Chapter 6. Concentration Inequalities

6.2: The Chernoff Bound

Slides (Google Drive) Alex Tsun Video (YouTube)

The more we know about a distribution, the stronger concentration inequality we can derive. We know that Markov's inequality is weak, since we only use the expectation of a random variable to get the probability bound. Chebyshev's inequality is a bit stronger, because we incorporate the variance into the probability bound. However, as we showed in the example in 6.1, these bounds are still pretty "loose". (They are tight in some cases though).

What if we know even more? In particular, its PMF/PDF and hence MGF? That will allow us to have an even stronger bound. The Chernoff bound is derived using a combination of Markov's inequality and moment generating functions.

6.2.1 The Chernoff Bound for the Binomial Distribution

Here is the idea for the Chernoff bound. We will only derive it for the Binomial distribution, but the same idea can be applied to any distribution.

Let X be any random variable. e^{tX} is always a non-negative random variable. Thus, for any $t > 0$, using Markov's inequality and the definition of MGF:

$$\begin{aligned}
\mathbb{P}(X \geq k) &= \mathbb{P}\left(e^{tX} \geq e^{tk}\right) && [\text{since } t > 0 \text{. if } t < 0 \text{, flip the inequality.}] \\
&\leq \frac{\mathbb{E}\left[e^{tX}\right]}{e^{tk}} && [\text{Markov's inequality}] \\
&= \frac{M_X(t)}{e^{tk}} && [\text{def of MGF}]
\end{aligned}$$

(Note that the first line requires $t > 0$, otherwise it would change to $\mathbb{P}\left(e^{tX} \leq e^{tk}\right)$. This is because $e^t > 1$ for $t > 0$ so we get something like 2^X which is monotone increasing. If $t < 0$, then $e^t < 1$ so we get something like 0.3^X which is monotone decreasing.)

Now the right hand side holds for (uncountably) infinitely many t. For example, if we plugged in $t = 0.5$ we might get $\frac{M_X(t)}{e^{tk}} = 0.53$ and if we plugged in $t = 3.26$ we might get 0.21. Since $\mathbb{P}(X \geq k)$ has to be less than **all** the possible values when plugging in different $t > 0$, it in particular must be less than the **minimum** of all the values.

$$\mathbb{P}(X \geq k) \leq \min_{t>0} \frac{M_X(t)}{e^{tk}}$$

This is good - if we can minimize the right hand side, we can get a very tight/strong bound.

We'll now focus our attention to deriving the Chernoff bound when X has a Binomial distribution. Everything above applies generally though.

Theorem 6.2.1: Chernoff Bound for Binomial Distribution

Let $X \sim \text{Bin}(n,p)$ and let $\mu = \mathbb{E}[X]$. For any $0 < \delta < 1$:
Upper tail bound:

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{3}\right)$$

Lower tail bound:

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

where $\exp(x) = e^x$.

The Chernoff bound will allow us to bound the probability that X is larger than some multiple of its mean, or less than or equal to it. These are the *tails* of a distribution as you go farther in either direction from the mean. For example, we might want to bound the probability that $X \geq 1.5\mu$ or $X \leq 0.1\mu$.

I think it's completely acceptable if you'd like to not read the proof, as it is very involved algebraically. You can still use the result regardless!

Proof of Chernoff Bound for Binomial.

If $X = \sum_{i=1}^n X_i$ where X_1, X_2,...,X_n are iid variables, then since the MGF of the (independent) sum equals the product of the MGFs. Taking our general result from above and using this fact, we get:

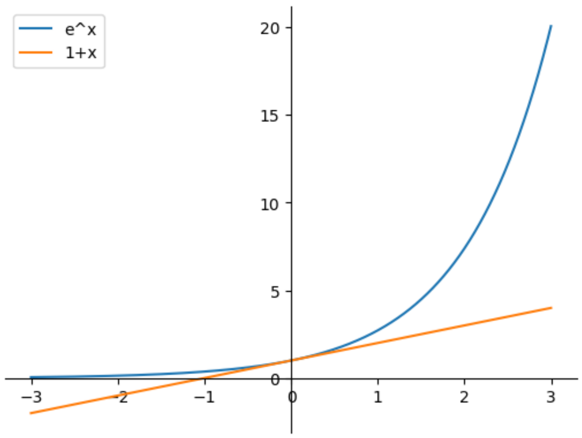
$$\mathbb{P}(X \geq k) \leq \min_{t>0} \frac{M_X(t)}{e^{tk}} = \min_{t>0} \frac{\prod_{i=1}^n M_{X_i}(t)}{e^{tk}}$$

Let's derive a Chernoff bound for $X \sim \text{Bin}(n,p)$, which has the form $\mathbb{P}(X \geq (1+\delta)\mu)$ for $\delta > 0$. For example with $\delta = 4$, you may want to bound $\mathbb{P}(X \geq 5\mathbb{E}[X])$.

Recall $X = \sum_{i=1}^n X_i$ where $X_i \sim \text{Ber}(p)$ are iid, with $\mu = \mathbb{E}[X] = np$.

$$\begin{aligned}
M_{X_i}(t) &= \mathbb{E}\left[e^{tX_i}\right] && [\text{def of MGF}] \\
&= e^{t\cdot 1}p_{X_i}(1) + e^{t\cdot 0}p_{X_i}(0) && [\text{LOTUS}] \\
&= pe^t + 1(1-p) && [X_i \sim \text{Ber}(p)] \\
&= 1 + p(e^t - 1) \\
&\leq e^{p(e^t-1)} && [1 + x \leq e^x \text{ with } x = p(e^t-1)]
\end{aligned}$$

See here for a pictorial proof that $1 + x \leq e^x$ for any real number x (just plot the two functions). Alternatively, use the Taylor series for e^x to argue this. We use this bound for algebra convenience coming up soon.



Now using the result from earlier and plugging in the MGF for the Ber(p) distribution, we get:

$$\begin{aligned}
\mathbb{P}(X \geq k) &\leq \min_{t>0} \frac{\prod_{i=1}^{n} M_{X_i}(t)}{e^{tk}} && \text{[from earlier]} \\
&= \min_{t>0} \frac{\left(e^{p(e^t-1)}\right)^n}{e^{tk}} && \text{[MGF of Ber}(p)\text{, } n \text{ times]} \\
&= \min_{t>0} \frac{e^{np(e^t-1)}}{e^{tk}} && \text{[algebra]} \\
&= \min_{t>0} \frac{e^{\mu(e^t-1)}}{e^{tk}} && [\mu = np]
\end{aligned}$$

For our bound, we want something like $\mathbb{P}(X \geq (1+\delta)\mu)$, so our $k = (1+\delta)\mu$. To minimize the RHS and get the tightest bound, the best bound we get is by choosing $t = \ln(1+\delta)$ after some terrible algebra (take the derivative and set to 0). We simply plug in k and our optimal value of t to the above equation:

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

Again, we wanted to choose t that minimizes our upper bound for the tail probability. Taking the derivative with respect to t tells us we should plug in $t = \ln(1+\delta)$ to minimize that quantity. This would actually be pretty annoying to plug into a calculator.

We actually can show that the final RHS is $\leq \exp\left(\frac{-\delta^2\mu}{2+\delta}\right)$ with some more messy algebra. Additionally, if we restrict $0 < \delta < 1$, we can simplify this even more to the bound provided earlier:

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(\frac{-\delta^2\mu}{3}\right)$$

The proof of the lower tail is entirely analogous, except optimizing over $t < 0$ when the inequality flips. It proceeds by taking $t = \ln(1-\delta)$.

We also get a lower tail bound:

$$\mathbb{P}(X \leq (1-\delta)\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} \leq \left(\frac{e^{-\delta}}{e^{-\delta+\frac{\delta^2}{2}}}\right)^{\mu} = \exp\left(\frac{-\delta^2\mu}{2}\right)$$

□

You may wonder, why are we bounding $\mathbb{P}(X \geq (1+\delta)\mu)$, when we can just sum the PMF of a binomial to get an exact answer? The reason is, it is very computationally expensive to compute the binomial PMF! For example, if $X \sim \text{Bin}(n = 20000, p = 0.1)$, then by plugging in the PMF, we get

$$\mathbb{P}(X = 13333) = \binom{20000}{13333} 0.1^{13333} 0.9^{20000-13333} = \frac{20000!}{13333!(20000-13333)!} 0.1^{13333} 0.9^{20000-13333}$$

(Actually, $n = 20000$ isn't even that large.) You have to multiply 20,000 numbers on the second two terms, and it multiplies to a number that is infinitesimally small. For the first term (binomial coefficient), computing 20000! is impossible - in fact, it is so large you can't even imagine. You would have to cleverly interleave multiplying a factorial vs the probability, to keep the value in an acceptable range for the computer. Then, sum up a bunch of these....

This is why we have/need the Poisson approximation, the Normal approximation (CLT), and the Chernoff bound for the Binomial!

Example(s)

Suppose $X \sim \text{Bin}(500, 0.2)$. Use Markov's inequality and the Chernoff bound to bound $P(X \geq 150)$, and compare the results.

Solution We have:

$$\mathbb{E}[X] = np = 500 \cdot 0.2 = 150$$
$$\text{Var}(X) = np(1-p) = 500 \cdot 0.2 \cdot 0.8 = 80$$

Using Markov's Inequality:

$$P(X \geq 150) \leq \frac{\mathbb{E}[X]}{150} = \frac{100}{150} \approx 0.6667$$

Using the Chernoff Bound (with $\delta = 0.5$):

$$P(X \geq 150) = P(X \geq (1+0.5) \cdot 100) \leq e^{-\frac{0.5^2 \cdot 100}{3}} \approx 0.00024$$

The Chernoff bound is much stronger! It isn't a fair comparison necessarily, because the Chernoff bound required knowing the MGF (and hence the distribution), whereas Markov only required knowing the mean (and that it was non-negative). □

These examples give you an overall comparison of all three inequalities we learned so far!

Example(s)

Suppose the number of red lights Alex encounters each day to work is on average 4.8 (according to historical trips to work). Alex really will be late if he encounters 8 or more red lights. Let X be the number of lights he gets on a given day.

1. Give a bound for $\mathbb{P}(X \geq 8)$ using Markov's inequality.
2. Give a bound for $\mathbb{P}(X \geq 8)$ using Chebyshev's inequality, if we also assume $\mathsf{Var}(X) = 2.88$.
3. Give a bound for $\mathbb{P}(X \geq 8)$ using the Chernoff bound. Assume that $X \sim \text{Bin}(12, 0.4)$ - that there are 12 traffic lights, and each is independently red with probability 0.4.

4. Compute $\mathbb{P}(X \geq 8)$ exactly using the assumption from the previous part.
5. Compare the three bounds and their assumptions.

1. Since X is nonnegative and we know its expectation, we can apply Markov's inequality:

$$\mathbb{P}(X \geq 8) \leq \frac{\mathbb{E}[X]}{8} = \frac{4.8}{8} = 0.6$$

2. Since we know X's variance, we can apply Chebyshevs inequality after some manipulation. We have to do this to match the form required:

$$\mathbb{P}(X \geq 8) \leq \mathbb{P}(X \geq 8) + \mathbb{P}(X \leq 1.6) = \mathbb{P}(|X - 4.8| \geq 3.2)$$

The reason we chose ≤ 1.6 is so it looks like $\mathbb{P}(|X - \mu| \geq \alpha)$. Now, applying Chebyshev's gives:

$$\leq \frac{\mathsf{Var}(X)}{3.2^2} = \frac{2.88}{3.2^2} = 0.28125$$

3. Actually, $X \sim \text{Bin}(12, 0.4)$ also has $\mathbb{E}[X] = np = 4.8$ and $\mathsf{Var}(X) = np(1-p) = 2.88$ (what a coincidence). The Chernoff bound requires something of the form $\mathbb{P}(X \geq (1+\delta)\mu)$, so we first need to solve for δ: $(1+\delta)4.8 = 8$ so that $\delta = 2/3$. Now,

$$\mathbb{P}(X \geq 8) = \mathbb{P}(X \geq (1 + 2/3) \cdot 4.8) \leq \exp\left(\frac{-(2/3)^2 4.8}{3}\right) \approx 0.4911$$

4. The exact probabiltity can be found summing the Binomial PMF:

$$\mathbb{P}(X \geq 8) = \sum_{k=8}^{12} \binom{12}{k} 0.4^k 0.6^{12-k} \approx 0.0573$$

5. Actually it's usually the case that the bounds are tighter/better as we move down the list Markov, Chebyshev, Chernoff. But in this case Chebyshev's gave us the tightest bound, even after being weakened by including some additional $\mathbb{P}(X \leq 1.6)$. Chernoff bounds will typically be better for farther tails - 8 isn't considered too far from the mean 4.8.

   It's also important to note that we found out more information progressively - we can't blindly apply all these inequalities every time. We need to make sure the conditions for the bound being valid are satisfied.

   Even our best bound of 0.28125 was 5-6x larger than the true probability of 0.0573.