5 Chernoff Bound and Hoeffding Inequality

If we apply Markov inequality to

$$\Pr[f(X) \geq f(t)]$$

with $f(x) = \mathrm{e}^{\alpha x}$ where $\alpha > 0$, then the bound amounts to bound $\mathbb{E}[\mathrm{e}^{\alpha X}]$ which is the *moment generating function* of X.

In case $\mathbb{E}[\mathrm{e}^{\alpha X}]$ can be well bounded, we obtain sharp concentration bounds.

5.1 Chernoff Bound

We first consider a simple case: the random variable X can be written as the sum of independent Bernoulli variables. In this case the moment generating function is easy to estimate.

Theorem 5.1. Chernoff Bound
Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \sim \mathrm{Ber}(p_i)$ for each $i = 1, 2, \ldots, n$. Let $X = \sum_{i=1}^n X_i$ and denote $\mu \triangleq \mathbb{E}[X] = \sum_{i=1}^n p_i$, we have

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

If $0 < \delta < 1$, then we have

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

Proof. For every $\lambda > 0$, we have

$$\Pr[X \geq (1+\delta)\mu] = \Pr\left[\mathrm{e}^{\lambda X} \geq e^{\alpha(1+\delta)\mu}\right] \leq \frac{\mathbb{E}[\mathrm{e}^{\lambda X}]}{\mathrm{e}^{\lambda(1+\delta)\mu}}.$$

Now we need to estimate the moment generating function $\mathbb{E}[\mathrm{e}^{\lambda X}]$. Since $X = \sum_{i=1}^n X_i$ is the sum of independent Bernoulli variables, we have

$$\mathbb{E}[\mathrm{e}^{\lambda X}] = \mathbb{E}\left[\mathrm{e}^{\lambda \sum_{i=1}^n X_i}\right] = \mathbb{E}\left[\prod_{i=1}^n \mathrm{e}^{\lambda X_i}\right] = \prod_{i=1}^n \mathbb{E}[\mathrm{e}^{\lambda X_i}].$$

Since $X_i \sim \mathrm{Ber}(p_i)$, we can compute $\mathbb{E}[\mathrm{e}^{\lambda X_i}]$ directly:

$$\mathbb{E}[\mathrm{e}^{\lambda X_i}] = p_i \mathrm{e}^{\lambda} + (1-p_i) = 1 + (\mathrm{e}^{\lambda} - 1)p_i \leq \exp\left((\mathrm{e}^{\lambda} - 1)p_i\right).$$

Therefore,

$$\mathbb{E}[\mathrm{e}^{\lambda X}] \leq \prod_{i=1}^n \exp((\mathrm{e}^{\lambda} - 1)p_i) = \exp\left((\mathrm{e}^{\lambda} - 1)\sum_{i=1}^n p_i\right) = \exp\left((\mathrm{e}^{\lambda} - 1)\mu\right).$$

Thus, we have

$$\Pr[X \le (1+\delta)\mu] \le \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda(1+\delta)\mu}} \le \left(\frac{\exp(e^{\lambda}-1)}{\exp(\lambda(1+\delta))}\right)^{\mu}$$

for any $\lambda > 0$. We can choose λ so as to minimize $\frac{\exp(e^{\lambda}-1)}{\exp(\lambda(1+\delta))} = \exp(e^{\lambda}-1-\lambda(1+\delta))$. To this end, we let

$$\frac{d}{d\lambda}\left(e^{\lambda}-1-\lambda(1+\delta)\right) = e^{\lambda}-(1+\delta) = 0\,,$$

which gives $\lambda = \ln(1+\delta)$. Hence, we conclude that

$$\Pr[X \ge (1+\delta)\mu] \le \left(\frac{\exp(e^{\lambda}-1)}{\exp(\lambda(1+\delta))}\right)^{\mu} \le \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}. \qquad \square$$

The following form of Chernoff bound is more convenient to use (but weaker):

Theorem 5.2. Eeay form of Chernoff bound
For any $0 < \delta < 1$,

$$\Pr[X \ge (1+\delta)\mu] \le \exp\left(-\frac{\delta^2}{3}\mu\right)$$
$$\Pr[X \le (1-\delta)\mu] \le \exp\left(-\frac{\delta^2}{2}\mu\right)$$

For any $t \ge 2e\mu$,

$$\Pr[X \ge t] \le 2^{-t}\,.$$

Proof. We only prove the first case. It suffices to verify that for $0 < \delta < 1$, we have

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

Taking logarithm of both sides, this is equivalent to

$$\delta - (1+\delta)\ln(1+\delta) \le -\frac{\delta^2}{3}$$

Let $f(\delta) = \delta - (1+\delta)\ln(1+\delta) + \frac{\delta^2}{3}$ and note that

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

Then for $0 < \delta < 1/2$, $f''(\delta) < 0$, and for $1/2 < \delta < 1$, $f''(\delta) > 0$. Therefore, $f'(\delta)$ first decreases and then increases in $[0,1]$. Also note that $f'(0) = 0$, $f'(1) < 0$ and $f'(\delta) \le 0$ when $0 \le \delta \le 1$. Therefore $f(\delta) \le f(0) = 0$. $\square$

Example 5.1.
If we toss a fair coin n times, the average number of heads is $n/2$. We want to determine the value δ such that with a constant probability (say 99%), the total number of heads is in the interval of $[(1-\delta)\frac{n}{2}, (1+\delta)\frac{n}{2}]$.
Let X denote the total number of heads, and $X_i \sim \text{Ber}\left(\frac{1}{2}\right)$ be the indicator of whether the i-th toss gives a head. Then by Chernoff bound, we have

$$\Pr\left[\left|X - \frac{n}{2}\right| \geq \delta \cdot \frac{n}{2}\right] \leq 2\exp\left(-\frac{\delta^2}{3} \cdot \frac{n}{2}\right) \leq 0.01$$

So it suffices to choose $\delta = \Omega\left(\frac{1}{\sqrt{n}}\right)$. If we would like to improve the accuracy (with high probability, say $1 - 1/n$), we need to add a $\sqrt{\ln n}$ factor to δ:

$$\Pr\left[X - \frac{n}{2} \geq \frac{1}{2}\sqrt{6n\ln n}\right] \leq \exp\left(-\frac{1}{3} \cdot \frac{n}{2} \cdot \frac{6\ln n}{n}\right) = \frac{1}{n}.$$

Remark 5.1.
We can compare Chernoff bound with *Markov's inequality* and *Cheybyshev's inequality*:

Chernoff bound: $\Pr\left[X \geq \frac{3}{4}n\right] \leq \exp\left(-\frac{1}{3} \cdot \frac{n}{2} \cdot \frac{1}{4}\right) = \mathrm{e}^{-n/24}$

Markov's inequality: $\Pr\left[X \geq \frac{3}{4}n\right] \leq \frac{n/2}{3n/4} = \frac{2}{3}$

Cheybyshev's inequality: $\Pr\left[X \geq \frac{3}{4}n\right] \leq \Pr\left[\left|X - \frac{n}{2}\right| \geq \frac{n}{2}\right] \leq \frac{n/4}{(n/4)^2} = \frac{4}{n}$

Example 5.2.
Consider the m-balls-into-n-bins model again. Let X_i be the number of balls in the i-th bin. It is easy to see that X_i is the sum of m Bernoulli variables with $p = 1/n$.

- If $m = n$, then $\mu = \mathbb{E}[X_i] = 1$. So for any $t > 1$,

  $$\Pr[X_i \geq t] \leq \frac{\mathrm{e}^t}{t^t}.$$

  Choosing $t = \frac{\mathrm{e}\ln n}{\ln\ln n}$, we obtain $\Pr[X_i \geq t] \leq n^{-2}$.
- If $m \geq n\ln n$, then $\mu = \frac{m}{n} \geq \ln n$. So for $t = 2\mathrm{e}\mu$, we have

  $$\Pr[X_i \geq t] \leq 2^{-t} \leq \frac{1}{n^2}.$$

5.2 Hoeffding's Inequality

One of annoying restrictions of Chernoff bound is that each X_i needs to be a Bernoulli random variable. Hoeffding's inequality generalizes Chernoff bound by allowing X_i to follow any distribution, provided its value is almost surely bounded.

Theorem 5.3. Hoeffding inequality
Let $X_1, \ldots, X_n$ be independent random variables where each $X_i \in [a_i, b_i]$ for certain $a_i \le b_i$ with probability 1. Assume $\mathbb{E}[X_i] = \mu_i$ for every $1 \le i \le n$. Let $X = \sum_{i=1}^n X_i$ and $\mu \triangleq \mathbb{E}[X] = \sum_{i=1}^n \mu_i$, then

$$\Pr[|X - \mu| \ge t] \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

for all $t \ge 0$.

We learnt from the proof of the Chernoff bound that the key to establish concentration inequalities of this form is to obtain a nice upper bound on the moment generating function.

We first consider a simple case. Suppose that $X \in [0,1]$ is a random variable with expectation p. Note that for any $x \in [0,1]$, by Jensen's inequality we have

$$\mathrm{e}^{\lambda x} \le x\mathrm{e}^{\lambda} + (1-x)\mathrm{e}^{0}$$

for every $\lambda \ge 0$. It follows that

$$\mathbb{E}\left[\mathrm{e}^{\lambda X}\right] \le \mathbb{E}[X]\mathrm{e}^{\lambda} + (1 - \mathbb{E}[X]) = p\mathrm{e}^{\lambda} + (1-p)\,.$$

Since we would like to bound $\Pr[|X - \mathbb{E}[X]| \ge t]$, it will be easier if we consider $Y = X - \mathbb{E}[X]$. In this case we have $Y \in [-p, 1-p]$ and $\mathbb{E}[Y] = 0$. We would like to give an upper bound of $\mathbb{E}\left[\mathrm{e}^{\lambda Y}\right]$ in general (not depending on p):

$$\mathbb{E}\left[\mathrm{e}^{\lambda Y}\right] \le p\mathrm{e}^{\lambda(1-p)} + (1-p)\mathrm{e}^{-\lambda p} \triangleq \mathrm{e}^{\varphi(\lambda)}\,,$$

where

$$\varphi(\lambda) = \ln\left(p\mathrm{e}^{\lambda(1-p)} + (1-p)\mathrm{e}^{-\lambda p}\right).$$

We now give an upper bound for $\varphi(\lambda)$. Using Taylor's theorem with Lagrange remainder, we know that

$$\varphi(\lambda) = \varphi(0) + \varphi'(0)\lambda + \frac{\varphi''(\xi)}{2}\lambda^2\,,$$

where $\xi \in [0, \lambda]$. Since

$$\varphi'(\lambda) = \frac{(1-p)p(\mathrm{e}^{\lambda} - 1)}{p\mathrm{e}^{\lambda} + (1-p)}$$
$$\varphi''(\lambda) = \frac{(1-p)p\mathrm{e}^{\lambda}}{\left(p\mathrm{e}^{\lambda} + (1-p)\right)^2}\,,$$

we have

$$\varphi(0) = \varphi'(0) = 0 \qquad 0 \le \varphi''(\xi) \le \frac{1}{4}\,,$$

which yields

$$\varphi(\lambda) \le \frac{\lambda^2}{8}$$

for all $\lambda \ge 0$. For general range $[a,b]$, let $Y = \frac{X}{a+b}$ and $\lambda = \gamma(a+b)$, then it gives the following key lemma.

Lemma 5.4. Hoeffding's lemma
Let X be a random variable with $\mathbb{E}[X] = 0$ and $X \in [a, b]$. Then it holds that

$$\mathbb{E}\left[e^{\gamma X}\right] \leq \exp\left(\frac{\gamma^2 (b-a)^2}{8}\right) \text{ for all } \gamma \in \mathbb{R}.$$

Armed with Hoeffding's lemma, it is routine to prove Hoeffding's inequality.

Proof of Hoeffding's inequality. First note that we can assume $\mathbb{E}[X_i] = 0$ and therefore $\mu = 0$ (if not so, replace X_i by $X_i - \mathbb{E}[X_i]$). By symmetry, we only need to prove that $\Pr[X \geq t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$. Since

$$\Pr[X \geq t] \overset{\lambda \geq 0}{=} \Pr\left[e^{\lambda X} \geq e^{\lambda t}\right] \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda t}}$$

and

$$\mathbb{E}\left[e^{\lambda X}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{\lambda X_i}\right] \leq \prod_{i=1}^n \exp\left(\frac{\lambda^2 (b_i - a_i)^2}{8}\right),$$

(applying Hoeffding's lemma for each $\mathbb{E}[e^{\lambda X_i}]$), we obtain

$$\begin{aligned}\Pr[X \geq t] &\leq \frac{\prod_{i=1}^n \exp\left(\frac{\lambda^2 (b_i - a_i)^2}{8}\right)}{e^{\lambda t}} \\ &= \exp\left(\frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2 - \lambda t\right)\end{aligned}$$

for all $\lambda \geq 0$. By choosing

$$\lambda = \frac{4t}{\sum_{i=1}^n (b_i - a_i)^2},$$

it follows that

$$\Pr[X \geq t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \qquad \square$$

Example 5.3.
Let $\mathbf{A}$ be an $n \times m$ matrix with entries in $\{0, 1\}$. Columns represent *individuals* and rows represent *features*, and $A_{i,j}$ indicates whether individual j has feature i. The **set balancing** problem asks to partition the set of individuals into two classes X and Y such that each feature is as balanced as possible between the classes (i.e., for every feature i, the number of individuals with i in $X \approx$ the number of individuals with i in Y).
Let $b \in \{-1, 1\}^m$ denote the partition vector. Our goal is to minimize the discrepancy

$$\|\mathbf{A}b\|_\infty = \max_i |(\mathbf{A}b)_i|$$

where $(\mathbf{A}b)_i = \sum_{j \in X} A_{i,j} - \sum_{j \in Y} A_{i,j}$. Choose $b \in \{-1, 1\}^m$ uniformly at random. Determine t such that $\Pr[\|\mathbf{A}b\|_\infty \geq t] = o(1)$. Note that

$$\Pr[\|\mathbf{A}b\|_\infty \geq t] = \Pr[\max_i |A_i b| \geq t] \leq n \Pr[|A_i b| \geq t].$$

For any $t \leqslant \|A_i\|_1 \leqslant n$, we have

$$\Pr[|A_i b| \geq t] = \Pr\left[\left|\sum A_{i,j} b_j\right| \geq t\right] \leqslant 2e^{-t^2/2m}.$$

Let $t = \sqrt{4m \ln n}$. Thus we have $\Pr[|A_i b| \geq t] \leq 2n^{-2}$.

It is instructive to compare Hoeffding and Chernoff when X_i's are independent Bernoulli variables. Formally, let $X_1, \ldots, X_n$ be i.i.d. random variables where $X_i \sim \text{Ber}(p)$ for all $i = 1, \ldots, n$. Set $X = \sum_{i=1}^{n} X_i$ and denote $\mathbb{E}[X] = np$ by μ. For $t = \delta\mu$, by Hoeffding's inequality, we have

$$\Pr[|X - \mu| \geq t] \leq 2\exp\left(-2\delta^2 p^2 n\right).$$

By Chernoff Bound, we have

$$\Pr[|X - \mu| \geq t] \leq 2\exp\left(-\frac{1}{3}\delta^2 pn\right).$$

Comparing the exponent, it is easy to see that for $p > 1/6$, Hoeffding's inequality is tighter up to a certain constant factor. However, for smaller p, Chernoff bound is significantly better than Hoeffding's inequality, as its dependency to p is linear.

The following simple example demonstrates the difference. Suppose we have a box of N balls. Among them pN are red and $(1-p)N$ are blue. We draw a random ball from this box, record its color and put it back. The problem is in how many rounds we are sure about the value $\widehat{p}$ (which is the percentage of red balls we record) we guess is within the range $(1 \pm 0.01)p$. The rounds required is $\Omega(1/p)$ if we apply Chernoff bound, and $\Omega(1/p^2)$ if we apply Hoeffding's inequality.

5.3 Poisson Tail Bound

Theorem 5.5.
Let X be a Poisson random variable with rate λ (i.e., $\Pr[X = k] = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}$), then

$$\forall t > \lambda \qquad \Pr[X \geq t] \leq \mathrm{e}^{-\lambda}\frac{(\mathrm{e}\lambda)^t}{t^t}$$

$$\forall t < \lambda \qquad \Pr[X \leq t] \leq \mathrm{e}^{-\lambda}\frac{(\mathrm{e}\lambda)^t}{t^t}$$

Proof. Suppose $t > \lambda$ w.l.o.g. Then $\forall \gamma > 0$, we have

$$\Pr[X \geq t] = \Pr\left[\mathrm{e}^{\gamma X} \geq \mathrm{e}^{\gamma t}\right] \leq \frac{\mathbb{E}[\mathrm{e}^{\gamma X}]}{\mathrm{e}^{\gamma t}} = \mathrm{e}^{\lambda(\mathrm{e}^{\gamma}-1)-\gamma t},$$

where

$$M_X(\gamma) \triangleq \mathbb{E}\left[\mathrm{e}^{\gamma X}\right] = \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}\mathrm{e}^{\gamma k} = \mathrm{e}^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda\mathrm{e}^{\gamma})^k}{k!} = \mathrm{e}^{\lambda(\mathrm{e}^{\gamma}-1)}.$$

Let $\gamma = \ln\frac{t}{\lambda} > 0$, we have

$$\Pr[X \geq t] \leq \mathrm{e}^{\lambda(\mathrm{e}^{\gamma}-1)-\gamma t} = \frac{\mathrm{e}^{-\lambda}(\mathrm{e}\lambda)^t}{t^t}. \qquad \square$$

We now can analyze the limit of coupon collector problem.

Theorem 5.6.
$\lim_{n\to\infty} \Pr[X > n\ln n + cn] = 1 - \mathrm{e}^{-\mathrm{e}^{-c}}$.

The Poisson approximation explains why the limit is $1 - \mathrm{e}^{-\mathrm{e}^{-c}}$. However, applying the approximation directly cannot establish the equality. Recall that, in the proof of

Poisson approximation, we only consider the event that $\sum Y_i = m$, which is too loose in this case.

Fix $c > 0$ to be a constant. Let $m = n \ln n + cn$, $\lambda = m/n$ and $Y_1, \ldots, Y_n \sim \text{Pois}(\lambda)$ be n independent Poisson variables. Denote $\mathcal{E}$ the event that at least Y_i is zero. Recall that

$$\lim_{n\to\infty} \Pr[\mathcal{E}] = 1 - \mathrm{e}^{-\mathrm{e}^{-c}} \quad \text{and} \quad \Pr[X > n\ln n + cn] = \Pr\Big[\mathcal{E} \mid \sum_{i=1}^{n} Y_i = m\Big]$$

Define $Y = \sum_{i=1}^{n} Y_i$. We analyze the conditional probability more precisely:

$$\begin{aligned}\Pr[\mathcal{E}] &= \sum_{k=0}^{\infty} \Pr[\mathcal{E} \mid Y = k] \cdot \Pr[Y = k] \\ &= \sum_{k=m-t}^{m+t} \Pr[\mathcal{E} \mid Y = k] \cdot \Pr[Y = k] + \Pr[\mathcal{E} \mid |Y - m| > t] \cdot \Pr[|Y - m| > t],\end{aligned}$$

where $t = \sqrt{2m \ln m}$. Our goal is to show that

$$\lim_{n\to\infty} \Pr[\mathcal{E}] - \Pr[\mathcal{E} \mid Y = m] = 0.$$

To this end, we claim that

$$\Pr[|Y - m| > t] = o(1) \quad \text{and} \quad \Pr[\mathcal{E} \mid Y = m] - \Pr[\mathcal{E} \mid Y = k] = o(1)$$

for all $m - t \le k \le m + t$.

The first equation is an immediate corollary of Poisson tail bound. Note that $Y \sim \text{Pois}(m)$. Applying $\ln(1+x) \ge x - x^2/2$, we obtain

$$\begin{aligned}\Pr[|Y - m| > t] &\le 2\mathrm{e}^t \left(\frac{m}{m+t}\right)^{m+t} = 2\exp(t - (m+t)\ln(1 + t/m)) \\ &\le 2\exp\left(t - (m+t)\left(\frac{t}{m} - \frac{t^2}{2m^2}\right)\right) \\ &= 2\exp\left(-\frac{t^2}{2m}\left(1 - \frac{t}{m}\right)\right) = m^{-(1-\sqrt{2\ln m/m})} = o(1).\end{aligned}$$

To prove the second claim, note that

$$\Pr[\mathcal{E} \mid Y = k] = \Pr[\text{there exists empty bins after throwing } k \text{ balls}].$$

So $\Pr[\mathcal{E} \mid Y = k]$ is decreasing, and we only need to show

$$\Pr[\mathcal{E} \mid Y = m - t] - \Pr[\mathcal{E} \mid Y = m + t] = o(1).$$

Suppose we throw balls one by one. This is precisely the probability that there exists empty bins after throwing $m - t$ balls but no empty bins after throwing an additional $2t$ balls. Since the probability of covering the empty bin is at most $1/n$ for each ball, this event happens with probability at most $2t/n$ by union bound, which is $o(1)$.

Overall, we conclude that

$$\lim_{n\to\infty} \Pr[X > n\ln n + cn] = \lim_{n\to\infty} \Pr[\mathcal{E} \mid Y = m] = \lim_{n\to\infty} \Pr[\mathcal{E}] = 1 - \mathrm{e}^{-\mathrm{e}^{-c}}.$$

Remark 5.2.
Actually, let Z_n be the number of empty bins after throwing $m = n\ln n + cn$ balls into n bins. We can show that

$$Z_n \xrightarrow{D} \text{Pois}(\mathrm{e}^{-c})$$