COMS 4773 Spring 2024 HW 3 (due Apr. 3 at noon)

Problem 1.

(a) Suppose the hypothesis classes $\mathcal{H}_1 \subseteq \{0,1\}^{\mathcal{X}}$ and $\mathcal{H}_2 \subseteq \{0,1\}^{\mathcal{X}}$, respectively, have VC dimensions d_1 and d_2. Prove that the VC dimension of $\mathcal{H}_1 \cup \mathcal{H}_2$ is at most $d_1 + d_2 + 1$.

(b) Define, for each $d \in \mathbb{N}$, a hypothesis class $\mathcal{H}_d \subset \{0,1\}^{\mathcal{X}}$ defined on $\mathcal{X} = \mathbb{N}$ such that:

- $\mathcal{H}_d$ has VC dimension d, and
- for all $n \in \mathbb{N}$ and all distinct $x_1, \dots, x_n \in \mathcal{X}$, the number of behaviors of $\mathcal{H}_d$ on $x_{1:n}$ is

$$|\mathcal{H}_d(x_{1:n})| = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}.$$

Prove that your choice of $\mathcal{H}_d$ satisfies these properties.

(This shows that the bound from Sauer's lemma can be tight for some hypothesis classes of a given VC dimension.)

Problem 2. Recall that $\mathsf{LTF}_d := \{h_{w,b} : w \in \mathbb{R}^d, b \in \mathbb{R}\}$, the class of linear threshold functions in $\mathbb{R}^d$, where

$$h_{w,b}(x) = \operatorname{sign}(\langle x, w \rangle + b) \quad \text{for all } x \in \mathbb{R}^d.$$

In this problem, you will show that for any $n \geq 2$ points $x_1, \dots, x_n \in \mathbb{R}^2$, the set of behaviors of LTF_2 on these points,

$$\mathsf{LTF}_2(x_{1:n}) = \{(h(x_1), \dots, h(x_n)) : h \in \mathsf{LTF}_2\},$$

has cardinality $O(n^2)$. Note that LTF_d has VC dimension $d + 1$, so Sauer's lemma only guarantees

$$|\mathsf{LTF}_2(x_{1:n})| \leq \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \binom{n}{3} = \Theta(n^3).$$

So this will show that the upper-bound provided by Sauer's lemma is *not* tight for all hypothesis classes of a given VC dimension.

(a) Prove that if $x_1, \dots, x_n$ are n arbitrary points in $\mathbb{R}^2$, then there are at most $2(n-1)$ "behaviors" $a = (a_1, \dots, a_n)$ realized by LTFs $h_{w,b}$ such that $\langle x_1, w \rangle = -b$.

Hint: Consider lines that pass through x_1 and x_i for $i = 2, \dots, n$. Then consider lines that pass through x_1 and the angle between two "adjacent" lines of the previous type. What are the different behaviors that these lines determine?

(b) Use the result from Part (a) to prove that $|\mathsf{LTF}_2(x_{1:n})| \leq 2n(n-1) + 1$ for any $x_1, \dots, x_n \in \mathbb{R}^2$.

Problem 3. Recall that for any $A \subseteq \mathbb{R}^n$, we define

$$\mathrm{Rad}_n(A) := \mathbb{E}_\sigma \sup_{a \in A} \langle \sigma, a \rangle_n$$

where σ is a random vector distributed uniformly in $\{-1, 1\}^n$ (i.e., the coordinates of σ are iid Rademacher random variables), and $\langle \cdot, \cdot \rangle_n$ is the *normalized* inner product

$$\langle u, v \rangle_n := \frac{1}{n} \sum_{i=1}^{n} u_i v_i.$$

(a) Let A and B be arbitrary subsets of $\{0, 1\}^n$. Define

$$A \odot B := \{a \odot b : a \in A, b \in B\},$$

and

$$A + B := \{a + b : a \in A, b \in B\},$$

where $u \odot v$ denotes the element-wise product of u and v (i.e., $w = u \odot v \in \mathbb{R}^n$ means $w_i = u_i v_i$ for all $i \in [n]$). Prove that

$$\mathrm{Rad}_n(A \odot B) \leq \mathrm{Rad}_n(A + B).$$

Hint: Consider using the Lipschitz contraction property of Rademacher averages.

(b) There is an alternative (and somewhat more standard) definition of Rademacher average, which we shall write as $\mathrm{Rad}_n^*(A)$:

$$\mathrm{Rad}_n^*(A) := \mathbb{E}_\sigma \sup_{a \in A} |\langle \sigma, a \rangle_n|.$$

Many of the properties of Rad_n also hold for Rad_n^*, up to some minor changes. One is the Lipschitz contraction property. In this problem, you will prove a simplified version of it. Suppose $L \geq 0$ and that $\phi \colon \mathbb{R} \to \mathbb{R}$ is an L-Lipschitz function satisfying $\phi(0) = 0$. For any $A \subseteq \mathbb{R}^n$, define

$$\phi(A) := \{(\phi(a_1), \ldots, \phi(a_n)) : (a_1, \ldots, a_n) \in A\}.$$

Prove that, for any $A \subseteq \mathbb{R}^n$,

$$\mathrm{Rad}_n^*(\phi(A)) \leq 2L \, \mathrm{Rad}_n^*(A).$$

Hint: There is a direct proof of this property, but I think it is quite messy, and it is much easier to leverage the Lipschitz contraction property of Rad_n. Start by proving the following intermediate equation and inequality (with ϕ as above):

$$\begin{aligned} \mathrm{Rad}_n^*(\phi(A)) &= \mathrm{Rad}_n((\phi(A) \cup \{0\}) \cup (-\phi(A) \cup \{0\})) \\ &\leq \mathrm{Rad}_n(\phi(A) \cup \{0\}) + \mathrm{Rad}_n(-\phi(A) \cup \{0\}). \end{aligned}$$

Problem 4. In this problem, you will prove a generalization guarantee for soft-margin support vector machine (SVM). Given training data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, 1\}$ and regularization parameter $\lambda > 0$, the soft-margin SVM classifier is the (homogeneous) linear threshold function $h_{\hat{w}_\lambda} \colon \mathbb{R}^d \to \{-1, 1\}$ defined by

$$h_{\hat{w}_\lambda}(x) = \operatorname{sign}(\langle x, \hat{w}_\lambda \rangle) \quad \text{for all } x \in \mathbb{R}^d,$$

where $\hat{w}_\lambda = \hat{w}_\lambda((x_1, y_1), \ldots, (x_n, y_n)) \in \mathbb{R}^d$ is the solution to the minimization problem

$$\min_{w \in \mathbb{R}^d} \frac{\lambda}{2} \|w\|_2^2 + \frac{1}{n} \sum_{i=1}^n \max\{0, 1 - y_i \langle x_i, w \rangle\}.$$

The goal is to prove the following guarantee.

Proposition 1. *Let $(X, Y) \sim P$ for a probability distribution P on $B^d \times \{-1, 1\}$, where $B^d := \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ is the unit ball in $\mathbb{R}^d$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an iid sample from P. There exists a universal constant $C > 0$ such that, for any $\delta \in (0, 1)$,*

$$\Pr\left[\mathcal{R}(\hat{w}_\lambda) \leq \min_{w \in \mathbb{R}^d} \left[\mathcal{R}(w) + \frac{\lambda}{2}\|w\|_2^2\right] + C\left(\sqrt{\frac{1}{\lambda n}} + \sqrt{\frac{\log(1/\delta)}{\min\{1, \lambda\} n}}\right)\right] \geq 1 - \delta,$$

where $\mathcal{R}(w) := \mathbb{E} \max\{0, 1 - Y\langle X, W \rangle\}$ for $w \in \mathbb{R}^d$.

(a) Prove that $\|\hat{w}_\lambda\|_2 \leq \sqrt{2/\lambda}$.

(b) For any $r \geq 0$, let $B^d(r) := \{w \in \mathbb{R}^d : \|w\|_2 \leq r\}$ be the ball of radius r in $\mathbb{R}^d$. Prove that for any $x_1, \ldots, x_n \in B^d$,

$$\operatorname{Rad}_n(A) \leq \frac{r}{\sqrt{n}},$$

where $A := \{(\langle x_1, w \rangle, \ldots, \langle x_n, w \rangle) : w \in B^d(r)\}$.

(c) Prove Proposition 1.

Hint: Use the following decompositon. For any $w \in \mathbb{R}^d$,

$$\begin{aligned}
\mathcal{R}(\hat{w}_\lambda) - \mathcal{R}(w) &= \mathcal{R}(\hat{w}_\lambda) - \mathcal{R}_n(\hat{w}_\lambda) \\
&\quad + \mathcal{R}_n(w) - \mathcal{R}(w) \\
&\quad + \left[\mathcal{R}_n(\hat{w}_\lambda) + \frac{\lambda}{2}\|\hat{w}_\lambda\|_2^2\right] - \left[\mathcal{R}_n(w) + \frac{\lambda}{2}\|w\|_2^2\right] \\
&\quad + \frac{\lambda}{2}\|w\|_2^2 - \frac{\lambda}{2}\|\hat{w}_\lambda\|_2^2
\end{aligned}$$

where $\mathcal{R}_n(w) := \frac{1}{n}\sum_{i=1}^n \max\{0, 1 - Y_i\langle X_i, w \rangle\}$ for $w \in \mathbb{R}^d$.