CPSC 532D: Assignment 3 – due Tuesday, 8 Nov 2022, **11:59pm**

As before: use LaTeX, either with the template I give or your own document if you prefer.

You can do this with a partner if you'd like (there's a "find a group" post on Piazza), but please **make sure you understand everything you're submitting** – don't just split an assignment in half. If you do parts of the assignment with a partner and parts separately, submit separate solutions, and say in each part you did together who you did it with.

If you look stuff up anywhere other than in **SSBD, MRT, Telgarsky, or Wainwright**, **cite your sources**: just say in the answer to that question where you looked. If you ask anyone else for help, **cite that too**. Please do not look at solution manuals / search for people proving the things we're trying to prove / etc. If you accidentally come across a solution while looking for something related, still write the argument up in your own words, link to wherever you found it, and be clear about what happened.

1 Monotonicity [10 points]

(a) [3 points] Prove that if $\mathcal{H} \subseteq \mathcal{H}'$, then $\text{VCdim}(\mathcal{H}) \leq \text{VCdim}(\mathcal{H}')$.

Answer: TODO

(b) [3 points] Prove that if $\mathcal{H} \subseteq \mathcal{H}'$, then $\text{Rad}(\mathcal{H}|_S) \leq \text{Rad}(\mathcal{H}'|_S)$.

Answer: TODO

(c) [4 points] Comment on how we should expect parts (a) and (b) to affect the generalization loss of running ERM in $\mathcal{H}$ versus $\mathcal{H}'$, that is, $L_{\mathcal{D}}(\text{ERM}_{\mathcal{H}}(S))$ versus $L_{\mathcal{D}}(\text{ERM}_{\mathcal{H}'}(S))$ for a fixed n. What other factors are at play?

Answer: TODO

2 Threshold functions [20 points]

Recall our old friend, the class of threshold functions on $\mathbb{R}$:

$$\mathcal{H} = \{x \mapsto \mathbb{1}(x \leq \theta) : \theta \in \mathbb{R}\}.$$

We showed in class (notes 6.1.1) that the VC dimension of $\mathcal{H}$ is 1: it can shatter a set of size one (a single point), but it cannot shatter any set of size two (since it can't label the left point 0 and the right point 1).

(a) [5 points] Use Sauer-Shelah (Lemma 6.8) and the (nicer) Corollary 6.9 to give two upper bounds on the growth function $\Pi_{\mathcal{H}}(n)$.

Answer: TODO

(b) [5 points] Directly derive the exact value of the growth function $\Pi_{\mathcal{H}}$ from its definition. How tight are the upper bounds from part (a)?

Answer: TODO

(c) [5 points] Plug the previous parts in to upper bound the empirical Rademacher complexity $\text{Rad}(\mathcal{H}|_S)$ for an S containing n distinct real numbers. *You should give multiple bounds here, one per distinct bound from the previous parts.*

Answer: TODO

(d) [5 points] Give the asymptotic value of $\text{Rad}(\mathcal{H}|_S)$ for an S containing n distinct real numbers. Your answer might look something like "$\text{Rad}(\mathcal{H}|_S) = 7n + \mathcal{O}(1)$," with a justification. To be clear, this means that $7n - a_n \leq \text{Rad}(\mathcal{H}|_S) \leq 7n + a_n$ for some $a_n = \mathcal{O}(1)$. How does it compare to the bound from part (c)?

Hint: Imagine playing a (pretty boring) betting game where you bet \$1 whether a coin I'm flipping comes up heads or tails. Since all physical coin flips are unbiased, you have a 50-50 shot of getting it right. The distribution of how much money I owe you is known as a simple random walk. *Your expected winnings at any time t are always 0 (it's the sum of a bunch of mean-zero variables), but if you play for a while and then go back and conveniently "forget" the record of flips after a certain point, the expected* maximum *value achieved at any point during a walk of length n turns out to be $\sqrt{\frac{2n}{\pi}} - \frac{1}{2} + \mathcal{O}(n^{-\frac{1}{2}})$, per (4) and (7) of the linked paper.*

Answer: TODO

3 Piecewise-constant functions [20 + 10 challenge + 5 bonus points]

Let $a = (a_1, a_2, \ldots, a_k, 0, 0, \ldots)$ be an eventually-zero sequence with entries $a_i \in \{0, 1\}$. Then define a hypothesis $h_a : \mathbb{R}_{>0} \to \{0, 1\}$ by

$$h_a(x) = a_{\lceil x \rceil} = \begin{cases} a_1 & \text{if } 0 < x \leq 1 \\ a_2 & \text{if } 1 < x \leq 2 \\ \vdots & \end{cases}.$$

Consider the hypothesis class of all such functions: $\mathcal{H} = \{h_a : \forall i \in \mathbb{N}, a_i \in \{0, 1\} \text{ and } a \text{ is eventually zero}\}$.

(a) [5 points] Show $\text{VCdim}(\mathcal{H}) = \infty$.

Answer: TODO

(b) [8 points] Give an example of a "nontrivial" distribution $\mathcal{D}_x$ on $\mathbb{R}_{>0}$ where, for some $n < \text{VCdim}(\mathcal{H})$, samples $S_x \sim \mathcal{D}_x^n$ have probability zero of being shattered by $\mathcal{H}$. *"Nontrivial" is of course a judgement call, but as an example, point masses at a single point are trivial, while, say, truncated normal distributions are not trivial.* Thus prove that, for any $\mathcal{D}$ with this x marginal $\mathcal{D}_x$, ERM over $\mathcal{H}$ (ε, δ)-competes with the best hypothesis in $\mathcal{H}$ for that $\mathcal{D}$ with some finite sample complexity, rather than the infinite sample complexity that would be implied by the VC bound.

Answer: TODO

(c) [7 points] Write $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \cdots$ for each $\mathcal{H}_k$ of a finite VC dimension, and write down an explicit SRM algorithm that nonuniformly learns $\mathcal{H}$. *By "an explicit algorithm," I mean to expand out things like the uniform convergence bound for $\mathcal{H}_k$; it's okay to write something as an argmin over $\mathcal{H}$ (like in notes (7.2) if you say what k_h is for a given h and give the value of the Rademacher complexity term), or to just appeal to the SRM algorithm pseudocode from the notes (as long as you say what's in each $\mathcal{H}_k$, what the ε_k functions are, and how to compute the stopping condition).*

Answer: TODO

(d) [5 bonus points] **Bonus question:** Suppose that instead of eventually-zero sequences, we allowed all possible sequences a, e.g. the a that infinitely alternates between 0 and 1 could be an option. Is this bigger $\mathcal{H}'$ nonuniformly learnable?

Answer: TODO

(e) [7 points] **Challenge question:** Prove that, for *any* $\mathcal{D}_x$, $\mathbb{E}_{S_x \sim \mathcal{D}_x^n} \text{Rad}(\mathcal{H}|_{S_x}) \to 0$ as $n \to \infty$. *Hint: One way to do it (there's probably more than one): first, reduce to the "ceiled" distribution over $\mathbb{N}$ instead of over $\mathbb{R}_{>0}$, and use Corollary 4.8 to reduce to a bound in terms of $\mathbb{E}\, M/n$, where M is the number of unique integers you've seen in S. Then prove that $\mathbb{E}\, M = o(n)$ for any distribution over $\mathbb{N}$.*

Answer: TODO

(f) [3 points] **Challenge question:** An absentminded professor made the following argument on the *final exam* for a course:

> *If a hypothesis class has $\mathbb{E}_{S_x \sim \mathcal{D}_x^n} \text{Rad}(\mathcal{H}|_{S_x}) \to 0$ for all $\mathcal{D}_x$, then for all realizable $\mathcal{D}$,*
>
> $$L_{\mathcal{D}}(\hat{h}_S) \leq \mathop{\mathbb{E}}_{S_X \sim \mathcal{D}_x^n} \text{Rad}(\mathcal{H}|_{S_x}) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}} \to 0.$$
>
> *Thus, by the "fundamental theorem of statistical learning," $\mathcal{H}$ must have finite VC dimension.*

Clearly this argument is wrong, since it puts parts (a) and (e) in contradiction. What was her mistake?

Answer: TODO

4 Generalization bound for a simple neural network [40 points]

Based on MRT exercise 3.11.

Here is a class of neural networks mapping $\mathbb{R}^d$ to $\mathbb{R}$, with one hidden layer of width m and activations $\phi : \mathbb{R} \to \mathbb{R}$ a 1-Lipschitz function (e.g. the ReLU or sigmoid function):

$$\mathcal{H} = \left\{ x \mapsto \sum_{j=1}^{m} w_j \phi(u_j^\mathsf{T} x) : \|w\|_1 \le \nu,\ \|u_j\|_2 \le \Lambda \text{ for each } j \in [m] \right\}.$$

We haven't used this in this class yet, but recall that $\|w\|_1 = \sum_{j=1}^m |w_j|$. Λ and ν are hyperparameters defining how complex the class is allowed to be.

(a) [10 points] Show that $\mathrm{Rad}(\mathcal{H}|_{S_x}) = \frac{\nu}{n} \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}} \left[\sup_{\|u\|_2 \le \Lambda} \left| \sum_{i=1}^n \epsilon_i \phi(u^\mathsf{T} x_i) \right| \right]$.

Answer: TODO

This variant of Talagrand's contraction lemma works for any $\mathcal{H}$ and ρ-Lipschitz function Φ:[1]

$$\frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}} \left[\sup_{h \in \mathcal{H}} \left| \sum_{i=1}^n \epsilon_i \Phi(h(x_i)) \right| \right] \le \frac{2\rho}{n} \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}} \left[\sup_{h \in \mathcal{H}} \left| \sum_{i=1}^n \epsilon_i h(x_i) \right| \right] + \frac{|\Phi(0)|}{\sqrt{n}}. \tag{*}$$

(b) [10 points] Use (*) to upper bound $\mathrm{Rad}(\mathcal{H}|_{S_x})$ in terms of $\mathrm{Rad}(\mathcal{H}'|_{S_x})$, for

$$\mathcal{H}' = \left\{ x \mapsto s\,(u^\mathsf{T} x) : \|u\|_2 \le \Lambda,\ s \in \{-1, +1\} \right\} = \left\{ x \mapsto u^\mathsf{T} x : \|u\|_2 \le \Lambda \right\}.$$

Answer: TODO

(c) [10 points] Bound $\mathbb{E}_{S \sim \mathcal{D}_x^n} \mathrm{Rad}(\mathcal{H}'|_S)$, and thereby $\mathbb{E}_{S \sim \mathcal{D}_x^n} \mathrm{Rad}(\mathcal{H}|_S)$. You'll need an assumption on $\|x\|$ from $\mathcal{D}_x$ to do this; be clear what you're assuming.

Answer: TODO

(d) [10 points] Give an expression for ε such that $\Pr_{S \sim \mathcal{D}^n} \left(\sup_{h \in \mathcal{H}} (L_\mathcal{D}(h) - L_S(h)) \le \varepsilon \right) \ge 1 - \delta$, where you'll need to choose some loss function ℓ for inside L, and be clear about any additional assumptions. Please choose a *specific* loss function ℓ; if it's not one we've used in class, be sure to justify it as a "'reasonable" loss function.

There are several valid approaches here. Try to pick a reasonable set of assumptions and loss function; something like "all the ys are equal to 0" is not reasonable. Your bound should have $\varepsilon \to 0$ as $n \to \infty$ with all other parameters fixed.

Answer: TODO

[1] The MRT exercise claims that, for any $\mathcal{H}$ and ρ-Lipschitz Φ,

$$\frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}} \left[\sup_{h \in \mathcal{H}} \left| \sum_{i=1}^n \epsilon_i \Phi(h(x_i)) \right| \right] \le \frac{\rho}{n} \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}} \left[\sup_{h \in \mathcal{H}} \left| \sum_{i=1}^n \epsilon_i h(x_i) \right| \right]. \tag{wrong}$$

This would be true if you dropped the absolute value bars, but the version as stated is not true. For a counterexample, consider the (not very interesting) hypothesis class $\mathcal{H} = \{x \mapsto 0\}$ and the function $\Phi(x) = C$ which, as it ignores its argument, is L-Lipschitz for *any* $L \ge 0$. The LHS of (wrong) is $|C|$ times $\mathbb{E}_\epsilon \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \right| > 0$. The RHS of (wrong), though, is exactly 0, since $h(x_i) = 0$ for all x_i. This example also shows that assuming a symmetric $\mathcal{H}$ would not be enough to fix (wrong): you would need $\Phi \circ \mathcal{H}$ to be symmetric as well.