AI2613 Stochastic Processes Homework 3

Xiangyuan Xue (521030910387)

1 Maximal Inequality

(a) Since $\{Z_t\}_{t\geq 0}$ is a martingale with respect to $\{F_t\}_{t\geq 1}$, we have

$$\mathbf{E}\left[Z_{t+1}|F_t\right] = Z_t$$

By the property of conditional expectation

$$\mathbf{E}\left[Z_{t+1}Z_t\right] = \mathbf{E}\left[\mathbf{E}\left[Z_{t+1}Z_t|F_t\right]\right] = \mathbf{E}\left[Z_t \cdot \mathbf{E}\left[Z_{t+1}|F_t\right]\right] = \mathbf{E}\left[Z_t^2\right]$$

Therefore, for any $n \in \mathbb{N}$ we have

$$\begin{aligned}
\sum_{k=1}^{n} \mathbf{E}\left[(Z_k - Z_{k-1})^2\right] &= \sum_{k=1}^{n} \mathbf{E}\left[Z_k^2\right] + \sum_{k=1}^{n} \mathbf{E}\left[Z_{k-1}^2\right] - 2\sum_{k=1}^{n} \mathbf{E}\left[Z_k Z_{k-1}\right] \\
&= \sum_{k=1}^{n} \mathbf{E}\left[Z_k^2\right] + \sum_{k=1}^{n} \mathbf{E}\left[Z_{k-1}^2\right] - 2\sum_{k=1}^{n} \mathbf{E}\left[Z_{k-1}^2\right] \\
&= \sum_{k=1}^{n} \mathbf{E}\left[Z_k^2\right] - \sum_{k=1}^{n} \mathbf{E}\left[Z_{k-1}^2\right] \\
&= \mathbf{E}\left[Z_n^2\right] - \mathbf{E}\left[Z_0^2\right]
\end{aligned}$$

(b) Note that $Z'_t = Z_{t\wedge\tau}$ where $t \wedge \tau = \min\{t, \tau\} \leq t$, which indicates that Z'_t is F_t-measurable. By the property of conditional expectation

$$\begin{aligned}
\mathbf{E}\left[Z'_{t+1}|F_t\right] &= \mathbf{E}\left[Z'_t + \left(Z'_{t+1} - Z'_t\right)|F_t\right] \\
&= \mathbf{E}\left[Z'_t + \mathbb{I}\left[\tau \geq t+1\right] \cdot \left(Z_{t+1} - Z_t\right)|F_t\right] \\
&= Z'_t + \mathbb{I}\left[\tau \geq t+1\right] \cdot \mathbf{E}\left[Z_{t+1} - Z_t|F_t\right] \\
&= Z'_t + \mathbb{I}\left[\tau \geq t+1\right] \cdot \left(\mathbf{E}\left[Z_{t+1}|F_t\right] - Z_t\right) \\
&= Z'_t
\end{aligned}$$

Therefore, $\{Z'_t\}_{t\geq 0}$ is a martingale with respect to $\{F_t\}_{t\geq 1}$.

(c) Let τ be a stopping time where

$$\tau = \begin{cases} \min\limits_{1\leq k\leq n} \left\{k : |S_k| \geq \lambda\right\}, & \max\limits_{1\leq k\leq n} |S_k| \geq \lambda \\ n, & \max\limits_{1\leq k\leq n} |S_k| < \lambda \end{cases}$$

which indicates the smallest k such that $|S_k| \geq \lambda$. Notice that

$$\mathbf{Pr}\left[\max_{1\leq k\leq n} |S_k| \geq \lambda\right] = \mathbf{Pr}\left[|S_\tau| \geq \lambda\right]$$

Since $\mathbf{E}\left[X_i\right] = 0$, it holds that

$$\mathbf{E}\left[S_i\right] = \mathbf{E}\left[\sum_{k=1}^{i} X_k\right] = \sum_{k=1}^{i} \mathbf{E}\left[X_k\right] = 0$$

By Chebyshev's inequality

$$\mathbf{Pr}\left[|S_\tau| \geq \lambda\right] = \mathbf{Pr}\left[|S_\tau - E\left[S_\tau\right]| \geq \lambda\right] \leq \frac{\mathbf{D}\left[S_\tau\right]}{\lambda^2} = \frac{\mathbf{E}\left[S_\tau^2\right] - \mathbf{E}^2\left[S_\tau\right]}{\lambda^2} = \frac{\mathbf{E}\left[S_\tau^2\right]}{\lambda^2}$$

By the property proved in (a), it holds that

$$\mathbf{E}\left[S_\tau^2\right] = \mathbf{E}\left[S_\tau^2\right] - \mathbf{E}\left[S_0^2\right] = \sum_{k=1}^{\tau} \mathbf{E}\left[(S_k - S_{k-1})^2\right] \leq \sum_{k=1}^{n} \mathbf{E}\left[X_k^2\right]$$

Therefore, for any $\lambda > 0$ we have

$$\mathbf{Pr}\left[\max_{1\leq k\leq n} |S_k| \geq \lambda\right] \leq \frac{1}{\lambda^2} \sum_{k=1}^{n} \mathbf{E}\left[X_k^2\right]$$

2 Biased Random Walk

(a) According to the definition of $\{S_t\}_{t\geq 0}$, we have

$$\begin{aligned}
\mathbf{E}\left[S_{t+1}|X_1, X_2, \ldots, X_t\right] &= \mathbf{E}\left[S_t + X_{t+1} + 2p - 1|X_1, X_2, \ldots, X_t\right] \\
&= S_t + 2p - 1 + \mathbf{E}\left[X_{t+1}|X_1, X_2, \ldots, X_t\right] \\
&= S_t + 2p - 1 + \mathbf{E}\left[X_{t+1}\right] \\
&= S_t + 2p - 1 + \left[p \cdot (-1) + (1-p) \cdot 1\right] \\
&= S_t
\end{aligned}$$

Therefore, $\{S_t\}_{t\geq 0}$ is a martingale with respect to $\{X_t\}_{t\geq 1}$.

(b) According to the definition of $\{P_t\}_{t\geq 0}$, we have

$$\begin{aligned}
\mathbf{E}\left[P_{t+1}|X_1, X_2, \ldots, X_t\right] &= \mathbf{E}\left[P_t \cdot \left(\frac{p}{1-p}\right)^{X_{t+1}} \middle| X_1, X_2, \ldots, X_t\right] \\
&= P_t \cdot \mathbf{E}\left[\left(\frac{p}{1-p}\right)^{X_{t+1}}\right] \\
&= P_t \cdot \left[p \cdot \left(\frac{1-p}{p}\right) + (1-p) \cdot \left(\frac{p}{1-p}\right)\right] \\
&= P_t
\end{aligned}$$

Therefore, $\{P_t\}_{t\geq 0}$ is a martingale with respect to $\{X_t\}_{t\geq 1}$.

(c) Consider the biased random walk as a Markov chain, which is obviously finite, irreducible and aperiodic. Therefore, any state is positive recurrent and $\mathbf{E}[\tau] < \infty$. Note that

$$\begin{aligned} |P_{t+1} - P_t| &= \left| \left(\frac{p}{1-p}\right)^{Z_{t+1}} - \left(\frac{p}{1-p}\right)^{Z_t} \right| \\ &\le \left(\frac{p}{1-p}\right)^{Z_{t+1}} + \left(\frac{p}{1-p}\right)^{Z_t} \\ &\le 2 \cdot \max\left\{ \left(\frac{p}{1-p}\right)^{-a}, \left(\frac{p}{1-p}\right)^{b} \right\} \end{aligned}$$

Thus, $|P_{t+1} - P_t|$ is bounded by a constant. By optional stopping theorem, it holds that

$$\mathbf{E}[P_\tau] = \mathbf{E}[P_0] = 1$$

which indicates that

$$\begin{cases} \mathbf{Pr}[Z_\tau = -a] + \mathbf{Pr}[Z_\tau = b] = 1 \\ \mathbf{Pr}[Z_\tau = -a] \cdot \left(\frac{p}{1-p}\right)^{-a} + \mathbf{Pr}[Z_\tau = b] \cdot \left(\frac{p}{1-p}\right)^{b} = 1 \end{cases}$$

which yields

$$\begin{cases} \mathbf{Pr}[Z_\tau = -a] = \dfrac{1 - \left(\frac{p}{1-p}\right)^{b}}{\left(\frac{p}{1-p}\right)^{-a} - \left(\frac{p}{1-p}\right)^{b}} \\ \mathbf{Pr}[Z_\tau = b] = \dfrac{\left(\frac{p}{1-p}\right)^{-a} - 1}{\left(\frac{p}{1-p}\right)^{-a} - \left(\frac{p}{1-p}\right)^{b}} \end{cases}$$

Note that

$$|S_{t+1} - S_t| = |X_{t+1} + 2p - 1| \le |X_{t+1}| + |2p - 1| \le 2$$

Thus, $|S_{t+1} - S_t|$ is bounded by a constant. By optional stopping theorem, it holds that

$$\mathbf{E}[S_\tau] = \mathbf{E}[S_0] = 0$$

which indicates that

$$\begin{aligned} \mathbf{E}[S_\tau] &= \mathbf{E}\left[\sum_{i=1}^{\tau} (X_i + 2p - 1)\right] \\ &= \mathbf{E}[Z_\tau + \tau(2p-1)] \\ &= \mathbf{E}[Z_\tau] + \mathbf{E}[\tau](2p-1) \\ &= \mathbf{Pr}[Z_\tau = -a] \cdot (-a) + \mathbf{Pr}[Z_\tau = b] \cdot b + \mathbf{E}[\tau](2p-1) \\ &= 0 \end{aligned}$$

which yields

$$\mathbf{E}[\tau] = \frac{a + b - a\left(\frac{p}{1-p}\right)^{b} - b\left(\frac{p}{1-p}\right)^{-a}}{(2p-1)\left[\left(\frac{p}{1-p}\right)^{-a} - \left(\frac{p}{1-p}\right)^{b}\right]}$$

The result holds for $p \ne \frac{1}{2}$. It is trivial that $\mathbf{E}[\tau] = ab$ when $p = \frac{1}{2}$.

3 Learning Theory

(a) Since $\sup_{h\in\mathcal{H}} |L(h) - L_S(h)| \le \frac{\varepsilon}{2}$, it holds that $|L(h^*) - L_S(h^*)| \le \frac{\varepsilon}{2}$ and $\left|L(\hat{h}) - L_S(\hat{h})\right| \le \frac{\varepsilon}{2}$.

Since $\hat{h} = \arg\min_{h\in\mathcal{H}} L_S(h)$, it holds that $L_S(\hat{h}) \le L_S(h^*)$. Hence

$$\begin{aligned} L(\hat{h}) - L(h^*) &= L(\hat{h}) - L_S(\hat{h}) + L_S(\hat{h}) - L_S(h^*) + L_S(h^*) - L(h^*) \\ &\le \left|L(\hat{h}) - L_S(\hat{h})\right| + |L(h^*) - L_S(h^*)| + \left[L_S(\hat{h}) - L_S(h^*)\right] \\ &\le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} + 0 \\ &= \varepsilon \end{aligned}$$

Therefore, it holds that $L(\hat{h}) \le L(h^*) + \varepsilon$.

(b) Consider $\mathrm{Rep}(S) = \sup_{h\in\mathcal{H}} [L(h) - L_S(h)]$ as a function of S. When a single sample is inserted into or removed from S, the function $\mathrm{Rep}(S)$ changes by at most $\frac{1}{m}$. Hence, $\mathrm{Rep}(S)$ satisfies $\frac{1}{m}$-Lipschitz condition. By McDiarmid's inequality

$$\mathbf{Pr}\left[\mathrm{Rep}(S) - \mathbf{E}_{S\sim\mathcal{X}^m}\left[\mathrm{Rep}(S)\right] \ge t\right] \le 2e^{-2mt^2}$$

For any $\delta \in (0,1)$, let $2e^{-2mt^2} = \delta$, we have $t = \sqrt{\frac{1}{2m}\log\frac{2}{\delta}}$. Hence, with probability at least $1-\delta$, it holds that

$$\mathrm{Rep}(S) - \mathbf{E}_{S\sim\mathcal{X}^m}\left[\mathrm{Rep}(S)\right] \le \sqrt{\frac{1}{2m}\log\frac{2}{\delta}}$$

Therefore, with probability at least $1-\delta$, for any $h\in\mathcal{H}$, it holds that

$$\begin{aligned} L(h) - L_S(h) &\le \mathrm{Rep}(S) \\ &\le \mathbf{E}_{S\sim\mathcal{X}^m}\left[\mathrm{Rep}(S)\right] + \sqrt{\frac{1}{2m}\log\frac{2}{\delta}} \\ &\le 2\cdot\mathbf{E}_{S\sim\mathcal{X}^m}\left[R(S)\right] + \sqrt{\frac{1}{2m}\log\frac{2}{\delta}} \end{aligned}$$

(c) Consider $R(S) = \frac{1}{m}\cdot\mathbf{E}_{\sigma\in\{-1,1\}^m}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^m \sigma_i\cdot\mathbb{I}\left[h(x_i)\ne l(x_i)\right]\right]$ as a function of S. When a single sample is inserted into or removed from S, the function $R(S)$ changes by at most $\frac{1}{m}$. Hence, $R(S)$ satisfies $\frac{1}{m}$-Lipschitz condition. By McDiarmid's Inequality

$$\mathbf{Pr}\left[\mathbf{E}_{S\sim\mathcal{X}^m}\left[R(S)\right] - R(S) \ge t\right] \le 2e^{-2mt^2}$$

For any $\delta\in(0,1)$, let $2e^{-2mt^2} = \frac{\delta}{2}$, we have $t = \sqrt{\frac{1}{2m}\log\frac{4}{\delta}}$. Hence, with probability at least $1-\frac{\delta}{2}$, it holds that

$$\mathbf{E}_{S\sim\mathcal{X}^m}\left[R(S)\right] \le R(S) + \sqrt{\frac{1}{2m}\log\frac{4}{\delta}}$$

According to (b), with probability at least $1-\frac{\delta}{2}$, it holds that

$$L(h)-L_S(h)\leq 2\cdot\mathbf{E}_{S\sim\mathcal{X}^m}\left[R(S)\right]+\sqrt{\frac{1}{2m}\log\frac{4}{\delta}}$$

Hence, with probability at least $\left(1-\frac{\delta}{2}\right)^2\geq 1-\delta$, for any $h\in\mathcal{H}$, it holds that

$$L(h)-L_S(h)\leq 2\cdot R(S)+3\cdot\sqrt{\frac{1}{2m}\log\frac{4}{\delta}}$$

Further, with probability at least $1-\frac{\delta}{2}$, it holds that

$$L(\hat{h})-L_S(\hat{h})\leq 2\cdot R(S)+3\cdot\sqrt{\frac{1}{2m}\log\frac{8}{\delta}}$$

Since $h^*=\arg\min_{h\in\mathcal{H}}L(h)$ does not depend on S, we can consider $L_S(h^*)$ as a function of S where $\mathbf{E}_{S\sim\mathcal{X}^m}\left[L_S(h^*)\right]=L(h^*)$. When a single sample is inserted into or removed from S, the function $L_S(h^*)$ changes by at most $\frac{1}{m}$, so $L_S(h^*)$ satisfies $\frac{1}{m}$-Lipschitz condition. By McDiarmid's Inequality

$$\mathbf{Pr}\left[L_S(h^*)-L(h^*)\geq t\right]\leq 2e^{-2mt^2}$$

Thus, with probability at least $1-\frac{\delta}{2}$, it holds that

$$L_S(h^*)-L(h^*)\leq\sqrt{\frac{1}{2m}\log\frac{4}{\delta}}$$

Hence, with probability at least $\left(1-\frac{\delta}{2}\right)^2\geq 1-\delta$, it holds that

$$\begin{aligned}L(\hat{h})-L(h^*)&=\left[L(\hat{h})-L_S(\hat{h})\right]+\left[L_S(\hat{h})-L_S(h^*)\right]+\left[L_S(h^*)-L(h^*)\right]\\&\leq 2\cdot R(S)+3\cdot\sqrt{\frac{1}{2m}\log\frac{8}{\delta}}+0+\sqrt{\frac{1}{2m}\log\frac{4}{\delta}}\\&\leq 2\cdot R(S)+4\cdot\sqrt{\frac{1}{2m}\log\frac{8}{\delta}}\end{aligned}$$

Therefore, for any $\delta\in(0,1)$, with probability at least $1-\delta$, it holds that

$$L(\hat{h})\leq L(h^*)+2\cdot R(S)+5\cdot\sqrt{\frac{1}{2m}\log\frac{8}{\delta}}$$

References

[1] Chandra Nair. Probability Theory. The Chinese University of Hong Kong.

[2] Feng Yu. Martingales. University of Bristol.

[3] Shalev-Shwartz S, Ben-David S. Understanding machine learning: From theory to algorithms[M]. Cambridge university press, 2014.