\ln k}$ part is really quite small, and the rest of the deviation, $\sqrt{n/k}$, is only the square-root of the average load. So, in a concrete setting with $n = 10^6$ and $k = 10^3$, the average load is

$$n/k = 1000,$$

and our deviation here is

$$c = 3\sqrt{\ln 1000}\sqrt{1000} \approx 250.$$

So our result will show, "With high probability, the *maximum* load is at most 1250." We'll see what "high probability" means next lecture.