Chernoff Bounds

Piyush Patil

February 21, 2017

In the last chapter, we looked at some preliminary methods for bounding the tail distributions. In particular, we looked at the Markov inequality, which uses the expectation of a random variable to bound its tail distribution, and Chebyshev's inequality, which uses the variance of a random variable to achieve a tighter bound. In general, the moments of a random variable completely determine the underlying probability distribution of the random variable, so it makes sense that if we make use of more probabilistic moments we can achieve tighter bounds on tail distributions. Chernoff bounds are such bounds, and are extremely powerful, providing exponentially decreasing bounds on the tail distribtion. In this chapter, we'll start off by discussing the moments of a random variable and how to manipulate them, and then apply the theory to derive Chernoff bounds for common distributions.

1 Moment Generating Functions

Clearly, there are probability distributions with the same expectation and variance but with completely different values. This shows that the expectation and variance alone are, unsurprisingly, not enough to fully characterize the distribution. As we introduce more and more moments, we add further constraints on the shape of the distribution and hence its values, converging to the underlying distribution of the random variable in question.
In general, generating functions are a way of treating real sequences, by viewing them as coefficients of a series. Specifically, the exponential generating function of a sequence $(a_n)_{n\in\mathbb{N}}$ is defined

$$EG(a_n) = \sum_{n\in\mathbb{N}} a_n x^n$$

Here, x is an indetermine, making the function a formal series. The moment generating function is simply the exponential generating function of the moments of a random variable.

(Definition) Moment generating function: *The **moment generating function** of a random variable X is defined*

$$M_X(t) = \mathbb{E}[e^{tX}]$$

This function is often viewed as the Taylor series defined by the moments of a random variable about zero. This is because of the convenient property

$$M_X(t) = \sum_{n\in\mathbb{N}} \frac{t^n}{n!}\mathbb{E}[X^n]$$

that we obtain by examining the Taylor series of the exponential function; this means (this relies on the assumption that expectation and differentiation are commutative operators)

$$\mathbb{E}[X^n] = M_X^{(n)}(0)$$

In other words, the moment generating function neatly captures all the moments of X. Knowing this, it isn't surprising that the moment generating function turns out to be an alternate description of a probability distribution, as opposed to the density function; the two are equivalent and can be used to derive each other, and are merely alternate formulations of the same distribution. However, it should be noted that the moment generating function is not guaranteed to converge everywhere, and moreover not every random variable has a well defined moment generating function.

(Theorem) Moment generating functions fully characterize random variables: *Let X and Y be two random variables. If*

$$\exists \delta \in \mathbb{R}^+ \text{ s.t. } \forall t \in [-\delta, \delta] : M_X(t) = M_Y(t)$$

then X and Y have the same distribution.

- *Proof*: TODO

Because the moment generating function is defined in terms of the exponential function, it shares its property of turning addition into multiplication with respect to independent random variables X and Y:

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

2 Deriving Chernoff Bounds

We can now introduce Chernoff bounds, which are obtained by applying Markov's inequality to the object of the moment generating function of a random variable X: e^{tX}. If we do so, we find

$$\forall a : \Pr(X \geq a) = \Pr(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}}$$

Since this holds for any positive t, it follows that

$$\Pr(X \leq a) \leq \min_{t<0} \frac{\mathbb{E}[e^{tX}]}{e^{ta}}$$

Bounds derived from this approach are known as *Chernoff bounds*. The most common application of Chernoff bounds are to the tail distribution of the sum of Bernoulli indicator variables (variables which are one when some binary event occurs, or zero otherwise). Such sums are known as *Poisson trials*, and are a generalization of Bernoulli trials (flipping a biased coin n times) in that each indicator variable can have a different distribution. Thus, the Chernoff bound for Poisson trials will also apply to Bernoulli trials, which are a special case. Let X_i, for $1 \leq i \leq n$ be independent indicator variables which are, respectively, 1 with probability p_i and 0 otherwise. Defining X as the sum of these variables, it follows that

$$X = \sum_{i=1}^{n} X_i \rightarrow \mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p_i$$

To find a Chernoff bound, we'll need to first identify the moment generating function for each X_i:

$$M_{X_i}(t) = \mathbb{E}[e^{tX_i}]p_i e^t + (1 - p_i) = 1 + p_i \cdot (e^t - 1) \leq e^{p_i \cdot (e^t - 1)}$$

It follows that since the X_i are independent,

$$M_X(t) = \prod_{i=1}^{n} M_{X_i}(t) \leq \prod_{i=1}^{n} e^{p_i \cdot (e^t - 1)} = \exp\left((e^t - 1)\sum_{i=1}^{n} p_i\right) = \exp((e^t - 1)\mathbb{E}[X])$$

Now that we've bounded the moment generating function, we can develop Chernoff bounds on X.

(Theorem) Chernoff bounds above the mean on Poisson trials: *Let $X_1, \cdots, X_n$ be Poisson trials with $\Pr(X_i = 1) = p_i$, and let X be their sum. Letting $\mu = \mathbb{E}[X]$, the following Chernoff bounds hold.*

1. $$\forall \delta \in \mathbb{R}^+ : \Pr(X \geq (1 + \delta)\mu) < \left(\frac{e^\delta}{(1 + \delta)^{1+\delta}}\right)^\mu$$

2. $$\forall \delta \in (0, 1] : \Pr(X \geq (1 + \delta)\mu) \leq \exp\left(-\frac{\mu\delta^2}{3}\right)$$

3. $$\forall R \geq 6\mu : \Pr(X \geq R) \leq 2^{-R}$$

• *Proof*: The first bound is the strongest, and we can use it to derive the other two, so we'll prove that one first. Applying Markov's inequality to an arbitrary positive t,

$$\Pr(X \geq (1 + \delta)\mu) = \Pr(e^{tX} \geq e^{t(1+\delta)\mu}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{t(1+\delta)\mu}} \leq \frac{e^{(e^t - 1)\mu}}{e^{t(1+\delta)\mu}}$$

The first Chernoff bound follows by setting $t = \log(1 + \delta)$:

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

The second bound is equivalent to showing that, for $0 < \delta \leq 1$,

$$\frac{e^\delta}{(1+\delta)^{1+\delta}} \leq \exp(\frac{-\delta^2}{3})$$

if we raise both side to μ. This relation in turn is equivalent to

$$f(\delta) := \delta - (1+\delta)\log(1+\delta) + \frac{\delta^2}{3} \leq 0$$

if we take the logarithm of both sides. We can prove this inequality with basic calculus: notice that

$$f'(\delta) = -\log(1+\delta) + \frac{2\delta}{3}$$

and

$$f''(\delta) = -\frac{1}{1+\delta} + \frac{2}{3}$$

On the interval $(0, 1]$, $f''(\delta)$ is negative on the first half and positive on the second half, which means $f'(\delta)$ decreases monotonically on the first half, then increases monotonically on the second half. But $f'(0) = 0$ and $f'(1) < 0$, which means there's no way the function can become positive. Thus,

$$f(\delta) \leq 0$$

which proves the second bound. The third inequality follows easily by, for arbitrary $R \geq 6\mu$, defining δ such that $R = (1+\delta)\mu$, since then

$$\Pr(X \geq R) = \Pr(X \geq (1+\delta)\mu) \leq \left(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right)^\mu \leq \left(\frac{e}{1+\delta}\right)^{(1+\delta)\mu} \leq \left(\frac{e}{6}\right)^R \leq 2^{-R}$$

□

We obtain symmetric bounds for deviations below the mean.

(Theorem) Chernoff bounds below the mean on Poisson trials: *Let $X_1, \cdots, X_n$ be independent Poisson trials with $\Pr(X_i = 1) = p_i$, and let X be their sum. Letting $\mu = \mathbb{E}[X]$, the following Chernoff bounds hold.*

1. $$\Pr(X \leq (1-\delta)\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^\mu$$

2. $$\Pr(X \leq (1-\delta)\mu) \leq \exp\left(-\frac{\mu\delta^2}{2}\right)$$

• *Proof*: TODO

A corollary is that for independent Poisson trials $X_1, \cdots, X_n$ where $\Pr(X_i) = p_i$,

$$\forall \delta \in (0,1) : \Pr(|X - \mu| \geq \delta\mu) \leq 2e^{-\frac{\mu\delta^2}{3}}$$

where X is the sum of the X_i and $\mu = \mathbb{E}[X]$. This simpler form is a more commonly used Chernoff bound. In practice, we often don't have the exact value for $\mathbb{E}[X]$, but we can still obtain upper and lower bounds if we can find some $\mu \geq \mathbb{E}[X]$ or $\mu \leq \mathbb{E}[X]$ and apply the above Chernoff bound for values above the mean or values below the mean, respectively.

Let's now move on to the next important application of Chernoff bounds. One of the most prominent and foundational problems in statistics is to use samples from some underlying distribution to understand the nature of the distribution. As a special case, suppose the underlying distribution is Bernoulli with unknown parameter p. We can think of the underlying distribution as a collection of infinitely many values, each of which is in one of two classes, with probability p. Thus, p is the proportion of the values which fall into the first class. If we can obtain n sample values, we can find the proportion of the samples which fall into the first class. Let q be this sample proportion; intuitively, as we sample more and more, we'd expect q to approach p, ie

$$\lim_{n\to\infty} q = p$$

where q is the sample proportion for n samples. One way to express this intuition is with the idea that given any margin of error about p, we can ensure that q fits in that margin of error given a sufficiently large number of samples. We formalize

this idea as follows.

(Definition) Confidence interval: *A* $1-\gamma$ ***confidence interval*** *for parameter* p *is an interval* $[q-\delta, q+\delta]$ *such that*

$$\Pr(p \in [q-\delta, q+\delta]) \geq 1-\gamma$$

Intuitively, a confidence interval is simply an estimate q for the parameter p with a neighborhood that p has a $1-\gamma$ probability of being in. We can thus use q as an estimate for p, within the margin of error δ of course, if we're willing to accept an error rate of γ for not capturing p in our margin of error. Of course, we can trivially find such intervals for any given γ if we make δ large enough; we're interested in confidence intervals with minimal interval size 2δ and error rate γ. As we'd expect intuitively, it's difficult to minimize both these quantities simultaneously - the two are a trade-off, since the lower we want our error to be, the larger the interval required to capture p, and conversely the smaller the interval is, the more likely that we missed p.
Let's derive the trade-off between the interval size and error probability. Assuming we've drawn n samples from the underlying distribution uniformly at random, let $X = nq$ be the number of samples which fall into the first class. Then X has a binomial distirbution with parameters n and p, ie

$$\mathbb{E}[X] = np$$

If p isn't captured in a confidence interval, ie $p \notin [q-\delta, q+\delta]$, then either $p < q-\delta$ or $p > q+\delta$. In the first case, this means $X = nq > n \cdot (p+\delta) = (np)(1+\frac{\delta}{p}) = (1+\frac{\delta}{p})\mathbb{E}[X]$. In the second case, $X = nq < n \cdot (p-\delta) = (1-\frac{\delta}{p})\mathbb{E}[X]$. We can now apply the Chernoff bounds for values above and below the mean, to the two cases, obtaining

$$\begin{aligned}\Pr(p \notin [q-\delta, q+\delta]) &= \Pr(p < q-\delta) + \Pr(p > q+\delta) = \Pr\left(X < \mathbb{E}[X] \cdot \left(1-\frac{\delta}{p}\right)\right) + \Pr\left(X > \mathbb{E}[X] \cdot \left(1+\frac{\delta}{p}\right)\right) \\ &< \exp\left(-\frac{\mathbb{E}[X]\delta^2}{3p^2}\right) + \exp\left(-\frac{\mathbb{E}[X]\delta^2}{2p^2}\right) = \exp\left(-\frac{n\delta^2}{3p}\right) + \exp\left(-\frac{n\delta^2}{2p}\right)\end{aligned}$$

Of course, this isn't very useful since we don't know the value of p; can use the fact that $p \leq 1$ to obtain the weaker bound

$$\gamma = \Pr(p \notin [q-\delta, q+\delta]) < \exp\left(-\frac{n\delta^2}{2}\right) + \exp\left(-\frac{n\delta^2}{3}\right)$$

Thus, the larger we smaller we force our confidence interval to be, by decreasing δ, the larger the error probability γ becomes.

3 Stronger Chernoff Bounds in Special Cases

Let's look at some common special cases in which we can obtain strong Chernoff bounds on random variables. First, let's consider the sum of independent random variables, each of which is either 1 or -1 with equal probability. Letting $X_1, \cdots, X_n$ be the random variables and X be their sum, so that $\Pr(X_i = 1) = \Pr(X_i = -1) = \frac{1}{2}$, we have

$$\forall a > 0 : \Pr(X \geq a) \leq \exp\left(-\frac{a^2}{2n}\right)$$

To prove this, we turn to the moment generating function of any of the X_i:

$$\forall t > 0 : \mathbb{E}[e^{tX_i}] = \frac{1}{2}e^t + \frac{1}{2}e^{-t} = \sum_{k \in N} \frac{t^k}{k!} + \sum_{k \in \mathbb{N}} (-1)^k \frac{t^k}{k!} = \sum_{k \in \mathbb{N}} \frac{t^{2k}}{(2k)!} \leq \sum_{k \in \mathbb{N}} \frac{\left(\frac{t^2}{2}\right)^k}{k!} = e^{\frac{t^2}{2}}$$

where we've used the Taylor series of the exponential function. It follows that

$$\mathbb{E}[e^{tX}] = \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}] \leq e^{\frac{nt^2}{2}}$$

We can apply Markov's inequality as follows

$$\Pr(X \geq a) = \Pr(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}} \leq \exp\left(\frac{nt^2}{2} - ta\right)$$

This holds for any positive t. Given a positive a, the result follows by setting $t = \frac{a}{n}$:

$$\Pr(X \geq a) \leq e^{-\frac{a^2}{2n}}$$

□By symmetry, we can prove the same bound for $\Pr(X \leq -a)$, so it follows as a corollary that

$$\Pr(|X| \geq a) \leq 2e^{-\frac{a^2}{2n}}$$

We can adapt this result for fair Bernoulli random variables Y_i, which assume the values 0 and 1 with equal probability, using the transformation

$$Y_i = \frac{X_i + 1}{2}$$

Letting Y be the sum of the Y_i, and $\mu = \mathbb{E}[X] = \frac{n}{2}$, we have for any positive a

$$\Pr(Y \geq \mu + a) = \Pr(\frac{1}{2}X + \mu \geq \mu + a) = \Pr(X \geq 2a) \leq e^{-\frac{2a^2}{n}}$$

Letting $a = \delta\mu = \frac{\delta n}{2}$, we have

$$\Pr(Y \geq (1+\delta)\mu) = \Pr(X \geq 2\delta\mu) \leq e^{-\frac{2\delta^2\mu^2}{n}} = e^{-\delta^2\mu}$$

Comparing this to the Chernoff bound for values above the mean that we found earlier, the above bound has a constant factor of one in the exponent, rather than a third, and is stronger. Similarly, we can prove

$$\Pr(Y \leq (1-\delta)\mu) \leq e^{-\delta^2\mu}$$