CPSC 532D, Fall 2022: Assignment 2
due Wednesday, 12 October 2022, **12:00 noon**

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 if so; feel free to delete whatever boilerplate you want. Or answer in a fresh document.

You can do this with a partner if you'd like (there's a "find a group" post on Piazza), but please **make sure you understand everything you're submitting** – don't just split an assignment in half. If you do parts of the assignment with a partner and parts separately, submit separate solutions, and say in each part you did together who you did it with.

If you look stuff up anywhere other than in **SSBD, MRT, Telgarsky, or Wainwright**, **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.

Submit your answers as a single PDF on Gradescope: link and login instructions on the Canvas site, which you should now be able to get to even if you're not yet officially enrolled yet. 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 names 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`).

Some notes on concentration inequalities

Here are some definitions and results from class which may be useful on this assignment.

Definition 1. *A random variable X with mean $\mu = \mathbb{E}[X]$ is called* sub-Gaussian with variance parameter $\sigma \geq 0$, *written $X \in \mathcal{SG}(\sigma)$, if $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{1}{2}\lambda^2\sigma^2}$ for all $\lambda \in \mathbb{R}$.*

We motivated this definition by noting that a Gaussian $\mathcal{N}(\mu, \sigma^2)$ is $\mathcal{SG}(\sigma)$.

Notice that if $\sigma_1 < \sigma_2$, then anything that's $\mathcal{SG}(\sigma_1)$ is also $\mathcal{SG}(\sigma_2)$.

Proposition 2 (Hoeffding's lemma)**.** *A real-valued random variable bounded in $[a, b]$ is $\mathcal{SG}\left(\frac{b-a}{2}\right)$.*

Proposition 3. *If $X_1 \in \mathcal{SG}(\sigma_1)$ and $X_2 \in \mathcal{SG}(\sigma_2)$ are independent, $X_1 + X_2 \in \mathcal{SG}(\sqrt{\sigma_1^2 + \sigma_2^2})$.*

Proof. $$\mathbb{E}[e^{\lambda(X_1+X_2-\mathbb{E}[X_1+X_2])}] = \mathbb{E}[e^{\lambda(X_1-\mathbb{E}[X_1])}]\,\mathbb{E}[e^{\lambda(X_2-\mathbb{E}[X_2])}] \leq e^{\frac{1}{2}\lambda^2\sigma_1^2}\,e^{\frac{1}{2}\lambda^2\sigma_2^2} = e^{\frac{1}{2}\lambda^2\left(\sqrt{\sigma_1^2+\sigma_2^2}\right)^2}.$$ □

Proposition 4. *If $X \in \mathcal{SG}(\sigma)$, then $aX \in \mathcal{SG}(|a|\sigma)$ for any $a \in \mathbb{R}$.*

Proof. $$\mathbb{E}[e^{\lambda(aX-\mathbb{E}[aX])}] = \mathbb{E}[e^{(a\lambda)(X-\mathbb{E}\,X)}] \leq e^{\frac{1}{2}(a\lambda)^2\sigma^2} = e^{\frac{1}{2}\lambda^2(|a|\sigma)^2}.$$ □

Proposition 5 (Markov)**.** *If X is a nonnegative-valued random variable,* $\Pr(X \geq t) \leq \frac{1}{t}\,\mathbb{E}\,X$.

Proof. Take expectations of both sides of $t\mathbb{1}(X \geq t) \leq X$. □

Proposition 6 (Chernoff, sub-Gaussians)**.** *If $X \in \mathcal{SG}(\sigma)$, then* $\Pr(X \geq \mathbb{E}\,X + t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$ *for $t \geq 0$.*

Proof. Note that $\Pr(X \geq \mathbb{E}\,X + t) = \Pr(\exp(\lambda(X - \mathbb{E}\,X)) \geq \exp(\lambda t))$ for any $\lambda \in \mathbb{R}$. Applying Markov's inequality gives an upper bound of $\exp(-\lambda t)\,\mathbb{E}\exp(\lambda(X - \mathbb{E}\,X)) \leq \exp(\frac{1}{2}\lambda^2\sigma^2 - \lambda t)$. Plug in $\lambda = t/\sigma^2$. □

Since $-X$ is also $\mathcal{SG}(\sigma)$ by Proposition 4, the same bound holds for a lower deviation $\Pr(X \leq \mathbb{E}\,X - t)$. A union bound then immediately gives $\Pr(|X - \mu| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$.

Proposition 7 (Hoeffding)**.** *If $X_1, \dots, X_n$ are independent and each $\mathcal{SG}(\sigma_i)$ with mean μ_i, for all $\varepsilon \geq 0$*

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n X_i \geq \frac{1}{n}\sum_{i=1}^n \mu_i + \varepsilon\right) \leq \exp\left(-\frac{n^2\varepsilon^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

Proof. By Propositions 3 and 4, $\frac{1}{n}\sum_{i=1}^n X_i \in \mathcal{SG}\left(\frac{1}{n}\sqrt{\sum_{i=1}^n \sigma_i^2}\right)$. Then apply Proposition 6. □

If the X_i have the same mean $\mu_i = \mu$ and parameter $\sigma_i = \sigma$, this becomes

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n X_i \geq \mu + \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right), \tag{Hoeffding}$$

which can also be stated as that, with probability at least $1 - \delta$,

$$\frac{1}{n}\sum_{i=1}^n X_i \leq \mu + \sigma\sqrt{\frac{2}{n}\log\frac{1}{\delta}}. \tag{Hoeffding'}$$

1 A question that's probably more or less about PAC learning [30 points]

(a) [10 points] Show that if A is an algorithm that agnostically PAC learns a hypothesis class $\mathcal{H}$, it's also a (realizable) PAC learner for $\mathcal{H}$.

Answer: TODO

(b) [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 $n(\varepsilon, \delta)$ such that for all $n \geq n(\varepsilon, \delta)$, $\Pr_{S \sim \mathcal{D}^n}(L > \varepsilon) < \delta$. *This is the kind of guarantee we've been using for PAC learning.*

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

Answer: TODO

(c) [10 points] Say we have a *countable* hypothesis set $\mathcal{H}$, and we have a distribution p assigning probability mass $p(h)$ to each hypothesis h in $\mathcal{H}$. Assume a loss bounded in $[0, 1]$. Use Hoeffding's inequality to prove that the "Bayesian-ish" bound

$$L_{\mathcal{D}}(h) \leq L_S(h) + \sqrt{\frac{1}{2n}\left[\log \frac{1}{p(h)} + \log \frac{1}{\delta}\right]}$$

holds uniformly over all $h \in \mathcal{H}$ with probability at least $1 - \delta$. *Because this is uniform, we could then plug it in to get a particular bound on our generalization error after training based on $L_S(\hat{h})$ and $p(\hat{h})$.*

Hint: In the finite case, we effectively allotted an error probability of $\delta/|\mathcal{H}|$ to each hypothesis having high error and applied (Hoeffding'). *Do something slightly different here.*

Answer: TODO

2 Sums, means, and maxes of sub-Gaussians [60 points]

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

(a) [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 $\mathcal{SG}(\sqrt{2}\sqrt{\sigma_1^2 + \sigma_2^2})$.

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

(b) [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

(c) [15 points] Try constructing a version of (Hoeffding) that doesn't assume independence based on applying part (a) or (b) rather than Proposition 3. How much worse is what you just got than (Hoeffding)? How does what you got compare to the best such inequality you could possibly hope to get, with any approach?

Answer: TODO

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

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

Hint: Bound $\exp(\lambda\,\mathbb{E}\max_i X_i)$ *in terms of something that only depends on* n*,* 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* n*,* σ*, and* λ*. Then optimize* λ *in that bound to get something only depending on* n *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.[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 Challenge problem: Bernstein-based rates [10 points]

I'm going to start introducing "challenge problems" on the homeworks. These will be notably more difficult than the other problems, but worth at most 10 points, so that if you just totally ignore these but do well on everything else, you can still absolutely get an A in the course (final average 85-89%). To have a good chance at an A+, however, you'll have to do at least some of the challenge problems. (I might adjust this scheme over the course of the term if it seems necessary.)

This problem will consider arbitrary loss functions bounded in $[0, 1]$, and finite hypothesis classes $\mathcal{H}$.

Recall in class that we showed two different kinds of rates for ERM: given n samples, with probability at least $1 - \delta$, ERM achieves excess error, $L_{\mathcal{D}}(\hat{h}_S) - \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(\mathcal{H})$, of

- $\frac{1}{n}\left[\log|\mathcal{H}| + \log\frac{1}{\delta}\right]$ in the realizable setting, where there's an h^* with $L_S(h^*) = 0$
- $\sqrt{\frac{2}{n}\left[\log|\mathcal{H}| + \log\frac{2}{\delta}\right]}$ in the agnostic setting.

This difference between $1/n$ and $1/\sqrt{n}$ is a big deal (often called a "fast rate" versus a "slow rate"), and it's pretty annoying that we have to decide in advance whether we think the problem is realizable, and if it's "almost realizable" then we don't get any "partial credit." In this problem, you'll prove a version of this bound that unifies the two rates.

One way to do this is based on *Bernstein's inequality*:

Proposition 8 (Bernstein, bounded variables)**.** *Let $X_1, \dots, X_n$ be independent random variables with means $\mu_i \in \mathbb{R}$, variances v_i*[3]*, and almost surely bounded in $[a, b]$. Then*

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\left(\frac{1}{n}\sum_{i=1}^{n} v_i\right) + \frac{2}{3}(b-a)\varepsilon}\right). \tag{Bernstein}$$

(a) [4 points] Use Proposition 8 to show that for a fixed h, it holds with probability at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^n$ that

$$L_S(h) \leq L_{\mathcal{D}}(h) + \frac{C_1 \log\frac{1}{\delta}}{n} + \sqrt{\frac{C_2 \log\frac{1}{\delta}}{n} L_{\mathcal{D}}(h)}. \tag{*}$$

for some (simple) universal constants C_1, C_2; give values for those constants.

Also show that, separately, it holds with probability at least $1 - \delta$ that

$$L_S(h) \geq L_{\mathcal{D}}(h) - \frac{C_1 \log\frac{1}{\delta}}{n} - \sqrt{\frac{C_2 \log\frac{1}{\delta}}{n} L_{\mathcal{D}}(h)}. \tag{**}$$

The same argument can show both; you don't need to write it all out twice, but we'll need both directions.

You don't need to do this part to do the next one; you can just write that in terms of C_1 and C_2.

Hint: This is not *an exact inverse of the Bernstein probability bound, the way that* (Hoeffding') *is for* (Hoeffding)*; we're being a little loose here to get a simpler form.*

Answer: TODO

Now on to the bound. Let $\hat{h}_S$ denote an ERM, and let h^* be an optimal hypothesis from a finite $\mathcal{H}$, $h^* \in \arg\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$, with loss $L^* = L_{\mathcal{D}}(h^*)$.[4]

[3] We're not calling this σ_i^2 to avoid confusion with the sub-Gaussian parameter, which this is emphatically not; Bernstein's inequality is more closely connected to something called sub-exponential variables.

[4] A minimizer is guaranteed to exist, since $\mathcal{H}$ is finite, but at least in my proof it doesn't actually matter that h^* be minimal; you could plug any hypothesis you like into the bound.

(b) [6 points] Prove a bound on $L_{\mathcal{D}}(\hat{h}_S) - L^*$ in terms of L^*, $|\mathcal{H}|$, and n of the form

$$L_{\mathcal{D}}(\hat{h}_S) \le L^* + \mathcal{O}\left(\frac{1}{n} \log \frac{|\mathcal{H}| + 1}{\delta} + \sqrt{\frac{L^*}{n} \log \frac{|\mathcal{H}| + 1}{\delta}} \right).$$

For full credit, use explicit constants in your answer, not $\mathcal{O}$.

You can assume that $\frac{1}{n} \log \frac{|\mathcal{H}|+1}{\delta} = o(1)$, i.e. a term with that squared will be "lower-order."

Hint: In my solution, $\mathcal{H}$ and δ only appear in the form $\log \frac{|\mathcal{H}|+1}{\delta}$; the 1 isn't some constant hidden by $\mathcal{O}$, it's just a 1.

Hint: Recall that since $\hat{h}_S$ is an ERM, $L_S(\hat{h}_S) \le L_S(h^*)$.

Hint: After doing the things in the hints above, you'll probably get something of the form $L_{\mathcal{D}}(\hat{h}_S) \le \beta\sqrt{L_{\mathcal{D}}(\hat{h}_S)} + \gamma$, where β and γ depend on all the other parameters of the problem. Think about what that equation tells us about $L_{\mathcal{D}}(\hat{h}_S)$, and make your middle school algebra teacher proud.

Answer: TODO