$
\def*#1{\mathbf{#1}} \def+#1{\mathcal{#1}}
\def-#1{\mathrm{#1}}\def^#1{\mathbb{#1}}\def!#1{\mathtt{#1}}
\newcommand{\norm}[1]{\left\Vert#1\right\Vert}
\newcommand{\abs}[1]{\left\vert#1\right\vert}
\newcommand{\set}[1]{\left{#1\right}}
\newcommand{\tuple}[1]{\left(#1\right)} \newcommand{\eps}{\varepsilon}
\newcommand{\inner}[2]{\langle #1,#2\rangle} \newcommand{\tp}{\tuple}
\renewcommand{\mid}{;\middle\vert;} \newcommand{\cmid}{,:,}
\newcommand{\numP}{#\mathbf{P}} \renewcommand{\P}{\mathbf{P}}
\newcommand{\defeq}{\triangleq} \renewcommand{\d}{,-d}
\newcommand{\ol}{\overline}
\newcommand{\Pr}[2][]{\mathbf{Pr}_{#1}\left[#2\right]}
\newcommand{\E}[2][]{\mathbf{E}_{#1}\left[#2\right]}
\newcommand{\Var}[2][]{\mathbf{Var}_{#1}\left[#2\right]}
\renewcommand{\emptyset}{\varnothing}
$

[Solution of Homework 3]

Problem 1 (A maximal inequality)

Let $\set{Z_t}_{t\ge 0}$ be a martingale with respect to a filtration $\set{+F_t}_{t\ge 1}$.

:::info
(a) Prove that for any $n\in^N$,
$
\sum_{k=1}^n\E{(Z_k-Z_{k-1})^2} = \E{Z_n^2} - \E{Z_0^2}.
$
:::
Proof. Note that for $k\geq 1$,
\begin{align}
\E{(Z_k-Z_{k-1})^2}&=\E{\E{(Z_k-Z_{k-1})^2\mid +F_{k-1}}}\\
&=\E{\E{Z_k^2+Z_{k-1}^2-2Z_kZ_{k-1}\mid +F_{k-1}}}\\
&=\E{\E{Z_k^2+Z_{k-1}^2-2Z_{k-1}^2\mid +F_{k-1}}}\\
&=\E{Z_k^2-Z_{k-1}^2}.
\end{align}
Therefore,
$
\sum_{k=1}^n\E{(Z_k-Z_{k-1})^2} =\sum_{k=1}^n\E{Z_k^2-Z_{k-1}^2}=\sum_{k=1}^n \E{Z_k^2}-\E{Z_{k-1}^2}=\E{Z_n^2} - \E{Z_0^2}.
$

:::info
(b) Let τ be a stopping time for the martingale $\set{Z_t}_{t\ge 0}$. Define another sequence $\set{Z'_t}_{t\ge 0}$ as
$
Z'_t =
\begin{cases}
Z_t & \mbox{ if }t<\tau;\\

Z\tau & \mbox{ if }t\ge \tau.
\end{cases}
$
Prove that $\set{Z_t'}{t\ge 0}$ is also a martingale.
:::
Proof.
\begin{align}
\E{Z_t'\mid +F_{t-1}}&=\E{Z_t'\cdot *1[\tau<t]+ Z_t'\cdot *1[\tau\geq t]\mid +F_{t-1}}\
&=\E{Z_{\tau}\cdot *1[\tau<t]\mid +F_{t-1}}+ \E{Z_t\cdot *1[\tau\geq t]\mid +F_{t-1}}\
&=\E{Z_{\tau}\mid +F_{t-1}}\cdot *1[\tau<t]+ \E{Z_t\cdot \mid +F_{t-1}}*1[\tau\geq t]\
&=Z_{\tau}\cdot *1[\tau\leq t-1] + Z_{t-1}\cdot*1[\tau> t-1]=Z_{t-1}'
\end{align}

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

Prove that for every $\lambda>0$,
$
\Pr{\max_{1\le k\le n} \abs{S_k}\ge \lambda} \le \frac{1}{\lambda^2}\sum_{k=1}^n
\E{X_k^2}.
$
:::
Proof.
Let $S_0=0$. Since $\E{S_t\mid S_0,S_1,\dots, S_{t-1}}=\E{S_{t-1}+X_t\mid S_0,S_1,\dots,
S_{t-1}} = S_{t-1},$ $\set{S_t}{t\geq 0}$ is a martingale. Let $\tau\triangleq \min\tp{n,
\min{t\leq n}\set{t\mid \abs{S_t}\geq \lambda}}$. By definition, τ is a stopping time for
$\set{S_t}{t\geq 0}$. We define another sequence $\set{S_t'}{t\geq 0}$ as
$
S't =
\begin{cases}
S_t & \mbox{ if }t<\tau;\
S\tau & \mbox{ if }t\ge \tau.
\end{cases}
$
Then by the Chebyshev's inequality, we have
\begin{align}
\Pr{\max_{1\le k\le n} \abs{S_k}\ge \lambda}=\Pr{\abs{S_n'}\geq \lambda}\leq
\frac{\Var{S_n'}}{\lambda^2}=\frac{\E{\tp{S_n'-\E{S_n'}}^2}}{\lambda^2}.
\end{align}
From (b) we know that $\set{S_i'}{i\geq 0}$ is a martingale. Therefore, $\E{S_n'}=S_0'=0$.
From (a), we have
$
\E{(S_n')^2}=\E{(S_0')^2}+\sum{k=1}^n\E{(S_k'-S'{k-1})^2}.
$
Note that for each $k\geq 1$,
$
\E{(S_k'-S'{k-1})^2}=\E{(S_k-S_{k-1})^2\cdot *1[\tau\geq k] + 0\cdot *1[\tau<k]}\leq
\E{(S_k-S_{k-1})^2}=\E{X_k^2}.
$
Therefore we have
$

\Pr{\max_{1\le k\le n} \abs{S_k}\ge \lambda}\leq \frac{\E{\tp{S_n'-\E{S_n'}}^2}}{\lambda^2}\leq \frac{1}{\lambda^2}\sum_{k=1}^n \E{X_k^2}.
$

Problem 2 (Biased random walk)

We study the biased random walk in this exercise. Let $Z_t=\sum_{i=1}^tX_i$ where each $X_i\in\set{-1,1}$ is independent, and satisfies $\Pr{X_i=-1}=p\in(0,1)$.
:::info
(a) Define $S_t=\sum_{i=1}^t(X_i+2p-1)$. Show that $\set{S_t}*{t\ge 0}$ is a martingale.*
:::
Proof.
\begin{align}*
\E{S_t\mid X_1,X_2,\dots,X{t-1}}&=\E{S_{t-1}+X_t+2p-1\mid X_1,X_2,\dots,X_{t-1}}\
&=S_{t-1} + 2p-1 + \E{X_t\mid X_1,X_2,\dots,X_{t-1}}\
&=S_{t-1} + 2p-1 + (-p)+1-p=S_{t-1}.
\end{align*}
Therefore $\set{S_t}*{t\ge 0}$ is a martingale with regard to $\set{X_t}*{t\ge 0}$.

:::info
(b) Define $P_t=\tp{\frac{p}{1-p}}^{Z_t}$. Show that $\set{P_t}*{t\ge 0}$ is a martingale.*
:::
Proof.
\begin{align}*
\E{P_t\mid X_1,X_2,\dots,X{t-1}}&=\E{\tp{\frac{p}{1-p}}^{X_t}\cdot \tp{\frac{p}{1-p}}^{Z_{t-1}}\mid X_1,X_2,\dots,X_{t-1}}\
&=\tp{\frac{p}{1-p}}^{Z_{t-1}} \cdot \tp{p\tp{\frac{p}{1-p}}^{-1} + (1-p)\tp{\frac{p}{1-p}}}\
&=P_{t-1}
\end{align*}
Therefore $\set{P_t}*{t\ge 0}$ is a martingale with regard to $\set{X_t}*{t\ge 0}$.

:::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 $\E{\tau}$.
:::
Solution.
Note that in a time period of $T=a+b$, if the man walks towards the same direction, he must have stopped. This happends w.p. $\min \set{{p}^{a+b}, (1-p)^{a+b}}$. W.l.o.g., assume $p< \frac{1}{2}$. Therefore,
$
\Pr{\tau\geq k\cdot T}\leq \tp{1-p^{a+b}}^k.
$
This indicates that $\E{\tau}< \infty$. We also have that $\E{\abs{S_t-S_{t-1}}\mid +F_{t-1}}\leq 2-2p$. By the ost, we have $\E{S_{\tau}}=\E{S_0}=0$. Let $P_a=\Pr{Z_{\tau}=-a}$ and $P_b=\Pr{Z_{\tau}=b}=1-P_a$. Then we have $-aP_a+bP_b+(2p-1)\E{\tau}=0$. Sequentially, $P_a=\frac{b+(2p-1)\E{\tau}}{a+b}$ and $P_b=\frac{a-(2p-1)\E{\tau}}{a+b}$

Similarly, $\E{P_{\tau}}=\E{P_0}=1$. That is, $P_a\tp{\frac{p}{1-p}}^{-a}+P_b\tp{\frac{p}{1-p}}^b=1$. Therefore we have
$
\tp{b+(2p-1)\E{\tau}}\tp{\frac{p}{1-p}}^{-a}+ \tp{a-(2p-1)\E{\tau}}\tp{\frac{p}{1-

p}}^b=a+b.
$
This yields $\E{\tau}=\frac{a\tp{1-\tp{\frac{p}{1-p}}^b}+b\tp{1-\tp{\frac{p}{1-p}}^{-a}}}{(2p-1)\tp{\tp{\frac{p}{1-p}}^{-a}-\tp{\frac{p}{1-p}}^b}}.$

Problem 3 (Learning theory)

A simple mathematical model for Machine Learning is as follows:

- There is a finite set $+X$ of domain.
- Each data point $x\in +X$ is associated with a label $\ell(x)\in\set{0,1}$.
- The *training data* $S=\set{(x_1,\ell(x_1)),(x_2,\ell(x_2)),\dots,(x_m,\ell(x_m))}$ is a collection of pairs in $+X\times\set{0,1}$, usually known by the learner.
- There is a class $+H$ of *hypothesis* where each $h\in +H$ is a function from $+X$ to $\set{0,1}$.
- Let $h^* = \argmin_{h\in+H}\sum_{x\in+X}*1[h(x)\ne\ell(x)]$ 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 $+X$ and a class of hypothesis $+H$.

Let $h:+X\to\set{0,1}$ be a function. Define the *average loss* $L(h)$ as
$
L(h) \defeq \frac{1}{\abs{+X}} \sum_{x\in +X}*{1}[h(x)\ne \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 $S=\set{(x_1,\ell(x_1)),\dots,(x_m,\ell(x_m))}$, we can also define the *average loss* $L_S(h)$ of h on S as
$
L_S(h) \defeq \frac{1}{\abs{S}} \sum_{(x,\ell(x))\in S}*{1}[h(x)\ne \ell(x)].
$

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

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

:::info

(a) Suppose the training set S satisfies
$
\sup_{h\in+H} \abs{L(h) - L_S(h)}\le \frac{\eps}{2}.
$

Let $\widehat h = \arg\min_{h\in+H}\sum_{(x,\ell(x))\in S} *1[h(x)\ne \ell(x)]$. Prove that
$
L(\widehat h) \le L(h^*)+\eps.*
$
:::
Proof.
Note that
$
L(\widehat h)\leq L_S(\widehat h) + \abs{L(\widehat h) - L_S(\widehat h)}\leq L_S(\widehat h)

+\frac{\eps}{2}
$
and similarly
$
L_S(h^)\leq L(h^*)+\frac{\eps}{2}.
$

Since $\widehat h = \arg\min_{h\in+H}\sum_{(x,\ell(x))\in S} *1[h(x)\ne \ell(x)]$, we have $L_S(\widehat h)\leq L_S(h^)$. Therefore, we have
$
L(\widehat h)\leq L_S(\widehat h) +\frac{\eps}{2}\leq L_S(h^) +\frac{\eps}{2}\leq L(h^*)+\eps.
$

We can define the notion of *representativeness* of S as
$
!{Rep}(S)\defeq \sup_{h\in+H} \tp{L(h) - L_S(h)}.
$

A natural question that arises is how to estimate $!{Rep}(S)$ 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+H}\tp{L_{S_1}(h)-L_{S_2}(h)}
$
on average.

This motivates the so-called *Rademacher complexity* $R(S)$ for a training set $S=\set{(x_1,\ell(x_1)),\dots,(x_m,\ell(x_m))}$:
$
R(S)\defeq \frac{1}{m}\E[\sigma\in\set{1,-1}^m]
{\sup_{h\in+H}\sum_{i=1}^m\sigma_i\cdot*1[h(x_i)\ne \ell(x_i)]}.
$

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

:::success
Theorem.
$
\E[S\sim +X^m]{!{Rep}(S)} \le 2\cdot \E[S\sim +X^m]{R(S)}.
$
:::

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

:::

In the following, we assume the theorem.

:::info
(b) Assume $S\sim +X^m$. Prove that for any $\delta\in (0,1)$, with probability at least $1-\delta$, for all $h\in +H$, it holds that
$
L(h)-L_S(h) \le 2\cdot \E[S\sim +X^m]{R(S)} + \sqrt{\frac{1}{2m}\log\frac{2}{\delta}}.
$
:::
Proof.
Let $\tp{X_1,\ell\tp{X_1}}$, $\tp{X_2,\ell\tp{X_2}}$, $\dots$, $\tp{X_m,\ell\tp{X_m}}$ be the m samples that form S. $!{Rep}$ is a function that maps these m samples to a real number. It is easy to verify that $!{Rep}$ is $\frac{1}{m}$-Lipschitz.
From the McDiarmid's inequality, we have that
$
\Pr{!{Rep}(S)-\E[S\sim +X^m]{!{Rep}(S)}\geq \sqrt{\frac{1}{2m}\log\frac{2}{\delta}}}\leq 2\exp\set{-\frac{2\frac{1}{2m}\log\frac{2}{\delta}}{m\cdot\frac{1}{m^2}}}=\delta.
$
Therefore, w.p. at least $1-\delta$, for all h
\begin{align}
L(h)-L_S(h) &\leq \E[S\sim +X^m]{!{Rep}(S)} + \sqrt{\frac{1}{2m}\log\frac{2}{\delta}}\
&\leq 2\cdot \E[S\sim +X^m]{R(S)} + \sqrt{\frac{1}{2m}\log\frac{2}{\delta}}.
\end{align}

:::info
(c) Assume $S\sim +X^m$. 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) \le L(h^) + 2\cdot R(S) + 5\sqrt{\frac{1}{2m}\log\frac{8}{\delta}}.
$
:::
Proof.
Note that
$
L(\widehat h) - L(h^) =L(\widehat h)- L_S(\widehat h)+L_S(\widehat h) - L(h^) \leq L(\widehat h)- L_S(\widehat h)+L_S(h^) - L(h^*).
$
We first bound the term $L(\widehat h)- L_S(\widehat h)$. From (b), we have that w.p. at least $1-\frac{\delta}{4}$,
$
L(\widehat h)- L_S(\widehat h)\leq 2\cdot \E[S\sim +X^m]{R(S)} + \sqrt{\frac{1}{2m}\log\frac{8}{\delta}}.
$
Note that $R(S)$ is also $\frac{1}{m}$-Lipshitz since if we change one item in S, the value of $\sum_{i=1}^m\sigma_i\cdot*1[h(x_i)\ne \ell(x_i)]$ changes at most 1 for any h and σ. Therefore, by the McDiarmid's inequality, we have that
$
\Pr{R(S)-\E[S\sim +X^m]{R(S)}\leq -\sqrt{\frac{1}{2m}\log\frac{8}{\delta}}}\leq \frac{\delta}{4}.
$

Then we bound $L_S(h^) - L(h^)$. By the Hoeffding's inequality,
$
\Pr{L_S(h^) - L(h^)\geq \sqrt{\frac{1}{2m}\log\frac{8}{\delta}}}\leq 2\exp\set{-

\frac{\frac{2m^2}{2m}\log\frac{8}{\delta}}{m}}=\frac{\delta}{4}.
$
Combining these together, by the union bound, w.p. at least δ, we have
$
L(\widehat h) \le L(h^*) + 2\cdot R(S) + 5\sqrt{\frac{1}{2m}\log\frac{8}{\delta}}.
$

