Chernoff Bound and its Applications

Qiyu Liu

Hong Kong University of Science and Technology

qliuau@cse.ust.hk

July 3, 2018

Overview

1. Chernoff Bound
   - Chernoff Bound 1
   - Chernoff Bound 2

2. Applications of Chernoff Bound
   - Sampling/Polling
   - Chernoff Bound + Union Bound
   - Load Balancing
   - Balls & Bins

Theorem (Markov Inequality)

For non-negative random variable X and non-negative value t, it holds

$$\Pr[X \geq t] \leq \frac{E[X]}{t}.$$

Theorem (Chebyshev's Inequality)

For arbitrary random variable X and non-negative value t, it holds

$$\Pr[|X - E[X]| \geq t] \leq \frac{Var[X]}{t^2}.$$

Revisit of Probabilistic Inequalities – Cont.

The intrinsic of probabilistic inequalities is trying to show that it is unlikely a random variable X is far away from its expectation. You can make better and better statements to this effect the "more you know" about X. If you know nothing about X, you can't really say anything. If you know that X is non-negative, Markov's Inequality tells you it's unlikely to be far bigger than its expectation. If you manage to compute $Var[X]$, Chebyshev's Inequality tells you the chance that X is t standard deviations or more from its expectation is at most $1/t^2$.

Chernoff Bound 1

Theorem (Chernoff Bound 1 – one side)

Let $X \sim \mathit{Binomial}(n, 1/2)$*. Then for any* $0 \leq t \leq \sqrt{n}$*,*

$$\Pr\left[X - \frac{n}{2} \geq t\frac{\sqrt{n}}{2}\right] \leq e^{-t^2/2}$$

$$\Pr\left[X - \frac{n}{2} \leq -t\frac{\sqrt{n}}{2}\right] \leq e^{-t^2/2}.$$

Corollary (Chernoff Bound 1 – two side)

$$\Pr\left[|X - \frac{n}{2}| \geq t\frac{\sqrt{n}}{2}\right] \leq 2e^{-t^2/2}.$$

Chernoff Bound 1 – Cont.

- Chernoff Bounds can be used when you know that X is the sum of many independent random variables.
- Actually, for distribution Binomial$(n, 1/2)$, its expectation is $n/2$ and its standard deviation is $\sqrt{n}/2$, which means the sum of a set of i.i.d. random variables is **exponentially** unlikely to be t standard deviations away from its expectation.

Chernoff Bound 1 – Example

We use an example to show Chernoff Bound is much tighter than Chebyshev's Inequality. Let's say $X \sim \text{Binonial}(n, 1/2)$, that is, $E[X] = n/2$ and $Var[X] = n/4$. By using Chebyshev's Inequality, we have

$$\Pr[|X - n/2| \geq 5\sqrt{n}] \leq \frac{Var[X]}{25n} = \frac{1}{100}.$$

However, by using Chernoff Bound, it holds that,

$$\Pr[|X - n/2| \geq 5\sqrt{n}] \leq 2e^{-t^2/2} = 2e^{-50} \approx 3.86 \times 10^{-22}.$$

Chernoff Bound 1 – Example Cont.

Let's see the real value of the probability $\Pr[|X - n/2| \geq 5\sqrt{n}]$. Set n to 300, then,

$$\begin{aligned}\Pr[|X - n/2| \geq 5\sqrt{n}] &= \Pr[|X - 150| \geq 86] \\ &= \Pr[64 \leq X \leq 236] \\ &= \sum_{i=64}^{236} \binom{300}{i} \frac{1}{2^{300}} \\ &\approx 7.89 \times 10^{-31}.\end{aligned}$$

From this example, we can see Chernoff Bound is much tighter than Chebyshev's Inequality!

Chernoff Bound 1 – Proof

Note that the proof shown below is quite tricky but easy to understand.

To prove the Chernoff Bound, is is equivalent to prove $\Pr[Y \geq t\sqrt{n}]$ where $Y = Y_1 + \cdots + Y_n$ and Y_i is ± 1 with probability $1/2$ each. Define $Z_i = (1+\lambda)^{Y_i}$ where λ is a small value. Thus,

$$Z_i = \begin{cases} 1+\lambda & \text{with probability } 1/2 \\ \frac{1}{1+\lambda} & \text{with probability } 1/2 \end{cases}$$

Notice that,

$$\begin{aligned} E[Z_i] &= \frac{1}{2}(1+\lambda) + \frac{1}{2}\frac{1}{1+\lambda} \\ &= 1 + \frac{\lambda^2}{2(1+\lambda)} \\ &\leq 1 + \frac{\lambda^2}{2}. \end{aligned}$$

Chernoff Bound 1 – Proof Cont.

Define $Z = Z_1 Z_2 \cdots Z_n$. Notice that the following events are equivalent,

$$Y \geq t\sqrt{n} \Longleftrightarrow (1+\lambda)^Y \geq (1+\lambda)^{t\sqrt{n}} \Longleftrightarrow Z \geq (1+\lambda)^{t\sqrt{n}}$$

Thus, we consider using existing probability inequality. Since Y_is are independent, all of the Z_is are independent. Thus, $E[Z] = E[Z_1]E[Z_2]\cdots E[Z_n]$. Besides, all Z_is are non-negative and Z is non-negative, which means we can use Markov Inequality, that is,

$$\begin{aligned}
\Pr[Y \geq t\sqrt{n}] &= \Pr[Z \geq (1+\lambda)^{t\sqrt{n}}] \\
&\leq \frac{E[Z]}{(1+\lambda)^{t\sqrt{n}}} \\
&\leq \frac{(1+\frac{\lambda^2}{2})^n}{(1+\lambda)^{t\sqrt{n}}}
\end{aligned}$$

Chernoff Bound 1 – Proof Cont.

By taking $\lambda = t/\sqrt{n}$,

$$
\begin{aligned}
\Pr[Y \geq t\sqrt{n}] &\leq \frac{(1+\frac{\lambda^2}{2})^n}{(1+\lambda)^{t\sqrt{n}}} \\
&\leq \frac{(1+\frac{t^2}{2n})^n}{(1+\frac{t}{\sqrt{n}})^{t\sqrt{n}}} \\
&\leq \frac{\exp(\frac{t^2}{2})}{\exp(t^2)} = \exp(-\frac{t^2}{2}).
\end{aligned}
$$

Thus we complete the proof. Note that it seems like we use the approximation $1 + x \approx e^x$ to both nominator and denominator, but it is actually true.

Chernoff Bound 2

Chernoff Bound 1 tells us that if X is the sum of many independent Bernoulli(1/2)'s, it's extremely unlikely that X will deviate even a little bit from its mean. Let's rephrase the above a little. Taking $t = \epsilon\sqrt{n}$ in Chernoff Bound, we get

$$\left.\begin{array}{l} \Pr[X \geq (1+\epsilon)(n/2)] \\ \Pr[X \leq (1-\epsilon)(n/2)] \end{array}\right\} \leq \exp(-\epsilon^2 n/2).$$

This inequality depicts the relative relationship between X and its expectation $E[X]$, which is more useful in algorithm analysis cases.

Chernoff Bound 2 – Generalized Form

Theorem (Chernoff Bound 2 – one side)

Let $X_1, \cdots, X_n$ *be* n *independent random variables (need not to be identically distributed). Assume* $0 \leq X_i \leq 1$ *always for each* i*. Let* $X = X_1 + \cdots + X_n$*. Denote* $\mu = E[X] = E[X_1] + \cdots + E[X_n]$*. Then for any* $\epsilon \geq 0$,

$$\Pr[X \geq (1+\epsilon)\mu] \leq \exp\left(-\frac{\epsilon^2}{2+\epsilon}\mu\right)$$

$$\textit{and, } \Pr[X \leq (1-\epsilon)\mu] \leq \exp\left(-\frac{\epsilon^2}{2}\mu\right).$$

Note that the above Chernoff Bound is not **symmetric** and you cannot make it symmetric!

Chernoff Bound 2 – Generalized Form Cont.

In practical usage, ϵ is usually set to a small value, that is to say, $\epsilon \leq 1$. Thus,

$$\Pr[X \geq (1+\epsilon)\mu] \leq \exp\left(-\frac{\epsilon^2}{3}\mu\right)$$

$$\text{and,}\ \ \Pr[X \leq (1-\epsilon)\mu] \leq \exp\left(-\frac{\epsilon^2}{2}\mu\right).$$

The two side Chernoff Bound is also easy to get by using the union bound:

Corollary (Chernoff Bound 2 – two side)

$$\Pr[|X-\mu| \geq \epsilon\mu] \leq 2\exp\left(-\frac{\epsilon^2}{2+\epsilon}\mu\right).$$

Comparison Between Chernoff Bound 1 and 2

Chernoff Bound 2 is more generalized than 1, then, why we need Chernoff Bound 1?

The reason is that for the sum of Bernoulli variables, the Chernoff Bound 1 is tighter than Chernoff Bound 2. Let's see an example where $X \sim \text{Binomial}(n, 1/2)$. The Chernoff Bound 1 tells us:

$$\Pr[X \leq n/2 - t\sqrt{n}/2] \leq \exp(-t^2/2).$$

However, by setting $\epsilon = t/\sqrt{n}$ in Chernoff Bound 2,

$$\begin{aligned}\Pr[X \leq n/2 - t\sqrt{n}/2] &= \Pr[X \leq (1-\epsilon)\mu] \\ &\leq \exp(-t^2/4)\end{aligned}$$

which is the **square root** of the result in Chernoff Bound 1.

Discussion about Chernoff Bound 2

Question: What if we don't have $0 \leq X_i \leq 1$?

Sometimes they might satisfy, say, $0 \leq X_i \leq 10$. In this case, the trick is to simply define $Y_i = X_i/10$, and $Y = Y_1 + \cdots + Y_n$. Denote $\mu_X = E[X]$ and $\mu_Y = E[Y]$. Then, we apply Chernoff Bound 2 on the Y_is,

$$\begin{aligned}
\Pr[X \geq (1+\epsilon)\mu_X] &= \Pr[X/10 \geq (1+\epsilon)\mu_X/10] \\
&= \Pr[Y \geq (1+\epsilon)\mu_Y] \\
&\leq \exp\left(-\frac{\epsilon^2}{2+\epsilon}\mu_Y\right) = \exp\left(-\frac{\epsilon^2}{2+\epsilon}\frac{\mu_X}{10}\right).
\end{aligned}$$

So you lose a factor of 10 inside the
nal exponential probability bound, but you still get something pretty good.

Sampling/Polling

One very important usage of Chernoff Bound is to analyze the sampling. The following statement is common:

The sampling estimator is accurate to within $\pm 2\%$ *with probability 95%.*

The probability 95% is also known as *confidence*. In the learning theory community, a similar phrase is *Probably Approximately Correct*, i.e., probably (with chance at least 95% over the choice of people), the empirical average is approximately (within $\pm 2\%$, say) correct (vis-a-vis the true fraction of the population).

Sampling/Polling – Cont.

Let's consider a typical sampling problem. Let the true fraction of the population that approves of the president be p. This is the "correct answer" that we are trying to elicit. To estimate p, we ask n uniformly chosen people for their opinion and let each person be chosen independently. Let X_i be the indicator random variable that the *ith* person we ask approves of the president. Notice that, $X_i \sim \text{Bernoulli}(p)$ and $X_1, \cdots, X_n$ are independent. Let $X = X_1 + \cdots + X_n$ and $\bar{X} = X/n$. $\bar{X}$ is the estimator of p.

Question: How large does n have to be so that we get good "accuracy" with high "confidence"? Formally, to guarantee the following inequality to be satisfied,

$$\Pr[|\bar{X} - p| \le \theta] \ge 1 - \delta,$$

how large do we have to make n? (n should be a function of θ and δ)

Applying the two side Chernoff Bound 2, we get

$$\begin{aligned}\Pr[|\bar{X} - p| \geq \epsilon p] &= \Pr[|X - np| \geq \epsilon np] \\ &\leq 2\exp\left(-\frac{\epsilon^2}{2+\epsilon}np\right).\end{aligned}$$

Let $\epsilon = \theta/p$,

$$\Pr[|\bar{X} - p| \geq \theta] \leq 2\exp\left(-\frac{\theta^2/p^2}{2+\theta/p}np\right) = 2\exp\left(-\frac{n\theta^2}{2p+\theta}\right).$$

Let the RHS less than δ, i.e.,

$$2\exp\left(-\frac{n\theta^2}{2+\theta}\right) \leq 2\exp\left(-\frac{n\theta^2}{2p+\theta}\right) \leq \delta,$$

which gives the result $n \geq \frac{2+\theta}{\theta^2}\ln\frac{2}{\delta}$.

Sampling/Polling – Analysis Cont.

Theorem (Sampling Theorem)

Suppose we use independent, uniformly random samples to estimate p, the fraction of a population with some property. If the number of samples n we use satisfies

$$n \geq \frac{2+\theta}{\theta^2} \ln \frac{2}{\delta},$$

then we can assert that the sample mean $\bar{X}$ satisfies

$$\bar{X} \in [p-\epsilon, p+\epsilon] \text{ with probability at least } 1-\delta.$$

- $[p-\epsilon, p+\epsilon]$ is the confidence interval.
- Usually we write $n = O(\frac{1}{\epsilon^2} \ln \frac{1}{\delta})$.
- $1/\epsilon^2$ is a fairly high price compared with $\ln \frac{1}{\delta}$.
- One beauty of the Sampling Theorem is that the number of samples n you need does not depend on the size of the total population.

Chernoff Bound + Union Bound

Union Bound: $\Pr[X_i \cup \cdots \cup X_n] \leq \sum_{i=1}^{n} \Pr[X_i]$.

The idea behind the Chernoff + Union Bound method is the following: The Chernoff Bound is extraordinarily strong, usually showing that the probability a certain "bad event" happens is extremely tiny. Thus, even if very many different bad events exist, if you bound each one's probability by something extremely tiny, you can afford to just add up the probabilities, i.e.,

$$\Pr[\text{anything bad at all}] = \Pr[\text{Bad}_1 \cup \cdots \cup \text{Bad}_L] \leq \sum_{i=1}^{L} \Pr[\text{Bad}_i]$$

By applying Chernoff Bound and get $\Pr[\text{Bad}_i] \leq$ something small, then,

$$\Pr[\text{anything bad at all}] \leq L \cdot \text{something small} = \text{bound}.$$

Load Balancing

Suppose we have k servers and n jobs. Assume all n jobs arrive very quickly, we assign each to a random server (independently), and the jobs take a while to process. What we are interested in is the *load* of the servers. Note that, we assume n is much bigger than k.

Question: The average "load", jobs per server, will of course be n/k. But how close to perfectly balanced will things be? In particular, is it true that the maximum load is not much bigger than n/k, with high probability?

Answer: Yes!

Load Balancing – Cont.

Let X_i be the number of jobs assigned to server i for $1 \leq i \leq k$. Notice that $X_i \sim \text{Binomial}(n, 1/k)$. $X_1, \cdots, X_k$ are **not** independent! The reason is that there is a constraint $\sum_{i=1}^{k} X_i = n$.

It is easy to know that each server is expected to have n/k jobs. But we are more interested in the *maximum* load among the k servers. Let $M = \max(X_1, \cdots, X_k)$. Our goal is to bound the probability like this:

$$\Pr[M \geq n/k + c] \leq \text{small}$$

where c cannot be very large.

Load Balancing – Cont.

Theorem

$$\Pr[M \geq n/k + 3\sqrt{\ln k}\sqrt{n/k}] \leq 1/k^2.$$

As the diagram of "Chernoff Bound + Union Bound", we first define the "bad event". Let's say $B_1, \cdots, B_k$ are k bad events as follows:

$$B_i = \text{``}X_i \geq n/k + c\text{''}.$$

Define $B = B_1 \cup \cdots \cup B_k$. B is the overall bad event. This is the event that at least one of the k servers gets load at least $n/k + c$. This is exactly the event we care about, i.e.,

$$B = \text{``}M \geq n/k + c\text{''}.$$

Our objective is to bound the probability that bad event B occurs.

Load Balancing – Cont.

By applying the union bound,

$$\begin{aligned}\Pr[M \geq n/k + c] &\leq \sum_{i=1}^{k} \Pr[B_i] \\ &\leq \sum_{i=1}^{k} \Pr[X_i \geq n/k + 3\sqrt{\ln k}\sqrt{n/k}].\end{aligned}$$

For each i, applying the Chernoff Bound 2, let $\epsilon = 3\sqrt{\ln k}/\sqrt{n/k}$,

$$\begin{aligned}\Pr[X_i \geq n/k + 3\sqrt{\ln k}\sqrt{n/k}] &= \Pr[X_i \geq (1+\epsilon)n/k] \\ &\leq \exp\left(-\frac{1}{3}\frac{n}{k}\frac{9\ln k}{n/k}\right) = \frac{1}{k^3}.\end{aligned}$$

Combing the union bound, $\Pr[M \geq n/k + c] \leq k\frac{1}{k^3} = \frac{1}{k^2}$. Q.E.D.

Discussion about Load Balancing

Question: Is this bound good? It seems $c = 3\sqrt{\ln k}\sqrt{n/k}$ is not very small!

We first see a specific case. In the case of $n = 10^6$, $k = 10^3$, the bound derived above can be translated into

$$\Pr[M \geq 1000 + 249.34] \leq 10^{-6}.$$

Consider the relative error parameter $\epsilon = 3\sqrt{k \ln k}/\sqrt{n}$. For a fixed k, ϵ converges to 0 with a factor of square root of n.

Balls & Bins

A more generalized model is "Balls & Bins" where there are n balls and m bins in total. Each ball is independently "thrown" into a uniformly random bin. The follows shows some example situations it models:

- **Load balancing**: Balls = jobs, Bins = servers. Now Balls and Bins models the random load balancing scenario we just discussed.
- **Data storage**: Balls =
les, Bins = disks.
- **Hashing**: Balls = data keys, Bins = hash table slots.
- **Routing**: Balls = connectivity requirements, Bins = paths in a network.

Balls & Bins

The common questions of "Balls & Bins" are summarized as follows:

- **Question:** What is the probability of a *collision* – i.e., getting more than 1 ball in some bin? This problem also known as the famous Birthday Problem.
- **Question:** How many balls are thrown before each bin has at least 1 ball? This is exactly what we're interested in in the Coupon Collector Problem.
- **Question:** How many balls end up in the bin with maximum load? This is what we just finished analyzing in the discussion of Randomized Load Balancing.

All these questions can be nicely answered by applying Chernoff Bound!

The End