CPSC 532D, Fall 2025: Assignment 2
due Wednesday, 15 October 2025, **12:00 noon**

You can do this assignment, and future ones, with a partner. **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 for anything content-related. 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 if you'd prefer.

Submit your answers as a single PDF on Gradescope: here's the link. 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).

Make sure you submit using Gradescope's group feature if you're submitting a joint assignment, and put both your names on the first page to be safe; if you did the assignment partially together and partially separately, hand in separate PDFs, and put a note on each question where you worked together like *I did this problem with Alice* so we don't think you cheated. :)

On the off chance something goes wrong, you can also email your assignment to me directly (`dsuth@cs.ubc.ca`).

1 Priors? In my frequentist analysis? [10 points]

Suppose we have a countable (maybe finite, maybe infinite) hypothesis set $\mathcal{H}$, and we assign some "prior probability" p_h to each $h \in \mathcal{H}$ such that each $p_h > 0$ and $\sum_{h \in \mathcal{H}} p_h \leq 1$. Assume a loss bounded in $[a, b]$.

Use Hoeffding's inequality to prove the "Bayesian-ish" bound

$$\Pr_{S \sim \mathcal{D}^m} \left(\forall h \in \mathcal{H}, \quad L_{\mathcal{D}}(h) - L_S(h) \leq (b-a) \sqrt{\frac{1}{2m} \left[\log \frac{1}{p_h} + \log \frac{1}{\delta} \right]} \right) \geq 1 - \delta.$$

Hint: It'll be pretty similar to the analogous step in the proof from lecture 2!

We could then use this to show a bound on ERM in the same way as always: by separately bounding the probability of $L_S(h^*)$ being very small, and adding the two together.

Answer: TODO

2 Sums, means, and maxes of subgaussians [50 points]

In this question, we're going to explore subgaussians and different versions of Hoeffding's inequality some more.

[2.1] [10 points] Let X_1 be $\mathcal{SG}(\sigma_1)$ and X_2 be $\mathcal{SG}(\sigma_2)$; **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 be $\mathcal{SG}(\sigma_1)$ and X_2 be $\mathcal{SG}(\sigma_2)$; **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 3.6 (which assumed independence). How much worse is what you just got than (Hoeffding') from the notes when the variables are actually independent, particularly in terms of its dependence on m? Could you have expected to get a better result, or can you construct a dependent example where this dependence on m is necessary?

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 subgaussianity. 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. 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.

Answer: TODO

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

3 Complexity: it's simpler than you think [30 points + 5 bonus points]

In class (or notes section 5), we defined the Rademacher complexity

$$\mathrm{Rad}(V) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim\mathrm{Unif}(\pm 1)^m} \sup_{v\in V} \frac{1}{m}\sum_{i=1}^{m} \sigma_i v_i = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim\mathrm{Unif}(\pm 1)^m} \sup_{v\in V} \frac{\boldsymbol{\sigma}\cdot v}{m}.$$

We also mentioned in class, for motivation, something that I called the "spherical complexity." letting $\mathbb{S} = \{w \in \mathbb{R}^m : \|w\| = 1\}$,

$$\mathrm{Sph}(V) = \mathop{\mathbb{E}}_{s\sim\mathrm{Unif}(\mathbb{S})} \sup_{v\in V} s\cdot v.$$

I mentioned these two are pretty close. We'll show now that $\mathrm{Sph}(V)$ is roughly proportional to $\sqrt{m}\,\mathrm{Rad}(V)$.

[3.1] [5 points] Argue directly from the definitions of Rad and Sph that the $\sqrt{m}$ scaling "makes sense." *This will probably be one or two lines.*

Answer: TODO

Let's write $a \odot b$ for the elementwise product of two vectors, $(a\odot b)_i = a_i b_i$, and $a \odot V = \{a \odot v : v \in V\}$.

It also may be helpful to note that $s \sim \mathrm{Unif}(\mathbb{S})$ is equivalent to taking $g \sim \mathcal{N}(0, I_m)$, meaning that each $g_i \sim \mathcal{N}(0,1)$ iid, and then setting $s = g/\|g\|$.

Define a random variable q by first taking a uniform vector $s \sim \mathrm{Unif}(\mathbb{S})$, then taking its elementwise absolute value: $q_i = |s_i|$. The rest of the questions will all have q follow this distribution.

[3.2] [10 points] Prove that $\mathrm{Sph}(V) = m\,\mathbb{E}_q\,\mathrm{Rad}(q\odot V)$.

Answer: TODO

[3.3] [5 points] Show that $\mathrm{Sph}(V) \le m\,(\mathbb{E}_q\|q\|_\infty)\,\mathrm{Rad}(V)$. *Recall that* $\|q\|_\infty = \max_i |q_i|$.

Answer: TODO

Let $g \sim \mathcal{N}(0, I_m)$ and $s = g/\|g\|$, so that $s \sim \mathrm{Unif}(\mathbb{S})$. Hence $s_i = \frac{g_i}{\sqrt{g_i^2 + \sum_{j\neq i} g_j^2}}$. For large m, the denominator will be very very likely to be dominated by the sum, meaning that $s_i \approx \frac{g_i}{\sqrt{m}}$. This suggests that, for large m, s_i should be roughly $\mathcal{SG}(C/\sqrt{m})$. In fact, this is correct (even for small m): using "concentration of Lipschitz functions on the sphere"[2], there exists some global constant $C > 0$ such that s_i is $\mathcal{SG}(C/\sqrt{m})$. (I found a paper claiming that $C < 560$, but didn't follow their references to check, and presumably that's extremely loose anyway.)

[3.4] [5 points] Let C be a constant such that s_i is $\mathcal{SG}(C/\sqrt{m})$ for each i. Prove that $\mathbb{E}_q\|q\|_\infty \le C\sqrt{\frac{2}{m}\log(2m)}$.

Hint: That form sure does look reminiscent of Question [2.4], doesn't it? You can use that result here.

Answer: TODO

[3.5] [5 bonus points] Show that $\mathrm{Sph}(V) \ge m(\mathbb{E}_q\, q_1)\,\mathrm{Rad}(V)$. *Here q_1 means the first coordinate of q.*

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

Answer: TODO

[3.6] [5 points] Show that $\mathop{\mathbb{E}}_q q_1 = \dfrac{\mathbb{E}_{x\sim\mathcal{N}(0,1)}|x|}{\mathbb{E}_{g\sim\mathcal{N}(0_m, I_m)}\|g\|}$.

Hint: Consider $\mathbb{E}|g_1|$, and think about the "opposite" of the argument above about constructing $s = g/\|g\|$.

[2] See, e.g., Theorem 5.1.3 (labeled with that quoted text) of the current draft of Vershynin, second edition.

Answer: TODO

We can just look up the mean of the chi distribution to get an exact formula for this ratio. The numerator is $\sqrt{2/\pi}$; the denominator is a ratio of gamma functions which is $\sqrt{m-1}\,(1-\mathcal{O}(1/m))$, giving that $\mathbb{E}\,q_1 \approx \sqrt{2/(\pi m)} \approx 0.8/\sqrt{m}$.

Thus, there are constants c, C such that $c\sqrt{m}\,\mathrm{Rad}(V) \le \mathrm{Sph}(V) \le C\sqrt{m}\log(2m)\,\mathrm{Rad}(V)$.

4 Limits of Learning Lipschitz Laws [10 challenge points]

So far, we've only seen covering number bounds based on covering norm balls in $\mathbb{R}^d$. Let's use an analogous argument with a different kind of result.

Let $\mathcal{H} = \{h : [0, C]^d \to \mathbb{R} : h(0) = 0, \|h\|_{\text{Lip}} \leq B\}$ for some $B \geq 0$, where the Lipschitz constant is with respect to the usual Euclidean norms. This is a nonparametric class that includes "a *lot*" of functions.

Consider the "sup-norm"/uniform norm defined as $\|f\|_\infty = \sup_x |f(x)|$, which induces a metric $\rho_{\mathcal{H}}(h, g) = \|h - g\|_\infty$. (Recall that $h - g$ is the function that maps x to $h(x) - g(x)$.) It can be shown[3] that the covering number of $\mathcal{H}$ with respect to this $\rho_{\mathcal{H}}$ satisfies

$$\log N(\mathcal{H}, \eta) \leq \left(\frac{aBC}{\eta}\right)^d \tag{1}$$

for some constant $a > 0$ and all η small enough that the right-hand side is at least 1. Compare this to the $d \log \frac{3B}{\eta}$ bound we saw for Euclidean balls.

Consider the absolute-value loss $\ell(h, (x, y)) = |h(x) - y|$, and suppose $\mathcal{D}$ is such that $\Pr_{(x,y)\sim\mathcal{D}}(|y| \leq Y) = 1$. Prove a high-probability bound on $\sup_{h\in\mathcal{H}} L_{\mathcal{D}}(\mathcal{H}) - L_S(h)$ with the best rate (in terms of m) you can find; it should depend on a, B, C, Y, d, m, and the error probability δ. Prove this bound with explicit constants, but then also summarize it in a $\mathcal{O}_p$ statement treating everything but m as a constant. Is the rate faster or slower than the logistic regression bound we saw in class?

Hint: This proof ends up kind of long (at least mine did). Split it into appropriate sub-parts, and maybe define some helper variables along the way so your expressions don't get too unwieldy (but then expand out the final answer). Feel free to make simplifications that make things look nicer at the cost of making the constants worse, but try to get the m dependence right.

Hint: It's not possible to find the exactly optimal choice of η here (when $d \geq 2$). You'll probably want to use $\sqrt{x + y} \leq \sqrt{x} + \sqrt{y}$ before picking η, which gives a nicer bound anyway.

Hint: The reverse triangle inequality is often useful.

You don't have to repeat any portion of the argument which is verbatim identical to the notes, but you can. If you're not, be very clear about exactly what you've changed.

Answer: TODO

[3] Example 5.10 of Wainwright's book shows a lower bound for $d = 1$, $C = 1$ and points towards how to do the upper bound. Just afterwards, he mentions the $d > 1$ case is analogous. *(Unfortunately, he only states it in $\asymp$ notation and I'm not totally sure whether the constant there is allowed to depend on d or not. The version of (1) is definitely valid – see e.g. Lemma 6 here which bounds a more general case with explicit constants – but I'm not certain it's necessary.)* To generalize to $C \neq 1$, consider that if $h : [0, C]^d \to \mathbb{R}$ is B-Lipschitz, then $x \mapsto h(Cx)$ is a $[0, 1]^d \to \mathbb{R}$ function which is BC-Lipschitz.