CPSC 532D, Fall 2023: Assignment 2
due Wednesday, 11 October 2023, **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 Concentrating on concentric circles [40 points]

Based in part on SSBD exercise 3.3.

In this problem, we'll show that a particular infinite hypothesis class can be PAC-learned with a "direct" proof. *It's pretty annoying to show general bounds on this problem with covering numbers, but once we cover VC dimension we'll return see a more direct way to show (agnostic) PAC learning.*

Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0, 1\}$, and let $\mathcal{H}$ be the class of indicator functions for circles around the origin – that is, $\mathcal{H} = \{h_r : r \in \mathbb{R}_{\geq 0}\}$, where $h_r(x) = \mathbb{1}_{[\|x\| \leq r]}$ (a function which is 1 if $\|x\| \leq r$, 0 otherwise). Use 0-1 loss.

For a given sample S, let $r_S = \max_{i:y_i=1} \|x_i\|$, and use $\hat{h}_S$ to denote h_{r_S}, the indicator function of a circle with radius r_S, the tightest circle containing all of the positive training points.

To start with, let's assume *realizability*: that there is an $h^* \in \mathcal{H}$ such that $L_{\mathcal{D}}(h^*) = 0$.

(1.1) [5 points] Show that $\hat{h}_S$ is an empirical risk minimizer for the hypothesis class $\mathcal{H}$.

Answer: TODO

(1.2) [20 points] Prove that $\hat{h}_S$ PAC-learns this $\mathcal{H}$, with sample complexity $\frac{1}{\varepsilon} \log \frac{1}{\delta}$.

Hint: Three steps: first, what makes a hypothesis have high error in this setting? Next, what would S have to look like in order to get one of those "bad" hypotheses? Last, how likely is it to see an S like that?

Hint: A frequently useful inequality is that $1 - a \leq \exp(-a)$.

Hint: If you're stuck and want to see something similar-ish (but a little more complicated), check out Example 2.4 of MRT, which is also Exercise 2.3 of SSBD.

Answer: TODO

Now let's make things a little harder on our learner, by adding random noise. Rather than perfect realizability, let $\mathcal{D}$ be such that $\Pr(y = 1 \mid x) = \begin{cases} 1 - \eta & \text{if } h^*(x) = 1 \\ \eta & \text{if } h^*(x) = 0 \end{cases}$ for some $h^* \in \mathcal{H}$: that is, labels are randomly flipped with probability $\eta \in (0, \frac{1}{2})$. The learner knows the value of η, but not which points have been flipped.

(1.3) [5 points] Is $\hat{h}_S$ as described above still an ERM?

Answer: TODO

(1.4) [10 points] Ambitious Ambrose claims to have proven the following:

> For any $\varepsilon, \delta \in (0, 1)$ and $0 \leq \eta < \frac{1}{2}$, there is a function $m(\eta, \varepsilon, \delta)$ such that, for any $m \geq m(\eta, \varepsilon, \delta)$ and any $\mathcal{D}$ of the form above,
>
> $$\Pr_{S \sim \mathcal{D}^m}(L_{\mathcal{D}}(\hat{h}_S) > \eta + \varepsilon) \leq \delta.$$

This is followed by an unreadably long computer-assisted proof using both category theory and complicated partial differential equations. Without reading that proof, argue that Ambrose must be wrong: no such function $m(\eta, \varepsilon, \delta)$ can exist.

Answer: TODO

2 Sums, means, and maxes of sub-Gaussians [50 points]

In this question, we're going to explore sub-Gaussians and different versions of Hoeffding's some more.

A reminder that you may want to refer back to the lecture notes.

(2.1) [10 points] Let $X_1 \in \mathcal{SG}(\sigma_1)$ and $X_2 \in \mathcal{SG}(\sigma_2)$; **do not** assume independence. Show that $X_1 + X_2$ is $\mathcal{SG}(\sqrt{2}\sqrt{\sigma_1^2 + \sigma_2^2})$.

Hint: One form of the ever-useful Cauchy-Schwarz inequality is that $\mathbb{E}[XY] \leq \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}$, even if X and Y are dependent.

Answer: TODO

(2.2) [15 points] Let $X_1 \in \mathcal{SG}(\sigma_1)$ and $X_2 \in \mathcal{SG}(\sigma_2)$; **do not** assume independence. Show that $X_1 + X_2$ is $\mathcal{SG}(\sigma_1 + \sigma_2)$.

Hint: One way is to use Hölder's inequality: $\mathbb{E}[XY] \leq \mathbb{E}[X^p]^{1/p}\,\mathbb{E}[Y^q]^{1/q}$ for all $p, q \in [1, \infty]$ with $1/p + 1/q = 1$, i.e. $q = p/(p-1)$. Do this for a general p, see what you get, then find the optimal p.

Answer: TODO

(2.3) [10 points] Let $X_1, \ldots, X_m$ each be $\mathcal{SG}(\sigma)$ with mean μ, but do *not* assume independence. Construct a high-probability bound on their mean, $\Pr\left(\frac{1}{m}\sum_{i=1}^m X_i > \mu + \text{something}\right) \leq \delta$, using either Question (2.1) or (2.2) rather than the notes' Proposition 6. How much worse is what you just got than (Hoeffding') from the notes if the variables are actually independent? Is your result tight?

Hint: One of these results is much *easier to use than the other one.*

Answer: TODO

(2.4) [15 points] *So far, we've only looked at means of a bunch of random variables. But for uniform convergence, we care about the* worst-case *behaviour of errors. We're going to (or have already, depending on when you're reading this...) use the following result in a key way in class.*

Let $X_1, \ldots, X_m$ be zero-mean random variables that are each $\mathcal{SG}(\sigma)$; **do not** assume independence.[1] Prove that

$$\mathbb{E}\left[\max_{i=1,\ldots,m} X_i\right] \leq \sigma\sqrt{2\log(m)}.$$

Hint: Bound $\exp(\lambda\,\mathbb{E}\max_i X_i)$ in terms of something that only depends on m, σ, and λ, by rearranging into a form that lets you plug in the definition of sub-Gaussianity. Then turn that into a bound on $\mathbb{E}\max_i X_i$ in terms of m, σ, and λ. Then optimize λ in that bound to get something only depending on m and σ.

Hint: By Jensen's inequality, $\exp(\mathbb{E}\,Y) \leq \mathbb{E}\exp(Y)$.

Hint: One way to upper-bound the max of a bunch of nonnegative numbers is by their sum.[2]

Answer: TODO

[1] As far as I know, independence actually wouldn't help here.

[2] Although this might seem really loose, if the max is a lot bigger than the second-biggest number – e.g. because they're on an exponential scale – it's not too bad.

3 Deep networks [10 challenge points]

Consider a hypothesis class of deep networks of depth D, with inputs in $\mathbb{R}^d$:

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

where the σ_i are r-Lipschitz elementwise activation functions with $\sigma_i(0) = 0$. That is, $\sigma_i(v) = (\sigma_i(v_j))_{j=1}^{\dim v}$, where on the right-hand side σ_i is taking a scalar argument. The canonical example for σ_i is $\mathrm{ReLU}(x) = \max(0, x)$.

Here the weight matrices W_i are 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 arbitrary.

The constraints $W_i \in \mathcal{W}_i$ are intentionally left unspecified; you'll have to choose it in your proof. Don't choose a "trivial" constraint like $\mathcal{W}_i = \{0\}$; let's say at least that $\mathcal{W}_i$ must have nonzero volume in $\mathbb{R}^{d_i \times d_{i-1}}$. It should probably something like a bound on the Frobenius norm of each W_i, or on $\|w\|_4^{6i+12}$ for each column w of W_i, or something like that. (Most likely, you'll want to make the constraint the same for each layer.)

You can make a similar "reasonable" assumption on $\mathcal{D}$. Assuming $\|x\| \le C$ is reasonable; assuming x is almost surely constant is not. At a minimum, again, the support of x should have nonzero volume in $\mathbb{R}^d$. Assuming realizability is reasonable (if you'd like); assuming $y = 0$ is not.

Choose a *specific* loss function ℓ. Zero-one, logistic, and $l(\hat{y}, y) = |\hat{y} - y|^p$ for any $p \in [1, \infty)$ are all reasonable; if you'd like to use something else, that's potentially okay, just justify it as "reasonable."

Make all of these assumptions extremely clear in your answer.

Prove a bound on $L_{\mathcal{D}}(\hat{h}_S)$ for the ERM $\hat{h}_S$ under the conditions you assumed. It can be either an expectation or a high-probability bound, but the excess error should go to zero as $m \to \infty$ with other parameters fixed.

Reminder: don't just look up and ape an existing bound; prove it yourself based on the class material.

Hint: you may want to tackle this with covering numbers, or with Rademacher complexity (to be covered starting in lecture 5); I think they're about equal amounts of work here, but (like usual with "basic" covering arguments) the Rademacher result is somewhat better. In either case, you'll probably want to do induction on the layers.

If you're using covering numbers, you can use the following result without proof:

Proposition 3.1. *Let $\mathbb{F}$ be a Banach space. The size of a minimal η-cover of the radius-R ball in $\mathbb{F}$, $\{f \in \mathbb{F} : \|f\|_{\mathbb{F}} \le R\}$, with respect to the metric $\|f - f\|_{\mathbb{F}}$, satisfies $N_{\mathbb{F}}(R, \eta) \le (4R/\eta)^{\dim \mathbb{F}}$.*

Letting $\mathbb{F}$ be $\mathbb{R}^d$, you get a slightly worse version of the result proved in our covering number notes. But maybe you could define a different $\mathbb{F}$ so that $\ell(h, z)$ is Lipschitz with respect to that metric. . . . (You don't need to explicitly prove that $\mathbb{F}$ is complete, if it's clear that it's a vector space, the norm you define is clearly a norm, and it seems reasonable that the space would be complete.)

If you're using Rademacher complexity, please prove any properties you use that aren't in the notes. But you might find it useful to prove these properties:

- $\mathrm{Rad}(\mathrm{conv}(V)) = \mathrm{Rad}(V)$, *where* $\mathrm{conv}(V)$ *is the convex hull of* V, $\{\sum_{i=1}^{|V|} \alpha_i v_i : \alpha_i \ge 0, \sum_i \alpha_i = 1\}$.
- *Let* V_i *be such that* $\forall \boldsymbol{\sigma} \in \{-1, 1\}^m$, $\sup_{v \in V_i} v \cdot \boldsymbol{\sigma} \ge 0$*; for instance, this is true if* $0 \in V_i$*, or if* $\forall v \in V_i$, $-v \in V_i$. *Then* $\mathrm{Rad}\left(\cup_i V_i\right) \le \sum_i \mathrm{Rad}(V_i)$.

(If you're not sure how to prove them but use them in your result, you'll get only a small point penalty.)

Answer: TODO