I’m likely to re-use at least some problems from year to year, so if you’re currently in or likely to take a future version of the course, please do not look at this solutions file; it is academic dishonesty to do so.

Personally, I kind of understand the motivation to cheat sometimes in an undergrad course (although, obviously, please please don’t do it in undergrad courses, but like, I get why people would). But in a grad course... it’s really just not going to help you.

If you’re feeling super stressed / whatever and just need to get the assignment in, please write to me about an extension or re-weighting or some kind of route to making things work for you – I’m very willing to be flexible with this kind of thing in grad courses.
1 Concentrating on concentric circles [40 + 4 bonus points]

In this problem, we’ll show that a particular infinite hypothesis class can be PAC-learned with a “direct” proof. Based in part on SSBD exercise 3.3.

Let \( X = \mathbb{R}^2 \), \( Y = \{0, 1\} \), 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 \( h_r(x) = 1_{\{|x| \leq r\}} \) (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 \( L_D(h^*) = 0 \).

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

Answer: It’s enough to show that \( L_S(\hat{h}_S) = 0 \), since the 0-1 loss can’t be negative. For a positive example \((x,1) \in S\), we have that \(|x| \leq r_S\) by definition of \( r_S \), and so \( x \) is classified correctly.

Now, because of the realizability assumption, we know there is some \( r^* \) such that all negative examples have \(|x| > r^*\); moreover, we must have \( r_S \leq r^* \). Thus any negative example in \( S \) must have \(|x| > r^* \geq r_S \), and so it will be classified correctly as well.

(b) [20 + 4 bonus points] Prove that if we observe \( n \geq \frac{1}{\epsilon} \log \frac{1}{\delta} \) samples from a realizable \( D \), then the probability that \( L_D(\hat{h}_S) \geq \epsilon \) is at most \( \delta \).

You may assume that \( D_x \), the distribution of \( x_S \) sampled from \( D \), is continuous. For [4 bonus points], also handle the case where \( D_x \) is not continuous (e.g. if it has a point mass).

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: Again, realizability means that there is some \( r^* \) for which \( f \) agrees almost surely with \( h_{r^*} \).

Note that necessarily \( r^* \geq r_S \), and so \( h \) doesn’t predict 1 for anything that’s actually a 0; the only kind of mistake it can make is predicting 0 for things that are actually a 1, which happens exactly when \( r_S < |x| \leq r^* \). Thus, \( L_D(\hat{h}_S) = \Pr_{x \sim D_S}(r_S < |x| \leq r^*) \).

We want to consider the probability of hitting a “bad” hypothesis: one with \( L_D(h) > \epsilon \), or – upper-bounding very slightly because it’s more convenient – \( L_D(h) \geq \epsilon \). Now, let \( r_S \) be the largest possible radius such that \( \Pr_{x \sim D}(r_S < |x| \leq r^*) \geq \epsilon \). If \( D_x \) is continuous, there is an \( r_S \) where this is exactly equal. In general, there might not be (imagine putting 2\( \epsilon \) mass exactly at norm \( r^* \), but the function \( r \to \Pr_{x \sim D_S}(r_S < |x| \leq r^*) \) will be left-continuous with right limits (aka càdlàg, the opposite of a CDF), so there will be a unique maximum. Thus, the probability of getting a “bad” hypothesis is just

\[
\Pr(L_D(\hat{h}_S) > \epsilon) \leq \Pr(L_D(h_S) \geq \epsilon) = \Pr(r_S \leq r_S).
\]

Because \( r_S = \max_{i \in [n]} \|x_i\| < r^* \), and \( S \sim D^n \), we have

\[
\Pr(r_S \leq r_S) = \Pr(\forall i \in [n], \|x_i\| \notin (r_S, r^*)) = (1 - \Pr(\|x\| \in (r_S, r^*))^n.
\]

But we chose \( r_S \) such that the probability of falling in that interval is at least \( \epsilon \), and so we get

\[
\Pr(L_D(h_S) > \epsilon) \leq (1 - \epsilon)^n \leq \exp(-n\epsilon),
\]

and so if \( n \geq \frac{1}{\epsilon} \log(1/\delta) \) we can guarantee the probability of a bad hypothesis is at most \( \delta \).
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 \( \eta \), but not which points have been flipped.

**c** [5 points] Is \( \hat{h}_S \) still an ERM?

Answer: Not necessarily. For example, suppose that the labels we observe, sorted from smallest norm of \( x \) to largest, are + + + + - - - - - +. Then \( \hat{h}_S \) would classify all points as positive, for 50% training error, while choosing the threshold before the first - would give us only 10% error.

**d** [10 points] Ambitious Angus claims to have proven the following:

For any \( \varepsilon, \delta > 0 \) and \( 0 \leq \eta < \frac{1}{2} \), there is a function \( n_H(\eta, \varepsilon, \delta) \) such that, for any \( n \geq n(\eta, \varepsilon, \delta) \) and any \( \mathcal{D} \) of the form above,

\[
\Pr_{S \sim \mathcal{D}^n} (L_{\mathcal{D}}(\hat{h}_S) > \eta + \varepsilon) \leq 1 - \delta.
\]

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 Angus must be wrong: no such function \( n_H \) can exist.

Answer: To disprove Angus’s statement, it suffices to show one \( \mathcal{D} \) for which it cannot hold. So let’s just use a distribution where \( \|x\| \) is uniform from 0 to 2, with \( h^* = h_1 \) thresholding at a radius of 1. As \( n \to \infty \), approximately \( n/4 \) points will have norm greater than 1.5, and so approximately \( \eta n/4 \) points will have norm above 1.5 and a positive label. Thus \( \hat{h}_S \) will almost surely have a threshold at least 1.5 – but that means that its error rate must be at least 0.25, contradicting Angus’s claim that it will become arbitrarily close to 0.
2 Loss functions [35 points]

The general form of learning problems we’ll usually work with in this course is as follows: \( \mathcal{D} \) is some distribution over a space \( Z \), and \( \ell : \mathcal{H} \times Z \rightarrow \mathbb{R} \) is a loss function.

For example, classification problems as we’ve mostly considered so far are usually framed with empirical risk is

\[
\mathbb{E}_{S \sim \mathcal{D}^n} L_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, z_i) = \mathbb{E}_{z \sim \mathcal{D}} \ell(h, z) = L_D(h).
\]

(a) [10 points] Prove that, for any given \( h \in \mathcal{H} \), \( L_S(h) \) is unbiased: \( \mathbb{E} L_S(h) = L_D(h) \).

Answer:
\[
\mathbb{E}_{S \sim \mathcal{D}^n} L_S(h) = \mathbb{E}_{z \sim \mathcal{D}} \ell(h, z) = L_D(h).
\]

(b) [5 points] Prove that the zero-one loss for \( k \)-way classification (\( \mathcal{Y} = \{1, \ldots, k\} \)) is equal to one minus the accuracy (the portion of correct answers).

Answer: There are two extremely related interpretations of this (ambiguous) question: that the training accuracy is one minus the training loss, or that the generalization accuracy is one minus the true loss. I meant the second one, but either is fine.

\[
L_D(h) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbb{1}(h(x) \neq y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(h(x_i) \neq y_i) = 1 - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(h(x_i) = y_i).
\]

(c) [5 points] For the canonical ImageNet Large Scale Visual Recognition Challenge, images are given with one of a thousand possible labels, and one major way of evaluating those models is the top-5 accuracy: models can make 5 guesses at the label, and we count how often the correct label is one of those 5 guesses. Frame this in the language above: what kind of object does \( h(x) \) output, and what does \( \ell(h, (x, y)) \) look like?

Answer: One reasonable way is to let \( h(x) \) output five guesses, maybe a 5-tuple in \( \mathcal{Y}^5 \) or a set of size exactly 5 with elements in \( \mathcal{Y} \); then \( \ell(h, (x, y)) = \mathbb{1}(y \notin h(x)) \).

Another, more like what these methods typically actually do, would be to have \( h(x) \) output a 1000-dimensional probability vector, and have \( \ell \) check if \( y \) is one of the five largest elements of \( h(x) \). (You’d have to be a little careful about tie-breaking there, though.)

(d) [5 points] Semantic segmentation is a computer vision problem where we try to label each pixel of an image as belonging to one of several classes (“tree,” “street,” “dog,” etc.). Let \( S = ((x_1, y_1), \ldots, (x_n, y_n)) \) where \( x_i \) are the given input images and \( y_i \) their corresponding pixel labels. One typical evaluation metric is called mIoU (“mean intersection over union”), and is measured on a test set as follows:

\[
\frac{1}{\text{# of classes}} \sum_{k=1}^{\text{# of classes}} \frac{\text{# of pixels from all images predicted as } k \text{ with label } k}{\text{# of pixels from all images predicted as } k \text{ and/or with label } k}.
\]
For next time: explicitly name the mIoU as some function, say $Q_S$; be explicit about measuring it on a set $S$ in “all images” in the equation, etc. Also throw in a $1-$ so it’s a better “loss.” Argue that this metric cannot be expressed using the form of loss function above on the given $S$. (A formal proof isn’t necessary on this question, just a good intuitive argument.)

Answer: The numerators and denominators from the different images are all “mixed together”: there’s no way to break it up into a sum over images. Any reasonably convincing argument along those lines would get full points, but here’s one way to see this formally:

Consider a problem with $k = 2$.

Let $(x_1, y_1)$ be a 10-pixel image which are all class 1, where $h$ correctly predicts 9 pixels as class 1, and incorrectly predicts 1 pixel as class 2.

Let $(x_2, y_2)$ be a 10-pixel image which are all class 2, where $h$ correctly predicts 9 pixels as class 2, and incorrectly predicts 1 pixel as class 1.

The mIoU for $S_1 = ((x_1, y_1))$ is $\frac{1}{2} \times \frac{9}{10} + \frac{1}{2} \times \frac{0}{1} = \frac{9}{20} = 0.45$; it’s the same for $S_2 = ((x_2, y_2))$.

If the mIoU were expressible as the kind of loss we look at here, the mIoU for $S_{12} = ((x_1, y_1), (x_2, y_2))$ would have to also be 0.45.

But it’s actually $\frac{1}{2} \times \frac{9}{11} + \frac{1}{2} \times \frac{9}{11} = \frac{9}{11} \approx 0.81$!

(e) [5 points] Principal component analysis (PCA) is a common technique that can try to find an underlying low-dimensional structure by a linear mapping to a low-dimensional space: a data point $x \in \mathbb{R}^d$ is mapped to a latent code $z = Wx \in \mathbb{R}^k$, where $W \in \mathbb{R}^{k \times d}$ is a matrix with orthonormal rows ($WW^\top = I$) that we want to learn. To reconstruct a point from its latent code $z$, we take $W^\top z$. To find $W$, we minimize the squared reconstruction error on a training set:

$$\arg \min_{W:WW^\top = I} \sum_{i=1}^{n} ||W^\top Wx_i - x_i||^2.$$  \hspace{1cm} (PCA)

Frame PCA as an empirical risk minimization problem: what are the data domain $Z$, the sample $S$, the hypothesis class $H$, and the loss function $\ell : H \times Z \to \mathbb{R}$ such that the set of ERMs is exactly the set of solutions to (PCA)?

Answer: We can take $Z = \mathbb{R}^d$, a sample of $S = (x_1, \ldots, x_n)$, a hypothesis class of $H = \{W \in \mathbb{R}^{k \times d} : WW^\top = I\}$, and the loss function $\ell(W, x) = ||x - W^\top Wx||^2$. Then the IRM problem is exactly like (PCA) except that there’s an extra $\frac{1}{n}$, which doesn’t affect the arg min.

(f) [5 points] Frame the problem of fitting a Gaussian distribution to a set of scalar observations as loss minimization above: what are the data domain $Z$, the sample $S$, the hypothesis class $H$, and the loss function $\ell : H \times Z \to \mathbb{R}$ such that the ERM agrees with the maximum likelihood estimate?

Answer: (I should have said that the samples were independent here.)

The parameters of a Gaussian distribution on $\mathbb{R}$ are just the mean and standard deviation, so we can let $H = \{(\mu, \sigma) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}_{\geq 0}\}$. Our observations are also just numbers in $Z = \mathbb{R}$, so our sample is
\( S = (x_1, \ldots, x_n) \) (or call them \( z_i \), whatever). Now, remember that maximum likelihood is
\[
\begin{align*}
\arg\max_{\mu, \sigma} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) &= \arg\max_{\mu, \sigma} \log \left( \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \right) \\
&= \arg\max_{\mu, \sigma} \sum_{i=1}^{n} \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \right) \\
&= \arg\min_{\mu, \sigma} \frac{1}{n} \sum_{i=1}^{n} -\log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \right)
\end{align*}
\]
since the log is monotonic, and multiplying by a negative number flips maximization into minimization. This last term is ERM with the negative log-likelihood
\[
\ell((\mu, \sigma), x) = -\log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \right) = \frac{1}{2} \log(2\pi) + \log \sigma + \frac{(x_i - \mu)^2}{2\sigma^2}.
\]
Notice that nothing up to the last (unnecessary) simplification actually cared what the likelihood was; this works for any probabilistic model. (Also, the additive constant doesn’t do anything here; you could drop it.)
3 Bayes optimality [25 points]

A Bayes-optimal predictor is a predictor which achieves the lowest possible error for any function, regardless of a choice of hypothesis class or anything like that.\footnote{As usual in this course, I’m ignoring issues of measurability and so on; this should all be formalizable by being appropriately careful and using “disintegrations” of probability measures, etc, but for the purpose of this question you can just ignore such issues.}

We’ll consider loss functions of the form \( \ell(h, (x, y)) = \lambda(h(x), y) \), where \( h : \mathcal{X} \to \hat{\mathcal{Y}} \) and \( \lambda : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R} \).\footnote{This is often how loss functions are defined in the first place; there are a few cases in the course where the more general \( \ell \) form is more convenient, but for this question, the \( \lambda \) form is a little easier.}

(We often have \( \hat{\mathcal{Y}} = \mathcal{Y} \), as in binary classification, but not necessarily, as you may have seen in the previous question.)

A Bayes-optimal predictor has no pesky constraints on the form of function it’s going to be, so it can just give an arbitrary different prediction for each \( x \). Let \( F(x) \) denote the conditional distribution of \( y \) for a given \( x \) under \( D \): if \( D \) is deterministic, this won’t be a very interesting distribution (a point mass), but in general it might be more complicated.

(a) [10 points] Argue that if \( h \) and \( g \) are predictors such that for every \( x \),
\[
E_{y \sim F(x)} \lambda(h(x), y) \leq E_{y \sim F(x)} \lambda(g(x), y),
\]
then we necessarily have that \( L_D(h) \leq L_D(g) \).

Answer: Using the law of total expectation, we can see that
\[
L_D(h) = E_{(x,y) \sim D} \lambda(h(X), y) = E_{x \sim D_x} \left[ E_{y \sim F(x)} \lambda(h(x), y) \right] \leq E_{x \sim D_x} \left[ E_{y \sim F(x)} \lambda(g(x), y) \right] = L_D(g).
\]

(b) [5 points] Use the above formulation to argue that \( f_{D,0-1}(x) = \begin{cases} 1 & \text{if } \Pr_{y \sim F(x)}(y = 1) \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \) is Bayes-optimal for binary classification problems with 0-1 loss.

Answer: We have
\[
E_{y \sim F(x)} \lambda(\hat{y}, y) = \Pr_{y \sim F(x)}(y \neq \hat{y}) \times 1 + \Pr_{y \sim F(x)}(y = \hat{y}) \times 0 = \Pr_{y \sim F(x)}(y \neq \hat{y}),
\]
which is minimized by the choice above. (It would also be minimized by changing ≥ to >.)

(c) [5 points] Use the above formulation to derive the Bayes-optimal predictor for a binary classification problem with the loss of an “is this mushroom edible” classifier:
\[
\lambda(\hat{y}, y) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 0.01 & \text{if } \hat{y} = 0 \neq y = 1 \\ 1 & \text{if } \hat{y} = 1 \neq y = 0 \end{cases}.
\]
Answer: The expected loss here is

$$E_{y \sim \mathcal{F}(x)} \lambda(\hat{y}, y) = \begin{cases} \Pr_{y \sim \mathcal{F}(x)}(y = 1) \times 0.01 + \Pr_{y \sim \mathcal{F}(x)}(y = 0) \times 0 & \text{if } \hat{y} = 0 \\ \Pr_{y \sim \mathcal{F}(x)}(y = 1) \times 0 + \Pr_{y \sim \mathcal{F}(x)}(y = 0) \times 1 & \text{if } \hat{y} = 1 \end{cases}$$

Thus, we should pick $\hat{y} = 1$ only if

$$0.01 \Pr_{y \sim \mathcal{F}(x)}(y = 1) \geq \Pr_{y \sim \mathcal{F}(x)}(y = 0) = 1 - \Pr_{y \sim \mathcal{F}(x)}(y = 1)$$

i.e. $\Pr_{y \sim \mathcal{F}(x)}(y = 1) \geq \frac{1}{1.01} \approx 0.99$:

$$f_{D,\lambda} = \begin{cases} 1 & \text{if } \Pr_{y \sim \mathcal{F}(x)}(y = 1) > \frac{1}{1.01} \\ 0 & \text{otherwise} \end{cases}$$

**d** [5 points] Use the above formulation to argue that

$$f_{D, sq}(x) = E_{y \sim \mathcal{F}(x)} y$$

is Bayes-optimal for scalar regression problems with squared loss $\lambda(\hat{y}, y) = (\hat{y} - y)^2$.

Answer: We can see this by, for example,

$$\frac{d}{d\hat{y}} E_y (\hat{y} - y)^2 = \frac{d}{d\hat{y}} [\hat{y}^2 - 2\hat{y} E_y y + E_y y^2] = 2(\hat{y} - E_y y),$$

which is zero only if $\hat{y} = E_y y$. To check this point is actually a minimum, note that the second derivative is $2 > 0$ everywhere.