CPSC 532S: Assignment 2 – due Friday, 18 Feb 2022, 11:59pm

This assignment is split into four questions (the parts with big section headers; most have sub-parts). You can solve each question in groups of up to three. Groups don't need to be consistent between problems; you can do Q1 and Q2 alone, Q3 with Alice, and Q4 with Bob and Carlos if you want.

Please **do not** just split the questions up and do the parts separately. **If your name is on a solution, you are pledging that you contributed significantly to the solution and understand it fully.**

There is a separate Gradescope assignment for each problem; use the Gradescope groups feature to submit once and associate with each of you, but also put all of your names on the first page as a backup.

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. (Note that free Overleaf accounts can only share with one "named collaborator," but you can collaborate with more people by sending them an edit link. Make sure you only share the parts of the homework you're handing in together!)

Feel free to ask questions if you get stuck on things on Piazza (but remove any details about the actual answers... feel free to make a private post if that's tough). If you look stuff up anywhere other than in one of the two course textbooks, please **cite your sources**: just say in the answer to that question where you looked. (A link is fine, no need for a formal citation.) Please do not look at solution manuals or so on. 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.

If you like, 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 if so; feel free to delete whatever boilerplate you want (or not, I'm not printing them out). Or answer in a fresh document; just make it clear which question you're answering where.

If you're using a consistent group and want to write your answers in one document, you could split the PDF with e.g. `qpdf a2.pdf --pages . 2-3 -- q1.pdf` or through the GUI of a PDF viewer. Or you can upload the full file four times and just make sure you assign pages appropriately.

Submit your answers as a PDF on Gradescope: instructions on Piazza. You'll be prompted to mark where each sub-part is in your PDF; make sure you mark all relevant pages for each part. (This saves me a surprising amount of time in grading.) If something goes wrong, you can also email your assignment to me directly (`dsuth@cs.ubc.ca`).

1 Threshold functions [20 points + 10 bonus points]

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

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

Let's also restate the Sauer-Shelah lemma for clarity[1]:

Lemma 1.1 (Sauer-Shelah)**.** *If the VC dimension of a hypothesis class $\mathcal{H}$ is d, the growth function $\tau_{\mathcal{H}}$ satisfies (for any $n \in \mathbb{N}$)*

$$\tau_{\mathcal{H}}(n) \leq \sum_{i=0}^{d} \binom{n}{i}.$$

Moreover, for all $n \geq d$, we have (where $e = \exp(1) \approx 2.718$ is Euler's constant) $\tau_{\mathcal{H}}(n) \leq \left(\frac{en}{d}\right)^d$.

We showed in class 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) [7 points] Use the two parts of the Sauer-Shelah lemma to give two upper bounds on the growth function $\tau_{\mathcal{H}}(n)$.

(b) [7 points] Directly derive the exact value of the growth function $\tau_{\mathcal{H}}$ from its definition. How tight are the upper bounds from part (a)?

(c) [6 points] Use the previous parts to give an upper bound on the empirical Rademacher complexity $\hat{\mathfrak{R}}_{\mathrm{S}}(\mathcal{H})$.

(d) [BONUS: 10 points] Give the asymptotic value of $\hat{\mathfrak{R}}_{\mathrm{S}}(\mathcal{H})$ for an S with n distinct points. Your answer might look something like "$\mathfrak{R}_{\mathrm{n}}(\mathcal{H}) = 7n + \mathcal{O}(1)$," with a justification. How does it compare to the bound from part (c)?

Hint: It's not $7n + \mathcal{O}(1)$.

Hint: I couldn't think of a real hint for this one that doesn't basically give you the answer, which is why it's bonus points. But I will say that to get the exact asymptotics, I had to look up a math paper. You'll get almost all the points if you just don't know the constant, and a good chunk of bonus points for an approach that seems fruitful. Remember to cite your sources.

[1] SSBD sets the threshold for the second part to $n \geq d + 2$, but $n \geq d$ is sufficient; see the proof from MRT we saw in class.

2 Hyper-rectangles [20 points]

The class of axis-aligned hyper-rectangles in $\mathbb{R}^d$ is given by

$$\mathcal{H} = \{x \mapsto \mathbb{1}(a_1 \leq x_1 \leq b_1) \cdots \mathbb{1}(a_d \leq x_d \leq b_d) : a_1 \leq b_1, \ldots, a_d \leq b_d\}.$$

That is, a hypothesis h returns 1 if $x \in \mathbb{R}^d$ is inside the hyper-rectangle $[a_1, b_1] \times \cdots \times [a_d, b_d]$, and 0 otherwise.

Prove that $\mathrm{VCdim}(\mathcal{H}) = 2d$.

Hint: SSBD section 6.3.3 shows this for $d = 2$.

3 Monotonicity [20 points]

(a) [6 points] Prove that if $\mathcal{H} \subseteq \mathcal{H}'$, then $\mathrm{VCdim}(\mathcal{H}) \leq \mathrm{VCdim}(\mathcal{H}')$.

(b) [6 points] Prove that if $\mathcal{H} \subseteq \mathcal{H}'$, then $\hat{\mathfrak{R}}_{\mathrm{S}}(\mathcal{H}) \leq \hat{\mathfrak{R}}_{\mathrm{S}}(\mathcal{H}')$ and $\mathfrak{R}_{\mathrm{n}}(\mathcal{H}) \leq \mathfrak{R}_{\mathrm{n}}(\mathcal{H}')$.

(c) [8 points] Comment on how we should expect parts (a) and (b) to affect the generalization loss of running ERM in $\mathcal{H}$ versus $\mathcal{H}'$. What other factors are at play?

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):

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

Λ and ν are hyperparameters defining how complex the class is allowed to be.

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

The following is a form of Talagrand's contraction lemma that works for any $\mathcal{H}$ and L-Lipschitz function Φ:[2]

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

(b) [10 points] Use (*) to upper bound $\hat{\mathfrak{R}}_{\mathrm{S}}(\mathcal{H})$ in terms of $\hat{\mathfrak{R}}_{\mathrm{S}}(\mathcal{H}')$, for

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

(c) [10 points] Bound $\mathfrak{R}_{\mathrm{n}}(\mathcal{H}')$, and thereby $\mathfrak{R}_{\mathrm{n}}(\mathcal{H})$. You'll need an assumption on $\|x\|$ to do this; be clear what you're assuming.

Hint: We essentially did this in class (lecture 8, slide 11; note that I messed up a bit in class, but it's fixed on the posted slides).

(d) [10 points] Give an expression for ε such that $\Pr_{S \sim \mathcal{D}^n} \left(\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_S(h)) \le \varepsilon \right) \ge 1 - \delta$.

Choose some reasonable loss function ℓ for inside L, and be clear about any additional assumptions.

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.

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

$$\frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}} \left| \sum_{i=1}^{n} \sigma_i \Phi(h(x_i)) \right| \right] \le \frac{L}{n} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}} \left| \sum_{i=1}^{n} \sigma_i h(x_i) \right| \right]. \qquad \text{(wrong)}$$

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 $\mathbb{E}_{\boldsymbol{\sigma}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \right| = \sqrt{2/(\pi n)} + \mathcal{O}(n^{-3/2})$, as we discussed briefly in class. 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.