Selected problems for "Patterns, predictions, and actions: Foundations of machine learning"

Moritz Hardt Benjamin Recht

June 18, 2024

Abstract

These are selected problems for the textbook: Patterns, predictions, and actions: Foundations of machine learning. Hardt and Recht. Princeton University Press, 2022.

You can access the textbook at https://mlstory.org.

These exercises primarily come from teaching UC Berkeley's CS 281a in the Fall of 2019, 2020, and 2021. We are grateful to our graduate student instructors Mihaela Curmei, Sarah Dean, Frances Ding, Sara Fridovich-Keil, Wenshuo Guo, Chloe Hsu, Meena Jagadeesan, John Miller, Robert Netzorg, Juan C. Perdomo, and Vickie Ye, for their help in developing this resource.

Contents

1 Fundamentals of prediction

1.1 Problem

Let Y be a continuous random variable distributed over the closed interval [0,1]. Under the null hypothesis H_0, Y is uniform:

$$p_{Y|H}(Y|H_0) = \begin{cases} 1 & 0 \leq y \leq 1 \\ 0 & otherwise \end{cases}$$

Under the alternative hypothesis H_1, the conditional pdf of Y is as follows:

$$p_{Y|H}(Y|H_1) = \begin{cases} 2y & 0 \leq y \leq 1 \\ 0 & otherwise \end{cases}$$

The *a priori* probability that y is uniformly distributed is p.

1. **Find** the decision rule that minimizes the probability of error.

2. **Find** the closed form expression for the operating characteristic of the likelihood ratio test (LRT), i.e., P_D as a function of P_F for the LRT.

3. Suppose we require P_D to be at least $(1+\epsilon)P_F$, where $\epsilon > 0$ is a fixed constant. **Find** $P_D^{\max}(\epsilon)$, the maximal value of P_D that is achievable under this constraint.

1.2 Problem

A disease has two varieties: the "0" strain and the "1" strain, with *a priori* probabilities p_0 and p_1 respectively.

1. Initially, a rather noisy test was developed which strain is present for patients who are known to have one of the two varieties. The output of the test is the value y_1 of a random variable Y_1. Given that the strain is "0" ($H = H_0$), $Y_1 = 5 + Z_1$, and given that the strain is "1" ($H = H_1$), $Y_1 = 1 + Z_1$. Here Z_1 is a random variable indicating measurement noise which is independent of H and is Gaussian with $Z_1 \sim \mathcal{N}(0, \sigma^2)$. **Find the MAP decision rule** i.e., determine the set of observations y_1 for which the decision is $\hat{H} = H_1$. **Compute the error probabilities** $\mathbb{P}[\hat{H} = H_1 | H = H_0]$ and $\mathbb{P}[\hat{H} = H_0 | H = H_1]$.

2. Suppose a new medical procedure is devised with two observation random variables Y_1, Y_2 with Y_1 being the same as in the first part. And $Y_2 = Y_1 + Z_2$ where Z_2 is independent of both Z_1 and H and $Z_2 \sim \mathcal{N}(0, \sigma^2)$. **Find the MAP decision rule** for $\hat{H}$ in terms of the joint observation (y_1, y_2). **Compute the error probabilities** $\mathbb{P}[\hat{H} = H_1 | H = H_0]$ and $\mathbb{P}[\hat{H} = H_0 | H = H_1]$. Did the extra measurement improve the error probabilities? **Explain** why or why not?

3. Suppose in the test in the first part with a single observation y_1, the random variable Z_1 is uniformly distributed between 0 and 1 instead of being Gaussian. Again, **find** the MAP decision rule and error probabilities.

1.3 Problem

In this problem, we consider an automated resume screening tool which is used by a company to sort candidates based on whether or not they are predicted to be invited for an on site interview after an initial phone screen. Let the random variable X denote the features of a candidate's application and Y denote whether a candidate is invited for an on site interview, where $Y = 1$ indicates that an individual was invited.

1. Suppose that there are many qualified individuals looking for jobs and that paying recruiters to call applicants is expensive. As a result, it is comparatively half as costly for the company to miss a candidate who would have been invited on site than it is to spend time calling an individual who is not invited for an interview (i.e. for some $\alpha > 0$, $C_{10} = \alpha$, $C_{01} = \frac{\alpha}{2}$, and other costs are zero). **Show that the company's optimal decision rule for resume screening has the form**

$$s(x) = \mathbb{E}[Y|X = x] \geq t\,,$$

and find the value of t.

2. Now suppose that unemployment has gone down, and there are no longer many qualified candidates looking for jobs. As a result, it is instead twice as costly to miss good candidates than it is to call ones who are not invited for an interview (i.e. for some $\beta > 0$, $C_{10} = \beta$, $C_{01} = 2\beta$, and other costs are zero). **How does the optimal decision rule change?**

Suppose now that some score function $\hat{s}(x)$ has been estimated from historical data, and a threshold rule is applied to assign individuals the screening predictions $\widehat{Y} = 1$ for those who will be considered more closely by recruiters and $\widehat{Y} = 0$ for those who will not. In the United States, it is illegal to discriminate against job applicants on the basis of religion, and your job is to evaluate this tool with that in mind. Below is a table which shows the predictions and outcomes for applicants split by membership in a minority religious group, with $A = 1$ indicating that an individual is a member of this group and $A = 0$ indicating that they are not. We have data from 500 candidates in the religious group and $5,000$ candidates not in the religious group.

	$A=1$		
	$Y=0$	$Y=1$	
$\widehat{Y}=0$	360	40	400
$\widehat{Y}=1$	40	60	100
	400	100	

	$A=0$		
	$Y=0$	$Y=1$	
$\widehat{Y}=0$	4050	450	4500
$\widehat{Y}=1$	200	300	500
	4250	750	

3. With membership in the religious group as the sensitive attribute, **does this classifier satisfy independence? Does it satisfy sufficiency?** Justify your answer.

4. For the criteria that the classifier doesn't satisfy, **propose a group-dependent change to the threshold** that results in a classifier that does satisfy the criteria. You do not need to specify exact quantities, rather comparisons with the current threshold. You should not propose a trivial threshold that results in 0% or 100% acceptance rates.

5. **Compare and contrast** the value of the intervention you suggested in part 4 for the following two circumstances:

    - You learn that the historical data comes from a hiring manager who is a member of the religious group and has been heard telling fellow members that they have an "in" regardless of their qualifications.
    - You learn that there is a well regarded religious university nearby that sends the resumes of highly qualified students to the company. Historically, these candidates have highly relevant skill sets and make up a majority of applications from the religious group.

1.4 Problem

This question introduces the Hoeffding bound.

1. ϵ is a random variable that is either +1 or -1 with equal probability. We call such random variables Rademacher random variables. Show that ϵ is *sub-Gaussian* with parameter $\sigma = 1$. *i.e.*, show $\forall\lambda$:

$$\mathbb{E}[\exp(\lambda\epsilon)] \leq \exp(\sigma^2\lambda^2/2)$$

2. The result from part 1 can be generalized to show that any bounded random variable $X \in [a, b]$ is sub-Gaussian with $\sigma = (b-a)/2$. Using this fact, prove that:

$$P\left[\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \geq t\right] \leq \exp\left(\frac{-2t^2}{n(b-a)^2}\right)$$

   where X_i are independent random variables supported on $[a, b]$. This result is known as the Hoeffding bound.

1.5 Problem

Suppose we are deciding between two hypotheses $H \in \{H_0, H_1\}$ based on observation $y \in \mathcal{Y} \in \mathbb{R}^+$. The models under the two hypotheses are

$$H_0: \quad p_{Y|H}(y|H_0) = \begin{cases} e^{-y} & if\ y \geq 0 \\ 0 & otherwise, \end{cases}$$

$$H_1: \quad p_{Y|H}(y|H_1) = \begin{cases} 2e^{-2y} & if\ y \geq 0 \\ 0 & otherwise, \end{cases}$$

The prior beliefs are$\mathbb{P}(H = H_1) = p$ and $\mathbb{P}(H = H_0) = 1-p$. Associated with the possible decisions are the costs $C_{00} = C_{11} = 0$, and $0 \leq C_{01}$, $C_{10} \leq \infty$, where C_{ij} is the cost of deciding $\hat{H}(y) = H_i$ when the correct hypothesis is $H = H_j$.

1. The decision rule $\hat{H}(\cdot)$ that minimizes the expected cost takes the form:

$$\hat{H} = \begin{cases} H_0 & if\ y \geq \gamma \\ H_1 & if\ y < \gamma, \end{cases}$$

   Express γ in terms of C_{10}, C_{01} and p.

2. **Express** P_D as a function of P_F. Note that P_D and P_F are defined as before:

$$P_D = \mathbb{P}(\hat{H} = H_1 | H = H_1)$$
$$P_F = \mathbb{P}(\hat{H} = H_1 | H = H_0)$$

   In the remainder of the problem, consider minimizing expected cost over "3-way" decision rules, whereby, in addition to $\hat{H}(y) = H_0$ or $\hat{H}(y) = H_1$, one can decide $\hat{H}(y) =$'?' ("I don't know") for some value(s) of y. Let us denote the corresponding costs using $C_{?0}$ and $C_{?1}$ when the underlying hypotheses are H_0 and H_1, respectively. Assume the costs are chosen to satisfy $0 = C_{00} \leq C_{?0} \leq C_{10}$ and $0 = C_{11} \leq C_{?1} \leq C_{01}$, so that admitting "I don't know" is less costly than making a wrong decision but more costly than making a correct decision.

3. The optimal decision rule $\hat{H}_{3-way}(\cdot)$ in this case can be expressed in the form

$$\hat{H}_{3-way}(y) \begin{cases} H_0 & if\ r(y) \leq u\ and\ r(y) \leq v \\ H_1 & if\ r(y) \geq u\ and\ r(y) \geq w \\ \text{'?'} & if\ r(y) \geq v\ and\ r(y) \leq w \end{cases}$$

   where $r(y) = \frac{\pi_1(y)}{\pi_0(y)}$ with $\pi_0(y) = \mathbb{P}(H = H_0 | Y = y)$ and $\pi_1(y) = \mathbb{P}(H = H_1 | Y = y)$, and u, v, w are constants. **Express** u, v, w in terms of the costs $C_{ij} (i \in \{0, 1, \text{'?'}\}$ and $j \in \{0, 1\})$.

4. **Determine** whether the following (italicized) statement is **true or false**, and **justify** your answer:

For at least some value(s) of P_F, the optimal 3-way decision rule can achieve a greater P_D than that corresponding to the operating characteristic you found in part (b).

1.6 Problem

In this problem, we continue with the case study from the above criminal justice case study with conceptual questions regarding the claims made by both companies. In ProPublica's 2016 investigation, they claim that COMPAS exhibited racial bias against black individuals. Specifically, the investigation revealed a racial disparity in the *error rates* of the tool. A higher rate of black than white defendants designated as "high risk" did not recividate, while a higher rate of white than black defendants designated as "low risk" did.

Northpointe, the company that sells COMPAS, published a report in response, arguing that their risk socres are equally accurate and predictive for white and black defendants. In order to evaluate whether both of their claims are true, we analyze the allegations and response in the framework of our non-discrimination criteria: independence, separation, and sufficiency.

In this problem, we will view the problem as binary. We let the classifier $\hat{Y}$ be 1 if a defendant is "high risk" and 0 if they are "low risk" according to their COMPAS score. Let Y be the true outcome, 1 if an individual recidivated and 0 otherwise. Finally, let A be the race of the defendant. We will consider the nondiscrimination criteria of *independence*, *separation*, and *sufficiency*.

1. **Interpret the following statements from ProPublica as relations between conditional distributions of the classifier, outcome, and sensitive attribute:**

   (a) Black defendants who did not recidivate over within two years were nearly twice as likely to be misclassified than their white counterparts.

   (b) White defendants who re-offended within the next two years were mistakenly labeled "low risk" almost twice as often.

2. **Do ProPublica's statements imply anything about COMPAS with respect to the nondiscrimination criteria? If so, please explain the way(s) in which they relate to them.**

3. **Interpret the following statements from Northpointe as relations between conditional probabilities:**

   (a) In comparison with white defendants, a similar percentage of black defendants were labeled "higher risk" but did not re-offend.

   (b) A comparable percentage of black as white defendants were labeled "low risk" but did re-offend.

4. **How are NorthPointe's statements related to the nondiscrimination criteria?**

5. Suppose *sufficiency* is satisfied. We define p_a as the proportion of group a predicted to be "high risk," TPR_a as the true positive rate within group a, FPR_a as the false positive rate within group a, PPV as the positive preditive value, and NPV as the negative predictive value.

   Verify that the following relations are true:

$$\text{TPR}_a = \frac{\text{PPV} \cdot p_a}{\text{PPV} \cdot p_a + (1 - \text{NPV}) \cdot (1 - p_a)}, \text{FPR}_a = \frac{(1 - \text{PPV}) \cdot p_a}{(1 - \text{PPV}) \cdot p_a + \text{NPV} \cdot (1 - p_a)}, \quad (1.1)$$

for $a \in \{\text{black}, \text{white}\}$.

6. **Show that if sufficiency is exactly satisfied and recidivism rates differ between groups, the $p_{\text{black}} \neq p_{\text{white}}$.**

   Explain why 1.1 implies that the separation and sufficiency criteria cannot be simultaneously met when rates of recidivism differ among different groups.

7. **Conversely, show that for $\hat{Y}$ with nonzero true and false positive rates, if separation holds, there must be two groups with different positive predictive values.**

8. **In light of the above discussion, what can we say about ProPublica's statements? Could they have observed anything else?**

9. Consider the observation that individual judges can biased. For instance, judges are statistically more likely to give harsher sentences the week after unexpected football game losses for the state's college football team. [1] **In light of this fact, give a reason why we should or should not favor COMPAS as being more objective.**

[1] https://www.nber.org/papers/w22611

1.7 Problem

This problem is adapted from Chapter 2 of Fairness and Machine Learning: Limitations and Opportunities (fairmlbook.org).

Risk assessment is an important component of the criminal justice system. In the United States, judges set bail and decide pre-trial detention based on their assessment of the risk that a released defendant would fail to appear at trial or cause harm to the public. While *actuarial risk assessment* is not new in this domain, there is increasing support for the use of learned risk scores to guide human judges in their decisions. Proponents argue that machine learning could lead to greater efficiency and less biased decisions compared with human judgment. Critical voices raise the concern that such scores can perpetuate inequalities found in historical data, and systematically harm historically disadvantaged groups.

In this problem, we'll begin to scratch at the surface of the complex criminal justice domain. Our starting point is an investigation carried out by ProPublica[2] of a proprietary risk score, called COMPAS score. These scores are intended to assess the risk that a defendant will re-offend, a task often called recidivism prediction. Within the academic community, the ProPublica article drew much attention to the trade-off between separation and sufficiency that we saw earlier in the chapter.

We'll use data obtained and released by ProPublica as a result of a public records request in Broward Country, Florida, concerning the COMPAS recidivism prediction system [3]. The data is available at https://raw.githubusercontent.com/propublica/compas-analysis/master/compas-scores-tw
csv. Following ProPublica's analysis, we'll filter out rows where `days_b_screening_arrest` is over 30 or under -30, leaving us with $6,172$ rows (you'll need to download the data from the link above and apply this filter).[4]

1. Sufficiency

   (a) **Plot the fraction of defendants recidivating within two years** `(two_year_recid == 1)` **as a function of risk score** `(decile_score)`, for black defendants `(race == "African-American")` and white defendants `(race == "Caucasian")`.

   (b) Based on these plots, **does the risk score satisfy sufficiency across racial groups in this dataset**? This is somewhat subjective, since we want to allow for approximate equality between groups; justify your answer in a sentence or two. *Hint: confidence intervals on your plots may be useful.*

2. Separation

[2] Julia Angwin et al., "Machine Bias," ProPublica, May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[4] The pandas python library (https://pandas.pydata.org/) can be very helpful for handling operations on CSV-like datasets. We suggest you start there, using `pandas.read_csv()` to load in the data. For reporting results, you might find `pandas.DataFrame.to_latex()` helpful.

(a) **Plot the distribution of scores received by the positive class (recidivists) and the distribution of scores received by the negative class (non-recidivists)** for black defendants and for white defendants.

(b) Based on these plots, **does COMPAS achieve separation between the risk score and race**?

3. From scores to classifiers

   Now, we consider using these scores to create a classifier to guess whether an individual will recidivate. The classifier will take the form of a threshold on the COMPAS score.

   (a) **Can you find two thresholds (one for black defendants, one for white defendants) such that FPR and FNR are roughly equal for the two groups** (say, within 1% of each other)? This means that your classifier satisfies *separation*. Note: trivial thresholds of 0 or 11 don't count. *Hint: it may be helpful to plot ROC curves for each race.*

   (b) **Does the resulting classifier satisfy sufficiency**?

Note: you get to make some design decisions about how you produce your tests and interpret results in this problem. It's meant to be open ended, but please state and justify your decisions (e.g. confidence parameters on any confidence intervals), as well as give ample but concise justifications for any conclusions you reach from your analysis.

1.8 Problem

Consider two probability distributions P and Q and two players A and B. The game proceeds as follows: Player A flips a fair coin and denote the result as $C \in \{Heads, Tails\}$. Player B doesn't know the result and he's goal is to guess C. If the result is a head, player A draws a sample x according to the distribution P; otherwise player A draws it according to the distribution Q. Then A shows the sample to B.

Show that the probability that player B's guess is correct under he's best strategy is $\frac{1}{2}(1 + P(D) - Q(D))$, where $D = \{x \in \mathcal{X} : P(x) > Q(x)\}$. Notice that $P(D) - Q(D)$ is the total variational distance, and this illustrates an operational interpretation of it.

2 Supervised learning

2.1 Problem

We have seen an upper bound on the number of mistakes the perceptron algorithm makes on a dataset that is linearly separable by a margin. In this problem, we will analyze what happens when the dataset is not perfectly separable.

For a dataset S, $D(S) = \max_{(x,y)\in S} ||x||$ denotes its diameter and $\gamma(S) = \max_{||w||=1} y_i\langle w, x_i\rangle$ denotes its margin. Then, we know the perceptron makes at most $(2 + D(S)^2)/\gamma(S)^2$ mistakes on S.

1. **Construct a dataset** of size n that has diameter at most D and for which the **perceptron algorithm makes $\Omega(n)$ mistakes in expectation**.

2. In the previous part, we showed that the perceptron algorithm could make many mistakes when the linear separability by a margin condition is violated. **But if the violation is small, we may be able to still get a low mistake bound**. We can quantify the violation as the amount by which to move points to achieve separability by a margin.

   In the dataset $S = ((x_1, y_1), \ldots, (x_m, y_m))$, let u be any vector with $||u|| = 1$ and let $\gamma > 0$. Define the deviation of point i as $d_i = \max\{0, \gamma - y_i\langle u_i, x_i\rangle\}$. And let $\Delta(S; \gamma, u) = \sqrt{\sum_{i=1}^m d_i^2}$. **You will show that then the number of mistakes the perceptron makes on S is at most**

   $$\frac{\left(\sqrt{2 + D(S)^2} + \Delta(S; \gamma, u)\right)^2}{\gamma^2}.$$

   We will show this by a **reduction of the inseparable case to a separable case in a higher dimension**. Consider a higher dimensional dataset $S' = ((x'_1, y_1), \ldots, (x'_m, y_m))$, where each $x_i \in \mathbb{R}^{n+m}$. The first n coordinates of x'_i are the same as the first n coordinates of x_i. The $n + i^{\text{th}}$ coordinate of x'_i is δ (to be set later) and all other coordinates greater than n are zero.

   Extend u to $u' \in \mathbb{R}^{m+n}$ with the first n coordinates of u' equal to u/Z (value of Z to be chosen later) and the $n + i^{\text{th}}$ coordinate is $(y_i d_i)/(Z\delta)$.

$$x'_i = \begin{bmatrix} x_i \\ 0 \\ \vdots \\ \delta \\ \vdots \\ 0 \end{bmatrix} \leftarrow (n+i)^{\text{th}} \qquad u' = \begin{bmatrix} u/Z \\ (y_1 d_1)/(Z\delta) \\ \vdots \\ (y_i d_i)/(Z\delta) \\ \vdots \\ (y_m d_m)/(Z\delta) \end{bmatrix} \leftarrow (n+i)^{\text{th}}$$

(a) **Compute an upper bound on the diameter of** S' ($\max ||x'_i||$).

(b) **Show that** S' **is separable** by u' and compute the margin of separation.

(c) **Set values of** Z, δ **to optimize the mistake bound** on the extended dataset S'.

(d) Show that the predictions of the perceptron on the extended dataset S' are the same as in the original dataset S.

2.2 Problem

In this problem, we study how gradient descent can be used to solve *online learning* problems. Suppose we have a sequence of T examples $(x_1, y_1), \ldots, (x_T, y_T)$ with $x_t \in \mathbb{R}^d$ and $y_t \in \{-1, 1\}$ for all t. In contrast to batch learning, the examples *arrive one at a time*, and we try to predict y_t using models of the form $f_w(x_t) = \text{sign}(\langle w, x_t \rangle)$ for $w \in \mathbb{R}^d$.

We wish to find a sequence of weights $w_1, w_2, \ldots, w_T$ that minimizes the number of *errors* we make on the sequence, where we make an error at time t if $f_{w_t}(x_t) \neq y_t$.

Consider the loss function $\ell_t(w) = \max\{0, -y_t \langle w, x_t \rangle\}$. Suppose $w_1 = 0$, and, after observing example (x_t, y_t), we update w_t with a gradient step on ℓ_t:

$$w_{t+1} = w_t - \nabla \ell_t(w_t).$$

Define the *margin* γ of a linearly separable sequence $(x_1, y_1), \ldots, (x_T, y_T)$ as

$$\gamma = \max_{\|w\|=1} \min_{t \in \{1,\ldots,T\}} y_t \langle w, x_t \rangle.. \tag{2.1}$$

Assume $\gamma > 0$. For simplicity, also assume $\|x_t\|_2 \leq 1$ for all t. Let m denote the number of errors the sequence $\{w_t\}$ makes on the sequence $(x_1, y_1), \ldots, (x_T, y_T)$.

1. Compute a (sub)-gradient of ℓ_t for $t \in \{1, \ldots, T\}$.
2. Prove $\|w_{t+1}\|_2^2 \leq \|w_t\|_2^2 + \mathbb{I}\{f_{w_t}(x_t) \neq y_t\}$.
3. Prove $\|w_{T+1}\|_2^2 \leq m$.
4. Let w^* be a maximizer of (2.1). Prove $\gamma \mathbb{I}\{f_{w_t}(x_t) \neq y_t\} \leq \langle w^*, w_{t+1} - w_t \rangle$ for every t.
5. Summing the bound in part (d) over every t, prove $m\gamma \leq \langle w^*, w_{T+1} \rangle \leq \sqrt{m}$. *Hint: Use telescoping sums and* $w_1 = 0$.
6. Argue the number of errors $m \leq \frac{1}{\gamma^2}$.

2.3 Problem

Suppose you go to a casino which has $n \geq 2$ slot machines, where the payouts from the i-th slot machine are i.i.d. random variables with normal distribution $\mathcal{N}(\theta_i, 1)$, where θ_i are fixed real numbers. Assume that one of the slot machines has mean payout $\theta_{\max}$, and all other machines have mean payout $\theta_{\min} < \theta_{\max}$.

For a fixed error probability $\delta \in (0, 1)$, your goal is to identify the slot machine with the highest mean payout with probability $1 - \delta$. You are allowed to do this by pulling each slot machine T times, guessing the best slot machine based based on the $T \times n$ payouts observed. At the end, you want to ensure that your guess is correct with probability at least $1 - \delta$.

This problem will walk you through steps necessary to determine an upper bound on the number of pulls T you need to identify the best slot machine.

1. **Show that if X is a real valued random variable, you can bound $\mathbb{P}[X \geq t] \leq \mathbb{E}[X(X \geq t)]/t$** for all $t > 0$, where $(X \geq t) = 1$ if $X \geq t$, and 0 otherwise. You may assume X has a continuous density $p(x)$.

2. Let Z be distributed as $\mathcal{N}(0, 1)$. **Show that**

$$\forall t > 0, \quad \mathbb{P}[Z \geq t] \leq \frac{1}{\sqrt{2\pi}t} e^{-t^2/2}$$

   Use this to bound $\mathbb{P}[|Z| \geq t]$.

3. Let $Z_1, \ldots, Z_n$ be distributed $\mathcal{N}(0, 1)$ (not necessarily independent!), and let $n \geq 2$. **Show that for any $t \geq 1$**

$$\mathbb{P}[\max_i |Z_i| \geq t] \leq n \cdot \sqrt{\frac{2}{\pi}} e^{-t^2/2}.$$

4. Suppose $n = 2$. **Show that that, in order to identify the slot machine with the highest payout with probability $1 - \delta$, it suffices to take**

$$T \geq \max\left\{1, \frac{4\log(2/\delta)}{(\theta_{\max} - \theta_{\min})^2}\right\} \text{ samples.}$$

   You should use the inequalities developed earlier in the problem.

5. **Generalize the above result to $n \geq 2$ slot machines.** *Hint: when δ is a constant (say $\delta = 1/2$), there should be a $\log n$ somewhere in your answer.*

2.4 Problem

In this problem, we will train linear classifiers on census data to predict whether individuals have annual incomes above a certain level. To download these datasets, run `pip install folktables`. For documentation on the datasets and prediction tasks, please see

https://github.com/socialfoundations/folktables

1. Download the data source for 2018 in a one-year time horizon using

```
data_source = ACSDataSource(survey_year='2018', horizon='1-Year', survey='person')
acs_data = data_source.get_data(download=True)
```

   Use the following code to create a prediction problem which labels individuals based on whether their income exceeds 50,000:

```
ACSIncomeNew = folktables.BasicProblem(
features=[
    'AGEP',
    'COW',
    'SCHL',
    'MAR',
    'OCCP',
    'POBP',
    'RELP',
    'WKHP',
    'SEX',
    'RAC1P',
],
target='PINCP',
target_transform=lambda x: x > 50000,
group='RAC1P',
preprocess=adult_filter,
postprocess=lambda x: np.nan_to_num(x, -1),
)
```

2. Split the data so that 20% of the data is in the test set, and train a classifier using logistic regression. (Hint: we recommend that you use the sklearn package.) Report the true positive rate, the false positive rate, and the accuracy (i.e. the fraction of data labelled correctly).

3. Now, train classifiers for different income thresholds. That is, change the 50000 in part 2 to 10000, 20000, 75000, and 100000. How do the TPRs and FPRs change as you vary the classification criterion? Provide a possible explanation.

4. For the task with income threshold 50,000, try to beat the performance of logistic regression. (That is, try to get a higher accuracy than that achieved in part 2.) You can use any approach of your choice. Report the true positive rate, the false positive rate, and the accuracy that your approach achieves. **Extra credit:** try to beat 80% accuracy. If you achieve X%, you will receive $(X - 80)/2$ points of extra credit.

2.5 Problem

Suppose we have a sequence of n examples $(x_1, y_1), \ldots, (x_n, y_n)$ with $x_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$ for all i, and $\|x_i\| \leq 1$ for all i. Suppose that this data set is linearly separable with margin γ. That is, there exists a vector $w_\star$ with Euclidean norm 1 such that

$$\gamma = \min_i y_i w_\star^T x_i$$

Consider the empirical risk minimization problem with a hinge loss:

$$\text{minimize} \quad \tfrac{1}{n} \textstyle\sum_{i=1}^n \text{hinge}(w^T x_i, y_i)$$

where

$$\text{hinge}(z, y) = \max(1 - zy, 0)\,.$$

Suppose we run stochastic gradient descent on this problem with batch size 1 and step size 1. We say that iterate w_t makes a *mistake* on example (x_t, y_t) if $\text{hinge}(w_t^T x_t, y_t) > 0$. Assume we initialize $w_1 = 0$.

1. **Compute a subgradient** of the hinge loss at w_t for example (x_t, y_t).
2. Suppose (x_t, y_t) is sampled at iteration t of SGD. **Show that** if w_t does not make a mistake on (x_t, y_t), then $w_{t+1} = w_t$.
3. Suppose (x_t, y_t) is sampled at iteration t of SGD. **Show that** if w_t does makes a mistake on (x_t, y_t), then $\|w_{t+1}\|^2 \leq \|w_t\|^2 + 3$.
4. Also **show that** if a mistake is made at iteration t, $\gamma \leq w_\star^T (w_{t+1} - w_t)$.
5. Let m denote the total number of mistakes made up to round T. **Show that** $\|w_{T+1}\| \leq \sqrt{3m}$.
6. Use the previous derivations to **prove that** the total number of mistakes made by SGD in the first T iterations is at most $3/\gamma^2$.

3 Representations and features

3.1 Problem

1. Let A be a $d \times n$ matrix. For any $\mu > 0$, **show that** $(AA^\mathsf{T} + \mu I)^{-1}A = A(A^\mathsf{T}A + \mu I)^{-1}$.

2. Let $(x_1, y_1), \cdots, (x_n, y_n)$ be a sequence of data points. Each y_i is a scalar and each x_i is a vector in $\mathbb{R}^d$. Let $X = [x_1, \ldots, x_n]^\mathsf{T}$ and $Y = [y_1, \ldots, y_n]^\mathsf{T}$. Consider the *regularized* least squares problem:

$$\min_{w \in \mathbb{R}^d} \|Xw - Y\|_2^2 + \mu \|w\|_2^2$$

   Show that the optimum w_* is unique and can be written as the linear combination $w_* = \sum_{i=1}^n \alpha_i x_i$ for some scalars α. What are the coefficients α_i?
   Hint: you may find eigendecomposition useful for representing α_i

3. More generally, consider the general regularized empirical risk minimization problem

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \text{loss}(w^\mathsf{T} x_i, y_i) + \mu \|w\|_2^2$$

   where the loss function is convex in the first argument. **Prove that the optimal solution has the form $w_* = \sum_{i=1}^n \alpha_i x_i$.** If the loss function is not convex, does the optimal solution have the form $w_* = \sum_{i=1}^n \alpha_i x_i$? **Justify your answer**.

4 Optimization

4.1 Problem

Suppose we have an i.i.d. sequence of n examples $(x_1, y_1), \ldots, (x_n, y_n)$ with $x_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$ for all i, and $\|x_i\| \leq 1$ for all i. Suppose that this data set is linearly separable with margin γ. That is, there exists a vector $w_\star$ with Euclidean norm 1 such that

$$\gamma < y w_\star^T x$$

for any (x, y) that are sampled from the same distribution as (x_i, y_i).

Consider the empirical risk minimization problem with a hinge loss:

$$\text{minimize} \quad \tfrac{1}{n} \textstyle\sum_{i=1}^n \text{hinge}(w^T x_i, y_i)$$

where

$$\text{hinge}(z, y) = \max(1 - zy, 0)\,.$$

On the midterm we analyzed what happens when we run stochastic gradient descent on this problem with batch size 1 and step size 1. We say that iterate w_t makes a *mistake* on example (x_t, y_t) if $\text{hinge}(w_t^T x_t, y_t) > 0$. Assuming we initialize $w_1 = 0$, we showed that no matter how many iterations we ran for, the stochastic gradient algorithm made at most $3/\gamma^2$ total mistakes.

1. Define the 0-1 loss to be
$$\text{loss}_{01}(z, y) = \begin{cases} 1 & \text{if } \text{sign}(z) \neq y \\ 0 & \text{otherwise} \end{cases}.$$
Show that $\text{loss}_{01}(z, y) \leq \text{hinge}(z, y)$.

2. Suppose that (x, y) and (x_{n+1}, y_{n+1}) are sampled from the same distribution as (x_i, y_i) above. Let w_t denote the SGD solution that has been trained until we see no errors on the training set. **Show that**
$$\mathbb{P}[\text{sign}(w_t^T x) \neq y] = \mathbb{E}\left[\frac{1}{n+1} \sum_{i=1}^{n+1} \text{loss}_{01}({w^{(i)}}^T x_i, y_i)\right]$$
where $w^{(i)}$ is the SGD solution arrived at by training on
$$(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_{n+1}, y_{n+1})\,.$$

3. Use the above calculations and the mistake bound from the midterm to **prove that**
$$\mathbb{P}[\text{sign}(w_t^T x) \neq y] \leq \frac{3}{(n+1)\gamma^2}$$

4.2 Problem

Consider the prediction problem of mapping some input $x \in \mathbb{R}^d$ to output $y \in \{-1, 1\}$. A linear predictor is governed by a weight vector $w \in \mathbb{R}^d$, and we typically wish to choose w to minimize the cumulative loss over a set of training examples. Three popular loss functions for classification are defined (on a single example (x, y)) as follows:

[(a)]Squared loss: $\ell(w; x, y) = \frac{1}{2}(y - w^\top x)^2$. Hinge loss: $\ell(w; x, y) = \max\{1 - yw^\top x, 0\}$. Logistic loss: $\ell(w; x, y) = \log(1 + \exp(-yw^\top x))$.

In this problem, we study some of the properties of these loss functions. These functions are ubiquitous in machine learning, and it's important to get good intuition for them.

1. **Show that each of these three loss function is convex**. *Hint: Where possible, use the composition rules for convex functions.*

2. **Compute the (sub)-gradient of each of these three loss functions with respect to** w.

3. Suppose that $\|w\|_2 \leq B$ and $\|x\|_2 \leq C$ for some constants $B, C < \infty$. **Give bounds on the** ℓ_2**-norms** $\|\cdot\|_2$ **of the (sub)-gradients for each of these three losses**.

4.3 Problem

Derive the expression for following questions. Please show your steps.

1. Let $x, a \in \mathbb{R}^n$. Compute $\frac{\partial(x^T \mathbf{a})}{\partial x}$.
2. Let $A \in \mathbb{R}^{n \times n}, x \in \mathbb{R}^n$. Compute $\frac{\partial(x^T A x)}{\partial x}$.
3. Let A, X be $n \times n$ matrices. Compute $\frac{\partial \text{Trace(XA)}}{\partial X}$.
4. Let $g(x) = \sup_{\|z\|_2 \leq 1} x^T z$. Compute the derivative of g^2 where $g^2(x) = (g(x))^2$.
   Hint: first prove an upper bound on $g(x)$, then propose a choice of z that achieves the bound.

4.4 Problem

In the high-dimensional problems, there are usually an infinite number of possible models that perfectly fit the observed data. *When a problem has multiple solutions, different optimization algorithms can find entirely different solutions to the same problem.* Even though all of the solutions perfectly fit the training data, their generalization performance can be vastly different. In this problem, we explore this phenomenon for two widely used optimization algorithms: gradient descent and Adam.

Consider a linear, binary classification problem under the squared loss. Let $X \in \mathbb{R}^{n \times d}$ be an $n \times d$ matrix of features, $y \in \{-1, 1\}^n$ be the corresponding vector of labels, and $\theta \in \mathbb{R}^d$ be the parameter vector. We wish to minimize the empirical risk

$$R_S[\theta] = \frac{1}{2}\|X\theta - y\|_2^2. \tag{4.1}$$

Assume that the rows of X are linearly independent and that $d > n$.

1. Prove that there are infinite many $\theta \in \mathbb{R}^d$ such that $R_S[\theta] = 0$.

2. Gradient descent generates a sequence of points $\left\{\theta_k^{\text{gd}}\right\}$ according to:

$$\theta_{k+1}^{\text{gd}} = \theta_k^{\text{gd}} - \alpha_k \nabla R_S[\theta_k^{\text{gd}}]\,, \tag{4.2}$$

   where α_k is a fixed sequence of learning rates. Assume the sequence α_k is chosen so that gradient descent converges to a minimizer of the objective (4.1). (You don't need to show how to select α_k).

   Suppose we initialize $\theta_0^{\text{gd}} = 0$. **Show that gradient descent converges to the minimum Euclidean norm solution to $X\theta = y$.** This solution is also the *maximum margin* solution (though you don't need to show this).

3. Rather than use a fixed learning rate, Adam attempts to *adapt* the learning rate for each parameter using past gradient information. In particular, Adam generates a sequence of points $\left\{\theta_k^{\text{ad}}\right\}$ according to:

$$\theta_{k+1}^{\text{ad}} = \theta_k^{\text{ad}} - \alpha_k H_k^{-1} \nabla R_S\left[\theta_k^{\text{ad}}\right] + \beta_k H_k^{-1} H_{k-1}(\theta_k^{\text{ad}} - \theta_{k-1}^{\text{ad}}), \tag{4.3}$$

   where α_k and β_k are fixed sequences, and H_k is a positive definite, diagonal matrix

$$H_k = \text{diag}\left(\left\{\sum_{i=1}^{k} \eta_i g_i \circ g_i\right\}^{1/2}\right), \tag{4.4}$$

   where η_k is another fixed set of coefficients, $g_k = \nabla R_S[\theta_k^{\text{ad}}]$, and $\circ$ denotes an entry-wise product. (You do not need to show Adam can be written in this form). Assume the sequences $\alpha_k, \beta_k, \eta_k$ are chosen so that Adam converges to a minimizer of the objective (4.1). (You don't need to show how to choose these sequences).

Suppose there exists some scalar c such that $X\text{sign}(X^\top y) = cy$, and we initialize $\theta_0^{\text{ad}}, \theta_{-1}^{\text{ad}} = 0$. **Prove that Adam converges to the unique solution $\theta^{\text{ad}} \propto \textbf{sign}(X^\top y)$.**

Hint: Use induction to show every iterate satisfies $\theta_k^{\text{ad}} = \lambda_k sign(X^\top y)$ for some scalar λ_k.

4. Fix the labels $y \in \{-1, 1\}^n$, and let $X = [y; I_{n\times n}]$. Hence, only the first feature is discriminative, and the others are unrelated to the true label. **Compute the solutions found by running (a) gradient descent and (b) Adam on this problem instance.**

5. **Compare the relative weight the solutions found in part (d) place on the discriminative feature relative to the remaining features, i.e. compute $\frac{|\theta[1]|}{|\theta[i]|}$ for both gradient descent and Adam,** where $\theta[i]$ denotes the i-th coordinate of θ. Heuristically, which solution do you expect to generalize better to new data?

4.5 Problem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a β-smooth, but possibly *non-convex* function. Recall a function is β-smooth if the gradient map ∇f is β-Lipschitz. An ϵ-substationary point is any point x with $\|\nabla f(x)\|_2 \leq \epsilon$. In this problem, we show that gradient descent reaches an ϵ-substationary point in $O(1/\epsilon^2)$ steps. For consistency, the gradient descent update is

$$x_{t+1} = x_t - \eta \nabla f(x_t) \ ,$$

where η is some fixed step-size. For this problem, you may use without proof that if f is β-smooth, then for all $x, y \in \mathbb{R}^d$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|y - x\|_2^2 \ . \tag{4.5}$$

1. Show that for any $\eta > 0$,

$$f(x_{t+1}) \leq f(x_t) - \left(1 - \frac{\beta\eta}{2}\right) \eta \|\nabla f(x_t)\|_2^2.$$

2. Show, for a careful choice of η,

$$\|\nabla f(x_t)\|_2^2 \leq 2\beta(f(x_t) - f(x_{t+1})).$$

3. Suppose we run gradient descent for T steps starting from x_0, with η choosen as in part (b). Use the previous inequality to show

$$\sum_{t=0}^{T} \|\nabla f(x_t)\|_2^2 \leq 2\beta(f(x_0) - f(x^*)),$$

   where $x^* \in \text{argmin} f(x)$.

4. Prove

$$\min_{t=0,\ldots,T} \|\nabla f(x_t)\|_2 \leq \sqrt{\frac{2\beta}{T+1}(f(x_0) - f(x^*))}.$$

5. Suppose f is bounded, so $\max_x |f(x)| \leq R$. How many steps T are required before one of the iterates $x_0, x_1, \ldots, x_T$ is guaranteed to have $\|\nabla f(x_t)\|_2 \leq \epsilon$?

4.6 Problem

We say that a distribution $\mathcal{D}$ over examples $(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$ is linearly separable if there exists $w \in \mathbb{R}^d, b \in \mathbb{R}$ such that the classifier $f_{w,b}(x) = \text{sign}(\langle w, x \rangle + b)$ has zero true risk. More formally, a classifier has zero true risk if for $(x_i, y_i) \sim \mathcal{D}$ the event $f_{w,b}(x_i) = y_i$ occurs with probability 1.

1. **Prove or provide a counterexample**. If a distribution is linearly separable, then for all linear subspaces $V \subseteq \mathbb{R}^d$ of dimension at least one, the distribution is linearly separable after the data is projected onto V.

   We say that a distribution $\mathcal{D}$ is linearly separable after being projected to a linear subspace V if there exists $w \in \mathbb{R}^d, b \in \mathbb{R}$ such that the classifier $f_{w,b}(x) = \text{sign}(\langle w, Px \rangle + b)$ has zero risk where P is the orthogonal projection matrix onto an arbitrary subspace V.

2. **Prove or provide a counterexample**. If a distribution is linearly separable, then there exists an orthogonal projection P to a linear subspace V of dimension 1 such that the data is linearly separable after projection onto V.

3. **Prove or provide a counterexample**. If there exists an orthogonal projection P onto a subspace V such that $\mathcal{D}$ is linearly separable after the data is projected onto V, where $1 \leq dim(V) < d$, then there exists parameters $w^\star, b^\star$ such that $\mathcal{D}$ is linearly separable in $\mathbb{R}^d$.

4.7 Problem

In this problem, we want to find an optimal decision rule for giving loans to individuals. Suppose that the cost of granting a loan to an individual who defaults is α (i.e., $C_{10} = \alpha$) and the "cost" of granting a loan to an individual who repays is $-\beta$ since there is some revenue from interest (i.e. $C_{11} = -\beta$). There is no cost or profit associated with denying loans to individuals.

Let the variable Y represent repayment, with $Y = 1$ indicating that an individual repays, and $Y = 0$ indicating that they default. Let the variable X represent the observed data about individuals. (Note that this notation, particularly the meaning of Y, is slightly different from that in the WWS notes.)

1. **Show that the optimal decision rule for observations x has the form**

$$\mathbb{E}[Y|X = x] \geq t \,,$$

and find an expression for t. *Hint: you can pose the decision problem as a hypothesis testing problem with $H_i = \{Y = i\}$.*

2. Now we would like to assign each individual a score based on their features, $S = s(X)$. **What score function minimizes the squared error**:

$$\mathbb{E}_{X,Y}[(s(X) - Y)^2] \,?$$

Propose two distinct methods of estimating $s(x)$ from datapoints $\{(x_i, y_i)\}$?

5 Generalization

5.1 Problem

Suppose that we want to estimate the mean $\mu = \mathbb{E}X$ of a random variable X, given samples $X_1, \ldots, X_n$. In particular, we want to have a ε-accurate estimate $\hat{\mu}$ of the true value μ, that is a value $\hat{\mu}$ such that $|\hat{\mu} - \mu| \leq \varepsilon$ with high probability. Assume that the random variable has bounded variance, $(X) = \sigma^2$.

1. **Show** that a sample size of $n = \mathcal{O}(\sigma^2/\varepsilon^2)$ is sufficient to compute an ε-accurate estimate with probability at least $3/4$.

2. **Show** that a sample size of $n = \mathcal{O}(\sigma^2/\varepsilon^2 \log(1/\delta))$ is sufficient to compute an ϵ-accurate estimate with probability at least $1 - \delta$.

   Hint: consider the estimator devised by computing the median k means. $\mathrm{Y} = \mathrm{median}(\bar{X}_1, \ldots, \bar{X}_k)$ where each $\bar{X}_i$ is the mean of a (disjoint) set of n/k points. You may use without proof that if $Y_1, \ldots, Y_k$ are Bernoulli random variables, $P(|\frac{1}{k}\sum_{i=1}^{k} Y_i - \mathbb{E}Y| > t) \leq c_1 \exp\{-c_2 k t^2\}$ for some universal constants c_1, c_2.

5.2 Problem

Let S and S' be independent sets of n samples (whose elements are z_i and z'_i, respectively) from the same data distribution D, and let $S^{(i)}$ denote a hybrid sample set whose elements all come from S except for the ith sample, which comes from S'. A sample z_i consists of a data point $x_i \in \mathbb{R}^d$ and its label $y_i \in \{0, 1\}$, so we are doing binary classification. $A(S)$ denotes the algorithm A trained on the samples in S. Assume that the y_i's are drawn according to a Bernoulli distribution with parameter $\frac{1}{2}$, the distribution D is continuous (over x_i's) ,and the loss ℓ is the 0 - 1 loss.

Consider the following algorithm $A(S)$. $A(S)$ produces a classification function f that *memorizes* the training data S, such that $f(x_i) = y_i$ for all $x_i \in S$ and $f(v) = 0$ for all $v \notin S$.

1. **Prove** that the empirical risk $R_S(f) = 0$ but $\mathbb{E}[\Delta_{gen}(A(S))] = \frac{1}{2}$, where $\Delta_{gen}(A(S)) = R(f) - R_S(f)$ is the *generalization gap*.

   In other words, this classifier achieves perfect (zero) empirical risk, but its expected true risk is no better than random guessing.

2. Using this algorithm, **construct** a distribution D such that $\Delta_{sup}(A) = \mathbb{E}[\Delta_{gen}(A(S))]$, and **show** that your D satisfies this property. (All other components of the problem aside from the distribution of the x_is and y_is remain as originally defined.)
   Hint: Recall that $\Delta_{sup}(A) = \sup_{S,S'} \sup_{i \in [1,n]} |\ell(A(S), z'_i) - \ell(A(S^{(i)}, z'_i)|$, and in general we have that $\Delta_{sup}(A) \geq \mathbb{E}[\Delta_{gen}(A(S))]$.

5.3 Problem

In this problem, we review how algorithmic stability implies generalization and demonstrate how the stochastic gradient method is a stable algorithm. Seeing as how it is one of the most common algorithms to train machine learning models, it's nice to see that SGM enjoys nice statistical properties.

Given a set of n labeled examples $S = (z_1, \ldots, z_n)$ where each $z_i \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is an iid draw from some distribution $\mathcal{D}$, consider a decomposable objective function $f_S : \Omega \to \mathbb{R}$ which we are trying to minimize:

$$f_S(w) = \frac{1}{n} \sum_{i=1}^{n} f\left(w; z_i\right)$$

A randomized algorithm A, is *uniformly stable* if for all datasets S, S' that differ in at most one example.

$$\sup_z \mathbb{E}_A \left[f(A(S); z) - f\left(A\left(S'\right); z\right) \right] \leq \epsilon$$

Here, $A : S \to \Omega$ is a randomized function that takes in datasets S and returns a solution w. Note that the expectation is taken over the inner randomness of A and that the randomness in A is the same when run on S and S'.

[(a)]**Prove** that if an algorithm A is ϵ uniformly stable, its output has ϵ generalization error:

$$\left| \mathbb{E}_{S,A} \big[f_S[A(S)] - \mathbb{E}_{z \sim \mathcal{D}} f(A(S); z) \big] \right| \leq \epsilon$$

Now, assume that the function f is differentiable, convex, L Lipschitz, and β-smooth. Given a dataset S, at each time step t, the stochastic gradient method chooses an example $z_i \in S$ uniformly at random, $i \sim \text{Uniform}([n])$, and updates the solution according to the following:

$$w_{t+1} \leftarrow w_t - \alpha_t \nabla f(w_t; z_i)$$

[(a)]**Prove** that for an arbitrary $z \in Z$, $\left|f\left(w_T; z\right) - f\left(w'_T; z\right)\right| \leq L \cdot \delta_T$ where $\delta_t = ||w_T - w_{T'}||$. Let $w_T = A(S)$ and $w'_T = A(S')$, be the outputs of running SGM for T steps on S and S', where S and S' are datasets that differ in at most one data point z_i. Likewise, let w_t and w'_t be the intermediate solutions found by SGM when run on S and S', respectively. Assume that $\alpha_t \leq 2/\beta$. **Prove** that if at time step t, SGM samples the same example z_i when run on S and S' then $\delta_{t+1} \leq \delta_t$. That is:

$$||w'_{t+1} - w_{t+1}|| \leq ||w'_t - w_t||$$

Why is this statement not true if the examples z_i chosen at time t are different? Hint: Try to write out the difference between w'_{t+1} and w_{t+1} in terms on the stochastic gradient updates.

Use the fact that if a function g is β smooth and convex then its gradients satisfy:

$$\langle \nabla g(v) - \nabla g(w), v - w \rangle \geq \frac{1}{\beta} \|\nabla g(v) - \nabla g(w)\|^2$$

Prove that $\mathbb{E}[\delta_{t+1}] \leq \mathbb{E}[\delta_t] + \frac{2L\alpha_t}{n}$. Unroll the recursion to conclude that if $\delta_0 = 0$ then

$$\mathbb{E}[\delta_T] \leq \frac{2L}{n} \sum_{t=0}^{T-1} \alpha_t$$

Hint: Analyze the problem by separately considering the cases where the samples chosen by SGM are the same at time t and when they are not. Using the previous parts, **show** that the stochastic gradient method is uniformly stable with parameter:

$$\epsilon_{\text{stab}} \leq \frac{2L^2}{n} \sum_{t=0}^{T-1} \alpha_t$$

5.4 Problem

1. Show that for a nonnegative random variable X:

$$EX = \int_0^\infty P(X \geq t)dt$$

2. Prove that if a random variable has density uniformly bounded by 1, that is $P(X = x) \leq 1$ for all x, then

$$\mathbb{E}e^{-tX_i} \leq 1/t \text{ for all } t > 0$$

3. Using the previous part, to show that if $X_1, \ldots, X_n$ are iid from this distribution then:

$$P\bigg(\sum_{i=1}^n X_i \leq \varepsilon n\bigg) \leq (e\varepsilon)^n$$

Hint: write $\sum_{i=1}^n X_i \leq \varepsilon n$ as $-1/\varepsilon \sum_{i=1}^n X_i \geq -n$ and apply a Chernoff bound.

1. For X nonnegative $x = \int_0^x dt = \int_0^\infty 1\{t \leq x\}dt$, taking expectations on both sides, we get

$$\mathbb{E}X = \int_0^\infty P(X \geq t)dt$$

2.

$$\begin{aligned}
\mathbb{E}e^{-tX} &= \int_0^\infty e^{-tx} P(X = x)dx \\
&\leq \int_0^\infty e^{-tx}dx \\
&= 1/t
\end{aligned}$$

3. First, we write $\sum_{i=1}^n X_i \leq \varepsilon n$ as $-1/\varepsilon \sum_{i=1}^n X_i \geq -n$ and then we proceed by applying Markov's inequality.

$$\begin{aligned}
P(\frac{-1}{\varepsilon}\sum_{i=1}^n X_i \geq -n) &\leq P(e^{-1/\varepsilon \sum_{i=1}^n X_i} \geq e^{-n}) \\
&\leq \frac{\mathbb{E}[e^{-1/\varepsilon \sum_{i=1}^n X_i}]}{e^{-n}} \\
&= \prod_{i=1}^n \mathbb{E}[e^{-1/\varepsilon X_i}]e^n \\
&\leq (e\varepsilon)^n
\end{aligned}$$

5.5 Problem

Here, we analyze the relationship between the true risk of a classifier and its empirical risk as a function of the number of samples. In particular, we consider this for the simple setting of binary classification.

Assume that you have access to a data set S of m samples $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ where $x \in \mathbb{R}^d, y \in \{0, 1\}$ and each pair (x_i, y_i) is sampled i.i.d from the true distribution $\mathcal{D}$. Furthermore, assume that we are going to perform classification by choosing a hypothesis function (classifiers) from a *finite* set $\mathcal{H} = \{h |\ h : \mathbb{R}^d \to \{0, 1\}\}$, $|\mathcal{H}| < \infty$. Let $L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell_{01}(h, x, y)$ denote the empirical risk of the hypothesis function h under the 0-1 loss.[5] Let $L_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\ell_{01}(h, x, y)$ be the true risk of the classifier.

1. **Prove that if we observe $m \geq \frac{\log(c|\mathcal{H}|/\delta)}{2\varepsilon^2}$ samples from the distribution $\mathcal{D}$, then with probability $1 - \delta$, $|L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon$, $\forall h \in \mathcal{H}$, where c is some universal constant.**

   Hint: Think of the empirical risk as a random variable try to show that it cannot be too far away from its mean.

2. Using the previous statement, **prove that if $h^\star \in \operatorname{argmin}_{h\in\mathcal{H}} L_S(h)$, $L_S(h^\star) \leq \alpha$, and $m \geq \frac{\log(c|\mathcal{H}|/\delta)}{2\varepsilon^2}$ then with probability $1 - \delta$, $L_{\mathcal{D}}(h^\star) \leq \alpha + \varepsilon$.**

3. Let $\mathcal{H}$ consists of weight vectors w where $w \in \mathbb{R}^d$. This is an infinite hypothesis class. Let $\tilde{\mathcal{H}}$ denote the discretization of $\tilde{\mathcal{H}}$ to 32 bit floats. (e.g each entry of w is now represented as a 32 bit float)

   Prove that if $m \geq \frac{1}{2\varepsilon^2}\log(c/\delta) + \frac{16\log(2)d}{\epsilon^2}$ then $|L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon$, $\forall h \in \tilde{\mathcal{H}}$ with probability at least $1 - \delta$.

[5] Remember $\ell_{01}(h, x, y) = 0$ if $h(x) = y$ and is 1 otherwise

5.6 Problem

In this problem, it will be useful to recall the following definitions. *Empirical risk* is defined as:

$$R_S = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{A}(S), z_i)\,.$$

Average stability is defined as:

$$\Delta(A) = \mathbb{E}_{S,S'}\left[\frac{1}{n}\sum_{i=1}^{n}(\ell(A(S), z_i') - \ell(A(S^{(i)}), z_i'))\right]\,.$$

where S and S' are two sample sets of size n, and each point is sampled identically and independently. $\mathcal{A}(S)$ denotes the function produced by algorithm $\mathcal{A}$ on the training set S. For a data point $z_i = (x_i, y_i)$, $\ell(\mathcal{A}(S), z_i)$ denotes the loss incurred by $f = \mathcal{A}(S)$ on z_i. $S^{(i)}$ denotes a hybrid dataset consisting of $(z_1, z_2, ..., z_{i-1}, z_i', z_{i+1}, ..., z_n)$ where z_i' is the ith element of S'. Finally, *uniform stability* is defined as:

$$\Delta_{sup}(\mathcal{A}) = \sup_{S,S'}\sup_{z}|\ell(\mathcal{A}(S), z) - \ell(\mathcal{A}(S^{(i)}), z)|\,.$$

Here the supremum is over all data sets S, S', and data points z, and these need not be sampled i.i.d.

1. Let (x, y) pairs be sampled such that $y \in \{-1, 1\}$ and $\mathbf{Pr}[y = 1] = p_1$ and $\mathbf{Pr}[y = -1] = p_0$. Let $\mathcal{A}(S)$ denote the algorithm that always returns the value 1. That is, it returns a predictor f such that $f(x) = 1$ for all x. Using the logistic loss:

   $$\text{logistic}(f, z) = \log(1 + \exp(-yf(x))),$$

   Prove that $\mathcal{A}(S)$ is uniformly stable (i.e. has *uniform stability* equal to 0). **Prove** that $\mathbb{E}[R_S] \geq 1/3$ if $p_1 \leq 0.95$.

2. Let (x, y) pairs be sampled such that $y \in \{-1, 1\}$ is a random coin flip *sampled independently* from x. That is, y are random labels. Let $\mathcal{A}(S)$ return the function f that memorizes the training data, so that $f(x_i) = y_i$ on the training set and $f(x) = 0$ outside the training set. Using the hinge loss,

   $$\text{hinge}(f, z) = \max(1 - yf(x), 0),$$

   prove that $\mathbb{E}[R_S] = 0$. Also **prove** that the *average stability* of $\mathcal{A}(S)$ equals 1.

5.7 Problem

Thus far in the course, we have considered *classification* problems where given input $x \in \mathbb{R}^d$, we wish to predict output $y \in \{0, 1\}$. In this problem, we explore *regression*, where the output y is now a real-valued response $y \in \mathbb{R}$.

Suppose we have a fixed set of n inputs $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ and a true underlying parameter $\theta^* \in \mathbb{R}^d$ that governs the outputs. For each $i = 1, 2, ..., n$, we observe:

$$y_i = x_i^\top \theta^* + \epsilon_i \quad i = 1, \ldots, n,$$

where we assume the noise terms ϵ_i are i.i.d. with mean $\mathbb{E}[\epsilon_i] = 0$ and variance $\mathrm{Var}(\epsilon_i) = \sigma^2$. Conceptually, we wish to recover the parameter θ^* from noisy measurements y_i. This setting is sometimes called *fixed-design* linear regression because the features $x_1, \ldots, x_n$ are fixed.

To set up notation, let $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times d}$ and $Y = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$ be the data we observe at training time. Let $\epsilon = [\epsilon_1, \ldots, \epsilon_n]^\top \in \mathbb{R}^n$ be the noise, and let $\Sigma = \frac{1}{n} X^\top X \in \mathbb{R}^{d \times d}$ be the second moment matrix. Assume $\Sigma \succ 0$.

Our goal is to find an estimate $\hat{\theta}$ of the true parameter θ^* that predicts output Y well, as measured by the squared loss. Mathematically, our goal is to minimize the expected risk:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[(x_i^\top \theta - y_i)^2\right] = \frac{1}{n} \mathbb{E} \left\| X\theta - Y \right\|_2^2 .$$

Note the above expectation is only over Y since X is fixed.

To solve this problem, the *least-squares* estimator is a very common choice:

$$\hat{\theta}_{\text{ls}} = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \|X\theta - Y\|_2^2$$

In the subsequent parts, we analyze the least-squares estimator.

1. **Give a closed-form expression for $\hat{\theta}_{\text{ls}}$.**

2. Let's understand the expected risk $L(\theta)$ in more detail. **Show that, for any $\theta \in \mathbb{R}^d$, you can write**

   $$L(\theta) = \|\theta - \theta^*\|_\Sigma^2 + \sigma^2,$$

   where $\|\cdot\|_\Sigma$ is a Mahalanobis norm with $\|x\|_\Sigma^2 = x^\top \Sigma x$. **Then compute $L(\theta^*)$, and use it to compute the *excess risk* $L(\theta) - L(\theta^*)$.**

3. For the least-squares estimator, **show the excess risk is**

   $$L(\hat{\theta}_{\text{ls}}) - L(\theta^*) = \frac{1}{n} \mathbf{Tr}(\Pi \epsilon \epsilon^\top),$$

   where Π is a carefully chosen projection matrix.

4. Taking expectation over the training data, **argue that**

$$\mathbb{E}[L(\hat{\theta}_{\text{ls}}) - L(\theta^*)] = \frac{d\sigma^2}{n}.$$

Give intuition about this result. How does the expected excess risk scale as a function of the dimension d, the noise variance σ^2, and the number of samples n? How many samples n are required to obtain expected excess risk at most δ?

6 Deep learning

7 Datasets

7.1 Problem

Traditionally in empirical risk estimation, a holdout set is reserved for evaluation. However, if the holdout set is reused by analysts to adaptively evaluate their models, the statistical guarantees of the risk estimates using this holdout set degrade. In this problem, we analyze an algorithm that can mitigate this effect. Let $X \in \mathbb{R}^n$ be the data domain and $Y = \{0, 1\}$ the class labels. We consider a classifier $f : X \to Y$ and its performance on a loss function $l : Y \times Y \to [0, 1]$. Given a sample $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d. from an unknown distribution $\mathcal{D}$ over $X \times Y$, we recall the definition of the empirical loss

$$R_S(f) := \frac{1}{n} \sum_{i}^{n} l(f(x_i), y_i),$$

and the true loss

$$R_{\mathcal{D}}(f) := \mathbb{E}[l(f(x), y)].$$

1. We first consider a sequence of classifiers $f_1, \ldots, f_k$. The fundamental estimation problem is to compute the risk estimates $R_1, \ldots, R_k$ such that

   $$\mathbb{P}[\exists t \in [k] : |R_t - R_{\mathcal{D}}(f_t)| > \varepsilon] \leq \delta.$$

   Suppose our functions f_i are fixed independently of the sample S. Let us take $R_i = R_S(f_i)$. **Show that**

   $$\mathbb{P}[\exists t \in [k] : |R_t - R_{\mathcal{D}}(f_t)| > \epsilon] \leq 2ke^{-2\varepsilon^2 n}.$$

2. In the adaptive setting, however, f_i can be chosen as a function of the previous estimates and classifiers. Because of the dependence of classifiers on past classifiers and risk estimates, our tain bound from part 1 using empirical risk as our risk estimates no longer holds. Let us then relax our requirement that each risk estimate be close to the true risk, and instead minimize the leaderboard error:

   $$L(R_1, \ldots, R_k) := \max_{1 \leq i \leq k} | \min_{1 \leq j \leq i} R_{\mathcal{D}}(f_i) - R_t|.$$

   Give an intuitive explanation for what minimizing this new error would imply for the risk estimates.

3. To formalize the adaptivity of the classifiers, we might say that there now exists a mapping $\mathcal{A}$ such that for all $t \in [k]$,

   $$f_t = \mathcal{A}(f_1, R_1, \ldots, f_{t-1}, R_{t-1}).$$

   We want to reason about the possible $\{f_1, R_1, \ldots, f_k, R_k\}$ that can result from this scheme; to do so, we define a graph of random variables. **Recursively define a tree $\mathcal{T}$ of depth t**

of classifiers and risk estimates generated by this scheme. *(Hint: The nodes can be written as sets of* $(f_1, R_1, \ldots, f_k, R_k)$*.)*

4. Now consider the following algorithm to minimize $L(R_1, \ldots, R_k)$: For sample S and step size $\eta > 0$, we initialize $R_0 = \infty$. At each iteration i, we receive the next classifier f_i. If $R_S(f_i) < R_{i-1} - \eta$, we let $R_i = [R_S(f_i)]_\eta$, where $[x]_\eta$ indicates x rounded to the nearest integer multiple of η. Otherwise we keep our current estimate. **Prove that with this algorithm, the size of the tree** $\mathcal{T}$ **has at most** 2^B **nodes, where** $B = (1/\eta + 2)\log(4t/\eta)$. *(Hint: show how to uniquely encode each node in the tree using* B *bits of information.)*

5. Now show that for all $t \leq k$ and $\varepsilon > 0$,

$$\mathbb{P}[|\min_{1 \leq i \leq t} R_{\mathcal{D}}(f_i) - R_t| > \varepsilon + \eta] \leq \exp(-2\varepsilon^2 n + B + 1).$$

7.2 Problem

The *holdout method* is a common technique in machine learning to perform model selection. The method holds out a set S of n examples (x_i, y_i) sampled i.i.d. from a distribution $\mathcal{D}$ and uses this set to evaluate the performance of a proposed model. Concretely, for a classifier f, one uses the empirical risk $R_S[f]$ on the holdout set S as a proxy for the true model risk $R[f]$. Throughout, assume we have binary labels $y_i \in \{0, 1\}$, and we measure performance using the $0-1$ loss.

(a) Fix a classifier f. **Show if $n \geq \frac{\log(2/\delta)}{2\varepsilon^2}$, then with probability $1-\delta$, $|R_S[f] - R[f]| \leq \varepsilon$.**

Hint: Hoeffding's inequality.

(b) The popular ImageNet ILSVRC and Cifar10 datasets have, respectively, $n = 50,000$ and $n = 10,000$ images in the validation set. If we set $\delta = 0.05$ (corresponding to a 95% confidence interval), **evaluate the bound from part (a) for both ImageNet and Cifar10**.

Most machine learning workflows, however, do not evaluate a single classifier on the holdout set and then stop. Instead, after looking at the validation loss, you try to improve it by, for instance, changing the feature set, adding more layers, tweaking the optimization algorithm, etc. and then reevaluate the new model on the *same* validation set. In the remainder of this problem, we explore the potential pitfalls of *adaptively* interacting with the holdout set.

Henceforth, suppose our features are binary $x \in \{0, 1\}^d$, and suppose examples (x, y) are drawn from the uniform distribution on $\{0, 1\}^d \times \{0, 1\}$. Consider the following procedure:

(c) Compute R_S for single-feature classifiers $h_i(x) = x_i$ for $i = 1, \ldots, d$.

(d) Say a feature i is *informative* if $R_S[h_i] \leq \frac{1}{2} - \frac{1}{\sqrt{n}}$. Let I denote the set of informative classifiers h_i.

(e) Construct a classifier $\tilde{f}$ consisting of a majority vote of the informative classifiers

$$\tilde{f}(x) = \begin{cases} 1 & \sum_{i \in I} h_i(x) \geq \frac{|I|}{2} \\ 0 & \text{otherwise.} \end{cases}$$

This is a fairly natural procedure: we first attempt to predict y using a single feature, and then ensemble the classifiers that seem to give predictive power. However, we will show the estimated risk $R_S[\tilde{f}]$ can be arbitrarily far from the true risk $R[\tilde{f}]$ when d is large.

(f) First, **prove** $R[\tilde{f}] = \frac{1}{2}$. This means $\tilde{f}$ is no better than random guessing on new examples.

Hint: You can in fact prove that $R[f] = \frac{1}{2}$ for any arbitrary classifier f.

(g) (**Bonus**) Show the expected empirical risk of $\tilde{f}$ shrinks exponentially fast with the number of informative features $|I|$. **Prove** $\mathbb{E}_S[R_S[\tilde{f}] \mid |I| = k] \leq \exp\left(\frac{-2k}{n}\right)$.

Even if you don't solve the bonus exercise, we'll use the conclusion in the subsequent parts.

Hint: Use the fact that the coordinates are independent, so $1[x_i = y]$ *and* $1[x_j = y]$ *are independent for* $i \neq j$*, along with the observation* $\tilde{f}(x) \neq y$ *iff* $\sum_{i \in I} 1[x_i = y] < \frac{|I|}{2}$.

The remainder of the problem is devoted to showing $|I|$ is large with high probability.

(h) **Prove** each coordinate i is informative with constant probability, i.e. show $\mathbb{P}\{i \in I\} \geq c$ for some constant $c > 0$

Hint: First, argue $R_s[h_i]$ *follows a rescaled binomial distribution, and* $i \in I$ *if the binomial deviates from it's mean by 2 standard deviations. Then, show this event occurs with constant probability by approximating the binomial to a normal distribution. You don't need to be fully rigorous with the approximation.*

(i) **Prove** $\mathbb{E}[|I|] \geq cd$.

(j) **Prove** with probability $1 - \delta$, the number of informative features $|I| \geq \frac{cd}{2}$ for $d \geq \frac{2\log(1/\delta)}{c^2}$.

Hint: Use each coordinate is independent, so $1[x_i \neq y]$ *and* $1[x_j \neq y]$ *are independent for* $i \neq j$*, and then apply Hoeffding's inequality.*

(k) **Put parts (c)-(g) together to prove the following**: there exists a constant α, such that if $d \geq \alpha n$, then with probability at least $\frac{3}{4}$,

$$\left| R_S[\tilde{f}] - R[\tilde{f}] \right| \geq 0.49.$$

8 Causality

8.1 Problem

1. Consider the following example of studying the effect of drug X on the recovery rate. The recovery rates of 600 patients were recorded, and the patients were given access to the drug. A total of 300 patients chose to take the drug and 300 patients did not.

Table 1: Study into a new drug, with gender being taken into account

	Drug	No drug
Female	69/75 (92%)	200/231 (87%)
Male	164/225 (73%)	27/69 (68%)
Overall	233/300 (78%)	247/300 (82%)

   (a) Let R be a random variable that denotes the recovery (e.g. $R = 1$ is the event of recovery). Let X be a random variable that denotes taking the drug (e.g. $X = 1$ is the event of taking the drug). Let G be a random variable that denotes the gender, which takes values in $\{male, female\}$. **Use this notation to write down the observation about overall recovery rates**, and the observation about **recovery rates by gender** as inequalities of probabilities respectively.

   (b) **Write** $P(Y = 1|X = 1)$ in terms of $P(Y = 1|X = 1, G = i)$ for $i \in \{male, female\}$. Using the above equations, **explain why** the effect of the drug for the overall population seems at odds with the pattern in individual gender.

   (c) Suppose you knew an additional fact: Estrogen has a negative effect on recovery, so women are less likely to recover than men, regardless of the drug. In addition, as we can see from the data, women are significantly more likely to take the drug than men are. Using the given information, **draw the causal graph** between the variables X, Y, G. **Using the causal graph, explain why** the effect of the drug for the overall population seems at odds with the pattern in individual gender.

   (d) **Using the causal graph, explain why** a policy maker should recommend the drug or not.

2. Consider the same table, where instead of recording participants' gender, patients' post-treatment blood pressure were recorded. In this case, we know that the drug affects recovery by lowering the blood pressure of those who take it. But it also has a toxic effect.

   (a) Using the given information, **draw the causal graph** between the variables X, Y, Z, where $Z \in \{high, low\}$ denotes the post-treatment blood pressure. **Using the causal graph, explain why** the effect of the drug for the overall population seems at odds with the pattern in individual subgroups with different blood pressure levels.

Table 2: Study into a new drug, with post-treatment blood pressure being taken into account

	Drug	No drug
Low BP	69/75 (92%)	200/231 (87%)
High BP	164/225 (73%)	27/69 (68%)
Overall	233/300 (78%)	247/300 (82%)

(b) **Using the causal graph, explain why** a policy maker should recommend the drug or not.

3. Suppose apart from the record of the patients' post-treatment blood pressure, the pre-treatment blood pressure was also recorded, and the usage of the drug takes the pre-treatment blood pressure into account.

Table 3: Study into a new drug, with post-treatment blood pressure being taken into account

	Drug	No drug
High pre-BP, Low post-BP	60/65	195/200
Low pre-BP, Low post-BP	9/10	5/31
High pre-BP, High post-BP	140/150	7/9
Low pre-BP, High post-BP	24/75	40/60
Overall	233/300 (78%)	247/300 (82%)

(a) Using the given information, **draw the causal graph** between the variables X, Y, Z_1, Z_2, where $Z_1 \in \{high, low\}$, $Z_2 \in \{high, low\}$ denote the pre-treatment and post-treatment blood pressures respectively.

(b) **Explain** in this case, why should the drug be recommended or not.

8.2 Problem

In an attempt to estimate the effectiveness of a new drug, a randomized experiment is conducted. In all, 50% of the patients are assigned to receive the new drug and 50% to receive a placebo. A day before the actual experiment, a nurse hands out lollipops to some patients who show signs of depression, mostly among those who have been assigned to treatment the next day (i.e., the nurse's round happened to take her through the treatment-bound ward). Strangely, the experimental data revealed a Simpson's reversal: Although the drug proved beneficial to the population as a whole, drug takers were less likely to recover than nontakers, among both lollipop receivers and lollipop nonreceivers. Assuming that lollipop sucking in itself has no effect whatsoever on recovery, answer the following questions:

1. **Draw** a graph that captures the story. Explain the edges in you graph.
   Hint: Use the fact that receiving a lollipop indicates a greater likelihood of being assigned to drug treatment, as well as depression, which is a symptom of risk factors that lower the likelihood of recovery.

2. **Determine** which variables must be adjusted for in order to determine the effect of the drug on recovery.

3. **What are the adjustment formulas** for the effect of the drug on recovery?

4. **Repeat the previous parts** assuming that the nurse gave lollipops a day after the study, still preferring patients who received treatment over those who received placebo. You may assume that depression is a long-term condition.

8.3 Problem

In this problem we analyze the stability of causal inference with respect to small errors in the problem parameters. In general, we can think of causal inference from observational data as an identification map, ID that given a graph G and a joint distribution P over a set $V = \{X_1, \ldots, X_n\}$ of n variables, returns the distribution of a target variable Y after we intervene on a set of variables X.

$$\text{ID}(G, V, X, Y) := P(Y|do(X = x))$$

In practice, we never have access to the true distribution P over the variables in our system. Instead, we have access to some noise version of $\tilde{P}$ that we estimate by samples. Assuming that out estimates of the distribution have some error in them, how far can our estimated causal effects be? We show that if we can identify causal effects via the backdoor criterion, then the blowup in error from observational to interventional distribution is small.

Define two distributions $P, \tilde{P}$ to be ϵ-*close* if for all outcomes $\omega \in \Omega$:

$$-\epsilon \leq \log \frac{P(\omega)}{\tilde{P}(\omega)} \leq \epsilon$$

1. **Prove that if two joint distributions P and $\tilde{P}$ are ϵ-close then all conditional distributions $P(A = a \mid B = b)$, $\tilde{P}(A = a \mid B = b)$ are 2ϵ close for all disjoint subsets $A, B \subset V$.**

2. Assume there exists a subset $Z \subset V$, $Z \cap Y = Z \cap X = \emptyset$ such that the interventional distribution $P(Y \mid do(X = x))$ can be identified via the backdoor criterion by adjusting for Z. **Prove that if two joint distributions P and $\tilde{P}$ are ϵ-close then all interventional distributions $P(Y = y \mid do(X = x))$, $\tilde{P}(Y = y \mid do(X = x))$ are 3ϵ close.**

8.4 Problem

In this problem, we consider hypothetical lending scenarios. The label $Y \in \{-1, 1\}$ indicates whether or not the applicant repays their loan, and the bank is interested in predicting this quantity. At the same time, the bank must ensure that the predictions do not encode unfair discrimination with respect to national origin, as required by the Equal Credit Opportunity Act. In a vast simplification, let natural origin be denoted by $A \in \{0, 1, \ldots, K\}$.

We consider the following two scenarios:

1. **Scenario 1**: The institution builds a model using two features: the languages spoken at the applicant's home, L, and a measure of applicant's financial history H. We assume that these random variables are generated by the following structural equation model:

   - $A := U_1$, where $U_1 \sim \text{Unif}\{0, 1, \ldots, K\}$.
   - $L := A$
   - $Y := 2U_2 - 1$, where $U_2 \sim \mathcal{B}(\sigma(A))$ is a Bernoulli random variable and the sigmoid function is defined as $\sigma(x) = \frac{1}{1+e^{-x}}$.
   - $H := Y + U_3$ with $U_3 \sim \mathcal{N}(0, 2)$ is a normal random variable.

2. **Scenario 2**: The institution builds a model using a single feature: the applicant's annual income I. Assume I depends on the sensitive attribute, and the label Y in turn depends on I. Formally, suppose:

   - $A := U_4$, where $U_4 \sim \text{Unif}\{0, 1, \ldots, K\}$.
   - $I := A + U_5\mathcal{N}(1, 2) + (1 - U_5)\mathcal{N}(-1, 2)$, where $U_5 \sim \mathcal{B}(\sigma(A))$.
   - $Y := 2U_6 - 1$, where $U_6 \sim \mathcal{B}(\sigma(I))$.

In what follows, we will consider a general notion of *Bayes-optimal predictors*. This general notion includes is any function of the features that, when thresholded, results in optimal decisions for any cost function. Another way to state this is that Bayes-optimal predictors are any monotonic transformation of *Bayes-optimal scores*, $r(x) = \mathbb{E}[Y|X = x]$.

1. In this problem, we consider a variety of possible predictors. For Scenario 1, define

   $$R_1^\star = L + H, \quad \tilde{R}_1 = H\,.$$

   For Scenario 2, define

   $$R_2^\star = I, \quad \tilde{R}_2 = I - A\,.$$

   Draw the causal graphs for $(A, L, H, Y, R_1^\star, \tilde{R}_1)$ in Scenario 1 and $(A, I, Y, R_2^\star, \tilde{R}_2)$ in Scenario 2.

2. **Show that** $R_1^\star$ and $R_2^\star$ are Bayes-optimal predictors for the respective scenarios.

   Hint: In Scenario 1, the joint distribution can be factorized in the following manner: $\mathbb{P}(Y = y, A = a, H = h) = \mathbb{P}(A = a)\mathbb{P}(Y = y \mid A = a)\mathbb{P}(H = h \mid Y = y)$

3. **Prove that** the joint distributions of optimal scores, protected attributes, and outcomes are equal in both scenarios, i.e. prove $(R_1^*, A, Y) \stackrel{d}{=} (R_2^*, A, Y)$.

   Hint: the joint distributions can be factorized as $\mathbb{P}(R_i^*, A, Y) = \mathbb{P}(A)\mathbb{P}(R_i^\star \mid A)\mathbb{P}(Y \mid R_i^\star, A)$*, and equality can be verified for each term individually.*

4. **Prove** that alternate predictors $\tilde{R}_1$ and $\tilde{R}_2$ satisfy *separation* in each scenario, respectively. *Hint: You may want to use the result of part 3.*

5. In 1-2 sentences, **explain why this example poses difficulties** for using observational fairness criteria to audit decisions.

8.5 Problem

One of the main dilemmas in causal inference is that causal effects are in general not identifiable from observational data. Surprisingly, (or perhaps unsurprisingly) this problem persists even if we assume that the underlying causal relationships between different variables are linear!

Linear Gaussian models are one of the most often used class of causal models, especially in the social sciences like Economics and Psychology. In a Linear Gaussian model, the set of variables $V = \{X_1, \ldots, X_n\}$ are jointly Gaussian, that is $V \sim \mathcal{N}(\mu, \Sigma)$, and all the functional relationships between variables are linear. In particular, if we denote by $Pa(X_i)$ the set of parents of a variable X_i in a linear, Gaussian causal model G, the functional form for all variables $X_i \in V$ is the following, where α_j are coefficients in $\mathbb{R}$:

$$X_i = \sum_{Z_j \in Pa(X_i)} \alpha_j Z_j + U_i, \quad U_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

Here, U_z is an exogenous noise variable. Let $I(G)$ denote the set of independence statements implied by a causal model G. We say that two causal models G and G' are *observationally equivalent* if $I(G) = I(G')$. Since independence relationships for Gaussians are captured by the covariance structure, in the case of linear Gaussian models, we say that two causal models G and G' are observationally equivalent if they are *covariance equivalent*, that is the respective covariance matrices Σ and Σ' are equal.

Let G_1 be the following linear Gaussian model, where U_X, U_Y, U_Z are i.i.d $\mathcal{N}(0, 1)$

$$\begin{aligned} X &= U_x \\ Y &= X + U_y \\ Z &= Y + U_z \end{aligned}$$

Similarly, let G_2 be a model over the same set of variables with different functional relationships.

$$\begin{aligned} X &= \frac{1}{2}Y + U_x, \quad U_x \sim \mathcal{N}(0, \frac{1}{2}) \\ Y &= U_y, \quad U_y \sim \mathcal{N}(0, 2) \\ Z &= Y + U_z \quad U_z \sim \mathcal{N}(0, 1) \end{aligned}$$

1. **Write down the causal graphs for G_1, G_2.** This tool makes it easy to produce graphs for latex.

2. Suppose we are interested in the causal effect of X on Y, in particular define the treatment effect $\text{TE} := \mathbb{E}[Y \mid do(X = 1)] - \mathbb{E}[Y \mid do(X = 0)]$. **What is the treatment effect of X on Y in G_1? How about in G_2? Do they have the same sign?**

3. **Using the graphs from part 1, prove that $I(G_1) = I(G_2)$.**

4. **Prove G_1 and G_2 are covariance equivalent.** Observe that this implies that G_1 and G_2 are observationally equivalent.

8.6 Problem

At the beginning of the year, a boarding school offers its students a choice between two meal plans for the year: Plan A and Plan B. The students' weights are recorded at the beginning and the end of the year. To determine how each plan affects students' weight gain, the school hired two statisticians who, oddly, reached different conclusions. The first statistician calculated the difference between each student's weight in June (W_F) and in September (W_I) and found that the average weight gain in each plan was zero.

The second statistician divided the students into several subgroups, one for each initial weight, W_I. They found that for each initial weight, the final weight for Plan B is higher than the final weight for Plan A.

So, the first statistician concluded that there was no effect of diet on weight gain and the second concluded there was.

Figure 1: Scatter plot with students' initial weights on the x-axis and final weights on the y-axis. The vertical line indicates students whose initial weights are the same, and whose final weights are higher (on average) for Plan B compared with Plan A.

Figure 1 illustrates data sets that can cause the two statisticians to reach conflicting conclusions. Statistician 1 examined the weight gain $W_F - W_I$, which, for each student, is represented by the shortest distance to the 45° line. Indeed, the average gain for each diet plan is zero: the two groups are each situated symmetrically relative to the zero-gain line, $W_F = W_I$. Statistician 2, on the other hand, compared the final weights of Plan A students to those of Plan B students who entered school with the same initial weight W_0 and, as the vertical line in the figure indicates, Plan B students are situated above Plan A students along this vertical line. The same will be the case for any other vertical line, regardless of W_0.

1. **Draw** a causal graph representing the situation.
2. **Determine** which statistician is correct. **Explain** why.

3. **How is this example related to Simpson's paradox?**

8.7 Problem

Consider the following (very) over-simplified model of a student. Let A represent the student's interest in the material (0 for lack of interest and 1 for interest), let B represent the student's understanding of the material (0 for confusion and 1 for understanding), and let C represent the student's attendance at office hours (0 for absence and 1 for attendance).

We might model these variables using the following generative model, in which Z_i are independent Bernoulli random variables (taking value 0 or 1 with equal probability).

- A := Z_1
- B := $Z_2 Z_3$
- C := if B = 0 then Z_4, else if A = 1 then $Z_4 Z_5$, else 0

This model might be helpful if, say, we'd like to study the effect of a student's interest in the material (A) on their understanding of the material (B), and it's easiest to survey students who attend office hours (C).

1. (a) Imagine we could survey the entire student body to answer our question. **Compute** $\mathbb{P}[B]$ and $\mathbb{P}[B|A]$. What can you conclude about the relationship between A and B in this model?

   (b) Imagine instead that we can only survey students who attend office hours. **Compute** $\mathbb{P}[B|C]$ and $\mathbb{P}[B|A, C]$. Is this what you expected?

2. (a) **Draw** a graph representing the causal relationships between A, B, and C.

   (b) Does your graph contain a **mediator**? If so, state which variable mediates between which other variables, and why.

   (c) Does your graph contain a **collider**? If so, state which variable is a collider and why.

   (d) In light of your causal graph, interpret the probabilities you computed in part 1. Can your graph explain why you observed what you did?

The phenomenon you observed is referred to as *Berkson's paradox*.

8.8 Problem

Although structural causal models and potential outcomes are sometimes cast as at odds, there is no tension between the two formalisms, and it is useful to have familiarity with both approaches. In this problem, we explore the connection between both frameworks and show how common estimators can be derived in both frameworks.

Suppose we have a structural causal model for variables (Z, T, Y), where Z is a set of background variables, $T \in \{0, 1\}$ is a binary treatment, and $Y \in \mathbb{R}$ is the outcome of interest. We wish to estimate the *average treatment effect* (ATE) of treatment T on outcome Y.

Throughout this problem, suppose $X \subset Z$ is a set of variables that satisfies the backdoor criterion relative to the pair (T, Y). Let $Y_0(U)$ and $Y_1(U)$ denote the potential outcomes defined in terms of the structural causal model, where U denotes the exogenous variables.

1. In the language of structural causal models, the average treatment effect is

$$\text{ATE}_{\text{scm}} \triangleq \mathbb{E}[Y \mid \text{Do}(T = 1)] - \mathbb{E}[Y \mid \text{Do}(T = 0)].$$

   In the language of potential outcomes, the average treatment effect is

$$\text{ATE}_{\text{po}} \triangleq \mathbb{E}[Y_1(U) - Y_0(U)].$$

   Show these quantities are equivalent, i.e. **prove** $\text{ATE}_{\text{scm}} = \text{ATE}_{\text{po}}$.

   A classic estimator for estimating average treatment effects is *inverse-propensity* (IP) weighting. Concretely, given n i.i.d. samples (X^i, T^i, Y^i), the estimator is

$$\widehat{\text{ATE}}_{\text{ip}} = \frac{1}{n} \sum_{i=1}^{n} \frac{T^i Y^i}{e(X^i)} - \frac{(1 - T^i)Y^i}{1 - e(X^i)},$$

   where $e(x)$ is the *propensity score*, $e(x) = \mathbb{P}(T = 1 \mid X = x)$. (Normally, the propensity scores are also estimated from data, but for this problem we assume they are known for simplicity.) Assume the propensity scores $e(X) \in (0, 1)$ almost surely. This is sometimes called *positivity*.

   In the next two parts, we show $\widehat{\text{ATE}}_{\text{ip}}$ is an unbiased estimate of the average treatment effect using both potential outcomes and structural causal models.

2. If X satisfies the backdoor criterion relative to (T, Y), then the potential outcomes satisfy an assumption called *ignorability*, namely $\{Y_0(U), Y_1(U)\} \perp\!\!\!\perp T \mid X$. (You do not need to show this).

   Using ignorability, show

$$\mathbb{E}\left[\frac{TY}{e(X)}\right] = \mathbb{E}[Y_1],$$

   and, similarly, **show**

$$\mathbb{E}\left[\frac{(1 - T)Y}{1 - e(X)}\right] = \mathbb{E}[Y_0].$$

   Consequently, $\mathbb{E}[\widehat{\text{ATE}}_{\text{ip}}] = \text{ATE}_{\text{po}}$.

3. **Using the backdoor-criterion, show**

$$\mathbb{E}\left[\frac{TY}{e(X)}\right] = \mathbb{E}[Y \mid \mathrm{Do}(T=1)],$$

and similarly, **show**

$$\mathbb{E}\left[\frac{(1-T)Y}{1-e(X)}\right] = \mathbb{E}[Y \mid \mathrm{Do}(T=0)].$$

Consequently, $\mathbb{E}[\widehat{\mathrm{ATE}}_{\mathrm{ip}}] = \mathrm{ATE}_{\mathrm{scm}}$.

8.9 Problem

1. **Prove that the number of causal graphs over a set of n variables is greater than $2^{\Omega(n^2)}$**

2. Consider a causal model G over $V = \{X, Y, Z_1, \ldots Z_n\}$ where there is no unobserved confounding. More formally, assume that the causal graph is Markovian and that every exogenous noise variable U has outdegree 1.

   Prove that the number of distinct values for $P(Y = y \mid do(X = x))$ is at most 2^n.

   Observe how this implies that the number of causal effects is much (much) smaller than the number of causal graphs.

9 Causal inference in practice

10 Sequential decision making and dynamic programming

10.1 Problem

Recall the model of strategic classification, where an institution designs a classifier $x \mapsto f(x)$ which maps features x of individuals to labels $f(x) \in \{0, 1\}$. Associated with each individual is their true features x and label y. In the *static learning problem*, the goal is simply to find a classifier that predicts the correct label, i.e. $f(x) = y$.

In the *strategic learning problem*, the individual additionally has a cost function $c(x, x')$. Individuals change their features to x' to maximize their expected gain $f(x') - c(x, x')$. The goal of the institution is instead to design a classifier such that even after manipulation, the correct label is chosen, i.e. $f(x') = y$.

In this problem, we consider linear classifiers of the form

$$f_{w,b}(x) = \begin{cases} 1 & w^\top x \geq b \\ 0 & w^\top x < b \end{cases}$$

We can also write the classifier as an indicator function, $f_{w,b}(x) = \mathbf{1}\{w^\top x \geq b\}$. We consider quadratic costs to individuals,

$$c(x, x') = (x - x')^\top Q (x - x') \,,$$

for a symmetric and positive definite matrix Q.

1. The optimal response of an individual, $x'_\star$, is the value which maximizes their expected gain. **Show that for a fixed classifier**, the optimal response can be written as the minimizer of a constrained optimization problem with some additional logic, i.e.

$$x'_\star = \begin{cases} \arg\min_{x' \in \mathcal{C}} g(x') & \min_{x' \in \mathcal{C}} g(x') \leq d \\ x & \min_{x' \in \mathcal{C}} g(x') > d \end{cases} .$$

   What is the form of the objective $g(x')$, the constraint set $\mathcal{C}$, and the parameter d?

2. What is the optimal response of the individual? **Simplify the expression derived in the previous part by solving the constrained minimization problem**.

3. **Show that** if parameters $w_\star, b_\star$ achieve the best possible for the static learning problem, then $w_\star, b'$ achieve the best accuracy for the strategic learning problem. **What is the value of b'?**

10.2 Problem

Consider a hidden process that at each time step has a binary state $x_i \in \{0, 1\}$. Each state x_i depends on the previous states $x_1, \ldots x_{i-1}$ through the states before it. We observe the current state x_i via noisy binary y_i. These sequence models are known as hidden Markov models; our goal is to predict the next state and observation given the past observations. At every time step, x_i evolves to x_{i+1} according to probability matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{bmatrix} p_{0|0} & p_{0|1} \\ p_{1|0} & p_{1|1} \end{bmatrix} = \begin{bmatrix} \tilde{f}p_{\bullet|0} & \tilde{f}p_{\bullet|1} \end{bmatrix}$$

where $p_{j|k} = P(x_{i+1} = j | x_i = k)$.

Suppose that at every time step, our observations are binary, and we observe y_i given x_i according to probability matrix $\mathbf{Q}$:

$$\mathbf{Q} = \begin{bmatrix} q_{0|0} & q_{0|1} \\ q_{1|0} & q_{1|1} \end{bmatrix} = \begin{bmatrix} \tilde{f}q_{0|\bullet}^\top \\ \tilde{f}q_{1|\bullet}^\top \end{bmatrix},$$

where $q_{j|k} = P(y_i = j | x_i = k)$.

At iteration i, we would like to derive the posterior distribution over the hidden state x_i, given all measurements, $P(x_i | y_1, \ldots, y_i)$.

1. In this case, we can derive an explicit expression for the posterior distribution. First, **write the joint distribution at iteration i, $P(x_i, y_1, \ldots, y_i)$ in terms of the joint distribution at iteration $i-1$, $P(x_{i-1}, y_1, \ldots, y_{i-1})$ and $\mathbf{P}$ and $\mathbf{Q}$.** Henceforth, we refer to this joint distribution at iteration i as a vector $\tilde{f}m^{(i)}$ with $\tilde{f}m_j^{(i)} = P(x_i = j, y_1, \ldots, y_i)$.

2. **Use this relation to write the posterior distributions $P(x_i | y_1, \ldots, y_i)$ and $P(y_i + 1 | y_1, \ldots, y_i)$ at iteration i.**

3. **Use the above expressions to derive the Bayes optimal estimator $\hat{y}_{i+1}(y_1, \ldots, y_i)$ under the squared loss.** Please express your solution in terms of $\tilde{f}m$ and $\tilde{f}h$, where $\tilde{f}h$ is defined as $\tilde{f}h := \tilde{f}q_{1|\bullet}^\top \tilde{f}P$. $\tilde{f}h_0$ is the probability of observing $y_k = 1$ given that the previous state $x_{k-1} = 0$; $\tilde{f}h_1$ is the probability of observing $y_k = 1$ given that the previous state $x_{k-1} = 1$.

10.3 Problem

Consider a discrete-time linear dynamical system:

$$x_{k+1} = A_k x_k + B_k u_k$$

where x_k is the state vector and u_k is a control input vector. Assume that A_k and B_k are known. The system runs for $k = 0, 1, \ldots, N-1$. At each time step k, assume that we get to observe the state vector x_k.

We want to choose u_k at each time step k in order to minimize a quadratic running cost:

$$x_N^T Q_N x_N + \sum_{k=0}^{N-1} (x_k^T Q_k x_k + u_k^T R_k u_k)$$

where Q_k and R_k are known positive definite matrices. Throughout this question, if you need to invert a matrix you may assume it is invertible.

It may be helpful to define the optimal cost-to-go (or value function) from a state x_k:

$$J_k(x_k) = \min_{u_j, j \geq k} \left[x_N^T Q_N x_N + \sum_{i=k}^{N-1} (x_i^T Q_i x_i + u_i^T R_i u_i) \right]$$

which is the cost incurred by an optimal controller starting at x_k.

1. **Prove** (by induction) that the optimal u_k in this case is a linear feedback on the state x_k, i.e. $u_k = -K_k x_k$, and solve for the gain matrix K_k. *Hint:* Show that the $J_k(x_k)$ is a quadratic form.

Now imagine that our system is modified as follows:

$$x_{k+1} = A_k x_k + B_k u_k + w_k$$

where w_k is a noise vector. Assume that w_k is zero-mean, independent, and finite-variance.

Also, instead of observing the full state vector x_k, we observe an output z_k, defined as:

$$z_k = h_k(x_k) + v_k$$

where h_k is a known function and v_k is another zero-mean, independent, finite-variance noise vector. (w_i is independent of v_j for all i, j.) Our cost (to minimize) is now:

$$\mathbb{E}\left[x_N^T Q_N x_N + \sum_{k=0}^{N-1} (x_k^T Q_k x_k + u_k^T R_k u_k) \right]$$

At every time step, the controller has access to all previous outputs and control inputs. We denote this available information in a vector $I_k = [z_0, ..., z_k, u_0, ..., u_{k-1}]$. Our goal is to choose u_k (based on our information I_k) to minimize the expected cost.

2. **Prove** (by induction) that the optimal u_k in this case is a linear feedback on the expected state (given our available information), i.e. $u_k = -L_k \mathbb{E}[x_k | I_k]$. How does L_k compare to K_k from the previous part? *Hint:* Follow a similar procedure as in the previous part, but at each step take into account the available information I_k.

   This is the *separation principle*; you can separately solve for an optimal state estimate and an optimal controller (as if there were no measurement error), and this combined controller is optimal.

3. If w_k and v_k were Gaussian, and h_k were linear, how would you compute $\mathbb{E}[x_k | I_k]$ efficiently?

11 Reinforcement learning

11.1 Problem

Let $\mathcal{E}$ be an arbitrary set of bandits. Suppose you are given a policy $\mathcal{A}$ designed for $\mathcal{E}$ that accepts the horizon T as a parameter and has a regret guarantee of

$$R_T \leq f_T(\nu), \quad \forall \nu \in \mathcal{E}\ ,$$

where $f_T : \mathcal{E} \to [0, \infty)$ is a sequence of functions. The purpose of this exercise is to analyze a meta-algorithm based on the so-called *doubling trick* that converts a policy depending on the horizon to a policy with similar guarantees that does not. Let $T_1 < T_2 < T_3 < \dots$ be a fixed sequence of integers and consider the policy that runs $\mathcal{A}$ with horizon T_1 until round $t = \min\{T, T_1\}$. Then restarts the algorithm with horizon T_2 until $t = \min\{T, T_1 + T_2\}$. Then restarts again with horizon T_3 until $t = \min\{T, T_1 + T_2 + T_3\}$ and so-on. Note that t is the real time counter and is not reset on each restart.

1. Let $\ell_{\max} = \min\{\ell : \sum_{i=1}^{\ell} T_i \geq T\}$. **Prove that the regret of the meta-algorithm is at most**
$$R_T \leq \sum_{\ell=1}^{\ell_{\max}} f_{T_\ell}(\nu).$$

2. Suppose that $f_T(\nu) \leq \sqrt{T}$. **Show that** if $T_\ell = 2^{\ell-1}$, then the regret of the meta-algorithm is at most $R_T \leq C\sqrt{T}$, where $C > 0$ is a carefully chosen universal constant.

3. Suppose that $f_T(\nu) = g(\nu)\log(T)$ for some function $g : \mathcal{E} \to [0, \infty)$. **What is the regret of the meta-algorithm** if $T_\ell = 2^{\ell-1}$? Can you find a better choice of $(T_\ell)_\ell$?

4. In lecture, we discussed the *explore then commit* strategy for two armed bandits, and showed that it can achieve logarithmic regret for a fixed horizon T and correctly chosen exploration horizon $m \leq T$. **Extend this strategy using the doubling trick so that it no longer depends on the horizon, and bound the resulting regret**.

11.2 Problem

Consider the following multi-armed bandit setting:

- arms $\{1, 2, \ldots, k\}$
- arm i gives reward that is sub-Gaussian with parameter $\sigma = 1$ and mean μ_i
- without loss of generality, assume that arm 1 is optimal, with $\mu_1 = 0$
- $\Delta_i = -\mu_i$, the optimality gap for arm i

We get to proceed for a fixed horizon n, where in each step we choose one arm i to pull, and we receive a reward drawn i.i.d. from the distribution for that arm. Let $a_j \in \{1, \ldots, k\}$ be the arm we decide to pull in step $j = 1, \ldots, n$. We want to minimize our *expected regret*, which is $\sum_{j=1}^{n} \Delta_{a_j}$. Consider the following successive elimination algorithm:

Algorithm 1 Successive Elimination

SuccessiveEliminationk, n $A_1 \leftarrow \{1, 2, \ldots, k\}$ phase $\ell = 1, 2, 3, \ldots$ (continue as long as $\sum_\ell m_\ell |A_\ell| \leq n$) Choose each arm $i \in A_\ell$ exactly m_ℓ times $\hat{\mu}_{i,\ell} \leftarrow$ the average reward for arm i from phase ℓ only $A_{\ell+1} \leftarrow \left\{i : \hat{\mu}_{i,\ell} + 2^{-\ell} \geq \max_{j \in A_\ell} \hat{\mu}_{j,\ell}\right\}$

The algorithm proceeds in phases ℓ and maintains an active set A_ℓ of arms, with the intention that arm 1 (an optimal arm) is always in the active set. At each phase, all active arms are sampled equally, average rewards are estimated, and the least promising arms are eliminated from the active set for the next phase.

1. **Show** that for any $\ell \geq 1$,

$$\mathbb{P}[1 \notin A_{\ell+1}, 1 \in A_\ell] \leq k \exp\left(-\frac{m_\ell 2^{-2\ell}}{4}\right)$$

   Since $\mathbb{P}[A, B] \leq \mathbb{P}[A|B]$, it suffices to show the statement with the left hand side replaced by $\mathbb{P}[1 \notin A_{\ell+1} | 1 \in A_\ell]$: i.e. assuming that arm 1 was in A_ℓ, we want to bound the probability that arm 1 is rejected and not placed in $A_{\ell+1}$. The intent here is to make sure that with high probability, we do not reject the optimal arm.

2. **Show** that if $i \in [k]$ and $\ell \geq 1$ are such that $\Delta_i \geq 2^{-\ell}$, then

$$\mathbb{P}[i \in A_{\ell+1}, 1 \in A_\ell, i \in A_\ell] \leq \exp\left(-\frac{m_\ell (\Delta_i - 2^{-\ell})^2}{4}\right)$$

   Since $\mathbb{P}[A, B] \leq P[A|B]$, it suffices to show the statement with the left hand side replaced by $\mathbb{P}[i \in A_{\ell+1} | 1 \in A_\ell, i \in A_\ell]$: i.e. assuming that arm 1 and arm i were both in A_ℓ, and i is sufficiently suboptimal, we want to bound the probability that arm i makes it into $A_{\ell+1}$. The intent here is to make sure that with high probability, we reject arms that are sufficiently suboptimal.

3. Let $\ell_i = \min\{\ell \geq 1 : 2^{-\ell} \leq \Delta_i/2\}$ **Choose** m_ℓ in such a way that $\mathbb{P}(\text{exists } \ell : 1 \notin A_\ell) \leq 1/n$ and $\mathbb{P}(i \in A_{\ell_i+1}) \leq 1/n$

4. **Show** that the algorithm has regret at most

$$R_n \leq C \sum_{i:\Delta_i>0} (\Delta_i + \frac{1}{\Delta_i} \log n)$$

11.3 Problem

For this problem, we consider a setting where there are K actions $u_1, \ldots, u_K$. When we perform action u_k we get reward x_k which has mean μ_k and is bounded between 0 and 1.

1. First assume that $\mu_{k^*} = \max_{i \in [K]} \mu_i$ satisfies $\mu_{k^*} - \mu_j \geq \Delta$ for all $j \neq k^*$. **Design an algorithm** and **prove** that it identifies k^* with probability $1 - \delta$ after taking $\frac{2K}{\Delta^2} \log(\frac{K}{\delta})$ many actions.

2. Now assume that we want to optimize reward adaptively. Fix a horizon T. Let R_t be the random reward realized at time $t \in [T]$. Define the expected regret of an algorithm A to be

$$\text{Regret}(A) = T\mu_{i^*} - \sum_{t=1}^{T} \mathbb{E}[R_t]$$

   Design an algorithm A and **prove** that $\text{Regret}(A) \leq \mathcal{O}(\frac{K}{\Delta^2} \log T)$.

3. Assume that the horizon T is unknown. **Provide pseudocode** which extends the algorithm above to this setting.

12 Other Problems

12.1 Problem

Let $\{x_1, x_2, \ldots, x_n, \ldots\}$ be a sequence of independent random variables with identical means $\mathbb{E}[x_i] = m$ and variances $\text{var}(x_i) = \sigma^2$. Define the sample mean and sample mean-square of the first N of the x_i's as

$$\text{sample mean:} \qquad m_N = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\text{sample mean-square:} \qquad s_N^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2$$

1. **Find the mean and variance of the sample mean. Show that**

   $$\lim_{N\to\infty} \mathbb{E}[(m_N - m)^2] = 0$$

   and use this result to deduce that

   $$\lim_{N\to\infty} \mathbb{P}[|m_N - m| \geq \epsilon] = 0$$

   for any $\epsilon > 0$.

2. Suppose the x_i are zero-mean Gaussian random variables. Find the mean and variance of the sample mean-square. **Show that**

   $$\lim_{N\to\infty} \mathbb{E}[(s_N^2 - \sigma^2)^2] = 0$$

3. Suppose that the x_is are independent, zero-mean, Gaussian random variables. **Are m_N and s_N^2 Gaussian random variables? Explain your reasoning.**

12.2 Problem

Let X and Y be independent random variables with probability density functions

$$p_X(x) = \frac{1}{2}\delta(x+1) + \frac{1}{2}\delta(x-1),$$

$$p_Y(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right).$$

Let $Z = X + Y$ and $W = XY$.

1. **Find $p_Z(z)$, the probability density function of Z.**
2. **Find the conditional probability density functions $p_{Z|X}(z|x=-1)$ and $p_{Z|X}(z|x=1)$.**
3. **Are Y and W uncorrelated (i.e. $\mathbb{E}[YW] = 0$)? Are they independent (i.e. $p_{W|Y}(w|y) = p_W(w)$)?**

12.3 Problem

1. For $p \in (0, \infty)$, let $\|x\|_p = (\sum_{i=1}^d |x_i|^p)^{1/p}$, and let $\|x\|_\infty = \max_i |x_i|$. **Please prove the following:**

   (a) Show that $\|x\|_2$, $\|x\|_1$, and $\|x\|_\infty$ are norms.

   (b) Show that $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$.

   (c) Show that $\|x\|_2^2 \leq \|x\|_1 \|x\|_\infty$.

   (d) Show that $\|x\|_1 \leq \sqrt{d}\|x\|_2$ and $\|x\|_2 \leq \sqrt{d}\|x\|_\infty$.

2. For each the following functions $f(x) : \mathbb{R}^d \to \mathbb{R}$, **state if f is always, sometimes, or never a norm.** Provide a proof or counterexample, and if the answer is 'sometimes,' give a necessary and sufficient condition for f to be a norm.

   (a) $f(x) = \log \sinh \|x\|_2$.

   (b) $f(x) = \sum_{i=1}^d x_i^2$.

   (c) $f(x) = \|Ax\|$, where $\|x\|$ is a norm on $\mathbb{R}^d$, and $A \in \mathbb{R}^{d \times d}$.

   (d) $f(x) = \sqrt{x^\top \Sigma x}$ where Σ is symmetric and has strictly positive eigenvalues.

   (e) $f(x) = \sqrt{x^\top \Sigma x}$, where $\Sigma = \begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix}$, and $A \in \mathbb{R}^{n \times m}$ where $m + n = d$.

   (f) $f(x) = \sum_i \alpha_i |x_i|$, $\alpha_i \in \mathbb{R}$.

3. Consider a function $f(x) = \sup_{w \in \mathcal{C}} \langle w, x \rangle$, where $\mathcal{C} \subset \mathbb{R}^d$ with the following properties:

   - $x \in \mathcal{C}$ if and only if $-x \in \mathcal{C}$.
   - $\mathcal{C}$ is bounded; that is $\sup_{x \in \mathcal{C}} \|x\| < \infty$.
   - There exists an orthonormal basis $\{e_1, e_2, \ldots, e_d\} \subset \mathcal{C}$.

   Is $f(x)$ a norm? Provide a proof.

12.4 Problem

1. Let the vector space $\mathcal{V}$ be a general *inner product space* (i.e. not necessarily $\mathbb{R}^n$) and let x, y be elements of $\mathcal{V}$. Suppose we want to approximate x as a multiple of y, that is, let $\widehat{x} = ay$ for $a \in \mathbb{R}$ so that $\widehat{x}$ is, in some sense, as close as possible for x.

   (a) Let $e = x - \widehat{x} = x - ay$. **Show that**

   $$J = \|e\| = \sqrt{\langle e, e \rangle}$$

   is minimized over all possible values of a when

   $$\langle e, y \rangle = 0 \,.$$

   Hint: you may want to use the orthogonal decomposition theorem.

   (b) **Find an explicit formula for a in terms of inner products involving x and y.**

   (c) **Give explicit formulas for a in the following two cases:**

   i. $\mathcal{V} = \mathbb{R}^n$ and $\langle u, v \rangle = u^\top v$.

   ii. $\mathcal{V}$ is defined over scalar random variables, with $\langle U, V \rangle = \mathbb{E}[UV]$

2. Let X, Y, Z be zero-mean, unit-variance random variables which satisfy

   $$\mathrm{Var}(X + Y + Z) = 0 \,.$$

   Find the covariance matrix of X, Y, Z, i.e. find the matrix

   $$\begin{bmatrix} \mathbb{E}[X^2] & \mathbb{E}[XY] & \mathbb{E}[XZ] \\ \mathbb{E}[YX] & \mathbb{E}[Y^2] & \mathbb{E}[YZ] \\ \mathbb{E}[ZX] & \mathbb{E}[ZY] & \mathbb{E}[Z^2] \end{bmatrix}$$

   Hint: you may want to use vector space ideas.

12.5 Problem

1. Assume that M, N are symmetric positive semidefinite matrices of the same dimensions. **For each of the following statements, either prove that it is true, or give a counterexample**.

   (a) $B^\top MB$ is positive semidefinite for any matrix B with consistent dimensions.

   (b) If $M - N$ is positive semidefinite, then $\lambda_{\max}(M) \geq \lambda_{\max}(N)$ where $\lambda_{\max}$ denotes the largest eigenvalue.

   (c) $\mathbf{Tr}(MN) = 0$ if and only if $MN = 0$.

   (d) Let $\circ$ denote the Hadamard product. $M \circ N$ is positive semidefinite.

   (e) Suppose that M, N are also invertible. If $M-N$ is positive semidefinite, then $N^{-1}-M^{-1}$ is positive semidefinite.

2. A is symmetric in all parts.

   (a) Let A be a positive definite matrix. Prove that all eigenvalues of A are greater than zero.

   (b) Let A be a positive definite matrix. Prove that A is invertible. (Hint: Use the previous part.)

   (c) Let A be a positive semidefinite matrix. Find all $\gamma \in \mathbb{R}$ such that $A + \gamma I$ is positive definite.

   (d) Let A be a positive definite matrix. Prove that there exist n linearly independent vectors $x_1, x_2, ..., x_n$ such that $A_{ij} = x_i^\top x_j$. (Hint: Use the spectral theorem to find a matrix B such that $A = B^\top B$.)

   (e) Show that for symmetric matrix A, $\cosh(A) = \frac{e^A + e^{-A}}{2}$ is positive definite. Recall that $e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!}$.

3. Suppose A is an $m \times n$ matrix and $A = U\Sigma V^T$ is a singular value decomposition.

   (a) **What is the singular value decomposition of A^T?**

   (b) If $m = n$ and A is nonsingular, **what is the singular value decomposition of A^{-1}?**

   (c) If $m = n$ and A is skew-symmetric (i.e., $A = -A^T$), **show that the nonzero singular values of A come in pairs.**

4. Let p be a probability distribution on the interval $[0, 1]$. Let the kth *moment* of p be the expected value

$$\mu_k = \mathbb{E}[x^k] = \int_0^1 x^k p(x) dx \,.$$

   Prove that the $n \times n$ matrix H with entries $H_{ij} = \mu_{i+j}$ is positive semidefinite.

12.6 Problem

This problem exercises your knowledge of basic probability in the context of understanding why lots of training data helps improve the accuracy of learning things.

For each θ in the interval $(1/4, 3/4)$, define $f_\theta : [0,1] \to \{0,1\}$, such that

$$f_\theta(x) = \begin{cases} 1 \text{ if } x > \theta \\ 0 \text{ otherwise.} \end{cases}$$

We draw samples $X_1, X_2, \ldots, X_n$ uniformly at random and i.i.d. from the interval $[0,1]$. Our goal is to learn an estimate for θ from n random samples $(X_1, f_\theta(X_1)), (X_2, f_\theta(X_2)), \ldots, (X_n, f_\theta(X_n))$. We let

$$T_{min} = \max\left(\left\{\frac{1}{4}\right\} \cup \{X_i | f_\theta(X_i) = 0\}\right) ,$$
$$T_{max} = \min\left(\left\{\frac{3}{4}\right\} \cup \{X_i | f_\theta(X_i) = 1\}\right) .$$

We know that the true θ must be larger than T_{min} and smaller than T_{max}. Both T_{min} and T_{max} are random variables, and the gap between them represents the uncertainty we will have about the true θ given the training data that we have received.

1. Suppose that you would like to have an estimate for θ that has an accuracy of 2ϵ, with probability at least $1 - \delta$. **How large must the number of samples n be?** *Hint: you may want to compute the probabilities $\mathbb{P}(T_{max} - \theta > \epsilon)$ and $\mathbb{P}(\theta - T_{min} > \epsilon)$ as a function of ϵ.*

2. Instead of getting random samples $(X_i, f(X_i))$, suppose we were allowed to choose where to sample the function, but you have to choose all the sampling locations in advance. Propose a method to estimate θ. **How many samples suffice to achieve an estimate that is within an interval of size 2ϵ?** *Hint: You need not use a randomized strategy.*

3. Now suppose that you can pick where to sample the function adaptively – choosing where to sample the function in response to the previously observed values. Propose a method to estimate θ. **How many samples suffice to achieve an estimate that is within an interval of size 2ϵ?**

4. Compare the scaling of n with ϵ and δ in the three sampling approaches above: random, deterministic, and adaptive.

12.7 Problem

1. Consider the random variables X, Y whose joint density is given by

$$p_{X,Y}(x,y) = \begin{cases} 2 & x,y \geq 0 \text{ and } x+y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

It may be helpful to sketch the density over x and y.

(a) **Compute the covariance matrix**

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{XX} & \lambda_{XY} \\ \lambda_{YX} & \lambda_{YY} \end{bmatrix}$$

(b) Knowledge about Y generally gives us information about the random variable X, and vice versa. Suppose we want to estimate X based on knowledge of Y. In particular, we want to estimate X as an affine function of Y:

$$\hat{X}(Y) = aY + b,$$

where a, b are constants. **Select a and b such that**

$$\mathbb{E}[(\hat{X} - X)^2],$$

the expected mean squared error between X and its estimate $\hat{X}(Y)$, is minimized.

2. Consider the random variables X, Y whose joint density is given by

$$p_{X,Y}(x,y) = \begin{cases} 1 & 0 \leq x,y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Repeat the above steps in the previous problem part.

12.8 Problem

Assume that $X_1, \ldots, X_n$ are nonnegative, independent (but not identically distributed) random variables and have density bounded by 1. That is $P(X_i = x) \leq 1$ for all x and X_i. **Prove that the following inequality holds:**

$$P\left(\sum_{i=1}^{n} X_i \leq \varepsilon n\right) \leq (e\varepsilon)^n$$

12.9 Problem

This question introduces the Chernoff bound.

1. Let X be a random variable with $\mathbb{E}X^k < \infty \ \ \forall k = 1, 2, ...$ **Prove** that

$$P[X - \mathbb{E}X \geq t] \leq \inf_{\lambda > 0} \mathbb{E}[\exp(\lambda(X - \mathbb{E}X))] \exp(-\lambda t)$$

   This inequality is known as the Chernoff bound. It is a very general tool that often comes in handy.

2. A random variable X is called *sub-Gaussian* with parameter σ if its moment generating function decays at least as fast as that of a Gaussian, *i.e.*:

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}X))] \leq \exp(\sigma^2 \lambda^2 / 2)$$

   holds for any λ.

   Use the result of part 1 to **show** that if a random variable X is sub-Gaussian with parameter σ, it satisfies the tail bound:

$$P[X - \mathbb{E}X \geq t] \leq \exp(\frac{-t^2}{2\sigma^2})$$