[AI2613 Lecture 8] Doob Martingale, Azuma-Hoeffding, McDiarmid

April 24, 2022

1 *Hoeffding's Inequality*

Recall the Chernoff bound we discussed in Lecture 2.

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

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

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

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

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 2 (Hoeffding's Inequality) *Let $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ be i.i.d. random variables where $X_i \sim \mathtt{Ber}(p)$ for all $i = 1, \dots, n$. 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.

Before proving Theorem 2 in Section 3, we see a practical application of Hoeffding's inequality.

Example 1 (Meal Delivery) *During the quarantine of our campus, the professors deliver meals for students using their private cars or trikes. Then a practical problem is how to estimate the amount of meals on a trike conveniently (See the news).*

Assume there are n boxes of meal on the trike ($n \geq 200$ and is unknown for us). Let X_i be the weight of the i-th box of meal. Assume that $X_1, X_2, \dots, X_n$ are i.i.d. random variables, each $X_i \in [250, 350]$ (unit: gram) and $\mu = \mathbf{E}[X_i] = 300$. Let S be the total weight of the meal boxes on the trike, that is, $S = \sum_{i=1}^n X_i$. We can weigh the meal boxes and use $\hat{n} = \frac{S}{\mu}$ as an estimator for n. Now we compute how accurate this estimator is.

Let $\delta \geq 0$ be a constant. By Hoeffding's inequality,

$$\mathbf{Pr}[|\hat{n} - n| > \delta n] = \mathbf{Pr}[|S - \mu n| > \delta \mu n] \leq 2\exp\left\{-\frac{2\delta^2\mu^2 n^2}{\sum_{i=1}^n (350-250)^2}\right\}. \quad (1)$$

Plugging $\mu = 300$, $\delta = 0.05$ and $n \geq 200$ into Equation (1), by direct calculation, we have

$$\mathbf{Pr}[\hat{n} \in [0.95n, 1.05n]] \geq 1 - 2.4682 \times 10^{-4}.$$

2 *Concentration on Martingale*

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 $X_i \sim \mathrm{Ber}\left(\frac{r}{r+g}\right)$ 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.

2.1 Azuma-Hoeffding's Inequality

Theorem 3 (Azuma-Hoeffding's Inequality) *If $\{S_n\}_{n\geq 0}$ where $S_k = \sum_{i=0}^{k} X_i$ is a martingale w.r.t. $\{X_n\}_{n\geq 0}$ with $X_i \in [a_i, b_i]$ with probability 1, then*

$$\mathbf{Pr}\left[|S_n - S_0| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{\sum\limits_{i=1}^{n}(b_i - a_i)^2}\right).$$

The proof of this theorem is in Section 3. The requirement of martingale in Theorem 3 seems to be even harder to satisfy than the requirement of independence. However, in many cases, we can construct a doob martingale to apply the Azuma-Hoeffding's inequality.

Definition 4 (Doob Martingale, Doob Sequence) *Let $X_1, \dots, X_n$ be a sequence of (unnecessarily independent) random variables and $f(\overline{X}_{1,n}) = f(X_1, \dots, X_n) \in \mathbb{R}$ be a function. For $i \geq 0$, Let $Z_i \triangleq \mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i}\right]$. Then we call $\{Z_n\}_{n\geq 0}$ a Doob martingale or a Doob sequence.*

It is easy to verify that $\{Z_n\}_{n\geq 0}$ in Definition 4 is indeed a martingale w.r.t. $\{X_n\}$ by seeing

$$\mathbf{E}\left[Z_i \,\middle|\, \overline{X}_{1,i-1}\right] = \mathbf{E}\left[\mathbf{E}[f(\overline{X}_{1,n}) \,|\, \overline{X}_{1,i}] \,\middle|\, \overline{X}_{1,i-1}\right] = \mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i-1}\right] = Z_{i-1}.$$

Let $\mathcal{F} = \sigma(\overline{X}_{1,i})$. We can see that Z_i is $\mathcal{F}_i$ measurable by definition. Moreover, we know that $Z_0 = \mathbf{E}\left[f(\overline{X}_{1,n})\right]$ and $Z_n = f(\overline{X}_{1,n})$.

Recall the balls-in-a-bag problem we discussed above. Define $f : \mathbb{R}^n \to \mathbb{R}$ by letting $f(y_1, y_2, \dots, y_n) = \sum_{i=1}^{n} y_i$. Then in the drawing without replacement scenario, $Y = \sum_{i=1}^{n} Y_i = f(Y_1, Y_2, \dots, Y_n)$. Now we construct the Doob martingale for f.

Let $Z_i = \mathbf{E}\left[f(\overline{Y}_{1,n}) \,\middle|\, \overline{Y}_{1,i}\right]$. We know that $Z_0 = \mathbf{E}\left[f(\overline{Y}_{1,n})\right] = \mathbf{E}\left[Y\right] = n \cdot \frac{r}{r+g}$ and $Z_n = f(\overline{Y}_{1,n})$. Let $X_0 \triangleq Z_0$ and $X_i \triangleq Z_i - Z_{i-1}$ for $i \geq 1$. Then Z_n can be represented as $Z_n = \sum_{i=0}^{n} X_i$. In order to apply Azuma-Hoeffding, we need to bound the *width* of the martingale, i.e. $|X_i| = |Z_i - Z_{i-1}|$.

By definition,

$$Z_i - Z_{i-1} = \mathbf{E}\left[f(\overline{Y}_{1,n}) \,\middle|\, \overline{Y}_{1,i}\right] - \mathbf{E}\left[f(\overline{Y}_{1,n}) \,\middle|\, \overline{Y}_{1,i-1}\right].$$

If we use S_i to denote the number of red balls among the first i balls, namely $S_i = \sum_{j=1}^{i} Y_j$, then

$$\mathbf{E}\left[f(\overline{Y}_{1,n}) \,\middle|\, \overline{Y}_{1,i}\right] = \mathbf{E}\left[f(\overline{Y}_{1,n}) \,\middle|\, S_i\right] = S_i + (n-i)\cdot\frac{r - S_i}{g + r - i}.$$

Therefore $S_i = S_{i-1} + Y_i$ and

$$\begin{aligned} Z_i - Z_{i-1} &= \left(S_i + (n-i)\cdot\frac{r - S_i}{g+r-i}\right) - \left(S_{i-1} + (n-i+1)\cdot\frac{r - S_{i-1}}{g+r-i+1}\right) \\ &= \frac{g+r-n}{g+r-i}\left(Y_i + \frac{S_{i-1} - r}{g+r-i+1}\right). \end{aligned}$$

Note that $r \geq S_{i-1}$ and $g \geq (i-1) - S_{i-1}$, we have

$$Z_i - Z_{i-1} \leq \frac{g+r-n}{g+r-i}\left(1 + \frac{S_{i-1}-r}{g+r-i+1}\right) \leq \frac{g+r-n}{g+r-i} \leq 1,$$

$$Z_i - Z_{i-1} \geq \frac{g+r-n}{g+r-i}\left(\frac{S_{i-1}-r}{g+r-i+1}\right) \geq -\frac{g+r-n}{g+r-i} \geq -1.$$

Therefore $-1 \leq X_i \leq 1$ and we can apply Azuma-Hoeffding to $Z_n - Z_0$ to obtain

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

2.2 McDiarmids Inequality

The Doob sequence we used in the balls-in-a-bag example is a very powerful and general tool to obtain concentration bounds. For a model defined by n random variables $X_1, \ldots, X_n$ and any quantity $f(X_1, \ldots, X_n)$ that we want to estimate, we can apply the Azuma-Hoeffding inequality to the Doob sequence of f. As shown in the previous example, the quality of the bound relies on the *width* of the martingale, that is, the volume of $|Z_i - Z_{i-1}|$. To determine the width of each $|Z_i - Z_{i-1}|$ is relatively easy if the function f and the variables $\{X_i\}_{1\leq i\leq n}$ enjoy certain nice properties.

Definition 5 (c-Lipschitz Function) *A function $f(x_1, \cdots, x_n)$ satisfies c-Lipschitz condition if*

$$\forall i \in [n], \forall x_1, \cdots, x_n, \forall y_i: \quad |f(x_1, \cdots, x_i, \cdots, x_n) - f(x_1, \cdots, y_i, \cdots, x_n)| \leq c.$$

The McDiarmid's inequality is the application of Azuma-Hoeffding inequality to Lipschitz f and independent $\{X_i\}$.

Theorem 6 (McDiarmid's Inequality) *Let f be a function on n variables satisfying c-Lipschitz condition and $X_1, \cdots, X_n$ be n independent variables. Then we have*

$$\mathbf{Pr}\left[|f(X_1, \cdots, X_n) - \mathbf{E}[f(X_1, \cdots, X_n)]| \geq t\right] \leq 2e^{-\frac{2t^2}{nc^2}}.$$

Proof. We use f and $\{X_i\}_{i\geq 1}$ to define a Doob martingale $\{Z_i\}_{i\geq 1}$:

$$\forall i: Z_i = \mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i}\right].$$

Then

$$Z_i - Z_{i-1} = \mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i}\right] - \mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i-1}\right].$$

Next we try to determine the width of $Z_i - Z_{i-1}$. Clearly

$$Z_i - Z_{i-1} \geq \inf_x \left\{\mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i-1}, X_i = x\right] - \mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i-1}\right]\right\},$$

and

$$Z_i - Z_{i-1} \leq \sup_y \left\{\mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i-1}, X_i = y\right] - \mathbf{E}\left[f(\overline{X}_{1,n}) \,\middle|\, \overline{X}_{1,i-1}\right]\right\}.$$

The gap between the upper bound and the lower bound is

$$\sup_{x,y}\left\{\mathbf{E}\left[f(\overline{X}_{1,n})\,\middle|\,\overline{X}_{1,i-1},X_i=y\right]-\mathbf{E}\left[f(\overline{X}_{1,n})\,\middle|\,\overline{X}_{1,i-1},X_i=x\right]\right\}.$$

For every x, y and $\sigma_1,\dots,\sigma_{i-1}$,

$$\begin{aligned}
&\mathbf{E}\left[f(\overline{X}_{1,n})\,\middle|\,\bigwedge_{1\le j\le i-1}X_j=\sigma_j,X_i=y\right]-\mathbf{E}\left[f(\overline{X}_{1,n})\,\middle|\,\bigwedge_{1\le j\le i-1}X_j=\sigma_j,X_i=x\right]\\
=&\sum_{\sigma_{i+1},\dots,\sigma_n}\left(\mathbf{Pr}\left[\bigwedge_{i+1\le j\le n}X_j=\sigma_j\,\middle|\,\bigwedge_{1\le j\le i-1}X_j=\sigma_j,X_i=y\right]\cdot f(\sigma_1,\dots,\sigma_{i-1},y,\sigma_{i+1},\dots,\sigma_n)\right.\\
&\left.-\mathbf{Pr}\left[\bigwedge_{i+1\le j\le n}X_j=\sigma_j\,\middle|\,\bigwedge_{1\le j\le i-1}X_j=\sigma_j,X_i=x\right]\cdot f(\sigma_1,\dots,\sigma_{i-1},x,\sigma_{i+1},\dots,\sigma_n)\right)\\
\overset{(\heartsuit)}{=}&\sum_{\sigma_{i+1},\dots,\sigma_n}\mathbf{Pr}\left[\bigwedge_{i+1\le j\le n}X_j=\sigma_j\right]\cdot\left(f(\sigma_1,\dots,\sigma_{i-1},y,\sigma_{i+1},\dots,\sigma_n)-f(\sigma_1,\dots,\sigma_{i-1},x,\sigma_{i+1},\dots,\sigma_n)\right)\\
\overset{(\clubsuit)}{\le}&\ c.
\end{aligned}$$

where $(\heartsuit)$ uses independence of $\{X_i\}$ and $(\clubsuit)$ uses the c-Lipsichitz property of f.

Applying Azuma-Hoeffding, we have

$$\mathbf{Pr}\left[|Z_n-Z_0|\ge t\right]=\mathbf{Pr}\left[|f(X_1,\cdots,X_n)-\mathbf{E}\left[f(X_1,\cdots,X_n)\right]|\ge t\right]\le 2e^{-\frac{2t^2}{nc^2}}.$$

□

Then we examine two applications of McDiarmid's inequality.

Example 2 (Pattern matching) *Let $P\in\{0,1\}^k$ be a fixed string. For a random string $X\in\{0,1\}^n$, what is the expected number of occurrences of P in X?*

The expectation of occurrence times can be easily calculated using the linearity of expectations. We define n independent random variables $X_1,\cdots,X_n$, where X_i denotes i-th character of X. Let $Y=f(X_1,\cdots,X_n)$ be the number of occurrences of P in X. Note that there are at most $n-k+1$ occurrences of P in X, and we can enumerate the first position of each occurrence. By the linearity of expectation, we have

$$\mathbf{E}\left[f\right]=\frac{n-k+1}{2^k}.$$

We can then use McDarmid's inequality to show that f is well-concentrated. To see this, we note that variables in $\{X_i\}$ are independent and the function f is k-Lipschitz: If we change one bit of X, the number of occurrences changes at most k.

Therefore

$$\mathbf{Pr}\left[|Z_n-Z_0|\ge t\right]=\mathbf{Pr}\left[|f-\mathbf{E}\left[f\right]|\ge t\right]\le 2e^{-\frac{2t^2}{nk^2}}.$$

Another application of McDiarmid's Inequality is to establish the concentration of chromatic number for Erdős-Rényi random graphs $\mathcal{G}(n,p)$.

Example 3 (Chromatic Number of $\mathcal{G}(n,p)$) *Recall the notation $\mathcal{G}(n,p)$ specifies a distribution over all undirected simple graphs with n vertices. In the model, each of the $\binom{n}{2}$ possible edges exists with probability p independently.*

For a graph $G \sim \mathcal{G}(n,p)$, we use $\chi(G)$ to denote its chromatic number, the minimum number q so that G can be properly colored using q colors. There are different ways to represent G using random variables.

The most natural way is to introduce a variable X_e for every pair of vertices $e = \{u,v\} \subseteq V$ where $X_e = \mathbf{1}[\text{the edge } e \text{ exists in } G]$. Then $\{X_e\}$ are independent and the chromatic number can be written as a function $\chi(X_{e_1}, X_{e_2}, \dots, X_{e_{\binom{n}{2}}})$. It is easy to see that χ is 1-Lipschitz as removing to adding one edge can only change the chromatic number by at most one. So by McDarmid's inequality, we have

$$\mathbf{Pr}\left[|\chi - \mathbf{E}[\chi]| \geq t\right] \leq 2e^{-2t^2\binom{n}{2}^{-1}}.$$

However, this bound is not satisfactory as we need to set $t = \Theta(n)$ in order to upper bound the RHS by a constant.

We can encode the graph G in a more efficient way while reserving the Lipschitz and the independence property. Suppose the vertex set of G is $\{v_1, \dots, v_n\}$. We define n random variables $Y_1, \cdots, Y_n$, where Y_i encodes the edges between v_i and $\{v_1, \cdots, v_{i-1}\}$. Once $Y_1, \cdots, Y_n$ are given, the graph is known, so the chromatic number can be written as a function $\chi(Y_1, \dots, Y_n)$. Since Y_i only involves the connections between v_i and $v_1, \cdots, v_{i-1}$, the n variables are independent.

It is also easy to see that if X_i changes, the chromatic number changes at most one. Hence χ is 1-Lipschitz as well. Applying McDiarmid's inequality we have

$$\mathbf{Pr}\left[|\chi - \mathbf{E}[\chi]| \geq t\right] \leq 2e^{-\frac{2t^2}{n}}.$$

In this way, we only need $t = \Theta(\sqrt{n})$ to bound the RHS.

3 Proof

3.1 Proof of Theorem 2

First, we prove the following Hoeffding's lemma which will be the main technical ingredient to prove the inequality.

Lemma 7 *et X be a random variable with $\mathbf{E}[X] = 0$ and $X \in [a,b]$. Then it holds that*

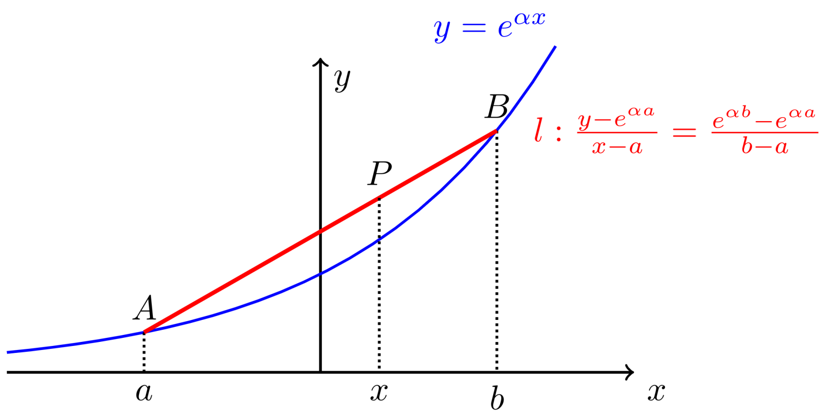
$$\mathbf{E}\left[e^{\alpha X}\right] \leq \exp\left(\frac{\alpha^2(b-a)^2}{8}\right) \text{ for all } \alpha \in \mathbb{R}.$$

Proof.

We first find a linear function to upper bound $e^{\alpha x}$ so that we could apply the linearity of expectation to bound $\mathbf{E}\left[e^{\alpha X}\right]$. By the convexity of the exponential function and as illustrated in the figure below, we have

$$e^{\alpha x} \leq \frac{e^{\alpha b} - e^{\alpha a}}{b-a}(x-a) + e^{\alpha a}, \quad \text{for all } a \leq x \leq b.$$

Thus,



$$\begin{aligned}
\mathbf{E}\left[e^{\alpha x}\right] &\leq \frac{e^{\alpha b} - e^{\alpha a}}{b-a}(-a) + e^{\alpha a} = \frac{-a}{b-a}e^{\alpha b} + \frac{b}{b-a}e^{\alpha a} \\
&= e^{\alpha a}\left(\frac{b}{b-a} - \frac{a}{b-a}e^{\alpha(b-a)}\right) \\
&= e^{-\theta t}(1-\theta+\theta e^{t}) && (\theta = -\frac{a}{b-a}, t = \alpha(b-a)) \\
&\triangleq e^{g(t)},
\end{aligned}$$

where

$$g(t) = -\theta t + \log\left(1-\theta+\theta e^{t}\right).$$

By Taylor's theorem, for every real t there exists a δ between 0 and t such that,

$$g(t) = g(0) + tg'(0) + \frac{1}{2}g''(\delta)t^2$$

Note that,

$$\begin{aligned}
g(0) &= 0; \\
g'(0) &= -\theta + \left.\frac{\theta e^t}{1-\theta+\theta e^t}\right|_{t=0} \\
&= 0; \\
g''(\delta) &= \frac{\theta e^t(1-\theta+\theta e^t) - \theta e^t}{(1-\theta+\theta e^t)^2} \\
&= \frac{(1-\theta)\theta e^t}{(1-\theta+\theta e^t)^2} \\
&= \frac{(1-\theta)\theta}{\theta^2 z + 2(1-\theta)\theta + \frac{(1-\theta)^2}{z}} && (z = e^t) \\
&\le \frac{(1-\theta)\theta}{2\theta(1-\theta) + 2(1-\theta)\theta} && (z > 0) \\
&= \frac{1}{4}.
\end{aligned}$$

Thus

$$g(t) \le 0 + t \cdot 0 + \frac{1}{2}t^2 \cdot \frac{1}{4} = \frac{1}{8}t^2 = \frac{1}{8}\alpha^2(b-a)^2.$$

Therefore, $\mathbf{E}\left[e^{\alpha x}\right] \le \exp\left(\frac{\alpha^2(b-a)^2}{8}\right)$ holds. □

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

Proof. [Proof of Theorem 2]

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

$$\mathbf{Pr}[X \ge t] \overset{\alpha \ge 0}{=} \mathbf{Pr}\left[e^{\alpha X} \ge e^{\alpha t}\right] \le \frac{\mathbf{E}\left[e^{\alpha X}\right]}{e^{\alpha t}}$$

and

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

applying Hoeffding's lemma for each $\mathbf{E}\left[e^{\alpha X_i}\right]$ yields

$$\mathbf{E}\left[e^{\alpha X_i}\right] \le \exp\left(\frac{\alpha^2(b_i-a_i)^2}{8}\right).$$

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

$$\begin{aligned}
\mathbf{Pr}[X \ge t] \le \frac{\prod_{i=1}^n \mathbf{E}\left[e^{\alpha X_i}\right]}{e^{\alpha t}} &\le \exp\left(-\alpha t + \frac{\alpha^2}{8}\sum_{i=1}^n (b_i-a_i)^2\right) \\
&= \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\right).
\end{aligned}$$

□

3.2 Proof of Theorem 3

Now we will sketch a proof of the Azuma-Hoeffding, which is quite similar to our proof of the Hoeffding inequality.

Proof. [Proof of Theorem 3]

Recall when we were trying to prove the Hoeffding inequality, the most difficult part is to estimate the term

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

In the case of Azuma-Hoeffding, we can use the property of martingales instead of independence to obtain a bound of this term. To see this, we have

$$\begin{aligned}\mathbf{E}\left[\prod_{i=1}^{n} e^{\alpha X_i}\right] &= \mathbf{E}\left[\mathbf{E}\left[\prod_{i=1}^{n} e^{\alpha X_i} \middle| \overline{X}_{n-1}\right]\right] \\ &= \mathbf{E}\left[\prod_{i=1}^{n-1} e^{\alpha X_i} \mathbf{E}\left[e^{\alpha X_n} \,\middle|\, \overline{X}_{n-1}\right]\right].\end{aligned}$$

The bounds then follows by an induction argument and a conditional expectation version of Hoeffding lemma:

$$\mathbf{E}\left[e^{\alpha X_n} \,\middle|\, \overline{X}_{n-1}\right] \le e^{-\frac{\alpha(b_i - a_i)^2}{8}}.$$

The proof is almost the same as our proof of Hoeffding lemma in the last lecture. □