[Homework 3] Martingale and Stopping Time

Problem 1 (A maximal inequality)

Let $\{Z_t\}_{t\geq 0}$ be a martingale with respect to a filtration $\{\mathcal{F}_t\}_{t\geq 1}$.
(a) Prove that for any $n \in \mathbb{N}$,

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

Proof:

For all $k \in 1, 2, \ldots, n$,

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

Sum both sides up, we have

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

Q.E.D.

(b) Let τ be a stopping time for the martingale $\{Z_t\}_{t\geq 0}$. Define another sequence $\{Z'_t\}_{t\geq 0}$ as

$$Z'_t = \begin{cases} Z_t & \text{if } t < \tau; \\ Z_\tau & \text{if } t \geq \tau. \end{cases}$$

Prove that $\{Z'_t\}_{t\geq 0}$ is also a martingale.

Proof:

For $t < \tau - 1$ and $t \geq \tau$,

$$Z'_{t+1} = \begin{cases} Z_{t+1}, \text{if } t < \tau - 1 \\ Z_\tau, \text{if } t \geq \tau \end{cases}, Z'_t = \begin{cases} Z_t, \text{if } t < \tau - 1 \\ Z_\tau, \text{if } t \geq \tau \end{cases}.$$

Then if $t < \tau - 1$, $\mathbf{E}\left[Z'_{t+1} | \overline{Z'_{0,t}}\right] = \mathbf{E}\left[Z_{t+1} | \overline{Z_{0,t}}\right] = Z_t = Z'_t$, if $t \geq \tau$, $\mathbf{E}\left[Z'_{t+1} | \overline{Z'_{0,t}}\right] = Z_\tau = Z'_t$.

For $t = \tau - 1$, since $\mathbf{E}\left[Z_{t+1} | \overline{Z_{0,t}}\right] = Z_t$, we have, for $i = 0, 1, \cdots, t$, $Z'_i = Z_i$, and $Z'_{t+1} = Z_\tau = Z_{t+1}$. Therefore, $\mathbf{E}\left[Z'_{t+1} | \overline{Z'_{0,t}}\right] = \mathbf{E}\left[Z_{t+1} | \overline{Z_{0,t}}\right] = Z_t = Z'_t$.

So $\{Z'_t\}_{t\geq 0}$ is also a martingale.

Q.E.D.

(c) Let $X_1, \ldots, X_n$ be independent random variables with $\mathbf{E}[X_i] = 0$ for every $i \in [n]$. Define $S_i = \sum_{k=1}^{i} X_k$ for every $i \in [n]$.

Prove that for every $\lambda > 0$,

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

Proof:

It's easy to show that $\{S_t\}_{t\geq 0}$ be a martingale with respect to a filtration $\{\mathcal{F}_t\}_{t\geq 1}$.

And define the stopping time τ as $\min\{n, min\{t||S_k| \geq \lambda\}\}$. Define $\{S'_t\}_{t\geq 0}$ as the same way in (b), we have $\{S'_t\}_{t\geq 0}$ is a martingale and $(S'_k - S'_{k-1})^2 \leq (S_k - S_{k-1})^2$ for all k. It is because for $k \leq \tau$, $LHS = RHS$, and for $k > \tau$, $LHS = 0$.

Therefore,

$$\begin{aligned}
\mathbf{Pr}\left[\max_{1\leq k\leq n} |S_k| \geq \lambda\right] &= \mathbf{Pr}\left[|S'_\tau| \geq \lambda\right] \\
&= \mathbf{Pr}\left[|S'_n| \geq \lambda\right] \\
&\leq \frac{\mathbf{E}\left[S'^2_n\right]}{\lambda^2} \\
&= \frac{\mathbf{E}\left[S'^2_n\right] - \mathbf{E}\left[S'^2_0\right]}{\lambda^2} \\
&= \frac{\sum_{k=1}^n \mathbf{E}\left[(S'_k - S'_{k-1})^2\right]}{\lambda^2} \\
&= \frac{\sum_{k=1}^n \mathbf{E}\left[(S_k - S_{k-1})^2\right]}{\lambda^2} \\
&\leq \frac{1}{\lambda^2}\sum_{k=1}^n \mathbf{E}\left[X_k^2\right].
\end{aligned}$$

Q.E.D.

Problem 2 (Biased random walk)

We study the biased random walk in this exercise. Let $Z_t = \sum_{i=1}^t X_i$ where each $X_i \in \{-1, 1\}$ is independent, and satisfies $\mathbf{Pr}\left[X_i = -1\right] = p \in (0, 1)$.
(a) Define $S_t = \sum_{i=1}^t (X_i + 2p - 1)$. Show that $\{S_t\}_{t\geq 0}$ is a martingale.

To show that $\{S_t\}_{t\geq 0}$ is a martingale, we'll show $\mathbf{E}\left[S_t | X_{\overline{1,t-1}}\right] = S_{t-1}$.

Firstly, we have $\mathbf{E}\left[X_t\right] = 1 \times \mathbf{Pr}\left[X_t = 1\right] + (-1) \times \mathbf{Pr}\left[X_t = -1\right] = 1 - 2p$.

Then,

$$
\begin{aligned}
\mathbf{E}\left[S_t|X_{\overline{1,t-1}}\right] &= \mathbf{E}\left[\sum_{i=1}^{t}(X_i+2p-1)|X_{\overline{1,t-1}}\right] \\
&= \mathbf{E}\left[\sum_{i=1}^{t-1}(X_i+2p-1)+X_t+2p-1|X_{\overline{1,t-1}}\right] \\
&= \mathbf{E}\left[S_{t-1}+X_t+2p-1|X_{\overline{1,t-1}}\right] \\
&= \mathbf{E}\left[S_{t-1}|X_{\overline{1,t-1}}\right]+\mathbf{E}\left[X_t|X_{\overline{1,t-1}}\right]+2p-1 \\
&= S_{t-1}+\mathbf{E}\left[X_t\right]+2p-1 \\
&= S_{t-1}
\end{aligned}
$$

Therefore, $\{S_t\}_{t\geq 0}$ is a martingale.

(b) Define $P_t = \left(\frac{p}{1-p}\right)^{Z_t}$. Show that $\{P_t\}_{t\geq 0}$ is a martingale.

To show that $\{P_t\}_{t\geq 0}$ is a martingale, we'll show $\mathbf{E}\left[P_t|X_{\overline{1,t-1}}\right] = P_{t-1}$.

$$
\begin{aligned}
\mathbf{E}\left[P_t|X_{\overline{1,t-1}}\right] &= \mathbf{E}\left[(\frac{p}{1-p})^{\sum_{i=1}^{t}X_i}|X_{\overline{1,t-1}}\right] \\
&= \mathbf{E}\left[(\frac{p}{1-p})^{\sum_{i=1}^{t-1}X_i}\times(\frac{p}{1-p})^{X_t}|X_{\overline{1,t-1}}\right] \\
&= \mathbf{E}\left[P_{t-1}\times(\frac{p}{1-p})^{X_t}|X_{\overline{1,t-1}}\right] \\
&= \mathbf{E}\left[P_{t-1}|X_{\overline{1,t-1}}\right]\mathbf{E}\left[(\frac{p}{1-p})^{X_t}|X_{\overline{1,t-1}}\right] \\
&= P_{t-1}\mathbf{E}\left[(\frac{p}{1-p})^{X_t}\right] \\
&= P_{t-1}(\frac{p}{1-p}\times\mathbf{Pr}\left[X_t=1\right]+\frac{1-p}{p}\times\mathbf{Pr}\left[X_t=-1\right]) \\
&= P_{t-1}
\end{aligned}
$$

Therefore, $\{P_t\}_{t\geq 0}$ is a martingale.

(c) Suppose the walk stops either when $Z_t=-a$ or $Z_t=b$ for some $a,b>0$. Let τ be the stopping time. Compute $\mathbf{E}\left[\tau\right]$.

If $p = \frac{1}{2}$, we have proved that $\mathbf{E}\left[\tau\right] = ab$ in class.

We focus on the case when $p \neq \frac{1}{2}$. Before calculating $\mathbf{E}\left[\tau\right]$, we first determine $\mathbf{Pr}\left[Z_\tau = -a\right]$, the probability that the man stops at position $-a$. Let $P_a \triangleq \mathbf{Pr}\left[Z_\tau = -a\right]$. we want to apply Optional Stopping Theorem to show $\mathbf{E}\left[S_\tau\right] = \mathbf{E}\left[S_0\right]$. In a time period of length $T = a + b$, if the man walks towards the same direction, he must have stopped, either at $-a$ or b, which happens with probability $(\frac{1}{p})^{-(a+b)}$(walk leftwards) and $(\frac{1}{1-p})^{-(a+b)}$(walk rightwards). Therefore, if we divide the time into consecutive periods in this manner, in expected finite time, we can meet some period when the event happened. Hence, $\mathbf{E}\left[\tau\right] < \infty$.

Moreover, we clearly have $\mathbf{E}\left[|S_{t+1} - S_t||\mathcal{F}_t\right] = \mathbf{E}\left[|X_{t+1} + 2p - 1||\mathcal{F}_t\right] < 2$ for every $0 \leq t < \tau$, so the third condition of OST holds, which implies that $\mathbf{E}\left[S_\tau\right] = \mathbf{E}\left[S_0\right] = 0$. Thus,

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

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

Similarly, we have $\mathbf{E}\left[|P_{t+1} - P_t||\mathcal{F}_t\right] = \mathbf{E}\left[|(\frac{p}{1-p})^{X_{t+1}} - 1||(\frac{p}{1-p})^{Z_t}||\mathcal{F}_t\right] \leq \frac{1}{1-p}(\frac{p}{1-p})^{max\{a,b\}}$ for every $0 \leq t < \tau$, so the third condition of OST holds, which implies that $\mathbf{E}\left[P_\tau\right] = \mathbf{E}\left[P_0\right] = 1$. Thus,

$$\begin{aligned}\mathbf{E}\left[P_\tau\right] &= \mathbf{E}\left[(\frac{p}{1-p})^{Z_\tau}\right] \\ &= E[Z_\tau] + (2p-1)\mathbf{E}\left[\tau\right] \\ &= (\frac{p}{1-p})^{-a}P_a + (\frac{p}{1-p})^{b}(1 - P_a) = 1.\end{aligned}$$

We get $P_a = \frac{(\frac{p}{1-p})^b - 1}{(\frac{p}{1-p})^b - (\frac{p}{1-p})^{-a}}$.

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

Problem 3 (Learning theory)

A simple mathematical model for Machine Learning is as follows:

- There is a finite set $\mathcal{X}$ of domain.
- Each data point $x \in \mathcal{X}$ is associated with a label $\ell(x) \in \{0,1\}$.
- The *training data* $S = \{(x_1, \ell(x_1)), (x_2, \ell(x_2)), \dots, (x_m, \ell(x_m))\}$ is a collection of pairs in $\mathcal{X} \times \{0,1\}$, usually known by the learner.
- There is a class $\mathcal{H}$ of *hypothesis* where each $h \in \mathcal{H}$ is a function from $\mathcal{X}$ to $\{0,1\}$.
- Let $h^* = \arg\min_{h \in \mathcal{H}} \sum_{x \in \mathcal{X}} \mathbf{1}[h(x) \neq \ell(x)]$ be the best hypothsis fitting the data. The goal of a learning algorithm is to find (or approximate) h^* provided the training data S.

Throughout this problem, we fix a domain $\mathcal{X}$ and a class of hypothesis $\mathcal{H}$.

Let $h : \mathcal{X} \to \{0,1\}$ be a function. Define the *average loss* $L(h)$ as

$$L(h) \triangleq \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathbf{1}[h(x) \neq \ell(x)].$$

That is, $L(h)$ is the ratio of data points that $h(\cdot)$ and $\ell(\cdot)$ do not match.

Given a training set $S = \{(x_1, \ell(x_1)), \dots, (x_m, \ell(x_m))\}$, we can also define the *average loss* $L_S(h)$ of h on S as

$$L_S(h) \triangleq \frac{1}{|S|} \sum_{(x,\ell(x)) \in S} \mathbf{1}[h(x) \neq \ell(x)].$$

Intuitively, a training set S is good if $L_S(h)$ is close to $L(h)$ for every $h \in \mathcal{H}$.

If $L_S(h)$ is close to $L(h)$, then a simple learning algorithm works well: choose the one performing best on S.

(a) Suppose the training set S satisfies

$$\sup_{h \in \mathcal{H}} |L(h) - L_S(h)| \leq \frac{\varepsilon}{2}.$$

Let $\widehat{h} = \arg\min_{h \in \mathcal{H}} \sum_{(x,\ell(x)) \in S} \mathbf{1}[h(x) \neq \ell(x)]$. Prove that

$$L(\widehat{h}) \leq L(h^*) + \varepsilon.$$

Proof:

$$L(\widehat{h}) \leq L_S(\widehat{h}) + \frac{\varepsilon}{2} \leq L_S(h^*) + \frac{\varepsilon}{2} \leq L(h^*) + \varepsilon.$$

Q.E.D.

We can define the notion of *representativeness* of S as

$$\texttt{Rep}(S) \triangleq \sup_{h \in \mathcal{H}} (L(h) - L_S(h)).$$

A natural question that arises is how to estimate $\texttt{Rep}(S)$ when only S is known. A heuristic approach would be to randomly split S into two sets, namely S_1 and S_2, which are then treated as the validation set and the training set respectively. Intuitively, a good S should have small

$$\sup_{h \in \mathcal{H}} (L_{S_1}(h) - L_{S_2}(h))$$

on average.

This motivates the so-called *Rademacher complexity* $R(S)$ for a training set $S = \{(x_1, \ell(x_1)), \cdots, (x_m, \ell(x_m))\}$:

$$R(S) \triangleq \frac{1}{m} \mathbf{E}_{\sigma \in \{1,-1\}^m} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i \cdot \mathbf{1}[h(x_i) \neq \ell(x_i)] \right].$$

An interesting fact in learning theory is the following relation between $\texttt{Rep}(S)$ and $R(S)$ when each data point S is sampled from $\mathcal{X}$ uniformly and independently at random (written as $S \sim \mathcal{X}^m$).

Theorem.

$$\mathbf{E}_{S \sim \mathcal{X}^m} [\texttt{Rep}(S)] \leq 2 \cdot \mathbf{E}_{S \sim \mathcal{X}^m} [R(S)].$$

(Optional) *Proof of Theorem:*

$$
\begin{aligned}
\mathbf{E}_{S\sim\mathcal{X}^m}\left[\mathtt{Rep}(S)\right] &= \mathbf{E}_{S\sim\mathcal{X}^m}\left[sup_{h\in\mathcal{H}}(L(h)-L_S(h))\right] \\
&= \mathbf{E}_{S\sim\mathcal{X}^m}\left[sup_{h\in\mathcal{H}}\mathbf{E}_{S'\sim\mathcal{X}^m}\left[(L_{S'}(h)-L_S(h))\right]\right] \\
&\leq \mathbf{E}_{S,S'\sim\mathcal{X}^m}\left[sup_{h\in\mathcal{H}}(L_{S'}(h)-L_S(h))\right] \\
&= \mathbf{E}_{S,S'\sim\mathcal{X}^m}\left[sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}(h(x_i')-h(x_i))\right] \\
&= \mathbf{E}_{S,S'\sim\mathcal{X}^m,\sigma\in\{-1,1\}^m}\left[sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h(x_i')-h(x_i))\right] \\
&\leq \mathbf{E}_{S',\sigma}\left[sup_h\frac{\sum_{i=1}^{m}\sigma_i h(x_i')}{m}\right]+\mathbf{E}_{S,\sigma}\left[sup_h\frac{\sum_{i=1}^{m}\sigma_i h(x_i)}{m}\right] \\
&= 2\cdot\mathbf{E}_{S\sim\mathcal{X}^m}\left[R(S)\right].
\end{aligned}
$$

Q.E.D.

(b) Assume $S\sim\mathcal{X}^m$. Prove that for any $\delta\in(0,1)$, with probability at least $1-\delta$, for all $h\in\mathcal{H}$, it holds that

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

Proof:

By Theorem, we want to prove that, for any $\delta\in(0,1)$, with probability at least $1-\delta$, for all $h\in\mathcal{H}$, it holds that

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

And for $\mathtt{Rep}(S)\triangleq\sup_{h\in\mathcal{H}}\left(L(h)-L_S(h)\right)$, the inequality below is stronger:

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

This inequality can be transformed to

$$\mathbf{Pr}\left[\mathtt{Rep}(S)-\mathbf{E}_{S\sim\mathcal{X}^m}\left[\mathtt{Rep}(S)\right]\geq\sqrt{\frac{1}{2m}\log\frac{2}{\delta}}\right]\leq\delta,$$

which is much more close to the form of McDiarmid's Inequality. It's clear that $\texttt{Rep}(S)$ is a function on m variables, and satisfies $\frac{1}{m} - \mathit{Lipschitz}$ condition because $\forall i \in [m], \forall x_1, \cdots, x_m, \forall y_i$, it holds that

$$\left|\texttt{Rep}(\overline{x_{1,i-1}}, x_i, \overline{x_{i+1,m}}) - \texttt{Rep}(\overline{x_{1,i-1}}, y_i, \overline{x_{i+1,m}})\right| \leq \frac{1}{m}.$$

Then by McDiarmid's Inequality, we have,

$$\mathbf{Pr}\left[\left|\texttt{Rep}(S) - \mathbf{E}_{S\sim\mathcal{X}^m}\left[\texttt{Rep}(S)\right]\right| \geq \sqrt{\frac{1}{2m}\log\frac{2}{\delta}}\right] \leq 2e^{-log\frac{2}{\delta}} = \delta.$$

Q.E.D.

In fact, Prof. Zhang have introduced only one of the form of McDiarmid's Inequality in class. However, there is another form which can lead to a better result for this question:

Theorem (Another Form of McDiarmid's Inequality)

Let f be a function on n variables satisfying $c - \mathit{Lipschitz}$ condition and $X_1, \cdots, X_n$ be n independent variables. Then we have

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

Using this form of McDiarmid's Inequality, we can prove that for any $\delta \in (0, 1)$, with probability at least $1-\delta$, for all $h \in \mathcal{H}$, it holds that

$$\texttt{Reg}(S) \leq \mathbf{E}_{S\sim\mathcal{X}^m}\left[\texttt{Reg}(S)\right] + \sqrt{\frac{1}{2m}\log\frac{1}{\delta}}.$$

(c) Assume $S \sim \mathcal{X}^m$. Let $\widehat{h}$ be the one defined in (a). Prove that for any $\delta \in (0, 1)$, with probablity at least $1 - \delta$, it holds that

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

Proof:

By the conlusion in (b),

$$L(\widehat{h}) - L_S(\widehat{h}) \leq 2 \cdot \mathbf{E}_{S\sim\mathcal{X}^m}[R(S)] + \sqrt{\frac{1}{2m}\log\frac{8}{\delta}} \quad (w.p.\ 1-\frac{\delta}{4}) \cdots (1)$$

Similarly, by McDiarmid's Inequality, we have

$$\begin{aligned} L_S(h^*) - L(h^*) &\leq \mathbf{E}_{S\sim\mathcal{X}^m}[L_S(h^*) - L(h^*)] + \sqrt{\frac{1}{2m}\log\frac{8}{\delta}} \\ &= \sqrt{\frac{1}{2m}\log\frac{8}{\delta}} \quad (w.p.\ 1-\frac{\delta}{4}) \cdots (2) \end{aligned}$$

and

$$\mathbf{E}_{S\sim\mathcal{X}^m}[R(S)] \leq R(S) + \sqrt{\frac{1}{2m}\log\frac{8}{\delta}} \quad (w.p.\ 1-\frac{\delta}{4}) \cdots (3)$$

It's clear that

$$L_S(\widehat{h}) - L_S(h^*) \leq 0 \cdots (4)$$

Add these up $((1)+(2)+2\times(3)+(4))$, we get that for any $\delta \in (0,1)$, with probablity at least $1-\delta$, it holds that

$$L(\widehat{h}) \leq L(h^*) + 2\cdot R(S) + 4\sqrt{\frac{1}{2m}\log\frac{8}{\delta}}.$$

Q.E.D.

Hint:

$$L(\widehat{h}) - L(h^*) = \left(L(\widehat{h}) - L_S(\widehat{h})\right) + \left(L_S(\widehat{h}) - L_S(h^*)\right) + \left(L_S(h^*) - L(h^*)\right)$$