Lecture 21: The Chernoff Bound

Anup Rao

February 26, 2018

We discuss the Chernoff Bound.

The central limit theorem is not always the most useful way to understand the distribution of the average of a number of independent samples from the same distribution. Although the CLT asserts that such an average converges to the normal distribution (after all he right scaling is done), it does not tell us how *fast* the convergence happens.

The Chernoff bound gives a much tighter control on the probability that a sum of independent random variables deviates from its expectation.

Although here we study it only for for the sums of bits, you can use the same methods to get a similar strong bound for the sum of independent samples for any real-valued distribution of small variance. We do not discuss the more general setting here.

There are several other kinds of bounds like *Hoeffding* bounds and *Azuma's inequality* that are closely related to Chernoff bounds.

Suppose $X_1, \ldots, X_n$ are independent random variables taking values in $\{0,1\}$, and let $X = X_1 + X_2 + \ldots + X_n$ be their sum, and $\mathbb{E}[X] = \mu$. There are many forms of the Chernoff bounds, but here we focus on this one:

Theorem 1. *Suppose* $0 < \delta$, *then*

$$p(X \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2 \mu}{2+\delta}},$$

and

$$p(X \leq (1-\delta)\mu) \leq e^{-\frac{\delta^2 \mu}{2}}.$$

You can combine both inequalities into one if you write it like this:

Theorem 2. *Suppose* $0 < \delta$, *then*

$$p(|X-\mu| > \delta\mu) \leq 2e^{-\frac{\delta^2 \mu}{2+\delta}}.$$

The proof is conceptually similar to the proof of Chebyshev's inequality—we use Markov's inequality applied to the right function of X. We will not do the whole proof here, but consider the random variable e^X.

We have

$$e^X = e^{X_1 + X_2 + \ldots + X_n} = e^{X_1} \cdot e^{X_2} \cdot e^{X_3} \ldots e^{X_n}.$$

Since $X_1, X_2, \ldots, X_n$ are mutually independent, this means that

$$\mathbb{E}\left[e^X\right] = \mathbb{E}\left[e^{X_1} \ldots e^{X_n}\right] = \mathbb{E}\left[e^{X_1}\right] \ldots \mathbb{E}\left[e^{X_n}\right].$$

Now we have

$$\mathbb{E}\left[e^{X_1}\right] = pe^1 + (1-p)e^0 = p(e-1) + 1 \leq e^{p(e-1)}.$$

since $1 + y \leq e^y$ for all y

So, by Markov's inequality,

$$p(X > an) = p(e^X > e^{an}) \leq \frac{\mathbb{E}\left[e^X\right]}{e^{an}} = e^{n(p(e-1)-a)},$$

which is exponentially small in n, when $a > p(e-1)$. The actual proof of the Chernoff bound comes from using calculus to determine the right constant to use instead of e in the above argument.

Example: Fair coin

Suppose you toss a fair coin 200 times. How likely is it that you see at least 120 heads?

The Chernoff bound says

$$\begin{aligned} p(X \geq 120) &= p(X \geq (1 + 20/100)100) \\ &\leq e^{-\frac{(1/5)^2)}{2+1/5} \cdot 100} = e^{-20/6} = 0.0356. \end{aligned}$$

Example: Polling

Suppose we want to conduct a national poll to estimate the fraction of people that support the Green party. We can just sample n uniformly random people and ask them if they support the party or not. What can we say about the accuracy of our poll?

To set this up, let us imagine that the true fraction of the country that supports the Green party is p. Let X_i be the indicator variable for whether or not the i'th person polled supports the Green party. Then $X_1, \ldots, X_n$ are independent Bernoulli variables, each of which is 1 with probability p. If we set $X = X_1 + X_2 + \ldots + X_n$, then we have $\mathbb{E}\left[X\right] = np$.

Suppose we want our estimate to be within θ of p. The Chernoff bound gives:

$$p(|X/n - p| > \delta p) = p(|X - pn| > \delta pn) \leq 2e^{-\frac{\delta^2 p}{2+\delta} \cdot n}.$$

Setting $\delta p = \theta$, we get

$$p(|X/n - p| > \theta) \leq 2e^{-\frac{\theta^2/p^2 \cdot p}{2+\theta/p} \cdot n} = 2e^{-\frac{\theta^2}{2p+\theta} \cdot n} \leq 2e^{-\frac{\theta^2}{2+\theta} \cdot n},$$

where in the last inequality we used the fact that $p \leq 1$.

So, if we want the probability of our estimate being off by θ to be at most ϵ, then we want

$$\begin{aligned} & 2e^{-\frac{\theta^2}{2+\theta}\cdot n} \leq \epsilon \\ \Rightarrow & e^{\frac{\theta^2}{2+\theta}\cdot n} \geq 2/\epsilon \\ \Rightarrow & \frac{\theta^2}{2+\theta}\cdot n \geq \ln(2/\epsilon) \\ \Rightarrow & n \geq \frac{2+\theta}{\theta^2}\cdot \ln(2/\epsilon). \end{aligned}$$

As long as n satisfies is large enough as above, we have that $p - \theta \leq X/n \leq p + \theta$ with probability at least $1 - \delta$.

The interval $[p - \theta, p + \theta]$ is sometimes called the *confidence interval*.

For example, if we want $\theta = 0.05$, and ϵ to be 1 in a hundred, we need to set $n \geq 4345$.

Example: Distributed Load Balancing

A common problem when handling a large website is *load balancing*. You have k servers dedicated to handling jobs, and you get $n \gg k$ jobs. How do you distribute the jobs?

Of course, you would like to distribute these jobs to the k servers as evenly as possible, but this is not as simple as it seems. The n jobs could be coming in a distributed fashion, so there is no single computer that knows how many requests are out there.

A simple solution is to just assign the requests to servers completely randomly. If we do this, we expect that each server will see n/k jobs on average. What can we say about the *maximum* load experienced by any one server?

Let $X_1, \ldots, X_k$ denote the number of jobs assigned to each of the servers. Then we see that each X_i is a binomial random variable, since each job is assigned to server i with probability $1/k$.

Are $X_1, \ldots, X_k$ independent?

Claim 3. *If $n > 9k \ln k$, then $p(X_i > n/k + 3\sqrt{n \ln k/k}) < 1/k^3$.*

To see the claim, we apply the Chernoff bound from Theorem 1 with $\delta = 3\sqrt{k \ln k/n} < 1$:

$$\begin{aligned} p(X_i > n/k + 3\sqrt{n \ln k/k}) &= p(X_i > n/k(1 + 3\sqrt{k \ln k/n})) \\ &\leq e^{-\frac{(3\sqrt{k \ln k/n})^2}{3}\cdot n/k} \\ &= e^{-3 \ln k} = 1/k^3. \end{aligned}$$

For example, if we have a thousand servers and a million jobs, this bound says that the probability that a single server sees more than $1000 + 3\sqrt{1000 \ln 1000} = 1249.38$ jobs is at most one in a billion!

By the union bound, the probability that any single server sees more than $n/k + 3\sqrt{n \ln k/k}$ jobs is at most $k \cdot 1/k^3 = 1/k^2$. This is still one in a million for the numbers we have picked.