Homework 3

Problem 1 Similarly as the Hoeffding's inequality, prove the following McDiarmid's inequality.

Theorem 1 (McDiarmid's Inequality) *Consider i.i.d. random variables* $X_1, \ldots, X_N \in \mathcal{X}$ *and a mapping* $\phi : \mathcal{X}^N \to \mathbb{R}$*. If* $\forall x_1, \ldots, x_N, x_i' \in \mathcal{X}$*,*

$$|\phi(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_N) - \phi(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_N)| \leq c,$$

then

$$\mathbb{P}\left(\phi(X_1, \ldots, X_N) - \mathbb{E}[\phi(X_1, \ldots, X_N)] \geq t\right) \leq \exp\left(\frac{-2t^2}{c^2 \cdot N}\right).$$

Like what we did in class for the Hoeffding's inequality, prove this theorem in the following two steps.

1. First prove the Hoeffding's Lemma that all bounded random variables are sub-Gaussian random variables. That is: if $X \in [a, b]$ with probability 1, where $R = b - a$, then

$$\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] \leq e^{\lambda^2 R^2/8}, \quad \forall \lambda \in \mathbb{R}.$$

2. Then define the following random variable:

$$Y_j = \mathbb{E}[\phi(X_1, \ldots, X_N)|X_1, \ldots, X_j] - \mathbb{E}[\phi(X_1, \ldots, X_N)|X_1, \ldots, X_{j-1}].$$

   When $j = 1$, define $\mathbb{E}[\phi(X_1, \ldots, X_N)|X_1, \ldots, X_{j-1}] = \mathbb{E}[\phi(X_1, \ldots, X_N)]$. Note that $\sum_{j=1}^N Y_j = \phi(X_1, \ldots, X_N) - \mathbb{E}[\phi(X_1, \ldots, X_N)]$. Also note that $\mathbb{E}[Y_j|X_1, \ldots, X_{j-1}] = 0 = \mathbb{E}[Y_i]$ by the tower rule of conditional expectations.

   Prove that Y_j is a bounded random variable for all $j = 1, \ldots, N$ and compute the bound. Then use the above Hoeffding's Lemma on Y_j conditional on $X_1, \ldots, X_{j-1}$, in conjunction with the property of conditional expectations to modify our proof for the Hoeffding's inequality and prove the McDiarmid's inequality.

Problem 2 Define the empirical Rademacher complexity of function class $\mathcal{F}$ over dataset S with $|S| = N$ to be

$$\hat{R}_N(loss(\mathcal{F})) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{z_i \in S} \sigma_i \cdot loss(f, z_i) \right],$$

for Rademacher random variables $\sigma = \{\sigma_1, \ldots, \sigma_N\} \in \{-1, +1\}^N$. Assume that

$$\sup_{f \in \mathcal{F}} |loss(f, z) - loss(f, z')| \leq b, \ \forall z, z' \sim D.$$

Use the McDiarmid's inequality (twice) to prove that for any function $f \in \mathcal{F}$, with at least $(1 - \delta)$ probability,

$$\mathcal{L}(f, D) \leq \mathcal{L}(f, S) + 2\hat{R}_N(loss(\mathcal{F})) + 3b\sqrt{\frac{\log(2/\delta)}{N}},$$

where the population risk $\mathcal{L}(f, D) = \mathbb{E}_{z \sim D}[loss(f, D)]$ and the empirical risk $\mathcal{L}(f, S) = \frac{1}{N} \sum_{z_i \in S} loss(f, z_i)$.

Machine Learning
HW - 3

Problem -1

1. We need to prove Hoeffding's lemma

1. Hoeffding's Lemma $\Rightarrow$

$$E\left[e^{\lambda(X - E[X])}\right] \leq e^{\lambda^2 R^2/8}$$

it can also be written as $\Rightarrow$

$$E\left[e^{sX}\right] \leq e^{s^2 (b-a)^2/8}$$

In particular, $X \sim \text{subG}\left(\frac{(b-a)^2}{4}\right)$

Proof :-

Let X be a random variable, such that $P(a \le X \le b) = 1$ for some $-\infty < a \le b < \infty$

$$\Rightarrow \operatorname{Var}[X] \le E[(X-c)^2] \quad - (1) \quad \forall c \in \mathbb{R}$$

[equality holds for $c = E[X]$]

If c is the mid point of a & b,

$$\Rightarrow |x - c| \le \frac{b-a}{2} \quad (2)$$

Using (1) and (2)

$$\Rightarrow \operatorname{Var}[X] \le E[(X-c)^2] \le \frac{(b-a)^2}{4}$$

$$\Rightarrow \operatorname{Var}[X] \le \frac{(b-a)^2}{4}$$

We want to bound $E[e^{s(X-E(X))}]$, which without loss of generality, we may assume $EX = 0$. Thus, we need to bound $E[e^{sX}]$

Let's instead consider logarithmic Moment Generating function :

$$\varphi(s) \triangleq \log E[e^{sX}]$$

$$\varphi'(s) = \frac{E[Xe^{sX}]}{E[e^{sX}]}, \quad \varphi''(s) = \frac{E[X^2 e^{sX}]}{E[e^{sX}]} - \left[\frac{E[Xe^{sX}]}{E[e^{sX}]}\right]^2$$

Now consider another random variable V, such that

$$E[f(U)] = \frac{E[f(x)e^{sx}]}{E[e^{sx}]}, \quad f: R \to R$$

To validate the legitimacy, we plug-in indicator function of any even A:

$$P(U \in A) = E[I_{\{U \in A\}}] = \frac{E[I_{\{x \in A\}} e^{sx}]}{E[e^{sx}]}$$

[applying exponential change of measure]

if $F[U] = U$

$$\Rightarrow E[U] = \frac{E[Xe^{sx}]}{E[e^{sx}]}$$

if $F(U) = U^2$

$$\Rightarrow E[U^2] = \frac{E[U^2 e^{sx}]}{E[e^{sx}]}$$

$$\Rightarrow \quad Var[U] = [E[U^2] - E[U]^2]$$

$$= \frac{E[x^2 e^{sx}]}{E[e^{sx}]} - \left[\frac{E[xe^{sx}]}{E[e^{sx}]}\right]^2$$

$$= \varphi''(s)$$

$$\Rightarrow \quad \varphi''(s) = Var[U] \le \frac{(b-a)^2}{4}$$

$$\Rightarrow \quad \varphi(s) = \int_0^s \int_0^t \varphi''(v)\, dv\, dt$$

$$\le \int_0^s \int_0^t \frac{(b-a)^2}{4}\, dv\, dt$$

$$\le \int_0^s \frac{(b-a)^2}{4} [t]\, dt$$

$$\le \frac{(b-a)^2 t^2}{8}\Big|_0^s \le \frac{s^2(b-a)^2}{8}$$

$$\Rightarrow \log E\left(e^{sx}\right) \leq \frac{s^2(b-a)^2}{8}$$

$$E\left[e^{sx}\right] \leq e^{\frac{s^2(b-a)^2}{8}}$$

$\because$ $E\left[e^{sx}\right]$ is bounded by $e^{\frac{s^2(b-a)^2}{8}}$

$$\Rightarrow E\left[e^{s(x-E[x])}\right] \leq e^{s^2R^2/8}$$

Hence Proved

(2) Given:

$$Y_j = E\left[\phi(X_1 \cdots X_N) \,[X_1 \cdots X_j]\right] - E\left[\phi(X_1 \cdots X_N)\right][X_1 \cdots X_{j-1}]$$

Proof:-

Let $U_j = \sup_u E(g \mid X_{1:j-1}, u) - E(g \mid X_{1:j-1})$

$L_j = \inf_l E(g \mid X_{1:j-1}, l) - E(g \mid X_{1:j-1})$

Now,

$$U_j - L_j \leq \sup_{l,u} E(g \mid X_{1:j-1}, u) - E(g \mid X_{1:j-1}, l)$$

$$\leq \sup_{l,u}\left(\int_{x_{j+1:n}} \left[g(X_{1:j-1}, u, x_{j+1:n}) - g(x_{1:j-1}, l, x_{j+1:n})\right] \cdot \prod_{i=j+1}^{n} f_{X_i}(x_i)\, dx_{j+1:n}\right)$$

$\because$ sup is convex, so from Jensen's inequality :-

$$\leq \int_{x_{j+1:n}} \sup_{l,u} [g(x_{1:j-1}, u, x_{j+1:n}) - g(x_{1:j-1}, l, x_{j+1:n})] \cdot \prod_{j=i+1}^{n} f_{X_i}(x_i)\, dx_{j+1:n})$$

$$\leq \int_{x_{j+1:n}} c_j \prod_{i=j+1}^{n} f(x_i)(x_i)\, dx_{j+1:n} \left[\because |g(x_1 \cdots x_N) - \phi(x_1 \cdots x_i' \cdots x_N)| \leq c \right]$$

$$= c_j$$

$\Rightarrow L_j \leq Y_j \leq U_j \Rightarrow Y_j$ is bounded

Applying Hoeffding's lemma, we get:

$$E[e^{tY_j} \mid X_1 \cdots X_{j-1}] \leq e^{t^2 c^2/8}$$

$\Rightarrow$ Bound

$$E(e^{t\sum_{j=1}^{n} Y_j}) = E(e^{t\sum_{j=1}^{n-1} Y_j} E(e^{tV_n} \mid X_1 \cdots X_{n-1}))$$

$$\leq E(e^{t\sum_{j=1}^{n-1} Y_j} \exp(t^2 c_n^2/8))$$

$$\leq \exp\left(\frac{1}{8} \sum_{j=1}^{n} t^2 c_j^2\right) \quad \text{— (a)}$$

$$P(\phi(X_1 \cdots X_n) - E(\phi(X_1 \cdots X_n)) > t)$$

$$= P\left(\sum_{j=1}^{n} Y_j \geq t\right)$$

$$= P\left(e^{t\sum_{j=1}^{n} Y_j} \geq e^{st}\right)$$

$$\leq \exp(-st)\, E\left(e^{s\sum_{j=1}^{n} Y_j}\right)$$

$$\leq \exp(-st) \exp\left(\frac{1}{8}\sum_{j=1}^{n} s^2 c_j^2\right)$$

$$= \exp\left(-st + \frac{1}{8}\sum_{j=1}^{n} s^2 c_j^2\right) = u$$

We need to minimize the expression : U, w.r.t S.

$$S = 4t \Big/ \sum_{j=1}^{n} c_j^2$$

Substituting the value we get :

$$= \exp\left(\frac{-2t^2}{NC^2}\right)$$

$$\Rightarrow P\left[\phi(X_1 \ldots X_n) - E[\phi(X_1 \ldots X_n)] \geq t\right]$$

$$\leq \exp\left(\frac{-2t^2}{N \cdot C^2}\right)$$

Hence Proved

Problem 2

Proof :- for a fixed function f, from the definition of supremum, we get :-

$$E_D[f(z)] \le \hat{E}_S[f(z)] + \sup_{h \in F} \left(E_D[h(z)] - \hat{E}_S[h(z)] \right) \quad \text{— (a)}$$

Let $\varphi(S) = \sup_{h \in F} \left(E_D[h(z)] - \hat{E}_S[h(z)] \right)$

We have been given:

$$\sup_{f \in F} |\mathrm{loss}(f, z) - \mathrm{loss}(f, z')| \le b \quad \forall z, z' \sim D$$

$\Rightarrow$ b is maximum change

$\Rightarrow$ change is occuring within an emperical avg, so its effect on the value of $\varphi(S)$ is scaled down by a factor of $1/N$, i.e b/N.

• So, we can apply McDiarmid's Inequiality, we get:

$$P[\varphi(S) - E[\varphi(S)] \ge t] \le 2e^{-2t^2 / \sum_{i=1}^{N} b^2/N^2} = 2e^{-t^2 N / b^2}$$

Now, setting the above probability to less than δ and solving for t, we get.

$$\Rightarrow \sqrt{\frac{b^2 \log(2/\delta)}{N}} \le t$$

$$\Rightarrow \varphi(S) \le E(\varphi(S)) + b\sqrt{\frac{\log(2/\delta)}{N}} \quad \text{— (b)}$$

Now bound of $\varphi(S)$

Let $\tilde{S} = \{\tilde{z}_1 \dots \tilde{z}_N\}$ be ghost sample, independently drawn identically to S.

Since, $E_{\tilde{S}}[\hat{E}_{\tilde{S}}[h(z)] | S] = E_D[h(z)]$ and

$E_{\tilde{S}}[\hat{E}_S[h(z)] | S] = E_S[h(z)]$, we can rewrite-

$$E_S\left[\sup_{h\in F}\left(E_D[h(z)] - \hat{E}_S[h(z)]\right)\right]$$

$$= E_S\left[\sup_{h\in F} E_{\tilde{S}}\left[\hat{E}_{\tilde{S}}[h(z)] - \hat{E}_S[h(z)] \,\middle|\, S\right]\right]$$

$$= E_S\left[\sup_{h\in F} E_{\tilde{S}}\left[\frac{1}{N}\sum_{i=1}^{N} h(\tilde{z}_i) - \frac{1}{N}\sum_{i=1}^{N} h(z_i) \,\middle|\, S\right]\right]$$

$$= E_S\left[\sup_{h\in F} E_{\tilde{S}}\left[\frac{1}{N}\sum_{i=1}^{N}\left(h(\tilde{z}_i) - h(z_i)\right) \,\middle|\, S\right]\right]$$

Since sup is a convex function we can apply Jensen's inequality to move the sup inside the expectation:

$$= E_S\left[\sup_{h\in F} E_{\tilde{S}}\left[\frac{1}{N}\sum_{i=1}^{N}\left(h(\tilde{z}_i) - h(z_i)\right) \middle| S\right]\right]$$

$$\le E_{S,\tilde{S}}\left[\sup \frac{1}{N}\sum_{i=1}^{N}\left(h(\tilde{z}_i) - h(z_i)\right)\right]$$

- Multiplying each term in the summation by a Rademacher variable σ_i will not change the expectation since $E[\sigma_i] = 0$ and also negating a Rademacher does not change its distribution.

Combining these facts:

$$\rightarrow E_{S,\tilde{S}}\left[\sup \frac{1}{N}\sum_{i=1}^{N}\left(h(\tilde{z}_i) - h(z_i)\right)\right] = E_{\sigma S\tilde{S}}\left[\sup_{h\in F}\frac{1}{N}\sum_{i=1}\sigma_i\left(h(\hat{z}_i) - h(z_i)\right)\right]$$

$$\le E_{\sigma S\tilde{S}}\left[\sup_{h\in F}\left(\frac{1}{N}\sum_{i=1}^{N}\sigma_i h(z_i)\right) + \sup_{h\in F}\left(\frac{1}{N}\sum_{i=1}^{N}-\sigma_i h(\tilde{z}_i)\right)\right]$$

$$= E_{\sigma S}\left[\sup_{h\in F}\frac{1}{N}\sum_{i=1}^{N}\sigma_i h(z_i)\right] + E_{\sigma,\hat{S}}\left[\sup_{h\in F}\frac{1}{N}\sum_{i=1}^{N}\sigma_i h(\tilde{z}_i)\right]$$

$$= 2R_N(F) \quad - \quad (C)$$

Let $\alpha(z_1, \dots, z_N) = \mathrm{Rad}_n(F, 2^N)$, such that $\alpha(z_1, \dots, z_N)$ change by at most b/N, if we change one observation.

Applying McDiarmid's inequality :

$$\Rightarrow |\hat{R}_N(F) - R_N(F)| \le \sqrt{\frac{b^2}{N}\log(2/\delta)}$$

$$\Rightarrow \sqrt{\frac{b^2}{N}\log(2/\delta)} \le \hat{R}_N(F) - R_N(F) \le \sqrt{\frac{b^2}{N}\log(2/\delta)}$$

$$\hat{R}_N(F) - b\sqrt{\frac{\log(2/\delta)}{N}} \le R_N(F) \le \tilde{R}_N(F) + b\sqrt{\frac{\log(2/\delta)}{N}} \quad (d)$$

Using equations a, b, c and d, we get :-

$$E_D[f(z)] \le \tilde{E}_S[f(s)] + 2\tilde{R}_N(F) + 3b\sqrt{\frac{\log(2/\delta)}{N}}$$

Hence Proved