CPSC 532D, Fall 2024: Assignment 3
due Friday, October 25 at 11:59pm

You can do this with a partner if you'd like (there's a "find a group" post on Piazza). **Read the website section on academic integrity** here for what you're allowed to do and not do; in particular, cite your sources (including people you talked to!) and don't use ChatGPT/etc. If you're not sure if something is okay, ask.

Prepare your answers to these questions using LaTeX; hopefully you're reasonably familiar with it, but if not, try using Overleaf and looking around for tutorials online. Feel free to ask questions if you get stuck on things on Piazza (but remove any details about the actual answers to the questions... make a private post if that's tough). If you prefer, the `.tex` source for this file is available on the course website, and you can put your answers in `\begin{answer} My answer here... \end{answer}` environments to make them stand out; feel free to delete whatever boilerplate you want. Or answer in a fresh document.

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 us a surprising amount of grading time).

Please **put your name 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 $\text{VCdim}(\mathcal{H}) \leq \text{VCdim}(\mathcal{H}')$.

Answer: TODO

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

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, $L_{\mathcal{D}}(\text{ERM}_{\mathcal{H}}(S))$ versus $L_{\mathcal{D}}(\text{ERM}_{\mathcal{H}'}(S))$ 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

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

Answer: TODO

2 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 (notes section 6.4.1.1) that $\text{VCdim}(\mathcal{H}) = 1$: 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).

[2.1] [5 points] Use Sauer-Shelah (Lemma 6.12), and also the simpler Corollary 6.10, to give two upper bounds on the growth function $\Gamma_{\mathcal{H}}(m)$.

Answer: TODO

[2.2] [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 [2.1]?

Answer: TODO

[2.3] [5 points] Plug the previous parts in to upper bound $\text{Rad}(\mathcal{H}|_{S_x})$ for an S containing m distinct real numbers. *You should give multiple bounds here: one for each bound, and one for the exact value of the growth function.*

Answer: TODO

[2.4] [5 points] Give the asymptotic value of $\text{Rad}(\mathcal{H}|_{S_x})$ for an S_x containing m distinct real numbers. Your answer might look something like "$\text{Rad}(\mathcal{H}|_{S_x}) = 7m + \mathcal{O}(1)$," with a justification. To be clear, this means that $7m - a_n \leq \text{Rad}(\mathcal{H}|_{S_x}) \leq 7m + a_n$ for some $a_m = \mathcal{O}(1)$. How does it compare to the bound from Question [2.3]?

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

3 Rademacher complexity of deep networks [50 points]

We're now going to prove a Rademacher complexity bound for deep networks. To do that, we're going to build up our repertoire of Rademacher properties a bit first.

Lemma 3.1. *Consider finitely many sets V_i such that for all $\boldsymbol{\sigma} \in \{-1,1\}^m$, it holds that $\sup_{v \in V_i} v \cdot \boldsymbol{\sigma} \geq 0$; for instance, this holds if $0 \in V_i$, or if for all $v \in V_i$ we also have $-v \in V_i$. Then $\mathrm{Rad}(\cup_i V_i) \leq \sum_i \mathrm{Rad}(V_i)$.*

[3.1] [10 points] Prove Lemma 3.1.

Answer: TODO

The convex hull of a set V is the set of all convex combinations of points in V:

$$\mathrm{conv}(V) = \bigcup_{k \geq 1} \left\{ \sum_{i=1}^k \alpha_i v_i : \alpha_i \geq 0; \sum_{i=1}^k \alpha_i = 1; v_1, \ldots, v_k \in V \right\}.$$

Lemma 3.2. *For any set V, $\mathrm{Rad}(\mathrm{conv}(V)) = \mathrm{Rad}(V)$.*

[3.2] [10 points] Prove Lemma 3.2.

Answer: TODO

Lemma 3.3. *For any set V, $\mathrm{Rad}\left(\left\{\sum_{i=1}^d w_i v_i : w_i \in \mathbb{R}, \sum_{i=1}^d |w_i| \leq B, v_i \in V\right\}\right) \leq B\,\mathrm{Rad}(V \cup (-V))$.*

[3.3] [10 points] Prove Lemma 3.3. *Hint: You might want to apply Lemmas 3.1 and 3.2.*

Answer: TODO

Now we're ready to bound a class of multilayer perceptrons (without bias terms because it makes things look a little cleaner – in practice, you should use bias terms!). Specifically,

$$\mathcal{H}_D = \{x \mapsto \sigma_D(W_D \sigma_{D-1}(\cdots \sigma_1(W_1 x) \cdots)) : W_1 \in \mathcal{W}_1, \ldots, W_D \in \mathcal{W}_D\}.$$

The σ_i are M_i-Lipschitz elementwise activation functions such that $\sigma_i(0) = 0$; for example, $\mathrm{ReLU}(x) = [\max(x_i, 0)]$. The W_i are matrices of shape $d_i \times d_{i-1}$, where the input dimension is $d_0 = d$, the output dimension is $d_D = 1$, and the in-between dimensions are some arbitrary, fixed sequence. The constraints are

$$\mathcal{W}_i = \left\{ W \in \mathbb{R}^{d_i \times d_{i-1}} : \forall j \in [d_i], \sum_{k=1}^{d_{i-1}} |W_{jk}| \leq B_i \right\}.$$

Since $\mathcal{H}_D$ has a nice recursive form, let's think about "peeling off" a layer at a time: bounding $\mathrm{Rad}(\mathcal{H}_D)$ in terms of $\mathrm{Rad}(\mathcal{H}_{D-1})$. To do this, recall that since we're dealing with a real-valued network, W_D is of shape $1 \times d_{D-1}$, and then notice that for $D \geq 2$,

$$\mathcal{H}_D \subseteq \left\{ x \mapsto \sigma_D\left(\sum_{j=1}^{d_{D-1}} (W_D)_j\, h_j(x) \right) : h_1, \ldots, h_{d_{D-1}} \in \mathcal{H}_{D-1}, W_D \in \mathcal{W}_D \right\}. \tag{3.1}$$

[3.4] [10 points] Prove that $\mathrm{Rad}(\mathcal{H}_D|_{S_x}) \leq 2 M_D B_D\, \mathrm{Rad}(\mathcal{H}_{D-1}|_{S_x})$.

Answer: TODO

If we define $\mathcal{H}_0$ in a way so that (3.1) also makes sense for $D = 1$, this leaves us with a bound of the form $\mathrm{Rad}(\mathcal{H}|_{S_x}) \leq \left(\prod_{i=1}^D (2M_i B_i)\right) \mathrm{Rad}(\mathcal{H}_0|_{S_x})$.

[3.5] [10 points] Give a definition of $\mathcal{H}_0$ so that (3.1) makes sense for $D = 1$. Bound $\text{Rad}(\mathcal{H}_0|_{S_x})$ under the assumption that $\max_{x \in S_x} \|x\|_p \leq C$, for some $p \in [1, \infty]$ of your choice. *Your bound should be $\mathcal{O}(1/\sqrt{m})$, treating everything but m as a constant.*

Answer: TODO

Armed with this bound, we can show generalization bounds for scalar-output MLPs in the same way as for anything else: for example, we can immediately get an expectation bound on $L_{\mathcal{D}}(\text{ERM}_{\mathcal{H}_D})$ for any Lipschitz loss, and if the loss is also bounded (either "naturally" or based on a bound of $|h(x)|$ as for logistic regression) then we can get a high-probability bound too. *(The bound won't be very good for very deep networks, though – it's exponential in the depth! It's possible to improve on this somewhat with fancier techniques, but if the $\mathcal{W}_i$ are all norm balls, a dependence on the product of those norms is unavoidable.)*

4 Gaussian complexity [10 challenge points]

We briefly mentioned the idea of Gaussian complexity in class:

$$\mathcal{G}(V) = \mathop{\mathbb{E}}_{s_i \sim \mathcal{N}(0,1)} \sup_{v \in V} \frac{1}{m} \sum_{i=1}^{m} s_i v_i = \mathop{\mathbb{E}}_{s \sim \mathcal{N}(0, I_m)} \sup_{v \in V} \frac{s \cdot v}{m}.$$

We also mentioned that Rademacher and Gaussian complexity are close. Let's prove that. The following results will be helpful in doing that.

If $X \sim \mathcal{N}(0,1)$, then we say that $|X| \sim \chi$. (You may be more familiar with the χ^2 distribution; this is its square root, with one degree of freedom.)

Lemma 4.1. *If $q \sim \chi$, $\mathbb{E}\, q = \sqrt{2/\pi}$.*

Proof. Use $\mathbb{E}\, q = \int_{-\infty}^{0} (-x) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \mathrm{d}x + \int_{0}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \mathrm{d}x = \frac{2}{\sqrt{2\pi}} \int_{-\infty}^{0} \left(-x e^{-\frac{1}{2}x^2}\right) \mathrm{d}x$. Noting that $\frac{\mathrm{d}}{\mathrm{d}x}\left(e^{-\frac{1}{2}x^2}\right) = -x e^{-\frac{1}{2}x^2}$, the integral is therefore $e^{-\frac{1}{2}0^2} - e^{-\frac{1}{2}(-\infty)^2} = 1 - 0 = 1$. □

It'll also help to write $a \odot b$ for the elementwise product of two vectors, $(a \odot b)_i = a_i b_i$.

[4.1] [3 challenge points] Prove that $\mathcal{G}(V) = \mathbb{E}_{q \sim \chi^m} \operatorname{Rad}(q \odot V)$, where $q \odot V = \{q \odot v : v \in V\}$.

Here χ^m means a vector of m independent χ variables; q is the same as taking the elementwise absolute value of $s \sim \mathcal{N}(0, I_m)$.

Answer: TODO

[4.2] [3 challenge points] Prove a bound of the form $\mathcal{G}(V) \le f(m) \operatorname{Rad}(V)$, where $f(m)$ depends only on m. Specify an explicit closed form for $f(m)$.

Answer: TODO

[4.3] [3 challenge points] Prove that $\operatorname{Rad}(V) \le \sqrt{\frac{\pi}{2}} \mathcal{G}(V)$.

Hint: Try using Jensen's inequality and brushing up on convex function properties.

Answer: TODO

[4.4] [1 challenge point] Give an example of a particular V where the previous bound is tight, i.e. $\operatorname{Rad}(V) = \sqrt{\frac{\pi}{2}} \mathcal{G}(V) > 0$.

Answer: TODO