Chernoff Bounds

This note introduces what are commonly called Chernoff bounds. Chernoff bounds are extremely powerful, giving exponentially decreasing bounds on the tail distribution. These bounds are derived by using Markov's inequality.

Often, our goal is to understand how collections of a large number of independent random variables behave. This is what the laws of large numbers reveal. In general, the idea is that the average of a large number of i.i.d. random variables will approach the expectation. Sometimes that is enough, but usually, we also need to understand how fast does the average converge to the expectation.

So far, our main tools have been the Markov and Chebyshev inequalities:

- Markov: If X is a non-negative random variable, then
$$Pr[X \geq a] \leq \frac{E[X]}{a}$$
- Chebyshev: If X_i are i.i.d. then
$$Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X_1]\right| \geq \varepsilon\right] \leq \frac{Var[X_1]}{n\varepsilon^2}$$

Our goal is to better bound the quantity $Pr[X \geq a]$, when X is the sum of i.i.d random variables X_i. We will use Markov's inequality, but this time, we will apply it not to the random variable X, but to e^{tX}, for an appropriate choice of t. Since X is the sum of i.i.d. random variables, i.e. $X = \sum_i X_i$, e^{tX} will be the product of e^{tX_i}. The product of small numbers decreases exponentially!

Let's see this formally:

Theorem 17.1. *Chernoff bounds Let $X_1, \ldots, X_n$ be independent indicator random variables such that $Pr[X_i = 1] = p_i$, and $Pr[X_i = 0] = 1 - p_i$. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = E[X]$. Then the following Chernoff bound holds: For any $\delta > 0$:*

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

Proof. Let's first apply Markov's inequality on X:

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

The crux of the proof is analyzing $E[e^{tX}]$. Let's first analyze $E[e^{tX_i}]$. e^{tX_i} is a random variable that takes value $e^{t\cdot 1}$ with probability p_i, and $e^{t\cdot 0} = 1$ with probability $1 - p_i$. Therefore,

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

where in the last inequality we used that for all y, $1+y \leq e^y$.

Now, for $E[e^{tX}]$ we have:

$$\begin{aligned} E[e^{tX}] &= E\left[e^{t\sum_{i=1}^n X_i}\right] = E\left[\prod_{i=1}^n e^{tX_i}\right] \\ &= \prod_{i=1}^n E\left[e^{tX_i}\right] \leq \prod_{i=1}^n e^{p_i(e^t-1)} \\ &= e^{\sum_i p_i(e^t-1)} = e^{(e^t-1)\sum_i p_i} = e^{(e^t-1)\mu} \end{aligned}$$

Plugging back in to equation 1 we have:

$$\begin{aligned} Pr[X \geq (1+\delta)\mu] &= Pr[e^{tX} \geq e^{t(1+\delta)\mu}] \\ &\leq \frac{E[e^{tX}]}{e^{t(1+\delta)\mu}} \\ &\leq \frac{e^{(e^t-1)\mu}}{e^{t(1+\delta)\mu}} = \left(\frac{e^{(e^t-1)}}{e^{t(1+\delta)}}\right)^{\mu} \end{aligned}$$

Since $\delta > 0$, we can set $t = \ln(1+\delta)$. Plugging in we get:

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

□

Using very similar proofs we can show the following variations:

- For any $\delta > 0$:
$$Pr[X \geq (1+\delta)\mu] \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$$
- For any $1 > \delta > 0$:
$$Pr[X \leq (1-\delta)\mu] \leq \left(\frac{e^{\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}$$
- For any $1 > \delta > 0$:
$$Pr[X \geq (1+\delta)\mu] \leq e^{-\frac{\mu\delta^2}{3}} \tag{2}$$
- For any $1 > \delta > 0$:
$$Pr[X \leq (1-\delta)\mu] \leq e^{-\frac{\mu\delta^2}{2}} \tag{3}$$
- For $R > 6\mu$:
$$Pr[X \geq R] \leq 2^{-R}$$

Estimating a parameter

Example taken from Mitzenmacher and Upfal's "Probability and Computing"

Suppose that we are interested in evaluating the probability that a particular gene mutation occurs in the population. Given a DNA sample, a lab test can determine if it carries the mutation. However, the test is expensive and we would like to obtain a relatively reliable estimate from a small number of samples.

Let p be the unknown value that we are trying to estimate. Assume that we have n samples and that $X = \tilde{p}n$ of these samples have the mutation. Given a sufficiently large number of samples, we expect the value p to be close to the sampled value $\tilde{p}$. We express this intuition using the concept of a confidence interval.

Definition 17.1. *A $1-\gamma$ confidence interval for a parameter p is an interval $[\tilde{p}-\delta, \tilde{p}+\delta]$ such that*

$$Pr(p \in [\tilde{p}-\delta, \tilde{p}+\delta]) \geq 1-\gamma.$$

Notice that, instead of predicting a single value for the parameter, we give an interval that is likely to contain the parameter. If p can take on any real value, it may not make sense to try to pin down its exact value from a finite sample, but it does make sense to estimate it within some small range.

Naturally we want both the interval size 2δ and the error probability γ to be as small as possible. We derive a trade-off between these two parameters and the number of samples n. In particular, given that among n samples (chosen uniformly at random from the entire population) we find the mutation in exactly $X = \tilde{p}n$ samples, we need to find values of δ and γ for which

$$Pr(p \in [\tilde{p}-\delta, \tilde{p}+\delta]) = Pr(np \in [n(\tilde{p}-\delta), n(\tilde{p}+\delta)]) \geq 1-\gamma$$

Now $X = \tilde{p}n$ has a binomial distribution with parameters n and p, so $E[X] = np$. If $p \notin [\tilde{p}-\delta, \tilde{p}+\delta]$ then we have one of the following two events:

1. if $p < \tilde{p}-\delta$, then $X = n\tilde{p} > np + n\delta = E[X]\left(1+\frac{\delta}{p}\right)$;
2. if $p > \tilde{p}+\delta$, then $X = n\tilde{p} < np - n\delta = E[X]\left(1-\frac{\delta}{p}\right)$;

We can apply Chernoff bounds 2 and 3 and compute:

$$\begin{aligned} Pr(p \notin [\tilde{p}-\delta, \tilde{p}+\delta]) &= Pr\left(X > np\left(1+\frac{\delta}{p}\right)\right) + Pr\left(X < np\left(1-\frac{\delta}{p}\right)\right) \\ &\leq e^{-\frac{np\left(\frac{\delta}{p}\right)^2}{3}} + e^{-\frac{np\left(\frac{\delta}{p}\right)^2}{2}} \\ &= e^{-\frac{n\delta^2}{3p}} + e^{-\frac{n\delta^2}{2p}} \end{aligned}$$

This bound is not really useful, since p is unknown. A simple solution it to use the fact that $p \leq 1$, and observe that the bound gets worse for larger p. Plugging in $p = 1$ we get:

$$Pr(p \notin [\tilde{p}-\delta, \tilde{p}+\delta]) \leq e^{-\frac{n\delta^2}{3}} + e^{-\frac{n\delta^2}{2}}$$

Setting $\gamma = e^{-\frac{n\delta^2}{3}} + e^{-\frac{n\delta^2}{2}}$ we get a tradeoff between the error probability γ, the number of samples n and δ (half the length of the confidence interval).

Example 1

Let's put the example above in numbers. Say we're interested in estimating a parameter p with 5% accuracy with probability at least 95%. Therefore, $\delta = 0.025$, and $\gamma = 0.05$. The number of samples necessary satisfies:

$$e^{-\frac{n\delta^2}{3}} + e^{-\frac{n\delta^2}{2}} \leq \gamma$$

$$e^{-\frac{n0.025^2}{3}} + e^{-\frac{n0.025^2}{2}} \leq 0.05$$

$$n \geq 15270$$

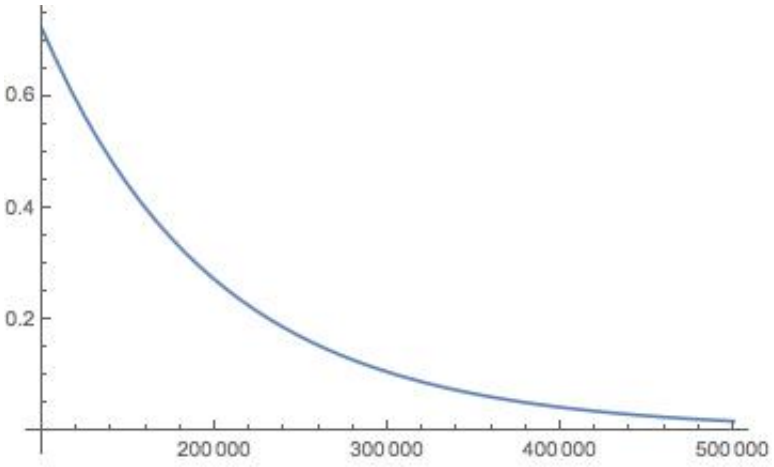
This means that if we want to run an opinion poll, where the answers are either yes or no, asking 16000 people is more than enough to get us within a 5% of the true answer, with probability at least 95%. Notice that this is independent of how big the population is! The only issue is making sure that the samples/random variables are independent.

Example 2

Let's fix a confidence interval size, say 1%, and see how the probability of being within the interval decreases, as n increases. The function we are interested is

$$f(n) = e^{-\frac{n\delta^2}{3}} + e^{-\frac{n\delta^2}{2}} = e^{-\frac{n0.005^2}{3}} + e^{-\frac{n0.005^2}{2}}$$

Here is a plot of this function, for n from $100,000$ to $500,000$:



For $100,000$ samples, our bound on the probability that we're in the interval is around 0.72. For $500,000$ samples, it drops to 0.017. For a million samples it drops to 0.00024.