CS7545, Spring 2023: Machine Learning Theory - Homework #2

Jacob Abernethy, Tyler LaBonte, and Yeojoon Youn — Due: Tuesday, February 21 at 11:59 p.m.

Homework Policy: Working in groups is fine, but *every student* must submit their own writeup, *i.e.,* write the solutions on their own and not submit a shared document. Please write the members of your group on your solutions. There is no strict limit to the size of the group but we may find it a bit suspicious if there are more than 4 to a team. Questions labelled with **(Challenge)** are not strictly required, but you'll get some participation credit if you have something interesting to add, even if it's only a partial answer.

1) **Properties of Rademacher complexity.** Suppose $A \subseteq \mathbb{R}^m$.

(a) Prove that $\mathfrak{R}(A+b) = \mathfrak{R}(A)$ where $A + b = \{a + b : a \in A\}$ for any $b \in \mathbb{R}^m$.

Using linearity of expectation,

$$\begin{aligned}
\mathfrak{R}(A+b) &= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{a' \in (A+b)} \frac{1}{m}\sum_{i=1}^m \sigma_i a'_i\right] && (1)\\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{a \in A} \frac{1}{m}\sum_{i=1}^m \sigma_i (a_i + b_i)\right] && (2)\\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{a \in A} \frac{1}{m}\sum_{i=1}^m \sigma_i a_i + \frac{1}{m}\sum_{i=1}^m \sigma_i b_i\right] && (3)\\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{a \in A} \frac{1}{m}\sum_{i=1}^m \sigma_i a_i\right] + \mathop{\mathbb{E}}_{\sigma}\left[\frac{1}{m}\sum_{i=1}^m \sigma_i b_i\right] && (4)\\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{a \in A} \frac{1}{m}\sum_{i=1}^m \sigma_i a_i\right] + \frac{1}{m}\sum_{i=1}^m b_i \mathop{\mathbb{E}}_{\sigma_i}[\sigma_i] && (5)\\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{a \in A} \frac{1}{m}\sum_{i=1}^m \sigma_i a_i\right] && (6)\\
&= \mathfrak{R}(A). && (7)
\end{aligned}$$

(b) Prove that $\mathfrak{R}(cA) = |c|\mathfrak{R}(A)$ where $cA = \{c \cdot a : a \in A\}$ for any $c \in \mathbb{R}$.

We have

$$\mathfrak{R}(cA) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{a' \in (cA)} \frac{1}{m}\sum_{i=1}^m \sigma_i a'_i\right] = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{a \in A} \frac{c}{m}\sum_{i=1}^m \sigma_i a_i\right]. \quad (8)$$

If $c \geq 0$ then

$$\mathop{\mathbb{E}}_{\sigma}\left[\sup_{a \in A} \frac{c}{m}\sum_{i=1}^m \sigma_i a_i\right] = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{a \in A} \frac{|c|}{m}\sum_{i=1}^m \sigma_i a_i\right] = |c| \mathop{\mathbb{E}}_{\sigma}\left[\sup_{a \in A} \frac{1}{m}\sum_{i=1}^m \sigma_i a_i\right]. \quad (9)$$

Otherwise if $c < 0$ then

$$\mathop{\mathbb{E}}_{\sigma}\left[\sup_{a\in A}\frac{c}{m}\sum_{i=1}^{m}\sigma_i a_i\right] = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{a\in A}\frac{-|c|}{m}\sum_{i=1}^{m}\sigma_i a_i\right] = |c|\mathop{\mathbb{E}}_{\sigma}\left[\sup_{a\in A}\frac{1}{m}\sum_{i=1}^{m}-\sigma_i a_i\right]. \tag{10}$$

But since σ_i and $-\sigma_i$ follow the same distribution, the right-hand side in either case is $|c|\mathfrak{R}(A)$.

(c) In lecture we proved the following one-sided generalization bound: for $\mathcal{F}$ containing functions $f : \mathcal{X} \to [0,1]$ and any $\delta > 0$, with probability at least $1-\delta$ over the choice of a sample $S \sim \mathcal{D}^m$, the following holds for all $f \in \mathcal{F}$:

$$L(f) \le \widehat{L}_S(f) + 2\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2m}}. \tag{11}$$

However, to show a bound on the estimation error of ERM we actually needed a two-sided bound, on $\sup_{f\in\mathcal{F}}\left|L(f) - \widehat{L}_S(f)\right|$. Use parts **(a)** and **(b)** to prove one.

Define $\mathcal{G} \coloneqq \{1-f : f \in \mathcal{F}\}$. By parts **(a)** and **(b)**, we have $\mathfrak{R}(\mathcal{G}) = |-1|\mathfrak{R}(\mathcal{F}) = \mathfrak{R}(\mathcal{F})$. Furthermore,

$$\begin{aligned} L(g) - \widehat{L}_S(g) &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}} g(x) - \frac{1}{m}\sum_{i=1}^{m} g(x_i) && (12)\\ &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[1-f(x)] - \frac{1}{m}\sum_{i=1}^{m}(1-f(x_i)) && (13)\\ &= (1 - \mathop{\mathbb{E}}_{x\sim\mathcal{D}} f(x)) - (1 - \frac{1}{m}\sum_{i=1}^{m} f(x_i)) && (14)\\ &= \frac{1}{m}\sum_{i=1}^{m} f(x_i) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}} f(x) && (15)\\ &= \widehat{L}_S(f) - L(f). && (16) \end{aligned}$$

Hence, with probability at least $1-\delta_1$,

$$\sup_{f\in\mathcal{F}} L(f) - \widehat{L}_S(f) \le 2\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{\log 1/\delta_1}{2m}}, \tag{17}$$

and with probability at least $1-\delta_2$,

$$\sup_{f\in\mathcal{F}} \widehat{L}_S(f) - L(f) = \sup_{g\in\mathcal{G}} L(g) - \widehat{L}_S(g) \le 2\mathfrak{R}(\mathcal{G}) + \sqrt{\frac{\log 1/\delta_2}{2m}}. \tag{18}$$

Taking a union bound with $\delta_1 = \delta_2 = \delta/2$, we have that with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}\left|L(f) - \widehat{L}_S(f)\right| \le 2\max(\mathfrak{R}(\mathcal{F}), \mathfrak{R}(\mathcal{G})) + \sqrt{\frac{\log 2/\delta}{2m}} = 2\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{\log 2/\delta}{2m}}. \tag{19}$$

2) **Rademacher complexity of ℓ_1-bounded neural networks.** Suppose the input space is $\mathcal{X} = \mathbb{R}^n$ and we have a training set $S = \{(x_i, y_i)\}_{i=1}^{m}$. Let $\phi : \mathbb{R} \to \mathbb{R}$ be an L-Lipschitz activation function such that

$\phi(0) = 0$ (for example, the ReLU function). Define the class of neural networks of depth $2 \le j \le D$ and width H with ℓ_1-bounded weights recursively as

$$\mathcal{F}_j \coloneqq \left\{ x \mapsto \sum_{k=1}^{H} w_k \phi(f_k(x)) : f_k \in \mathcal{F}_{j-1}, \|w\|_1 \le B_j \right\}. \tag{20}$$

Here, ϕ is applied elementwise, *i.e.,* $\phi(x) = (\phi(x_1), \dots, \phi(x_n))$.

(a) Define $\mathcal{F}_1 \coloneqq \{x \mapsto \langle w, x\rangle : \|w\|_1 \le B_1\}$ and suppose $\|x_i\|_\infty \le C$ for all $1 \le i \le m$. Prove that

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_1) \le B_1 C \sqrt{\frac{2\log 2n}{m}}. \tag{21}$$

Hint. Use Hölder's inequality and Massart's lemma.

Applying Hölder's inequality,

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{F}_1) &= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}_1} \frac{1}{m}\sum_{i=1}^{m} \sigma_i f(x_i)\right] && (22)\\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{w:\|w\|_1\le B_1} \frac{1}{m}\sum_{i=1}^{m} \sigma_i \langle w, x_i\rangle\right] && (23)\\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{w:\|w\|_1\le B_1} \frac{1}{m}\Big\langle w, \sum_{i=1}^{m} \sigma_i x_i\Big\rangle\right] && (24)\\
&\le B_1 \mathop{\mathbb{E}}_{\sigma}\left[\frac{1}{m}\Big\| \sum_{i=1}^{m} \sigma_i x_i\Big\|_\infty\right]. && (25)
\end{aligned}$$

For $1 \le j \le n$, let $a_j = (x_{1j}, \dots, x_{mj})$ and $A = \{a_1, \dots, a_n, -a_1, \dots, -a_n\}$. Then,

$$\Big\|\sum_{i=1}^{m} \sigma_i x_i\Big\|_\infty = \max_{1\le j\le n} \Big|\sum_{i=1}^{m} \sigma_i x_i\Big|_j = \max_{1\le j\le n} \Big|\sum_{i=1}^{m} \sigma_i x_{ij}\Big| = \sup_{a\in A} \sum_{i=1}^{m} \sigma_i a_i. \tag{26}$$

Hence,

$$\mathop{\mathbb{E}}_{\sigma}\left[\frac{1}{m}\Big\|\sum_{i=1}^{m} \sigma_i x_i\Big\|_\infty\right] = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{a\in A} \frac{1}{m}\sum_{i=1}^{m} \sigma_i a_i\right] = \mathfrak{R}(A). \tag{27}$$

Note that $\|a_j\| \le \sqrt{m}\max_i \|x_i\|_\infty$. By Massart's lemma,

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_1) \le B_1 \mathfrak{R}(A) \le B_1 C \sqrt{\frac{2\log 2n}{m}}. \tag{28}$$

(b) Prove that $\widehat{\mathfrak{R}}_S(\mathcal{F}_j) \le 2LB_j \widehat{\mathfrak{R}}_S(\mathcal{F}_{j-1})$ for $2 \le j \le D$. **Hint.** Use Hölder's inequality and Talagrand's contraction lemma. You may use the fact that if a function class $\mathcal{G}$ contains the zero function, then

$$\mathop{\mathbb{E}}_{\sigma}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\left|\sum_{i=1}^{m} \sigma_i g(x_i)\right|\right] \le 2\widehat{\mathfrak{R}}_S(\mathcal{G}). \tag{29}$$

We have

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_j) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}_j} \frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i)\right] \tag{30}$$

$$= \mathbb{E}_\sigma\left[\sup_{\substack{\|w\|_1\le B_j\\ f_k\in\mathcal{F}_{j-1}}} \frac{1}{m}\sum_{i=1}^m\sum_{k=1}^H \sigma_i w_k \phi(f_k(x_i))\right] \tag{31}$$

$$= \mathbb{E}_\sigma\left[\sup_{\substack{\|w\|_1\le B_j\\ f_k\in\mathcal{F}_{j-1}}} \frac{1}{m}\sum_{k=1}^H w_k \sum_{i=1}^m \sigma_i \phi(f_k(x_i))\right]. \tag{32}$$

Since $\phi(0) = 0$, every $\mathcal{F}_j$ contains the zero function. Applying Hölder's inequality and the hint,

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_j) \le \mathbb{E}_\sigma\left[\sup_{\substack{\|w\|_1\le B_j\\ f_k\in\mathcal{F}_{j-1}}} \frac{1}{m}\|w\|_1 \max_{1\le k\le H}\left|\sum_{i=1}^m \sigma_i \phi(f_k(x_i))\right|\right] \tag{33}$$

$$\le B_j\,\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}_{j-1}} \frac{1}{m}\left|\sum_{i=1}^m \sigma_i\phi(f(x_i))\right|\right] \tag{34}$$

$$= 2B_j\,\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}_{j-1}} \frac{1}{m}\sum_{i=1}^m \sigma_i\phi(f(x_i))\right]. \tag{35}$$

Define $A = \{(f(x_1),\ldots,f(x_m)) : f\in\mathcal{F}_{j-1}\}$. Applying Talagrand's contraction lemma,

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_j) \le 2B_j\,\mathbb{E}_\sigma\left[\sup_{a\in A}\frac{1}{m}\sum_{i=1}^m \sigma_i\phi(a_i)\right] = 2B_j\mathfrak{R}(\phi(A)) \le 2LB_j\mathfrak{R}(A) = 2LB_j\widehat{\mathfrak{R}}_S(\mathcal{F}_{j-1}). \tag{36}$$

(c) Use parts **(a)** and **(b)** to show an upper bound on the Rademacher complexity of $\widehat{\mathfrak{R}}_S(\mathcal{F}_D)$.

Solving the recurrence from part **(b)** and substituting the answer from part **(a)** gives

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_D) \le \prod_{j=2}^D 2LB_j\widehat{\mathfrak{R}}_S(\mathcal{F}_{j-1}) = (2L)^{D-1}\prod_{j=2}^D B_j\cdot\widehat{\mathfrak{R}}_S(\mathcal{F}_1) = (2L)^{D-1}\prod_{j=1}^D B_j\cdot C\sqrt{\frac{2\log 2n}{m}}. \tag{37}$$

(d) If we have time near the end of the semester, we will prove that the VC-dimension of depth D neural networks with piecewise linear activations and W weights is $\mathcal{O}(WD\log W)$. What are the advantages and disadvantages of your Rademacher complexity bound compared to this VC-dimension bound?

The advantage of the Rademacher complexity bound is that it does not depend on the number of weights W, but only on the weight norms B_j. Furthermore, the Rademacher complexity bound applies to any Lipschitz activation function ϕ with $\phi(0) = 0$ instead of only piecewise linear activations.

The disadvantage of the Rademacher complexity bound is that it is exponential in the depth instead of linear in the depth. Furthermore, the Rademacher complexity bound has a mild dependence on the input dimension n, while the VC-dimension bound is independent of the input dimension.

3) **Massart's lemma, take 2.** If you recall when we proved Massart's lemma, it looked suspiciously similar to the proof of Hoeffding's inequality – indeed, you can reduce it directly, with a slightly worse result.

Let $A \subseteq [-1,1]^m$ be a finite set. Prove the following via a reduction to Hoeffding's inequality:

$$\mathfrak{R}(A) = O\left(\sqrt{\frac{\log m + \log |A|}{m}}\right). \tag{38}$$

Hint. For any real random variable Z and any real t, we have $\mathbb{E}[Z] = \mathbb{E}[Z \cdot \mathbb{1}(Z \leq t)] + \mathbb{E}[Z \cdot \mathbb{1}(Z > t)]$.

Let $Z = \sup_{a \in A} \langle \sigma, a \rangle$ be a random variable over σ. Note that $a_i \in [-1,1]$ and so $Z \leq m$. For any $t \geq 0$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\sigma}[Z] &= \mathop{\mathbb{E}}_{\sigma}[Z \cdot \mathbb{1}(Z \leq t)] + \mathop{\mathbb{E}}_{\sigma}[Z \cdot \mathbb{1}(Z > t)] && (39)\\
&\leq \mathop{\mathbb{E}}_{\sigma}[t \cdot \mathbb{1}(Z \leq t)] + \mathop{\mathbb{E}}_{\sigma}[m \cdot \mathbb{1}(Z > t)] && (40)\\
&\leq t + m \Pr_{\sigma}[Z > t] && (41)\\
&= t + m \Pr_{\sigma}[\exists a \in A : \langle \sigma, a \rangle > t] && (42)\\
&\leq t + m \sum_{a \in A} \Pr_{\sigma}[\langle \sigma, a \rangle > t], && (43)
\end{aligned}$$

where the last inequality follows from a union bound. For $a \in A$, let $X_i = \sigma_i a_i$ and $X = \langle \sigma, a \rangle = \sum_{i=1}^m X_i$. Note $\mathbb{E}[X] = 0$. Because $|X_i| = |a_i| \leq 1$ and are independent, by Hoeffding's inequality,

$$\Pr_{\sigma}[\langle \sigma, a \rangle > t] = \Pr_{\sigma}[X - \mathbb{E}[X] > t] \leq e^{-\frac{t^2}{2m}}. \tag{44}$$

Therefore,

$$\mathop{\mathbb{E}}_{\sigma}[Z] \leq t + m \sum_{a \in A} e^{-\frac{t^2}{2m}} = t + m|A|e^{-\frac{t^2}{2m}}. \tag{45}$$

Setting $t = \sqrt{2m(\log m + \log |A|)}$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\sigma}[Z] &\leq \sqrt{2m(\log m + \log |A|)} + m|A|e^{-\frac{2m \log m|A|}{2m}} && (46)\\
&= \sqrt{2m(\log m + \log |A|)} + 1 && (47)\\
&= \mathcal{O}(\sqrt{m(\log m + \log |A|)}). && (48)
\end{aligned}$$

Dividing by m gives the result.

4) **Growth function.** In lecture we studied the growth function for classes of functions taking values in the set $\{-1,1\}$, but the same definition applies to classes of functions taking values in the finite set $\mathcal{Y}$. In this case, $\Pi_{\mathcal{H}}(m) \leq |\mathcal{Y}|^m$ (analogous to 2^m in the original setup).

(a) Let $\mathcal{H}_1 \subseteq \{h : \mathcal{X} \to \mathcal{Y}_1\}$ and $\mathcal{H}_2 \subseteq \{h : \mathcal{X} \to \mathcal{Y}_2\}$ be function classes and let $\mathcal{H}_3 \subseteq \{h : \mathcal{X} \times \mathcal{X} \to \mathcal{Y}_1 \times \mathcal{Y}_2\}$ such that $\mathcal{H}_3 = \{(h_1, h_2) : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$. Show that

$$\Pi_{\mathcal{H}_3}(m) = \Pi_{\mathcal{H}_1}(m) \cdot \Pi_{\mathcal{H}_2}(m). \tag{49}$$

For any $S = ((x_1, x'_1), \ldots, (x_m, x'_m)) \subseteq \mathcal{X} \times \mathcal{X}$,

$$\begin{aligned}
|\mathcal{H}_3|_S| &= |\{(h_3(x_1, x'_1), \ldots, h_3(x_m, x'_m)) : h_3 \in \mathcal{H}_3\}| && (50)\\
&= |\{((h_1(x_1), h_2(x'_1)), \ldots, (h_1(x_m), h_2(x'_m))) : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}| && (51)\\
&= |\{(h_1(x_1), \ldots, h_1(x_m)) : h_1 \in \mathcal{H}_1\}| \cdot |\{(h_2(x'_1), \ldots, h_2(x'_m)) : h_2 \in \mathcal{H}_2\}| && (52)\\
&= |\mathcal{H}_1|_S| \cdot |\mathcal{H}_2|_S| && (53)
\end{aligned}$$

Hence $\Pi_{\mathcal{H}_3}(m) = \Pi_{\mathcal{H}_1}(m) \cdot \Pi_{\mathcal{H}_2}(m)$.

(b) Let $\mathcal{H}_1 \subseteq \{h : \mathcal{X} \to \mathcal{Y}_1\}$ and $\mathcal{H}_2 \subseteq \{h : \mathcal{Y}_1 \to \mathcal{Y}_2\}$ be function classes and let $\mathcal{H}_3 \subseteq \{h : \mathcal{X} \to \mathcal{Y}_2\}$ such that $\mathcal{H}_3 = \{h_2 \circ h_1 : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$. Show that

$$\Pi_{\mathcal{H}_3}(m) \leq \Pi_{\mathcal{H}_1}(m) \cdot \Pi_{\mathcal{H}_2}(m). \tag{54}$$

For any $S = (x_1, \dots, x_m) \subseteq \mathcal{X}$,

$$\begin{aligned} \mathcal{H}_3|_S &= \{(h_3(x_1), \dots, h_3(x_m)) : h_3 \in \mathcal{H}_3\} && (55)\\ &= \{(h_2(h_1(x_1)), \dots, h_2(h_1(x_m))) : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\} && (56)\\ &= \bigcup_{u \in \mathcal{H}_1|_S} \{(h_2(u_1), \dots, h_2(u_m)) : h_2 \in \mathcal{H}_2\}. && (57) \end{aligned}$$

Thus,

$$\begin{aligned} |\mathcal{H}_3|_S| &\leq \sum_{u \in \mathcal{H}_1|_S} |\{(h_2(u_1), \dots, h_2(u_m)) : h_2 \in \mathcal{H}_2\}| && (58)\\ &\leq \sum_{u \in \mathcal{H}_1|_S} \Pi_{\mathcal{H}_2}(m) && (59)\\ &= |\mathcal{H}_1|_S| \cdot \Pi_{\mathcal{H}_2}(m) && (60)\\ &\leq \Pi_{\mathcal{H}_1}(m) \cdot \Pi_{\mathcal{H}_2}(m). && (61) \end{aligned}$$

Hence $\Pi_{\mathcal{H}_3}(m) \leq \Pi_{\mathcal{H}_1}(m) \cdot \Pi_{\mathcal{H}_2}(m)$.

5) **VC-dimension.**

(a) What is the VC-dimension of a union of k intervals on the real line?

The VC-dimension is $2k$. Suppose A is a set of $2k$ points in $\mathbb{R}$. For any $\{-1, 1\}$ labeling of A, we may cover all adjacent 1s with the same interval, and we only need a new interval after a -1 label. Since there can be at most k sets of adjacent 1s, A is shattered. On the other hand, any set of size $2k + 1$ cannot be shattered, because we cannot form the label assignment $1, -1, 1, -1, \dots, 1$.

(b) What is the VC-dimension of axis-aligned hyperrectangles in $\mathbb{R}^n$?

The VC-dimension is $2n$. Let A be the set of standard basis vectors for $\mathbb{R}^n$. Then, A is shattered because we can adjust the axes of the hyperrectangle individually to include or exclude each point as desired. On the other hand, any set of size $2n + 1$ cannot be shattered. To see this, consider finding the minimum and maximum values of the points across each dimension and constructing a hyperrectangle with these bounds. Then, since all the points are distinct, at least one point x must lie inside the hyperrectangle (or on its boundary, but not at a vertex). We cannot form the label assignment where every point is labelled 1 except for x which is labelled -1.

(c) A simplex in $\mathbb{R}^n$ is the intersection of $n + 1$ halfspaces (not necessarily bounded). Prove that the VC-dimension of simplices in $\mathbb{R}^n$ is $\mathcal{O}(n^2 \log n)$. **Hint.** Use the VC-dimension of halfspaces in $\mathbb{R}^n$.

Let $\mathcal{H}$ denote a hypothesis class with VC-dimension d and $\mathcal{S}$ denote the class of simplices in $\mathbb{R}^n$. Recall from the Sauer-Shelah lemma that $\text{VC}(\mathcal{H}) = d$ implies $\Pi_{\mathcal{H}}(m) \leq m^d$, and the definition of shattering m points is $\Pi_{\mathcal{H}}(m) = 2^m$.

Suppose $\mathcal{H}^{\cap k}$ is the intersection of k hypotheses from $\mathcal{H}$. Then since each hypothesis has at most $\Pi_{\mathcal{H}}(m)$ distinct labelings, we must have $\Pi_{\mathcal{H}^{\cap k}}(m) \leq (\Pi_{\mathcal{H}}(m))^k$ for any m. Hence, $\Pi_{\mathcal{H}^{\cap k}}(m) \leq m^{dk}$. To show $\mathrm{VC}(\mathcal{H}^{\cap k}) < m$ we can show $\Pi_{\mathcal{H}^{\cap k}}(m) < 2^m$, that is $m^{dk} < 2^m$. Taking logs, this is equivalent to $dk \log m < m$. Setting $m = 2dk \log dk$, we find $2dk \log dk < (dk)^2$, which is true when $dk > 4$. So $\mathrm{VC}(\mathcal{H}^{\cap k}) = \mathcal{O}(dk \log dk)$. Since a simplex in $\mathbb{R}^n$ is the intersection of $n+1$ halfspaces, and halfspaces have VC-dimension $n+1$, we obtain

$$\mathrm{VC}(\mathcal{S}) = \mathcal{O}((n+1)^2 \log(n+1)^2) = \mathcal{O}(n^2 \log n). \tag{62}$$

(d) Prove the best lower bound you can on the VC-dimension of simplices in $\mathbb{R}^n$ **(Challenge).**

A lower bound of $\Omega(n)$ can be obtained by noticing that simplices can shatter any $n+1$ affinely independent points. In particular, let S be the simplex with these points as its vertices. Then, any labeling of these points can be achieved by "wiggling" one of the halfspaces at each vertex v so that v is included or not included in the simplex. Formally, let x be some point strictly inside S and let $\epsilon > 0$ be small. Then for each vertex v labelled -1, pick one of the halfspaces H which intersect at v. Since a hyperplane in $\mathbb{R}^n$ is defined by n points, let H' be the halfspace formed by the $n-1$ other points forming H as well as the point $y = (1-\epsilon)v + \epsilon x$. The new simplex S' formed by using H' instead of H is still an intersection of $n+1$ halfspaces, and it contains all the original vertices except v.

A lower bound of $\Omega(n^2)$ can be found in Lemma 3.7 of this paper, and a (much harder) lower bound of $\Omega(n^2 \log n)$ was recently proved in this paper. Hence, the VC-dimension of the simplex is indeed $\Theta(n^2 \log n)$.

6) **Desymmetrization (Challenge).** Let $S = \{x_1, \ldots, x_m\} \sim \mathcal{D}^m$ and suppose $\mathcal{F}$ contains functions $f : \mathcal{X} \to [0,1]$. Prove the symmetrization lower bound, also called the desymmetrization inequality:

$$\frac{1}{2}\mathfrak{R}(\mathcal{F}) - \sqrt{\frac{\log 2}{2m}} \leq \mathop{\mathbb{E}}_{S}\left[\sup_{f\in\mathcal{F}} \left| L(f) - \widehat{L}_S(f) \right| \right]. \tag{63}$$

We have

$$\begin{aligned}
\mathfrak{R}(\mathcal{F}) &= \mathop{\mathbb{E}}_{S,\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i)\right] && (64)\\
&= \mathop{\mathbb{E}}_{S,\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i)\right] - \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i L(f)\right] + \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i L(f)\right] && (65)\\
&= \underbrace{\mathop{\mathbb{E}}_{S,\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i) - \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i L(f)\right]}_{\text{Term 1}} + \underbrace{\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i L(f)\right]}_{\text{Term 2}}. && (66)
\end{aligned}$$

Introducing a ghost sample S',

$$\begin{aligned}
\text{Term 1} &\le \mathop{\mathbb{E}}_{S,\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i(f(x_i)-L(f))\right] \tag{67}\\
&= \mathop{\mathbb{E}}_{S,\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i(f(x_i)-\mathop{\mathbb{E}}_{S'}\widehat{L}_{S'}(f))\right] \tag{68}\\
&= \mathop{\mathbb{E}}_{S,\sigma}\left[\sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{S'}\frac{1}{m}\sum_{i=1}^{m}\sigma_i(f(x_i)-f(x_i'))\right] \tag{69}\\
&\le \mathop{\mathbb{E}}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i(f(x_i)-f(x_i'))\right]. \tag{70}
\end{aligned}$$

By symmetrization,

$$\begin{aligned}
\text{Term 1} &\le \mathop{\mathbb{E}}_{S,S'}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}(f(x_i)-f(x_i'))\right] \tag{71}\\
&= \mathop{\mathbb{E}}_{S,S'}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}(f(x_i)-L(f)+L(f)+f(x_i'))\right] \tag{72}\\
&\le \mathop{\mathbb{E}}_{S}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}(f(x_i)-L(f))\right] + \mathop{\mathbb{E}}_{S'}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}(L(f)-f(x_i'))\right] \tag{73}\\
&= \mathop{\mathbb{E}}_{S}\left[\sup_{f\in\mathcal{F}} \widehat{L}_S(f)-L(f)\right] + \mathop{\mathbb{E}}_{S'}\left[\sup_{f\in\mathcal{F}} L(f)-\widehat{L}_{S'}(f)\right] \tag{74}\\
&\le 2\mathop{\mathbb{E}}_{S}\left[\sup_{f\in\mathcal{F}} \left|L(f)-\widehat{L}_S(f)\right|\right]. \tag{75}
\end{aligned}$$

For Term 2, note that $f(x)\in[0,1]$ implies $L(f)\in[0,1]$. Consider the expression $Q = L(f)\sum_{i=1}^{m}\sigma_i$. If the sum is positive, then Q is maximized when $L(f)=1$. Likewise, if the sum is negative, then Q is maximized when $L(f)=0$. Hence by Massart's lemma,

$$\text{Term 2} \le \mathop{\mathbb{E}}_{\sigma}\left[\max\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\cdot 0, \frac{1}{m}\sum_{i=1}^{m}\sigma_i\cdot 1\right)\right] = \mathop{\mathbb{E}}_{\sigma}\left[\max_{a\in(\vec{0},\vec{1})}\frac{1}{m}\sum_{i=1}^{m}\sigma_i a_i\right] \le \sqrt{\frac{2\log 2}{m}}. \tag{76}$$

The result follows by combining the upper bounds on Term 1 and Term 2.