CSCI 678: Theoretical Machine Learning
Homework 1

Fall 2024, Instructor: Haipeng Luo

This homework is due on **9/22, 11:59pm**. *See course website for more instructions on finishing and submitting your homework as well as the late policy. Total points: 50*

1. (**Rademacher complexity and Dudley entropy integral**) Consider inputs $x_1, \ldots, x_n \in \mathbb{R}^d$ and the linear class $\mathcal{F} = \left\{ f_\theta(x) = \langle \theta, x \rangle \mid \theta \in \mathbb{R}^d, \|\theta\|_2 \le b \right\}$.

   (a) (5pts) Prove the following:

$$\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F}; x_{1:n}) \le \frac{b}{n}\sqrt{\sum_{t=1}^{n} \|x_t\|_2^2}$$

   using the definition of Rademacher complexity directly (that is, without invoking its upper bounds in terms of covering numbers or other measures). Hint: you will need to use the inequality $\mathbb{E}[a] \le \sqrt{\mathbb{E}[a^2]}$ (which is a consequence of Jensen's inequality).

   Proof. We proceed as follows:

$$\begin{aligned}
\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F}; x_{1:n}) &= \frac{1}{n}\mathbb{E}_{\epsilon_{1:n}}\left[\sup_{\theta:\|\theta\|_2 \le b} \sum_{t=1}^{n} \epsilon_t \langle \theta, x_t \rangle\right] \\
&= \frac{1}{n}\mathbb{E}_{\epsilon_{1:n}}\left[\sup_{\theta:\|\theta\|_2 \le b} \left\langle \theta, \sum_{t=1}^{n} \epsilon_t x_t \right\rangle\right] \\
&= \frac{b}{n}\mathbb{E}_{\epsilon_{1:n}}\left[\left\|\sum_{t=1}^{n} \epsilon_t x_t\right\|_2\right] \\
&\le \frac{b}{n}\sqrt{\mathbb{E}_{\epsilon_{1:n}}\left[\left\|\sum_{t=1}^{n} \epsilon_t x_t\right\|_2^2\right]} && (\mathbb{E}[a] \le \sqrt{\mathbb{E}[a^2]}) \\
&= \frac{b}{n}\sqrt{\mathbb{E}_{\epsilon_{1:n}}\left[\sum_{t=1}^{n}\sum_{t'=1}^{n} \epsilon_t \epsilon_{t'} \langle x_t, x_{t'} \rangle\right]} \\
&= \frac{b}{n}\sqrt{\sum_{t=1}^{n} \|x_t\|_2^2}
\end{aligned}$$

   where the last step uses the fact that $\mathbb{E}[\epsilon_t \epsilon_{t'}]$ is 1 for $t = t'$ and 0 for $t \ne t'$. □

(b) (3pts) In Lecture 4, we will prove the following log covering number bound for this class: $\ln \mathcal{N}_2(\mathcal{F}|_{x_{1:n}}, \alpha) \leq \frac{b^2 \ln(2d) \sum_{t=1}^n \|x_t\|_2^2}{n\alpha^2}$. Use this bound and the Dudley entropy integral to prove

$$\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F}; x_{1:n}) \leq \widetilde{\mathcal{O}}\left(\frac{b}{n}\sqrt{\sum_{t=1}^n \|x_t\|_2^2}\right),$$

where the $\widetilde{\mathcal{O}}(\cdot)$ notation hides all logarithmic factors. (This bound is thus of the same order as the one from the last question.)

Proof. Since the dependence on α is $1/\alpha^2$ in the log covering number bound, we calculate the following integral:

$$\int_\alpha^1 \sqrt{\frac{1}{\delta^2}} d\delta = (\ln \delta)\Big|_\alpha^1 = \ln\left(\frac{1}{\alpha}\right).$$

Plugging this into Dudley entropy integral, we obtain

$$\widehat{\mathcal{R}}^{\text{iid}}(\mathcal{F}; x_{1:n}) \leq \min_{0\leq\alpha\leq 1}\left(4\alpha + \frac{12b\sqrt{\ln(2d)\sum_{t=1}^n \|x_t\|_2^2}}{n}\ln\left(\frac{1}{\alpha}\right)\right)$$

and complete the proof by picking $\alpha = 1/n$. □

2. (**Growth function and VC-dimension**)

   (a) Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{F} = \left\{f_{\theta,b}(x) = \text{sign}\left(\langle x, \theta\rangle + b\right) \mid \theta \in \mathbb{R}^d, b \in \mathbb{R}\right\}$ be the set of d-dimensional linear classifiers. Prove VCdim$(\mathcal{F}) = d + 1$ by following the two steps below.

      i. (4pts) Construct $d + 1$ points $x_1, \ldots, x_{d+1} \in \mathbb{R}^d$ and argue that for any labeling $y_1, \ldots, y_{d+1} \in \{-1, +1\}$, there exists $f \in \mathcal{F}$ such that $f(x_t) = y_t$ for all $t = 1, \ldots, d + 1$.

      Proof. For $t = 1, \ldots, d$, let $x_t = e_t$ be the basic vector such that the t-th coordinate is 1 (and all other coordinates are 0). Also let $x_{d+1} = 0$ be the all-zero vector. Then for any labeling $y_1, \ldots, y_{d+1} \in \{-1, +1\}$, let $\theta = \sum_{t=1}^{d}(y_t - y_{d+1})e_t$ and $b = y_{d+1}$, we have $f_{\theta,b}(x_t) = \text{sign}(y_t - y_{d+1} + y_{d+1}) = y_t$ for $t = 1, \ldots, d$ and also $f_{\theta,b}(x_{d+1}) = \text{sign}(b) = y_{d+1}$, finishing the proof. □

      ii. (6pts) Prove that for any $d + 2$ points $x_1, \ldots, x_{d+2} \in \mathbb{R}^d$, there exists a labeling $y_1, \ldots, y_{d+2} \in \{-1, +1\}$ such that no $f \in \mathcal{F}$ satisfies $f(x_t) = y_t$ for all $t = 1, \ldots, d + 2$. Hint: consider appending 1 to the end of each of the $d + 2$ points: $(x_1, 1), \cdots, (x_{d+2}, 1) \in \mathbb{R}^{d+1}$, and start with the fact that these $d + 2$ points must be linearly dependent (since they live in $\mathbb{R}^{d+1}$).

      Proof. Since $(x_1, 1), \cdots, (x_{d+2}, 1) \in \mathbb{R}^{d+1}$ are linearly dependent, we can assume without loss of generality that $(x_{d+2}, 1) = \sum_{t=1}^{d+1} c_t(x_t, 1)$ for some coefficients $c_1, \ldots, c_{d+1}$. By looking at the last coordinate, we must have $\sum_{t=1}^{d+1} c_t = 1$.

      Now, set $y_t = \text{sign}(c_t)$ for $t < d + 2$ and $y_{d+2} = -1$. Suppose that some $f_{\theta,b} \in \mathcal{F}$ realizes this labeling, then we have

$$0 > \langle \theta, x_{d+2}\rangle + b = \left(\sum_{t=1}^{d+1} c_t \langle \theta, x_t\rangle\right) + b = \sum_{t=1}^{d+1} c_t \left(\langle \theta, x_t\rangle + b\right) \geq 0,$$

      which is a contradiction. □

(b) (5pts) Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{F} = \{f_\theta(x) = \text{sign}(\sin(\theta x)) \mid \theta \in \mathbb{R}\}$. Prove that for any n, if $x_t = 2^{-2t}$, then $\mathcal{F}$ shatters the set $x_{1:n}$, which means $\text{VCdim}(\mathcal{F}) = \infty$. (Hint: for any labeling $y_{1:n}$, consider $\theta = \pi\left(1 + \sum_{i=1}^n (1-y_i)2^{2i-1}\right)$.)

Proof. As the hint suggested, let $\theta = \pi\left(1 + \sum_{i=1}^n (1-y_i)2^{2i-1}\right)$. We have for any t:

$$\begin{aligned}\theta x_t &= \pi\left(2^{-2t} + \sum_{i=1}^n \frac{1-y_i}{2}4^{(i-t)}\right)\\ &= \pi\left(2^{-2t} + \frac{1-y_t}{2} + \sum_{i=1}^{t-1}\frac{1-y_i}{2}4^{(i-t)} + \sum_{i=t+1}^{n}\frac{1-y_i}{2}4^{(i-t)}\right).\end{aligned}$$

Let $\epsilon = 2^{-2t} + \sum_{i=1}^{t-1}\frac{1-y_i}{2}4^{(i-t)}$. First, note that

$$\sin(\theta x_t) = \sin\left(\pi\left(\epsilon + \frac{1-y_t}{2}\right)\right),$$

because $\sum_{i=t+1}^{n}\frac{1-y_i}{2}4^{(i-t)}$ is always an even number. Next, note that $0 < \epsilon < 2^{-2t} + \sum_{j=1}^{\infty} 4^{-j} = 2^{-2t} + 1/3 < 1$. Therefore, when $y_t = -1$, we have

$$\sin(\theta x_t) = \sin\left(\pi\left(\epsilon + 1\right)\right) < 0,$$

and thus $\text{sign}(\sin(\theta x_t)) = y_t$; on the other hand, when $y_t = 1$, we have

$$\sin(\theta x_t) = \sin\left(\pi\epsilon\right) > 0,$$

and thus $\text{sign}(\sin(\theta x_t)) = y_t$ also holds. This completes the proof. □

3. (**Covering number**)

(a) In Proposition 2 of Lecture 3, via a volumetric argument we show that the linear class $\mathcal{F} = \left\{ f_\theta(x) = \langle \theta, x \rangle \mid \theta \in B_p^d \right\}$ for $\mathcal{X} = B_q^d$ and some $p \geq 1$ and $q \geq 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$ has bounded pointwise covering number: $\mathcal{N}(\mathcal{F}, \alpha) \leq \left(\frac{2}{\alpha} + 1\right)^d$ for any $0 \leq \alpha \leq 1$. Follow the two steps below to further show $\mathcal{N}(\mathcal{F}, \alpha) \geq \left(\frac{1}{2\alpha}\right)^d$.

i. (5pts) Given any pointwise α-cover $\mathcal{H} \subset [-1, +1]^{\mathcal{X}}$, construct a pointwise 2α-cover $\mathcal{H}' \subset \mathcal{F}$ so that $|\mathcal{H}'| \leq |\mathcal{H}|$ (note that $\mathcal{H}'$ has to be a subset of $\mathcal{F}$).

Proof. The pointwise α-cover $\mathcal{H}$ naturally groups the functions in $\mathcal{F}$ into different clusters where each cluster has the same representative. Formally, for each $h \in \mathcal{H}$, let $\mathcal{F}_h = \{f \in \mathcal{F} : |h(x) - f(x)| \leq \alpha, \forall\, x \in \mathcal{X}\}$ and f_h be an arbitrary function from $\mathcal{F}_h$. Then $\mathcal{H}' = \{f_h \in \mathcal{F} \mid h \in \mathcal{H}\}$ clearly has cardinality no more than that of $\mathcal{H}$ and also is a pointwise 2α-cover since for any $f \in \mathcal{F}$, with $h \in \mathcal{H}$ being the representative of f, we have

$$|f_h(x) - f(x)| \leq |f_h(x) - h(x)| + |h(x) - f(x)| \leq 2\alpha,$$

for any $x \in \mathcal{X}$ by the construction of f_h. □

ii. (6pts) Prove that if $\mathcal{H}' \subset \mathcal{F}$ is a pointwise 2α-cover of $\mathcal{F}$, then we must have $|\mathcal{H}'| \geq \left(\frac{1}{2\alpha}\right)^d$, which then implies $\mathcal{N}(\mathcal{F}, \alpha) \geq \left(\frac{1}{2\alpha}\right)^d$ as desired. Hint: use a similar volumetric argument.

Proof. Each function in $\mathcal{H}'$ is parameterized by a point in B_p^d. If for each of these points we put a small ball $2\alpha B_p^d$ centered at it, then it must "cover" the entire ball B_p^d, for otherwise there is a function $f \in \mathcal{F}$ that is not covered by $\mathcal{H}'$. So we have $|\mathcal{H}'|\mathrm{Vol}(2\alpha B_p^d) \geq \mathrm{Vol}(B_p^d)$. Using the fact $\mathrm{Vol}(2\alpha B_p^d) = (2\alpha)^d \mathrm{Vol}(B_p^d)$ and rearranging then finishes the proof. □

(b) Let $v_1, \ldots, v_d \in B_2^n$ be d points within the n-dimensional ℓ_2-norm unit ball and

$$\mathcal{S} = \left\{ \sum_{i=1}^{d} \beta_i v_i \;\middle|\; \beta_i \geq 0,\ \forall i,\ \text{and } \sum_{i=1}^{d} \beta_i \leq B \right\}$$

be the convex hull of these d points scaled by $B > 0$.

i. (5pts) Prove $\mathcal{N}_2(\mathcal{S}, \alpha) \leq \left(\frac{2B}{\sqrt{n}\alpha} + 1\right)^d$.

Proof. Note that in Proposition 2 of Lecture 3, we have constructed a set $\mathcal{C} \subset B_1^d$ of size at most $\left(\frac{2}{\alpha'} + 1\right)^d$ such that for any $\theta \in B_1^d$, we can find $c \in \mathcal{C}$ with $\|\theta - c\|_1 \leq \alpha'$. Now, set $\alpha' = \frac{\sqrt{n}\alpha}{B}$ and let

$$\mathcal{S}' = \left\{ \sum_{i=1}^{d} B c_i v_i \;\middle|\; (c_1, \ldots, c_d) \in \mathcal{C} \right\}.$$

The size of $\mathcal{S}'$ is clearly at most $\left(\frac{2B}{\sqrt{n}\alpha} + 1\right)^d$, and it remains to prove that it is an α-cover of $\mathcal{S}$ in terms of ℓ_2-norm. Indeed, for any $\sum_{i=1}^{d} \beta_i v_i \in \mathcal{S}$, let c be the closest point in $\mathcal{C}$ to $\frac{1}{B}(\beta_1, \ldots, \beta_d) \in B_1^d$, then we have

$$\begin{aligned}
\left\| \sum_{i=1}^{d} B c_i v_i - \sum_{i=1}^{d} \beta_i v_i \right\|_2 &= \left\| \sum_{i=1}^{d} (B c_i - \beta_i) v_i \right\|_2 \\
&\leq \sum_{i=1}^{d} |B c_i - \beta_i| \, \|v_i\|_2 && \text{(triangle inequality)} \\
&\leq B \sum_{i=1}^{d} \left| c_i - \frac{1}{B}\beta_i \right| && (\|v_i\|_2 \leq 1) \\
&\leq B\alpha' = \sqrt{n}\alpha,
\end{aligned}$$

which completes the proof. □

ii. Follow the steps below to prove a different covering number bound $\mathcal{N}_2(\mathcal{S},\alpha) \le d^{\frac{B^2}{n\alpha^2}}$.

A. (4pts) For any $v = \sum_{i=1}^d \beta_i v_i \in \mathcal{S}$, let $\beta = (\beta_1,\ldots,\beta_d)$ and define m i.i.d. random variables $u_1,\ldots,u_m$, each of which is $\|\beta\|_1 v_i$ with probability $\beta_i / \|\beta\|_1$ for $i = 1,\ldots,d$. Prove that the mean of these random variables is v and the variance of $u = \frac{1}{m}\sum_{j=1}^m u_j$ is bounded as:

$$\mathbb{E}\left[\|u-v\|_2^2\right] \le \frac{\|\beta\|_1^2}{m}.$$

Proof. By definition, the mean is simply $\sum_{i=1}^d \frac{\beta_i}{\|\beta\|_1} \cdot \|\beta\|_1 v_i = \sum_{i=1}^d \beta_i v_i = v$. For the second statement, proceed as follows:

$$\begin{aligned}\mathbb{E}\left[\|u-v\|_2^2\right] &= \frac{1}{m}\mathbb{E}\left[\|u_1-v\|_2^2\right] && (u_1,\ldots,u_m \text{ are i.i.d.})\\ &\le \frac{1}{m}\mathbb{E}\left[\|u_1\|_2^2\right]\\ &= \frac{1}{m}\sum_{i=1}^d \frac{\beta_i}{\|\beta\|_1}\cdot\|\beta\|_1^2\|v_i\|_2^2 \le \frac{\|\beta\|_1^2}{m}, && (\|v_i\|_2 \le 1)\end{aligned}$$

which finishes the proof. □

B. (7pts) Prove that the following is an α-cover of $\mathcal{S}$ with respect to ℓ_2-norm:

$$\mathcal{S}' = \left\{ \frac{B}{M}\sum_{i=1}^d m_i v_i \;\middle|\; \text{each } m_i \text{ is a nonnegative integer and } \sum_{i=1}^d m_i \le M \right\}$$

where $M = \frac{B^2}{n\alpha^2}$. (The statement $\mathcal{N}_2(\mathcal{S},\alpha) \le d^{\frac{B^2}{n\alpha^2}}$ then follows immediately.)

Proof. For any $v = \sum_{i=1}^d \beta_i v_i \in \mathcal{S}$, the conclusion in the last question implies that there exists one realization of $u = \frac{1}{m}\sum_{j=1}^m u_j$ such that $\|u-v\|_2 \le \frac{\|\beta\|_1}{\sqrt{m}}$, which is at most $\sqrt{n}\alpha$ if we set $m = \frac{B\|\beta\|_1}{n\alpha^2}$. Now, note that u can be rewritten as

$$u = \frac{1}{m}\sum_{j=1}^m u_j = \frac{\|\beta\|_1}{m}\sum_{i=1}^d m_i v_i$$

for some nonnegative integers $m_1,\ldots,m_d$ with $\sum_{i=1}^d m_i = m \le M$. Finally, using the definition of m and M shows

$$u = \frac{\|\beta\|_1}{m}\sum_{i=1}^d m_i v_i = \frac{n\alpha^2}{B}\sum_{i=1}^d m_i v_i = \frac{B}{M}\sum_{i=1}^d m_i v_i$$

and thus $u \in \mathcal{S}'$, finishing the proof. □