CS 7545: Machine Learning Theory Due March 1, 2023, 11:59pm

Problem Set 2 Solutions

Instructors: Jake Abernethy, Tyler LaBonte

Problem 1

Problem.

1. **Graded.** Is it possible for an ERM hypothesis $\hat{h} \in \mathcal{H}$ on a set $S \subseteq \mathcal{X}$ to have $\widehat{L}_S(\hat{h}) = 0$ but $L(\hat{h}) = 1$? Why or why not? Would this be overfitting or underfitting? What is the role of the complexity of $\mathcal{H}$ in this situation?

2. **Graded.** Let $\mathcal{H}$ be a hypothesis class, $\hat{h} \in \mathcal{H}$ be an ERM hypothesis for a sample S, and $h^\star = \arg\inf_{h\in\mathcal{H}} L(h)$. Show that

$$\mathbb{E}_{S\sim\mathcal{D}^m}[\widehat{L}_S(\hat{h})] \le L(h^\star) \le \mathbb{E}_{S\sim\mathcal{D}^m}[L(\hat{h})]. \tag{1}$$

3. **Ungraded, optional.** Here are some resources if you would like to study the proof of the no-free-lunch theorem. Most sources prove a simpler version; the one from class takes a bit more work. I highly recommend this illustrated proof. For a textbook version, see Section 5.1 "The No-Free-Lunch Theorem" in Understanding Machine Learning: From Theory to Algorithms. (Don't worry about the PAC-learning stuff in Corollary 5.2).

Solution.

1. For a space $\mathcal{X}$ with infinite cardinality (*e.g.*, the interval $[0,1]$) and a finite set S, we choose the hypothesis which labels points in S correctly but everything else incorrectly. The generalization error is 1 because S is a measure zero set in $\mathcal{X}$. This is the most extreme form of overfitting. Typically, functions like this will exist in only the most complex hypothesis classes (*e.g.*, the set of all functions), so restricting the complexity of the hypothesis class can prevent this type of overfitting.

2. By definition of ERM,

$$\mathbb{E}_{S\sim\mathcal{D}^m}[\widehat{L}_S(\hat{h})] = \mathbb{E}_{S\sim\mathcal{D}^m}[\inf_{h\in\mathcal{H}} \widehat{L}_S(h)] \le \mathbb{E}_{S\sim\mathcal{D}^m}[\widehat{L}_S(h^\star)] = L(h^\star). \tag{2}$$

By definition of $h^\star$,

$$L(h^\star) = \inf_{h\in\mathcal{H}} L(h) = \mathbb{E}_{S\sim\mathcal{D}^m}[\inf_{h\in\mathcal{H}} L(h)] \le \mathbb{E}_{S\sim\mathcal{D}^m}[L(\hat{h})]. \tag{3}$$

Problem 2

Problem. In lecture we showed the following one-sided uniform convergence generalization bound: for $\mathcal{H}$ containing functions $h : \mathcal{X} \to \{-1, 1\}$ such that $|\mathcal{H}| < \infty$ and any $\delta > 0$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$, the following holds for all $h \in \mathcal{H}$:

$$L(h) \le \widehat{L}_S(h) + \sqrt{\frac{\log|\mathcal{H}| + \log 1/\delta}{2m}}. \tag{4}$$

This bound shows that the estimation error of the empirical risk minimizer goes to zero with $1/\sqrt{m}$, which is often called the "slow rate". Interestingly, we did not use any properties of $\mathcal{H}$ besides its size. One may wonder whether there is any advantage to choosing a hypothesis class $\mathcal{H}'$ which is the same size as $\mathcal{H}$ but contains "better" functions. In fact, it turns out that if all the functions in $\mathcal{H}'$ have sufficiently low generalization error, we can achieve a "fast rate" of $1/m$.

In order to prove the fast rate bound, we need a more sophisticated concentration bound than Hoeffding's inequality. Let us state *Bernstein's inequality*: Let $Z_1, \ldots, Z_m$ be *i.i.d.* random variables with zero mean such that $|Z_i| \leq C$ and $\mathrm{Var}(Z_i) \leq D$ for all i. Then for all $\epsilon > 0$,

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m} Z_i \geq \epsilon\right] \leq \exp\left(\frac{-(m\epsilon^2)/2}{D + (C\epsilon)/3}\right). \tag{5}$$

1. **Graded.** Let $\mathcal{H}$ contain functions $h : \mathcal{X} \to \{-1, 1\}$ with $|\mathcal{H}| < \infty$. Suppose there exists a function $q : \mathbb{R} \to \mathbb{R}$ such that $L(h) \leq q(m)$ for any $h \in \mathcal{H}$ and $S \subseteq \mathcal{X}$ with $|S| = m$. Use Bernstein's inequality to prove that for any $\delta > 0$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$, the following holds for all $h \in \mathcal{H}$:

$$L(h) \leq \widehat{L}_S(h) + \sqrt{\frac{2(\log|\mathcal{H}| + \log 1/\delta)q(m)}{m}} + \frac{2(\log|\mathcal{H}| + \log 1/\delta)}{3m}. \tag{6}$$

   Hint. First, define the Z_i's and compute C and D. Then, the rest will be similar to the proof of our original bound, but with Bernstein instead of Hoeffding. It's OK if you get different constants than are listed here.

2. **Graded.** How should $q(m)$ scale in order to obtain the fast rate? It is sufficient to give an answer like $q(m) = \mathcal{O}(?)$ and explain your reasoning.

3. **Ungraded, optional.** A more general form of Bernstein's inequality holds for a very large class of distributions called *subexponential* distributions. These distributions are roughly characterized by having heavier tails than a Gaussian – they decay with e^{-x} instead of e^{-x^2} – and they come up often in machine learning theory. If you would like to learn more, read sections 2.7 and 2.8 of High-Dimensional Probability.

Solution.

1. Fix some $h \in \mathcal{H}$ and define $Z_i = L(h) - 1(h(x_i) \neq y_i)$. Note that $|Z_i| \leq 1$. We have

$$\mathbb{E}[Z_i] = L(h) - \mathbb{E}[1(h(x) \neq y)] = L(h) - L(h) = 0, \tag{7}$$

   and

$$\begin{aligned}
\mathrm{Var}(Z_i) &= \mathbb{E}[Z_i^2] - \mathbb{E}[Z_i]^2 &(8)\\
&= L(h)^2 - 2L(h)\mathbb{E}[1(h(x) \neq y)] + \mathbb{E}[1(h(x) \neq y)^2] &(9)\\
&= \mathbb{E}[1(h(x) \neq y)^2] - L(h)^2 &(10)\\
&\leq \mathbb{E}[1(h(x) \neq y)^2] &(11)\\
&= \mathbb{E}[1(h(x) \neq y)] &(12)\\
&= L(h) \leq q(m). &(13)
\end{aligned}$$

   Furthermore,

$$\frac{1}{m}\sum_{i=1}^{m} Z_i = \frac{1}{m}(mL(h) - \sum_{i=1}^{m} 1(h(x_i) \neq y_i)) = L(h) - \widehat{L}_S(h). \tag{14}$$

   Therefore, by Bernstein's inequality,

$$\Pr\left[L(h) - \widehat{L}_S(h) \geq \epsilon\right] = \exp\left(\frac{-(m\epsilon^2)/2}{q(m) + \epsilon/3}\right). \tag{15}$$

We can plug this into the proof of the original bound to see that

$$\Pr[\exists h \in \mathcal{H} : L(h) - \widehat{L}_S(h) \geq \epsilon] \leq \sum_{h \in \mathcal{H}} \Pr[L(h) - \widehat{L}_S(h) \geq \epsilon] \leq |\mathcal{H}| \exp\left(\frac{-(m\epsilon^2)/2}{q(m) + \epsilon/3}\right). \quad (16)$$

Setting the right-hand side to δ and solving for ϵ,

$$\frac{(m\epsilon^2)/2}{q(m) + \epsilon/3} = \log |\mathcal{H}| + \log 1/\delta \quad (17)$$

$$\implies \epsilon^2 - \frac{2(\log |\mathcal{H}| + \log 1/\delta)}{3m}\epsilon - \frac{2}{m}(\log |\mathcal{H}| + \log 1/\delta)q(m) = 0 \quad (18)$$

$$\implies \epsilon = \frac{(\log |\mathcal{H}| + \log 1/\delta)}{3m} + \sqrt{\left(\frac{(\log |\mathcal{H}| + \log 1/\delta)}{3m}\right)^2 + \frac{2}{m}(\log |\mathcal{H}| + \log 1/\delta)q(m)} \quad (19)$$

$$\implies \epsilon \leq \frac{2(\log |\mathcal{H}| + \log 1/\delta)}{3m} + \sqrt{\frac{2(\log |\mathcal{H}| + \log 1/\delta)q(m)}{m}}, \quad (20)$$

where in the last step we used $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$.

2. We require $q(m) = \mathcal{O}(1/m)$ so that the m in the denominator can be factored out of the root.

Problem 3

Problem. In this problem, we will prove a classical bound on the Rademacher complexity of 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 $\phi(0) = 0$ (*e.g.*, the ReLU function). Define the class of neural networks of depth $2 \leq j \leq D$ and width H with ℓ_1-bounded weights recursively as

$$\mathcal{F}_j := \left\{ x \mapsto \sum_{k=1}^{H} w_k \phi(f_k(x)) : f_k \in \mathcal{F}_{j-1}, \|w\|_1 \leq B_j \right\}. \quad (21)$$

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

1. **Graded.** Define $\mathcal{F}_1 := \{x \mapsto \langle w, x \rangle : \|w\|_1 \leq B_1\}$ and suppose $\|x_i\|_\infty \leq C$ for all $1 \leq i \leq m$. Prove that
$$\mathfrak{R}_S(\mathcal{F}_1) \leq B_1 C \sqrt{\frac{2 \log 2n}{m}}. \quad (22)$$
Hint. Use Hölder's inequality and Massart's lemma.

2. **Graded.** Prove that $\mathfrak{R}_S(\mathcal{F}_j) \leq 2LB_j \mathfrak{R}_S(\mathcal{F}_{j-1})$ for $2 \leq j \leq D$. **Hint.** Use Hölder's inequality and Talagrand's contraction lemma. You may use part **(4)** without proof.

3. **Graded.** Use parts **(1)** and **(2)** to show an upper bound on the Rademacher complexity of $\mathfrak{R}_S(\mathcal{F}_D)$. (You must use parts **(1)** and **(2)**).

4. **Ungraded, optional.** Prove that if a function class $\mathcal{G}$ contains the zero function, then
$$\mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i g(x_i) \right| \right] \leq 2\mathfrak{R}_S(\mathcal{G}). \quad (23)$$

Solution.

1. Applying Hölder's inequality,

$$\mathfrak{R}_S(\mathcal{F}_1) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}_1}\frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i)\right] \tag{24}$$

$$= \mathbb{E}_\sigma\left[\sup_{w:\|w\|_1\le B_1}\frac{1}{m}\sum_{i=1}^m \sigma_i\langle w, x_i\rangle\right] \tag{25}$$

$$= \mathbb{E}_\sigma\left[\sup_{w:\|w\|_1\le B_1}\frac{1}{m}\left\langle w, \sum_{i=1}^m \sigma_i x_i\right\rangle\right] \tag{26}$$

$$\le B_1\mathbb{E}_\sigma\left[\frac{1}{m}\left\|\sum_{i=1}^m \sigma_i x_i\right\|_\infty\right]. \tag{27}$$

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,

$$\left\|\sum_{i=1}^m \sigma_i x_i\right\|_\infty = \max_{1\le j\le n}\left|\sum_{i=1}^m \sigma_i x_i\right|_j = \max_{1\le j\le n}\left|\sum_{i=1}^m \sigma_i x_{ij}\right| = \sup_{a\in A}\sum_{i=1}^m \sigma_i a_i. \tag{28}$$

Hence,

$$\mathbb{E}_\sigma\left[\frac{1}{m}\left\|\sum_{i=1}^m \sigma_i x_i\right\|_\infty\right] = \mathbb{E}_\sigma\left[\sup_{a\in A}\frac{1}{m}\sum_{i=1}^m \sigma_i a_i\right] = \mathfrak{R}(A). \tag{29}$$

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

$$\mathfrak{R}_S(\mathcal{F}_1) \le B_1\mathfrak{R}(A) \le B_1 C\sqrt{\frac{2\log 2n}{m}}. \tag{30}$$

2. We have

$$\mathfrak{R}_S(\mathcal{F}_j) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}_j}\frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i)\right] \tag{31}$$

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

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

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

$$\mathfrak{R}_S(\mathcal{F}_j) \le \mathbb{E}_\sigma\left[\sup_{\substack{\|w\|_1\le B_j\\ f_k\in\mathcal{F}_{j-1}}}\frac{1}{m}\|w\|_1\max_{1\le k\le H}\left|\sum_{i=1}^m \sigma_i\phi(f_k(x_i))\right|\right] \tag{34}$$

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

$$= 2B_j\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}_{j-1}}\frac{1}{m}\sum_{i=1}^m \sigma_i\phi(f(x_i))\right]. \tag{36}$$

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

$$\mathfrak{R}_S(\mathcal{F}_j) \le 2B_j \mathbb{E}_\sigma \left[\sup_{a \in A} \frac{1}{m} \sum_{i=1}^m \sigma_i \phi(a_i) \right] = 2B_j \mathfrak{R}(\phi(A)) \le 2LB_j \mathfrak{R}(A) = 2LB_j \mathfrak{R}_S(\mathcal{F}_{j-1}). \quad (37)$$

3. Solving the recurrence from part **(2)** and substituting the answer from part **(1)** gives

$$\mathfrak{R}_S(\mathcal{F}_D) \le \prod_{j=2}^D 2LB_j \mathfrak{R}_S(\mathcal{F}_{j-1}) = (2L)^{D-1} \prod_{j=2}^D B_j \cdot \mathfrak{R}_S(\mathcal{F}_1) = (2L)^{D-1} \prod_{j=1}^D B_j \cdot C \sqrt{\frac{2 \log 2n}{m}}. \quad (38)$$

4. Let $A \subseteq \mathbb{R}$ such that $0 \in A$. Then,

$$\sup_{a \in A} |a| = \max(\sup_{a \in A} a, -\inf_{a \in A} a) \le \sup_{a \in A} a - \inf_{a \in A} a, \quad (39)$$

where $0 \in A$ is sufficient for the maximands to be non-negative. Therefore, since $\mathcal{G}$ contains the zero function, we have

$$\begin{aligned}
\mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \left| \frac{1}{m} \sum_{i=1}^m \sigma_i g(x_i) \right| \right] &\le \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^m \sigma_i g(x_i) - \inf_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^m \sigma_i g(x_i) \right] && (40) \\
&= \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^m \sigma_i g(x_i) \right] + \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^m -\sigma_i g(x_i) \right] && (41) \\
&= 2\mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^m \sigma_i g(x_i) \right] && (42) \\
&= 2\mathfrak{R}_S(\mathcal{G}), && (43)
\end{aligned}$$

where we used the fact that σ_i and $-\sigma_i$ have the same distribution.

Problem 4

Problem. Suppose $A \subseteq \mathbb{R}^m$.

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

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

3. **Graded.** In lecture we proved the following one-sided uniform convergence generalization bound: for $\mathcal{F}$ containing functions $f : \mathcal{X} \to [0, 1]$ and any $\delta > 0$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$, the following holds for all $f \in \mathcal{F}$:

$$L(f) \le \widehat{L}_S(f) + 2\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2m}}. \quad (44)$$

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 **(1)** and **(2)** to prove one. (You must use parts **(1)** and **(2)**).

4. **Challenge, optional, 1 point extra credit.** Let $S \sim \mathcal{D}^m$ and suppose $\mathcal{F}$ contains functions $f : \mathcal{X} \to [0, 1]$. Prove the symmetrization lower bound, also called the desymmetrization inequality:

$$\frac{1}{2}\mathfrak{R}(\mathcal{F}) - \sqrt{\frac{\log 2}{2m}} \le \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left| L(f) - \widehat{L}_S(f) \right| \right]. \quad (45)$$

Solution.

1. Using linearity of expectation,

$$\begin{aligned}
\mathfrak{R}(A+b) &= \mathbb{E}_\sigma\left[\sup_{a'\in(A+b)} \frac{1}{m}\sum_{i=1}^m \sigma_i a'_i\right] && (46)\\
&= \mathbb{E}_\sigma\left[\sup_{a\in A} \frac{1}{m}\sum_{i=1}^m \sigma_i (a_i+b_i)\right] && (47)\\
&= \mathbb{E}_\sigma\left[\sup_{a\in A} \frac{1}{m}\sum_{i=1}^m \sigma_i a_i + \frac{1}{m}\sum_{i=1}^m \sigma_i b_i\right] && (48)\\
&= \mathbb{E}_\sigma\left[\sup_{a\in A} \frac{1}{m}\sum_{i=1}^m \sigma_i a_i\right] + \mathbb{E}_\sigma\left[\frac{1}{m}\sum_{i=1}^m \sigma_i b_i\right] && (49)\\
&= \mathbb{E}_\sigma\left[\sup_{a\in A} \frac{1}{m}\sum_{i=1}^m \sigma_i a_i\right] + \frac{1}{m}\sum_{i=1}^m b_i \mathbb{E}_{\sigma_i}[\sigma_i] && (50)\\
&= \mathbb{E}_\sigma\left[\sup_{a\in A} \frac{1}{m}\sum_{i=1}^m \sigma_i a_i\right] && (51)\\
&= \mathfrak{R}(A). && (52)
\end{aligned}$$

2. We have

$$\mathfrak{R}(cA) = \mathbb{E}_\sigma\left[\sup_{a'\in(cA)} \frac{1}{m}\sum_{i=1}^m \sigma_i a'_i\right] = \mathbb{E}_\sigma\left[\sup_{a\in A} \frac{c}{m}\sum_{i=1}^m \sigma_i a_i\right]. \quad (53)$$

If $c \geq 0$ then

$$\mathbb{E}_\sigma\left[\sup_{a\in A} \frac{c}{m}\sum_{i=1}^m \sigma_i a_i\right] = \mathbb{E}_\sigma\left[\sup_{a\in A} \frac{|c|}{m}\sum_{i=1}^m \sigma_i a_i\right] = |c|\mathbb{E}_\sigma\left[\sup_{a\in A} \frac{1}{m}\sum_{i=1}^m \sigma_i a_i\right]. \quad (54)$$

Otherwise if $c < 0$ then

$$\mathbb{E}_\sigma\left[\sup_{a\in A} \frac{c}{m}\sum_{i=1}^m \sigma_i a_i\right] = \mathbb{E}_\sigma\left[\sup_{a\in A} \frac{-|c|}{m}\sum_{i=1}^m \sigma_i a_i\right] = |c|\mathbb{E}_\sigma\left[\sup_{a\in A} \frac{1}{m}\sum_{i=1}^m -\sigma_i a_i\right]. \quad (55)$$

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

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

$$\begin{aligned}
L(g) - \widehat{L}_S(g) &= \mathbb{E}_{x\sim\mathcal{D}} g(x) - \frac{1}{m}\sum_{i=1}^m g(x_i) && (56)\\
&= \mathbb{E}_{x\sim\mathcal{D}}[1 - f(x)] - \frac{1}{m}\sum_{i=1}^m (1 - f(x_i)) && (57)\\
&= (1 - \mathbb{E}_{x\sim\mathcal{D}} f(x)) - (1 - \frac{1}{m}\sum_{i=1}^m f(x_i)) && (58)\\
&= \frac{1}{m}\sum_{i=1}^m f(x_i) - \mathbb{E}_{x\sim\mathcal{D}} f(x) && (59)\\
&= \widehat{L}_S(f) - L(f). && (60)
\end{aligned}$$

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

$$\sup_{f\in\mathcal{F}} L(f) - \widehat{L}_S(f) \le 2\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{\log 1/\delta_1}{2m}}, \tag{61}$$

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

$$\sup_{f\in\mathcal{F}} \widehat{L}_S(f) - L(f) = \sup_{g\in\mathcal{G}} L(g) - \widehat{L}_S(g) \le 2\mathfrak{R}(\mathcal{G}) + \sqrt{\frac{\log 1/\delta_2}{2m}}. \tag{62}$$

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

$$\sup_{f\in\mathcal{F}} \left|L(f) - \widehat{L}_S(f)\right| \le 2\max(\mathfrak{R}(\mathcal{F}), \mathfrak{R}(\mathcal{G})) + \sqrt{\frac{\log 2/\delta}{2m}} = 2\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{\log 2/\delta}{2m}}. \tag{63}$$

4. We have

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

Introducing a ghost sample S',

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

By symmetrization,

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

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,

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

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

Problem 5

Problem. 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, $\Pi_{\mathcal{H}}(m) \le |\mathcal{Y}|^m$ (analogous to 2^m in the original setup).

1. **Graded.** 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
$$\Pi_{\mathcal{H}_3}(m) = \Pi_{\mathcal{H}_1}(m)\cdot\Pi_{\mathcal{H}_2}(m). \quad (77)$$
2. **Graded.** 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
$$\Pi_{\mathcal{H}_3}(m) \le \Pi_{\mathcal{H}_1}(m)\cdot\Pi_{\mathcal{H}_2}(m). \quad (78)$$
3. **Ungraded, optional.** Prove that **(2)** is tight, *i.e.*, exhibit $\mathcal{X},\mathcal{Y}_1,\mathcal{Y}_2,\mathcal{H}_1,\mathcal{H}_2,m$ such that $\Pi_{\mathcal{H}_3}(m) = \Pi_{\mathcal{H}_1}(m)\cdot\Pi_{\mathcal{H}_2}(m)$. **Hint.** You can take $|\mathcal{X}| = m = 1$.

Solution.

1. For any $S = ((x_1,x_1'),\ldots,(x_m,x_m')) \subseteq \mathcal{X}\times\mathcal{X}$,
$$\begin{aligned}
|\mathcal{H}_3|_S| &= |\{(h_3(x_1,x_1'),\ldots,h_3(x_m,x_m')) : h_3 \in \mathcal{H}_3\}| && (79)\\
&= |\{((h_1(x_1),h_2(x_1')),\ldots,(h_1(x_m),h_2(x_m'))) : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}| && (80)\\
&= |\{(h_1(x_1),\ldots,h_1(x_m)) : h_1 \in \mathcal{H}_1\}|\cdot|\{(h_2(x_1'),\ldots,h_2(x_m')) : h_2 \in \mathcal{H}_2\}| && (81)\\
&= |\mathcal{H}_1|_S|\cdot|\mathcal{H}_2|_S| && (82)
\end{aligned}$$
Hence $\Pi_{\mathcal{H}_3}(m) = \Pi_{\mathcal{H}_1}(m)\cdot\Pi_{\mathcal{H}_2}(m)$.
2. For any $S = (x_1,\ldots,x_m) \subseteq \mathcal{X}$,
$$\begin{aligned}
\mathcal{H}_3|_S &= \{(h_3(x_1),\ldots,h_3(x_m)) : h_3 \in \mathcal{H}_3\} && (83)\\
&= \{(h_2(h_1(x_1)),\ldots,h_2(h_1(x_m))) : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\} && (84)\\
&= \bigcup_{u\in\mathcal{H}_1|_S} \{(h_2(u_1),\ldots,h_2(u_m)) : h_2 \in \mathcal{H}_2\}. && (85)
\end{aligned}$$
Thus,
$$\begin{aligned}
|\mathcal{H}_3|_S| &\le \sum_{u\in\mathcal{H}_1|_S} |\{(h_2(u_1),\ldots,h_2(u_m)) : h_2 \in \mathcal{H}_2\}| && (86)\\
&\le \sum_{u\in\mathcal{H}_1|_S} \Pi_{\mathcal{H}_2}(m) && (87)\\
&= |\mathcal{H}_1|_S|\cdot\Pi_{\mathcal{H}_2}(m) && (88)\\
&\le \Pi_{\mathcal{H}_1}(m)\cdot\Pi_{\mathcal{H}_2}(m). && (89)
\end{aligned}$$
Hence $\Pi_{\mathcal{H}_3}(m) \le \Pi_{\mathcal{H}_1}(m)\cdot\Pi_{\mathcal{H}_2}(m)$.
3. Take $\mathcal{X} = \{a\}, \mathcal{Y}_1 = \{c,d\}, \mathcal{Y}_2 = \{e,f\}$. Define $\mathcal{H}_1 = \{a \mapsto c, a \mapsto d\}$ and $\mathcal{H}_2 = \{(c \mapsto e, d \mapsto f)\}$. Now $\mathcal{H}_3 = \{a \mapsto e, b \mapsto f\}$. However, $\Pi_{\mathcal{H}_1}(1) = 2$ and $\Pi_{\mathcal{H}_2}(1) = 1$, but $\Pi_{\mathcal{H}_3}(1) = 2$.

Problem 6

Problem.

1. **Graded.** What is the VC-dimension of a union of k intervals on the real line?
2. **Graded.** What is the VC-dimension of axis-aligned hyperrectangles in $\mathbb{R}^n$?
3. **Graded.** 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 $\mathbb{R}^n$.
4. **Challenge, optional, 1 extra credit point.** Prove the best lower bound you can on the VC-dimension of simplices in $\mathbb{R}^n$. You will receive the extra credit point if you either (i) prove a lower bound of $\Omega(n)$ and show a reasonable attempt at improving it, or (ii) prove a lower bound better than $\Omega(n)$.

Solution.

1. 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 $1, -1, 1, -1, \ldots, 1$.
2. 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.
3. 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 $\mathrm{VC}(\mathcal{H}) = d$ 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

   $$\mathrm{VC}(\mathcal{S}) = \mathcal{O}((n+1)^2 \log(n+1)^2) = \mathcal{O}(n^2 \log n). \tag{90}$$
4. 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 $\Theta(n^2 \log n)$.