CPSC 532D, Fall 2024: Assignment 2
due Friday, October 11 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 Questions you should probably get approximately correct [40 points]

(1.1) [8 points] Let $\mathcal{A}$ be an algorithm that agnostically PAC learns a hypothesis class $\mathcal{H}$. Show that $\mathcal{A}$ also (realizably) PAC learns $\mathcal{H}$.

Answer: TODO

(1.2) [7 points] Let $\mathcal{A} : \mathcal{Z}^* \to \mathcal{H}$ be an algorithm and ℓ a loss such that there is some function $\varepsilon : \mathbb{N} \times (0,1) \to \mathbb{R}$ such that for all $m \in \mathbb{N}$ and $\delta \in (0,1)$, for all $\varepsilon > \varepsilon(m, \delta)$, it holds for all $\mathcal{D}$ that

$$\Pr_{S \sim \mathcal{D}^m, \mathcal{A}} \left(L_\mathcal{D}(\mathcal{A}(S)) - \inf_{h \in \mathcal{H}} L_\mathcal{D}(h) \geq \varepsilon \right) \leq 1 - \delta,$$

where the randomness is over both the choice of sample set S and any internal randomness in the algorithm $\mathcal{A}$. Further suppose that $\varepsilon(m, \delta)$ is nonincreasing in m for each fixed $\delta \in (0,1)$, and that $\lim_{m \to \infty} \varepsilon(m, \delta) = 0$. Show that $\mathcal{A}$ agnostically PAC learns $\mathcal{H}$.

Answer: TODO

(1.3) [10 points] Let A be a learning algorithm, $\mathcal{D}$ a probability distribution, and let L denote the random variable $L_\mathcal{D}(A(S))$ for some loss function bounded in $[0, 1]$. Prove that the following two statements are equivalent:

1. For every $\varepsilon, \delta > 0$, there is some $m(\varepsilon, \delta)$ such that for all $m \geq m(\varepsilon, \delta)$, $\Pr_{S \sim \mathcal{D}^m}(L > \varepsilon) < \delta$.

2. $\lim_{m \to \infty} \mathbb{E}_{S \sim \mathcal{D}^m} L = 0$. *The expected loss goes to zero asymptotically.*

Answer: TODO

(1.4) [15 points] Consider data that is a set of binary attributes, $\mathcal{X} \subseteq \{(0,1)\}^d$ for $d \geq 2$, and has $\mathcal{Y} = \{0, 1\}$. A *binary decision tree* is a model that looks generally like

- If x_3, then:
  - If x_{12}, then:
    * Return 1
  - Otherwise (not x_{12}):
    * If x_1, then
      · Return 0
    * Otherwise (not x_1):
      · Return 1
- Otherwise (not x_3):
  - Return 0

Let $\mathcal{H}$ be the set of all such trees of depth at most $k \leq d$, and let $\hat{h}_S$ be any ERM in $\mathcal{H}_k$.[1] Bound $|\mathcal{H}_k|$ as a function of d and k, and hence show that ERM successfully PAC-learns this class for the zero-one loss based on results from class. State the error bound or sample complexity (either is fine) with explicit constants, but then give a $\mathcal{O}_p$ statement in terms of m, k, and d (treating δ as a constant).

You don't have to try to get the super-tightest-possible bound here, though of course you can if you want to. But also don't be absurdly loose (e.g. totally ignoring the depth-k limitation).

---

[1]This problem is NP-hard, but specialized algorithms can usually solve an *almost*-equivalent problem well in practice.

Hint: A perfect binary tree *of depth k has branches at every "interior" node, i.e. no "early returns." Such a tree has 2^k leaf nodes at the bottom. One thing you can try is to map $\mathcal{H}_k$ onto the set of perfect binary trees.*

Hint: In this kind of bound, it's okay to think of $\mathcal{H}$ as an input-output mapping: if two trees have different representations, but return the same value for every possible input in $\mathcal{X}$, then you can think of them as the same hypothesis, because their value of $L_{\mathcal{D}} - L_S$ must be the same.

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 $X_1 \in \mathcal{SG}(\sigma_1)$ 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 $\mathbb{E}[XY] \le \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}$, 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: $\mathbb{E}[XY] \le \mathbb{E}[X^p]^{1/p}\,\mathbb{E}[Y^q]^{1/q}$ 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, $\Pr\left(\frac{1}{m}\sum_{i=1}^{m} X_i > \mu + \text{something}\right) \le \delta$, 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.[2] Prove that

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

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, $\exp(\mathbb{E}Y) \le \mathbb{E}\exp(Y)$.

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

---

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

3   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 $\mathcal{H} = \{h : [0, C]^d \to \mathbb{R} : h(0) = 0, \|h\|_{\mathrm{Lip}} \leq B\}$ 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

$$\log N(\mathcal{H}, \eta) \leq \left(\frac{aBC}{\eta}\right)^d \tag{3.1}$$

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 (3.1) 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.