Lecture 22: Moment Generating Functions

Anup Rao

May 22, 2019

Last time, we stated and used the Chernoff-Hoeffding bound. Suppose $X_1, \dots, X_n$ are independent random variables taking values in between 0 and 1, and let $X = X_1 + X_2 + \dots + X_n$ be their sum, and $\mathbb{E}[X] = \mu$.

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:

If $X_1, \dots, X_n$ do not lie in between 0 and 1, you can always scale them so that they do, and then apply the bound. Careful though—scaling the random variables changes the value of μ as well!

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

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

When $\delta \leq 1$, we have $e^{-\frac{\delta^2 \mu}{2+\delta}} \leq e^{-\frac{\delta^2 \mu}{3}}$, which is easier to work with sometimes.

Distributed Load Balancing

A common problem when handling a high traffic 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. In addition, servers can be going offline and coming online at different times to account for maintenance.

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

Let $X_1, \dots, X_k$ 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 $X_1, \dots, X_k$ 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.

Intuition for the Proof of the Chernoff-Hoeffding bound

The proof of the bound 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 let us prove something weaker here. 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

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

Now suppose:

$$\mathbb{E}\left[e^{X_1}\right] = c,$$

for some constant c. Then, by Markov's inequality,

$$p(X > \alpha n) = p(e^X > e^{\alpha n}) \leq \frac{\mathbb{E}\left[e^X\right]}{e^{\alpha n}} = c^n / e^{\alpha n},$$

which is exponentially small in n, when $\alpha > \ln c$. The actual proof of the Chernoff-Hoeffding bound comes from using calculus to determine the right constant to use instead of e in the above argument.

Moment Generating Functions

THE OUTLINE OF THE PROOF WE SAW ABOVE naturally leads us to an important concept in probability theory—*moment generating functions*. Given a real valued random variable X, its moment generating function is a function $M : \mathbb{R} \to \mathbb{R}$ given by

$$M(t) = \mathbb{E}\left[e^{tX}\right].$$

This is a function that maps every number t to another number.

To explain the name, we first need to explain what a moment is. The moments of X are the numbers $\mathbb{E}[X], \mathbb{E}\left[X^2\right], \mathbb{E}\left[X^3\right], \ldots$. These are important statistics of a distribution. We have already seen that the first two moments determine the variance of X. The other moments provide lots of other information about X.

Now, by linearity of expectation and the Taylor series for e, we obtain:

$$\begin{aligned} M(t) &= \mathbb{E}\left[e^{tX}\right] \\ &= \mathbb{E}\left[1 + tX + (tX)^2/2 + (tX)^3/3! + \cdots\right] \\ &= 1 + \mathbb{E}[X] \cdot t + \mathbb{E}\left[X^2\right] \cdot t^2/2 + \mathbb{E}\left[X^3\right] \cdot t/3! + \cdots . \end{aligned}$$

And this means that the moment generating function *determines all the moments* of X. Indeed, to compute the k'th moment of X, you just need to take the k'th derivative of $M(t)$ with respect to t, and evaluate this at 0. For example, the 3'rd derivative of $M(t)$ is

$$\mathbb{E}\left[X^3\right] + \mathbb{E}\left[X^4\right] \cdot t + \mathbb{E}\left[X^5\right] \cdot t^2/2 + \cdots ,$$

so when $t = 0$, this is the same as $\mathbb{E}\left[X^3\right]$.

Moreover, we have an important fact:

Theorem 4. *If X, Y have the same moment generating function, then they have the same cumulative distribution function.*

In fact, this theorem is how the central limit theorem is proved. You can show that no matter what distribution you start with, the moment generating function of $X_1 + X_2 + \cdots + X_n$ converges to the moment generating function for the normal distribution.

Example: Adding two Independent copies

Suppose X, Y are independent with moment generating functions $M_x(t), M_y(t)$. Then the moment generating function of $X + Y$ is just

$M_x(t) \cdot M_y(t)$. This is because:

$$\begin{aligned}
\mathbb{E}\left[e^{t(X+Y)}\right] &= \mathbb{E}\left[e^{tX+tY}\right] \\
&= \mathbb{E}\left[e^{tX}e^{tY}\right] \\
&= \mathbb{E}\left[e^{tX}\right] \cdot \mathbb{E}\left[e^{tY}\right] && \text{since } X, Y \text{ are independent} \\
&= M_x(t) \cdot M_y(t).
\end{aligned}$$

Example: Poisson

Suppose X is poisson with parameter λ. Then the moment generating function of X is

$$\begin{aligned}
M(t) &= \mathbb{E}\left[e^{tX}\right] \\
&= \sum_{k=0}^{\infty} e^{tk} e^{-\lambda} (\lambda)^k / k! \\
&= e^{-\lambda} \sum_{k=0}^{\infty} (e^t \lambda)^k / k! \\
&= e^{-\lambda} e^{\lambda e^t} \\
&= e^{\lambda(e^t - 1)}.
\end{aligned}$$

So, for example, if X is Poisson with parameter λ_1 and Y is Poisson with parameter λ_2, then $X + Y$ has moment generating function $e^{\lambda_1(e^t-1)} e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}$. Since this is identical to the moment generating function of a Poisson with parameter $\lambda_1 + \lambda_2$, this proves that $X + Y$ has the same distribution as a Poisson with parameter $\lambda_1 + \lambda_2$, something you proved on your homework using the binomial identity.