[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 $\{\mathcal{F}_t\}_{t\geq 1}$.

:::info
(a) Prove that for any $n \in \mathbb{N}$,

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

:::

:::info
(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.
:::

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

Prove that for every $\lambda > 0$,

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

:::

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 $\mathbf{Pr}\left[X_i = -1\right] = p \in (0, 1)$.
:::info
(a) Define $S_t = \sum_{i=1}^{t}(X_i + 2p - 1)$. Show that $\{X_t\}_{t\geq 0}$ is a martingale.
:::

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

:::info
(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 $\mathbf{E}\left[\tau\right]$.
:::

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

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

Intuitively, a training set S is good if $L_S(h)$ is close to $L(h)$ for every $h \in \mathcal{H}$. As a result, we can define the notion of *representativeness* of S as

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

If $\mathtt{Rep}(S)$ is small, then a simple learning algorithm works well: choose the one performing best on S.

:::info

(a) Let $\widehat{h} = \arg\min_{h \in \mathcal{H}} \sum_{(x, \ell(x)) \in S} \mathbf{1}[h(x) \neq \ell(x)]$. Prove that if $\mathtt{Rep}(S) \leq \frac{\varepsilon}{2}$, then

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

:::

A natural question that arises is how to estimate $\mathtt{Rep}(S)$ 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

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

on average.

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

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

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

:::success
Theorem.

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

:::

::: spoiler Click if you are interested in a proof of this
别急
:::

In the following, we assume the theorem.

:::info
(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

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

:::

:::info
(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

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

:::