[AI2613 Lecture 6]: Concentration Inequalities, Martingale

April 6, 2023

1 Chernoff Bounds

Recall the Markov inequality and Chebyshev's inequality we introduced before. They are used to prove that a random variable is concentrated around its expectation.

If we apply Markov inequality to

$$\mathbf{Pr}\left[f(X) \geq f(t)\right]$$

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

When the random variable X can be written as the sum of independent Bernoulli variables, its moment generating function is easy to estimate and we obtain sharp concentration bounds.

Theorem 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 $\mu \triangleq \mathbf{E}[X] = \sum_{i=1}^{n} p_i$, we have*

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

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

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

Proof. We only prove the upper tail bound and the proof of lower tail bound is similar. For every $\alpha > 0$, we have

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

Therefore, we need to estimate the moment generating function $\mathbf{E}\left[e^{\alpha X}\right]$. Since $X = \sum_{i=1}^{n} X_i$ is the sum of independent Bernoulli variables, we have

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

Since $X_i \sim \mathrm{Ber}(p_i)$, we can compute $\mathbf{E}\left[e^{\alpha X_i}\right]$ directly:

$$\mathbf{E}\left[e^{\alpha X_i}\right] = p_i e^{\alpha} + (1-p_i) = 1 + (e^{\alpha}-1)p_i \leq e^{((e^{\alpha}-1)p_i)}.$$

Therefore,

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

Therefore,

$$\mathbf{Pr}\left[X \leq (1+\delta)\mu\right] \leq \frac{\mathbf{E}\left[e^{\alpha x}\right]}{e^{\alpha(1+\delta)\mu}} \leq \left(\frac{e^{(e^{\alpha}-1)}}{e^{(\alpha(1+\delta))}}\right)^{\mu}$$

Note that above holds for any $\alpha > 0$. Therefore, we can choose α so as to minimize $\frac{e^{(e^{\alpha}-1)}}{e^{(\alpha(1+\delta))}}$. To this end, we let $\left(\frac{e^{(e^{\alpha}-1)}}{e^{(\alpha(1+\delta))}}\right)' = 0$. This gives $\alpha = \log(1+\delta)$. Therefore

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

□

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

Corollary 2 *For any $0 < \delta < 1$,*

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

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

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

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

Taking logarithm of both sides, this is equivalent to

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

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 $f'(\delta) \leq 0$ when $0 \leq \delta \leq 1$. Therefore $f(\delta) \leq f(0) = 0$. □

Example 1 (Tossing p-coins) *. Consider a p-coin that we get a head with probability p when tossing it. If we toss a p-coin n times, the average number of heads is pn.We want to determine the value δ such that with high probability (say 99%), the total number of heads is in the interval of $[(1-\delta)pn, (1+\delta)pn]$. We use Chernoff bound to determine δ.*

Let X denote the total number of heads, and $X_i \sim \mathrm{Ber}(p)$ be the indicator of whether the i-th toss gives a head. Then by Chernoff bound, we have

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

So if p is a constant, it suffices to choose

$$\delta = \Omega\left(\frac{1}{\sqrt{n}}\right).$$

2 Hoeffding's Inequality

One of annoying restrictions of Chernoff bound is that each X_i needs to be a Bernoulli random variable. We first relax this requirement by introducing Hoeffding's inequality which allows X_i to follow any distribution, provided its value is almost surely bounded.

Theorem 3 (Hoeffding's Inequality) *Let $X_1, \ldots, X_n$ be independent random variables where each $X_i \in [a_i, b_i]$ for certain $a_i \leq b_i$ with probability 1. Let $X = \sum_{i=1}^n X_i$ and $\mu \triangleq \mathbf{E}[X] = \sum_{i=1}^n \mathbf{E}[X_i]$, then*

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

for all $t \geq 0$.

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 $X_i \sim \texttt{Ber}(p)$ for all $i = 1, \ldots, n$. Set $X = \sum_{i=1}^n X_i$ and denote $\mathbf{E}[X] = np$ by μ. By Hoeffding's inequality, we have

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

By Chernoff Bound, we have

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

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.

We consider the balls-in-a-bag problem. There are g green balls and r red balls in a bag. These balls are the all same except for the color. We want to estimate the ratio $\frac{r}{r+g}$ by drawing balls. There are two scenarios.

- Draw balls with replacement. Let $X_i = \mathbf{1}[\text{the } i\text{-th ball is red}]$. Let $X = \sum_{i=1}^n X_i$. Then clearly each $X_i \sim \text{Ber}\left(\frac{r}{r+g}\right)$ and $\mathbf{E}[X] = n \cdot \frac{r}{r+b}$.

  Since all X_i's are independent, we can directly apply Hoeffding's inequality and obtain

  $$\mathbf{Pr}[|X - \mathbf{E}[X]| \geq t] \leq 2\exp\left(-\frac{2t^2}{n}\right).$$

- Draw balls without replacement. Again we let $Y_i = \mathbf{1}[\text{the } i\text{-th ball is red}]$, then unlike the case of drawing with replacement, variables in $\{Y_i\}$ are dependent. Let $Y = \sum_{i=1}^n Y_i$. We first calculate $\mathbf{E}[Y]$.

  For every $i \geq 1$, $\mathbf{E}[Y_i]$ is the probability that the i-th draw is a red ball. Note that drawing without replacement is equivalent to first drawing a

uniform permutation of $r+g$ balls and drawing each ball one by one in that order. Therefore, the probabilty of $Y_i = 1$ is $\frac{r\cdot(r+g-1)!}{(r+g)!} = \frac{r}{r+g}$. So we have $\mathbf{E}[Y] = n \cdot \frac{r}{r+g}$.

However, since $\{Y_i\}$ are dependent, we cannot apply Hoeffding's inequality directly. This motivate us to generalize it by removing the requirement of independence.

3 Martingale

We develop the theory of martingale, which is a core concept in probability theory. We use martingale to get rid of the independence requirement in the concentration inequalities mentioned above.

Consider a fair gambling game in which the expected gain in each round is zero. As a result, regardless of how much one bets in each round, the money in expectation remains the same. The balances after each round form a *martingale*.

Definition 4 (Martingale) *Let $\{X_n\}_{n\geq 0}$ and $\{Z_n\}_{n\geq 0}$ be two sequences of random variables. Let $Z_n = \sum_{t=0}^{n} X_t$.*[1] *We say $\{Z_n\}_{n\geq 0}$ is a martingale w.r.t. $\{X_n\}_{n\geq 0}$ if*

$$\mathbf{E}[Z_{n+1} \mid X_0, X_1 \ldots, X_n] = Z_n.$$

[1] Consider the problem of fair gambling where X_n is the gain of n-th round and $Z_n = \sum_{t=1}^{n} X_n$. $\{Z_n\}_{n\geq 0}$ is not necessarily a Markov chain. The value X_n may depend on information before round $n-1$.

Sometimes we say a single sequence $\{X_n\}_{n\geq 0}$ is a martingale if it is a martingale w.r.t. itself. Formally, if for every $n \geq 0$, it holds that

$$\mathbf{E}[X_{n+1} \mid X_0, \ldots, X_n] = X_n.$$

For convenience, from now on we use $\overline{X}_{i,j} = (X_i, X_{i+1} \ldots, X_j)$ to simplify the notations. The conditional expectation $\mathbf{E}\left[Z_{n+1} \,\middle|\, \overline{X}_{0,n}\right]$ is equivalent to $\mathbf{E}\left[Z_{n+1} \,\middle|\, \sigma(\overline{X}_{0,n})\right]$ where $\sigma(\overline{X}_{0,n})$ is the σ-algebra generated by $X_0, \ldots, X_n$. The motivates us to define martingale in a more general way.

Definition 5 (Martingale (defined by filtration)) *Let $\{\mathcal{F}_n\}_{n\geq 0}$ be a sequence of σ-algebras. We call such σ-algebra sequence a filtration if it satisfies*

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n \subseteq \mathcal{F}_{n+1} \subseteq \cdots.$$

Given a filtration $\{\mathcal{F}_n\}_{n\geq 0}$, let $\{Z_n\}_{n\geq 0}$ be a stochastic process that Z_n is $\mathcal{F}_n$-measurable for every $n \geq 0$. Then we say $\{Z_n\}_{n\geq 0}$ is a martingale w.r.t. $\{\mathcal{F}_n\}_{n\geq 0}$ if for every $n \geq 0$

$$\mathbf{E}[Z_{n+1} | \mathcal{F}_n] = Z_n.$$

If $\mathbf{E}[Z_{n+1}|\mathcal{F}_n] \leq Z_n$ in Definition 5, we call $\{Z_n\}_{n\geq 0}$ a supermartingale w.r.t. $\{\mathcal{F}_n\}_{n\geq 0}$. Similarly, if $\mathbf{E}[Z_{n+1}|\mathcal{F}_n] \geq Z_n$, we call it a submartingale.

Example 2 (1-D Random Walk) *Consider a random walk on $\mathbb{Z}$ starting from 0. The probability to the left and the probability to the right are both $\frac{1}{2}$ at each step. Denote the action at the n-th step by a uniform random variable*

$X_n \in \{-1, +1\}$. Let $S_n = \sum_{k=0}^{n} X_k$. Then we can verify $\{S_n\}_{n\geq 0}$ is a martingale w.r.t. $\{X_n\}_{n\geq 0}$ (or w.r.t. $\{S_n\}_{n\geq 0}$) by noticing that

$$\mathbf{E}\left[S_{n+1} \,\middle|\, \overline{X}_{0,n}\right] = \mathbf{E}\left[S_n + X_{n+1} \,\middle|\, \overline{X}_{0,n}\right] = S_n + \mathbf{E}\left[X_{n+1} \,\middle|\, \overline{X}_{0,n}\right] = S_n.$$

More generally, if $\mathbf{E}\left[X_k \,\middle|\, \overline{X}_{0,n}\right] = \mu$, we define $Y_k = X_k - \mu$ and $S'_n \triangleq \sum_{k=0}^{n} Y_k = S_n - (n+1)\mu$. Then S'_n is a martingale w.r.t. $\{Y_n\}_{n\geq 0}$.

Example 3 *Consider a sequence of random variables $\{X_n\}_{n\geq 0}$ where $\mathbf{E}\left[X_n \,\middle|\, \overline{X}_{0,n-1}\right] = 1$ for all $n \geq 1$. Let $P_n = \prod_{k=0}^{n} X_k$. Then we can verify $\{P_n\}_{n\geq 0}$ is a martingale w.r.t. $\{X_n\}_{n\geq 0}$ by verifying that*

$$\mathbf{E}\left[P_{n+1} \,\middle|\, \overline{X}_{0,n}\right] = \mathbf{E}\left[P_n \cdot X_{n+1} \mid \bar{X}_{0,n}\right] = P_n \cdot \mathbf{E}\left[X_{n+1} \,\middle|\, \overline{X}_{0,n}\right] = P_n.$$

Example 4 (Galton-Watson Process) *Recall the Galton-Watson process we discussed in the last lecture. Suppose that all the individuals reproduce independently of each other and have the same offspring distribution. Let G_t be the number of individuals of the t-th generation. Each individual of generation t gives birth to a random number of children of generation $t+1$. Denote by $X_{t,k}$ the number of children of the k-th individual in the t-th generation. Assume that $X_{t,k}$ are i.i.d. and let $\mu \triangleq \mathbf{E}\left[X_{t,k}\right]$. Then we have $G_{t+1} = \sum_{k=1}^{G_t} X_{t,k}$. Thus,*

$$\mathbf{E}\left[G_{t+1} \mid G_t\right] = \mathbf{E}\left[\sum_{k=1}^{G_t} X_{t,k} \,\middle|\, G_t\right] = G_t \cdot \mathbf{E}\left[X_{t,1}\right] = \mu G_t.$$

Define $M_t = \mu^{-t} G_t$. Then

$$\mathbf{E}\left[M_{t+1} \mid G_t\right] = \mu^{-t-1}\mathbf{E}\left[G_{t+1} \mid G_t\right] = \mu^{-t} G_t = M_t.$$

That is, $\{M_t\}_{t\geq 0}$ is a martingale w.r.t. $\{G_t\}_{t\geq 0}$.

Example 5 (Pólya's urn) *Suppose there are some white balls and black balls in an urn. All of these balls are identical except the colors. Consider the following stochastic process: each round we pick a ball uniformly at random and observe its color; then we return the ball, and add an additional ball of the same color into the urn. We repeat the process, and our goal is to study the sequence of colors of the selected balls.*

Example 5 shows that X_n does not have to be i.i.d..

W.l.o.g. assume that we start from only one white ball and one black ball in the urn, and the index of rounds starts from 2. Then after round n, there are exactly n balls in the urn. Let X_n be the number of black balls after round n, and $Z_n = \frac{X_n}{n}$ is the ratio of black balls after round n. Clearly $Z_2 = \frac{1}{2}$. Then we have

$$\begin{aligned}\mathbf{E}\left[Z_{n+1} \mid \bar{X}_{2,n}\right] &= \frac{1}{n+1}\mathbf{E}\left[X_{n+1} \,\middle|\, \overline{X}_{2,n}\right] \\ &= \frac{1}{n+1}\left(Z_n(X_n+1) + (1-Z_n)X_n\right) = \frac{Z_n + X_n}{n+1} = Z_n.\end{aligned}$$

That is, $\{Z_n\}_{n\geq 2}$ is a martingale w.r.t. $\{X_n\}_{n\geq 2}$.

References