CPSC 532D, Fall 2023: Assignment 3
due Friday November 10th, **11:59 pm**

Use LaTeX, like usual.

You can do this with a partner if you'd like (there's a "find a group" post on Piazza). If so, **do not just split the questions up**; if you hand in an assignment with your name on it, you're pledging that you participated in and understand all of the solutions. (If you work with a partner on some problems and then end up doing some of them separately, hand in separate answers and put a note in each question saying whether you did it with a partner or not.)

*If you look stuff up anywhere other than in SSBD or MRT, **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. Also, please do* not *ask ChatGPT or similar models. It's okay to talk to others outside your group about general strategies – if so, just say who and for which questions – but **not** to sit down and do the assignment together.*

Submit your answers as a single PDF on Gradescope: here's the link. Make sure to use the Gradescope group feature if you're working in a group. *You'll be prompted to mark where each question is in your PDF; make sure you mark all relevant pages for each part (which saves a surprising amount of grading time).*

Please **put your name(s) on the first page** as a backup, just in case. If something goes wrong, you can also email your assignment to me directly (`dsuth@cs.ubc.ca`).

1 Monotonicity and model selection [20 points]

[1.1] [5 points] Prove that if $\mathcal{H} \subseteq \mathcal{H}'$, then $\mathrm{VCdim}(\mathcal{H}) \leq \mathrm{VCdim}(\mathcal{H}')$.

Answer: TODO

[1.2] [5 points] Prove that if $\mathcal{H} \subseteq \mathcal{H}'$, then $\mathrm{Rad}(\mathcal{H}|_S) \leq \mathrm{Rad}(\mathcal{H}'|_S)$.

Answer: TODO

[1.3] [5 points] Comment on how we should expect Questions [1.1] and [1.2] to affect the generalization loss of running ERM in $\mathcal{H}$ versus $\mathcal{H}' \supseteq \mathcal{H}$, that is, $L_{\mathcal{D}}(\mathrm{ERM}_{\mathcal{H}}(S))$ versus $L_{\mathcal{D}}(\mathrm{ERM}_{\mathcal{H}'}(S))$ for a fixed sample size m. What other factors are relevant to that comparison?

Answer: TODO

[1.4] [5 points] For any $\mathcal{H}$, show that

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m}[L_S(\mathrm{ERM}_{\mathcal{H}}(S))] \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \leq \mathop{\mathbb{E}}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(\mathrm{ERM}_{\mathcal{H}}(S))].$$

Answer: TODO

2 Rademacher complexity of weirder linear classes [20 points]

Consider $\mathcal{H}_{\|\cdot\| \le B} = \{x \mapsto w \cdot x : \|w\| \le B\}$, where throughout this question $\|w\|$ denotes a *generic* vector norm of w, not necessarily the Euclidean norm $\|w\|_2$. For example, we could use $\|w\|_1 = \sum_{j \in [d]} |w_j|$, $\|w\|_\infty = \max_{j \in [d]} |w_j|$, or $\|w\|_S = \sqrt{w^\mathsf{T} S w}$.

The *dual norm* of a norm $\|\cdot\|$ is given by

$$\|v\|^* = \sup_{\|w\| \le 1} v \cdot w.$$

For instance, for the Euclidean norm $\|w\|_2$, we have

$$\|v\|_2^* = \sup_{\|w\|_2 \le 1} v \cdot w \le \sup_{\|w\|_2 \le 1} \|v\|_2 \|w\|_2 = \|v\|_2,$$

using Cauchy-Schwarz; the inequality is actually an equality, achieved by picking $w = v / \|v\|_2$.

More generally, Hölder's inequality shows the dual norm of $\|w\|_p = (\sum_{j=1}^d |w_j|^p)^{1/p}$ is $\|\cdot\|_q$, where $\frac{1}{p} + \frac{1}{q} = 1$. Thus the dual of $\|\cdot\|_2$ is still $\|\cdot\|_2$, but also the dual of $\|\cdot\|_1$ is $\|\cdot\|_\infty$, and vice versa.

Consider for a general norm the function class

$$\mathcal{H}_{\|\cdot\| \le B} = \{x \mapsto x \cdot w : \|w\| \le B\}.$$

Recall that we bounded the Rademacher complexity of $\mathcal{H}_{\|\cdot\|_2 \le B}$ in Section 3.2 of the Rademacher notes.

[2.1] [5 points] Follow the same strategy to show that

$$\mathrm{Rad}(\mathcal{H}_{\|\cdot\| \le B}|_{S_x}) = \frac{B}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left\| \sum_{i=1}^m \sigma_i x_i \right\|^* .$$

Answer: TODO

We can use this result, as in SSBD Lemma 26.11, to see that for $\mathcal{H}_{\|\cdot\|_1 \le B}$ (corresponding to Lasso),

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left\| \sum_{i=1}^m \sigma_i x_i \right\|_\infty &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \max \left(\left| \left(\sum_{i=1}^m \sigma_i x_i \right)_1 \right|, \dots, \left| \left(\sum_{i=1}^m \sigma_i x_i \right)_d \right| \right) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \max \left(\sum_i \sigma_i (x_i)_1, \sum_i \sigma_i (-x_i)_1, \dots, \sum_i \sigma_i (x_i)_d, \sum_i \sigma_i (-x_i)_d, \right) \\
&= \mathrm{Rad}\left(\left\{ \left((x_1)_1, \dots, (x_m)_1 \right), \left((-x_1)_1, \dots, (-x_m)_1 \right), \dots, \left((x_1)_d, \dots, (x_m)_d \right), \left((-x_1)_d, \dots, (-x_m)_d \right) \right\} \right).
\end{aligned}$$

This last expression is the Rademacher complexity of a set of size $2d$. If $|(x_i)_j| \le C$ for all $i \in [m]$, $j \in [d]$, then each vector in this set has Euclidean norm at most $\sqrt{\sum_{i=1}^m C^2} = C\sqrt{m}$; thus, applying Massart's finite class lemma (Lemma 1 from the VC notes) gives that $\mathrm{Rad}(\mathcal{H}_{\|\cdot\|_1 \le B}|_{S_x}) \le BC\sqrt{2\log(2d)/m}$.

[2.2] [5 points] Bound $\mathrm{Rad}(\mathcal{H}_{\|\cdot\|_p \le B}|_{S_x})$ in terms of $\frac{1}{m}\sum_{i=1}^m \|x_i\|_2^2$ for general $p \ge 1$, with a bound that goes to zero as $m \to \infty$.

Hint: If $0 < a < b$, then Hölder's inequality implies $\|x\|_b \le \|x\|_a \le d^{\frac{1}{a} - \frac{1}{b}} \|x\|_b$ for $x \in \mathbb{R}^d$.

(This isn't the most natural bound; there should be one in terms of $\|x_i\|_q$ with $\frac{1}{p} + \frac{1}{q} = 1$, but honestly I couldn't figure it out right away.)

Answer: TODO

[2.3] [10 points] The *Mahalanobis norm* is $\|x\|_S = \sqrt{x^\mathsf{T} S x}$ for a strictly positive-definite matrix S. Show that $\|x\|_S^* = \|x\|_{S^{-1}}$, and bound $\operatorname{Rad}(\mathcal{H}_{\|\cdot\|_S \le B}|_{S_x})$ in terms of $\frac{1}{m}\sum_{i=1}^m \|x_i\|_{S^{-1}}$, with a bound that goes to zero as $m \to \infty$.

Hint: Recall that if S is strictly positive definite, there is a symmetric matrix $S^{\frac{1}{2}}$ such that $S = S^{\frac{1}{2}} S^{\frac{1}{2}}$.

Answer: TODO

3 Threshold functions [20 points]

This question is about the class of threshold functions on $\mathbb{R}$:

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

We showed in class (VC notes, section 4.1.1) that the VCdim($\mathcal{H}$) = 1: it can shatter a single point, but it cannot shatter any set of size two (since it can't label the left point 1 and the right point 0).

[3.1] [5 points] Use Sauer-Shelah (Lemma 11 in the notes), and also the simpler Corollary 9, to give two upper bounds on the growth function $\Gamma_{\mathcal{H}}(n)$.

Answer: TODO

[3.2] [5 points] Directly derive the exact value of the growth function $\Pi_{\mathcal{H}}$ from its definition. How tight are the upper bounds from Question [3.1]?

Answer: TODO

[3.3] [5 points] Plug the previous parts in to upper bound $\mathrm{Rad}(\mathcal{H}|_{S_x})$ for an S containing m distinct real numbers. *You should give multiple bounds here, one per distinct bound from the previous parts.*

Answer: TODO

[3.4] [5 points] Give the asymptotic value of $\mathrm{Rad}(\mathcal{H}|_{S_x})$ for an S_x containing m distinct real numbers. Your answer might look something like "$\mathrm{Rad}(\mathcal{H}|_{S_x}) = 7m + \mathcal{O}(1)$," with a justification. To be clear, this means that $7m - a_n \leq \mathrm{Rad}(\mathcal{H}|_{S_x}) \leq 7m + a_n$ for some $a_m = \mathcal{O}(1)$. How does it compare to the bound from Question [3.3]?

Hint: Imagine playing a (pretty boring) betting game where you bet \$1 whether a coin I'm flipping comes up heads or tails, with even odds. 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). If we play for a while, and then you conveniently "lose" the records of what happened after some time t that just so happens to be the best possible time for you to have forgotten, you'll probably be able to win some money: the expected* maximum *value achieved at any point during a simple random walk of length m turns out to be $\sqrt{\frac{2m}{\pi}} - \frac{1}{2} + \mathcal{O}(m^{-\frac{1}{2}})$. (This is from equations (4) and (7) of the linked paper.)*

Answer: TODO

4 Piecewise-constant functions [30 points + 4 challenge points]

Let $a = (a_1, a_2, \dots, a_k, 0, 0, \dots)$ 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: $\mathcal{H} = \{h_a : \forall i \in \mathbb{N}, a_i \in \{0, 1\} \text{ and } a \text{ is eventually zero}\}$. We'll use the 0-1 loss in this question.

[4.1] [5 points] Show $\mathrm{VCdim}(\mathcal{H}) = \infty$.

Answer: TODO

[4.2] [10 points] Give an example of a continuous distribution $\mathcal{D}_x$ on (a subset of) $\mathbb{R}_{>0}$ where, for some $m < \mathrm{VCdim}(\mathcal{H})$, samples $S_x \sim \mathcal{D}_x^m$ have probability zero of being shattered by $\mathcal{H}$. 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

[4.3] [10 points] Write $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \cdots$, where each $\mathcal{H}_k$ has 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 equation (2) of the SRM notes, 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

[4.4] [2 challenge points] **Challenge 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. Prove that this bigger $\mathcal{H}'$ is *not* nonuniformly learnable. *This implies a sort of no-free-lunch theorem for nonuniform learnability.*

Hint: Try a diagonalization argument.

Answer: TODO

[4.5] [2 challenge points] **Challenge question:** Prove that, for *any* $\mathcal{D}_x$, $\mathbb{E}_{S_x \sim \mathcal{D}_x^m} \mathrm{Rad}(\mathcal{H}|_{S_x}) \to 0$ as $m \to \infty$.

Hint: One way to do it (there's probably more than one): first, reduce to the "ceiled" distribution over $\mathbb{N}$ instead of over $\mathbb{R}_{>0}$. Then, letting Q_S denote the number of unique *integers you've seen in your sample, get a bound in terms of $\mathbb{E}\, Q_S/m$. Then prove that $\mathbb{E}\, Q_S = o(m)$ for any distribution over $\mathbb{N}$.*

Answer: TODO

[4.6] [5 points] An absentminded professor made the following argument on the *final exam* for a course:

If a hypothesis class has $\mathbb{E}_{S_x \sim \mathcal{D}_x^m} \mathrm{Rad}(\mathcal{H}|_{S_x}) \to 0$ for all $\mathcal{D}_x$, then for all realizable $\mathcal{D}$,

$$L_{\mathcal{D}}(\hat{h}_S) \leq \mathop{\mathbb{E}}_{S_X \sim \mathcal{D}_x^m} \mathrm{Rad}(\mathcal{H}|_{S_x}) + \sqrt{\frac{1}{2m} \log \frac{1}{\delta}} \to 0.$$

Thus, by the "fundamental theorem of statistical learning," $\mathcal{H}$ must have finite VC dimension.

Clearly this argument is wrong, since it puts Questions [4.1] and [4.5] in contradiction. What was her mistake?

Answer: TODO

5 Challenge: Rademacher lower bounds [6 challenge points]

Using the no-free-lunch theorem, we proved a lower bound on the ability of any algorithm to learn a binary classifier from $\mathcal{H}$ in 0-1 loss based on $\operatorname{VCdim}(\mathcal{H})$ (Theorem 3 in the no-free-lunch notes).

We didn't say anything about Rademacher lower bounds, though. In this challenge question, we'll explore what can and can't be said for lower bounds based on Rademacher complexity.

First, let $\mathcal{F}$ be some class of functions $\mathcal{Z} \to \mathbb{R}$. We're going to prove that, for any $\mathcal{D}$ over $\mathcal{Z}$,

$$\frac{1}{2}\left(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \sup_{f\in\mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^m f(z_i) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}} f(z)\right) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \sup_{f\in\mathcal{F}} \left(\mathop{\mathbb{E}}_{z\sim\mathcal{D}} f(z) - \frac{1}{m}\sum_{i=1}^m f(z_i)\right)\right) \\ \geq \frac{1}{2}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \operatorname{Rad}(\mathcal{F}|_S) - \frac{1}{2\sqrt{m}}\sup_{f\in\mathcal{F}}\left|\mathop{\mathbb{E}}_{z\sim\mathcal{D}} f(z)\right|. \tag{1}$$

The left-hand side here is the average of the two directions of one-sided uniform convergence. Recall that the left-hand side is upper-bounded by $2\,\mathbb{E}_{S\sim\mathcal{D}^m}\operatorname{Rad}(\mathcal{F}|_S)$: we bounded the $\mathbb{E}_z f(z) - \frac{1}{m}\sum_i f(z_i)$ term by this in the Rademacher notes, and examining the symmetrization argument shows that the same bound holds for the $\frac{1}{m}\sum_i f(z_i) - \mathbb{E} f(z)$ one as well.

Let's start by proving (1):

[5.1] [1 points] Let $\mathcal{F}' = \{z \mapsto f(z) - c_f : f \in \mathcal{F}\}$, where $c_f \in \mathbb{R}$ may differ for each f. Prove that $\operatorname{Rad}(\mathcal{F}'|_S) \leq \operatorname{Rad}(\mathcal{F}|_S) + \frac{1}{\sqrt{m}}\sup_{f\in\mathcal{F}}|c_f|$.

Answer: TODO

[5.2] [3 points] Prove (1).

Hint: Start by defining the centred class $\tilde{\mathcal{F}}_\mathcal{D} = \{z \mapsto f(z) - \mathbb{E}_{z\sim\mathcal{D}}[f(z)] : f \in \mathcal{F}\}$, and consider $\frac{1}{2}\mathbb{E}_S \operatorname{Rad}(\tilde{\mathcal{F}}_\mathcal{D}|_S)$. An appropriate application of Question [5.1] will show this is at least the right-hand side. To show it's at most the left-hand side, expand out the definition and follow essentially the same argument as the symmetrization proof from Section 2 of the Rademacher notes.

Answer: TODO

When $a \leq f(z) \leq b$ for all f, z, we can bound $\sup_{f\in\mathcal{F}}|\mathbb{E}_{z\sim\mathcal{D}} f(z)| \leq \max(|a|, |b|)$. Now, while the left-hand side of (1) doesn't change if we shift all of $\mathcal{F}$ by a constant, and recalling that $\operatorname{Rad}(V + \{w\}) = \operatorname{Rad}(V)$ the first-term of the right-hand side doesn't either, $\max(|a|, |b|)$ does. Thus, if we shift $\mathcal{F}$ so that $|f(z)| \leq \frac{1}{2}(b-a)$ for all f and z, we get that the average of the two directions of expected worst-case one-sided uniform convergence is at least

$$\frac{1}{2}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\operatorname{Rad}(\mathcal{F}|_S) - \frac{b-a}{4\sqrt{m}}.$$

Many sources in the literature consider two-sided uniform convergence, $\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m f(z_i) - \mathbb{E}_{z\sim\mathcal{D}} f(z)\right|$, rather than the one-sided convergence we've always used; the upper bound then looks like

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m f(z_i) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}} f(z)\right| \leq 2\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\operatorname{Rad}((\mathcal{F}\cup -\mathcal{F})|_S),$$

where $-\mathcal{F} = \{z \mapsto -f(z) : f \in \mathcal{F}\}$.[1] If $\mathcal{F} = -\mathcal{F}$, i.e. the function class is symmetric, this is the same bound as we got in the one-sided case, because then indeed the one-sided and two-sided cases are the same.

[1]Note that $\operatorname{Rad}((\mathcal{F}\cup -\mathcal{F})|_S) = \mathbb{E}_{\boldsymbol{\sigma}}\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i f(z_i)\right|$; the original definition of Rademacher complexity, which still appears in many sources, was $\operatorname{Rad}_{|\cdot|}(V) = \mathbb{E}_{\boldsymbol{\sigma}}\sup_{v\in V}|v\cdot\boldsymbol{\sigma}|/m$. The version we use, without the absolute value, has turned out to be preferable for a bunch of reasons.

Notice that the left-hand side of (1) is always at most $\mathbb{E}_{S\sim\mathcal{D}^m} \sup_{f\in\mathcal{F}} \left|\frac{1}{m}\sum_{i=1}^m f(z_i) - \mathbb{E}_{z\sim\mathcal{D}} f(z)\right|$. Thus, if $a \le f(z) \le b$ for all f and z, the same McDiarmid argument as in Theorem 8 of the Rademacher notes gives that

$$\Pr_{S\sim\mathcal{D}^m}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m f(z_i) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}} f(z)\right| \ge \frac{1}{2}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \operatorname{Rad}(\mathcal{F}|_S) - \frac{b-a}{\sqrt{m}}\left(\frac{1}{4} + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)\right) \ge 1-\delta \tag{2}$$
$$\Pr_{S\sim\mathcal{D}^m}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m f(z_i) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}} f(z)\right| \le 2\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \operatorname{Rad}((\mathcal{F}\cup -\mathcal{F})|_S) + \frac{b-a}{\sqrt{m}}\sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right) \ge 1-\delta.$$

So far we've only really looked at upper bounds on the Rademacher complexity. It's possible to get lower bounds, though; Question 3 has one example. Another is given by equation (D.24) of [MRT], which implies[2]

$$\text{for } \mathcal{H}_B = \{x \mapsto w\cdot x : \|w\| \le B\}, \quad \operatorname{Rad}(\mathcal{H}_B|_{S_x}) \ge \frac{B}{\sqrt{2m}} \cdot \sqrt{\frac{1}{m}\sum_{i=1}^m \|x_i\|^2}. \tag{3}$$

Jensen's inequality goes the wrong way to lower-bound bound the expected Rademacher complexity, but at least asymptotically we know that $\mathbb{E}\sqrt{\frac{1}{m}\sum_{i=1}^m\|x_i\|^2}$ converges to a nonzero constant as $m\to\infty$, as long as x is not almost surely zero.

The real problem, though, is that Talagrand's contraction lemma is only one way. Using (2), our lower bound on $\sup_{h\in\mathcal{H}} L_\mathcal{D}(h) - L_S(h)$ would depend on $\mathbb{E}_S \operatorname{Rad}((\ell\circ\mathcal{H})|_S)$, and it's not obvious how to lower-bound that by something depending on $\operatorname{Rad}(\mathcal{H}|_{S_x})$.

(It's also not obvious how to use a lower bound on $\sup_{h\in\mathcal{H}} L_\mathcal{D}(h) - L_S(h)$ to get a lower bound on any learning algorithm, even the ERM: maybe the h with small $L_S(h)$ all have small $L_\mathcal{D}(h) - L_S(h)$, but there are hypotheses where $L_\mathcal{D}(h)$ and $L_S(h)$ are both big and far away from each other.)

These problems are in fact not possible to fix in general:

[5.3] [2 points] Give an example of a problem (an $\mathcal{H}$, $\mathcal{D}$, and ℓ) where $\mathbb{E}_{S\sim\mathcal{D}^m}\operatorname{Rad}(\mathcal{H}|_{S_x}) \not\to 0$ as $m\to\infty$, and yet ERM can achieve arbitrarily small excess error with enough samples.

Answer: TODO

[2]This means that $\operatorname{Rad}(\mathcal{H}_B|_S)/\left(\frac{B}{\sqrt{m}}\sqrt{\frac{1}{m}\sum_{i=1}^m\|x_i\|^2}\right) \in \left[\frac{1}{\sqrt{2}}, 1\right] \subset [0.7, 1]$; that's pretty nice!