CS 189
Fall 2018

Introduction to Machine Learning

EXAMFINAL

After the exam starts, please write your student ID (or name) on **EVERY PAGE**.

There are **8** questions for a total of **29** parts. You may consult your sheet of notes. Calculators, phones, computers, and other electronic devices are not permitted. There are **30** pages on the exam. **Notify a proctor immediately if a page is missing.** You may, without proof, use theorems and lemmas that were proven in the notes and/or in lecture, unless we explicitly ask for a derivation. However, you must clearly state what theorem or lemma you are using and where/how you are using it.

Please write legibly if you want full credit on all problems.

You have 170 minutes.

PRINT and SIGN Your Name: _____, _____, _____
(last) (first) (signature)

PRINT Your Student ID: _____

Person before you: _____, _____
(name) (SID)

Person behind you: _____, _____
(name) (SID)

Person to your left: _____, _____
(name) (SID)

Person to your right: _____, _____
(name) (SID)

Seat Number: _____

Do not turn this page until your instructor tells you to do so.

1 Short answer (5 parts, 14 points)

For each of the following questions, write at most one sentence in answer.
If you write more than one sentence, we will only grade the first sentence.

(a) (4 points) For the following loss functions $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, **write a YES or a NO in the table** for whether it

(1) assigns a loss of 0 to any prediction with the right sign, i.e., for any $\hat{y}, y_* \in \mathbb{R}$,

$$\ell(\hat{y}, y_*) = 0 \text{ if } \operatorname{sign}(\hat{y}) = \operatorname{sign}(y_*) ,$$

(2) is bounded, i.e. $\max_{\hat{y} \in \mathbb{R}, y_* \in \mathbb{R}} \left|\ell(\hat{y}, y_*)\right| \leq C$ for some finite constant $C \geq 0$,

(3) is convex on $\mathbb{R}$ in its first argument, i.e. $\hat{y} \mapsto \ell(\hat{y}, y_*)$ is convex on $\mathbb{R}$,

(4) is differentiable everywhere on $\mathbb{R}$ in its first argument, i.e. $\hat{y} \mapsto \ell(\hat{y}, y_*)$ is differentiable on $\mathbb{R}$.

The losses are defined as follows (recall that both $\hat{y}$ and y_* are scalars):

- 0-1 loss: $\ell(\hat{y}, y_*) = \begin{cases} 1 & \text{if } \operatorname{sign}(\hat{y}) \neq \operatorname{sign}(y_*) \\ 0 & \text{otherwise} \end{cases}$
- squared loss: $\ell(\hat{y}, y_*) = (\hat{y} - y_*)^2$
- absolute loss: $\ell(\hat{y}, y_*) = |\hat{y} - y_*|$
- sigmoid loss: $\ell(\hat{y}, y_*) = \frac{1}{1+\exp(-\hat{y} \cdot y_*)}$
- hinge loss: $\ell(\hat{y}, y_*) = \max\{0, 1 - \hat{y} \cdot y_*\}$

The first column of answers is provided for you as an example.

	0-1 loss	squared loss	absolute loss	sigmoid loss	hinge loss
(1) sign	YES				
(2) bounded	YES				
(3) convex	NO				
(4) everywhere differentiable	NO				

Solution:

	0-1 loss	squared loss	absolute loss	sigmoid loss	hinge loss
(1) sign	YES	NO	NO	NO	NO
(2) bounded	YES	NO	NO	YES	NO
(3) convex	NO	YES	YES	NO	YES
(4) everywhere differentiable	NO	YES	NO	YES	NO

(b) (2 points) **Give an example of two random variables $X, Y \in \mathbb{R}$ that are uncorrelated, but *not* independent.**

Solution:

Many examples exist. A simple one is

$$X = \mathcal{N}(0,1)\ ,$$
$$Y = \begin{cases} X & \text{w.p. } 1/2 \\ -X & \text{w.p. } 1/2 \end{cases}\ .$$

(c) (2 points) **Give a necessary and sufficient condition on a matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$**, under which we can write $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ for $\mathbf{D} \in \mathbb{R}^{n \times n}$ diagonal and $\mathbf{U} \in \mathbb{R}^{n \times n}$ orthonormal.

Solution: The matrix $\mathbf{X}$ must be symmetric.

(d) (2 points) Recall that when describing the optimal regression weights $\mathbf{w}$ using ridge regression, the pseudo-inverse, and PCA, we can write the weights using the SVD of the feature matrix $\mathbf{X}$ as follows, with rescaling on the singular values s_i according to the spectral function $\rho(s_i)$:

$$\mathbf{w} = \sum_{i=1}^{n} \rho(s_i) \mathbf{v}_i \mathbf{u}_i^\top y$$

Figure 1 plots the spectral function $\rho(s_i)$ for (1) Pseudo-Inverse, (2) PCA, and (3) Ridge Regression. **In the boxes on Figure 1, label the graphs of the spectral function with the method that they correspond to**.

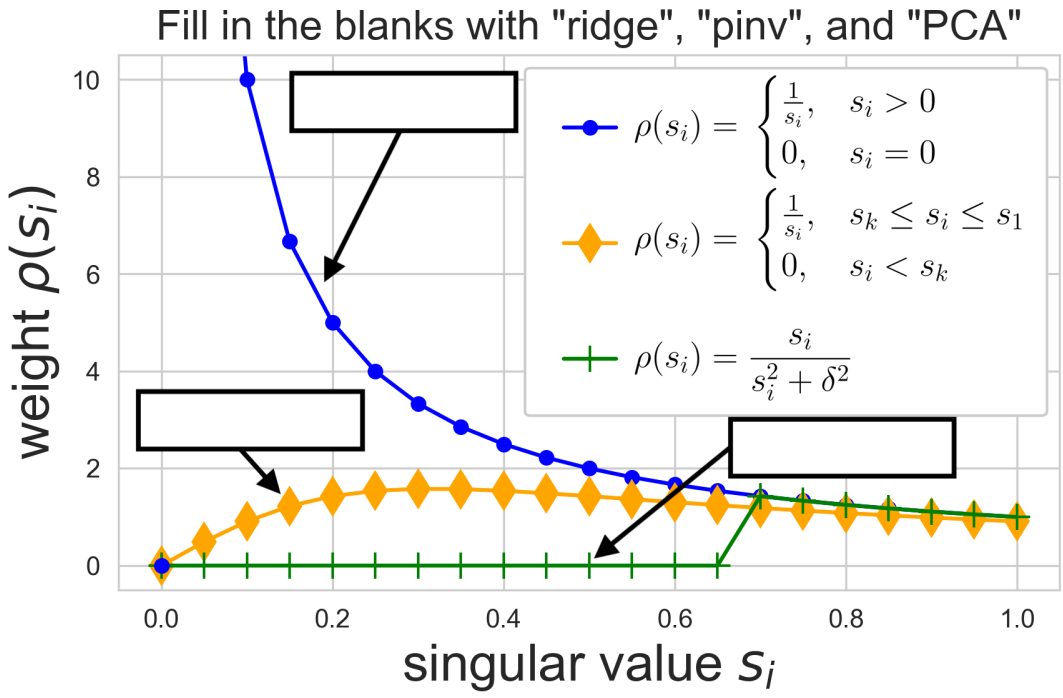


Figure 1: Fill in the boxes with “ridge”, “pinv”, and “PCA.”

Solution: Figure 2

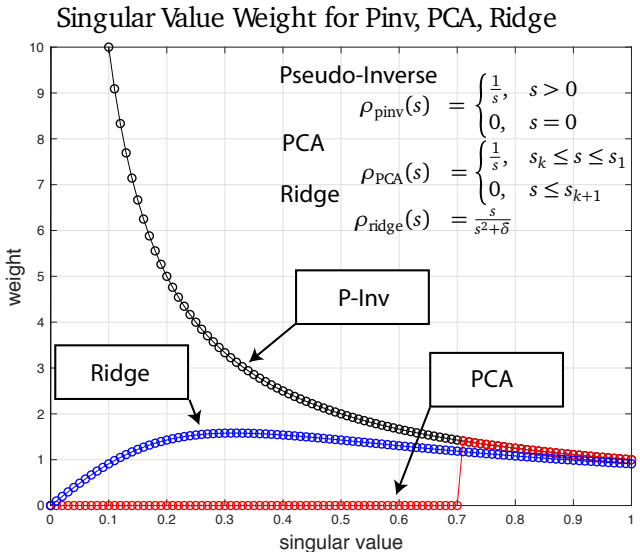


Figure 2: Solutions

(e) (4 points) Assume that $\mathbf{X} \in \mathbb{R}^{n \times d_1}$ and $\mathbf{Y} \in \mathbb{R}^{n \times d_2}$ are two full column-rank datasets ($n \geq d_1, n \geq d_2$). Recall that in the computations for canonical correlation analysis (CCA), we would like to maximize the correlation coefficient,

$$\rho_{\text{cca}}(\mathbf{X}, \mathbf{Y}) := \max_{\substack{\mathbf{u} \in \mathbb{R}^{d_1},\, \mathbf{v} \in \mathbb{R}^{d_2} \\ \mathbf{u} \neq 0, \mathbf{v} \neq 0}} \frac{\mathbf{v}^\top \mathbf{X}^\top \mathbf{Y} \mathbf{u}}{\sqrt{\mathbf{v}^\top \mathbf{X}^\top \mathbf{X} \mathbf{v} \cdot \mathbf{u}^\top \mathbf{Y}^\top \mathbf{Y} \mathbf{u}}}.$$

Now consider a full column-rank matrix $\mathbf{Z} \in \mathbb{R}^{n \times d}$ (i.e. $\text{rank}(\mathbf{Z}) = d$). **Compute the following correlation coefficients, and write down any corresponding pair of vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ that achieves the maximum.**

(1) $\rho_{\text{cca}}(\mathbf{Z}, \mathbf{Z})$

(2) $\rho_{\text{cca}}(\mathbf{Z}, -\mathbf{Z})$

Solution: Let $\mathbf{q}$ be any unit eigenvector of $\mathbf{Z}^\top \mathbf{Z}$. We have $\rho_{\text{cca}}(\mathbf{Z}, \mathbf{Z}) = 1$ with $\mathbf{u} = \mathbf{v} = \mathbf{q}$. Similarly, we have $\rho_{\text{cca}}(\mathbf{Z}, -Z) = 1$ with $\mathbf{u} = \mathbf{q}$ and $\mathbf{v} = -\mathbf{q}$.

2 Gaussian Kernels (3 parts, 12 points)

In this question, we will look at training a binary classifier with a Gaussian kernel. Specifically given a labelled dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \{\pm 1\}$ and a kernel function $k(\mathbf{x}_1, \mathbf{x}_2)$, we consider classifiers of the form:

$$\widehat{f}(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \mathbf{x})\right) ,$$

where we define $\text{sign}(u)$ to be 1 if $u \geq 0$ or -1 if $u < 0$. In order to choose the weights $\alpha_i, i = 1, ..., n$, we will consider the least-squares problem:

$$\boldsymbol{\alpha} \in \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \|\mathbf{K}\boldsymbol{\alpha} - \mathbf{y}\|_2^2 , \tag{1}$$

where $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j))_{i=1,j=1}^n$ is the kernel matrix and $\mathbf{y} = (y_1, ..., y_n)$ is the vector of labels. We will work with the Gaussian kernel. Recall that the Gaussian kernel with bandwidth $\sigma > 0$ is defined as:

$$k(\mathbf{x}_i, \mathbf{x}_j) := \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right) .$$

(a) (4 points) When the bandwidth parameter $\sigma \to 0$, observe that the off-diagonal entries of the kernel matrix $\mathbf{K}$ tend also to zero. Consider a two sample dataset S (i.e. $n = 2$) with $(x_1, y_1) = (1, 1)$ and $(x_2, y_2) = (-1, -1)$. **Assuming that as $\sigma \to 0$ the off-diagonal entries of $\mathbf{K}$ are equal to zero (and the diagonal entries unmodified), what is the optimal solution of $\boldsymbol{\alpha}$ for the optimization problem (1) and what is the resulting classifier $\widehat{f}(x)$?**

Solution: With our assumptions, the kernel matrix $\mathbf{K}$ is identity, so $\boldsymbol{\alpha}_\star = (+1, -1)$. By symmetry, the resulting $\widehat{f}(x) = \text{sign}(x)$.

(b) (4 points) Now we consider the regime when the bandwidth parameter $\sigma \to +\infty$. Observe in this regime, the off-diagonal entries of the kernel matrix $\mathbf{K}$ tend to one. Given a dataset S, suppose we solve the optimization problem (1) with all the off-diagonal entries of $\mathbf{K}$ equal to one (and the diagonal entries unmodified). **Prove that if the number of $+1$ labels in S equals the number of -1 labels in S, then $\boldsymbol{\alpha} = \mathbf{0}$ is an optimal solution of (1). What is the resulting classifier $\widehat{f}(x)$?**

Solution: With our assumptions, the kernel matrix $\mathbf{K} = \mathbf{1}\mathbf{1}^\top$. The normal equations for (1) are given by:

$$(\mathbf{1}\mathbf{1}^\top)^\top(\mathbf{1}\mathbf{1}^\top)\boldsymbol{\alpha} = (\mathbf{1}\mathbf{1}^\top)\mathbf{y} \iff n\mathbf{1}\mathbf{1}^\top\boldsymbol{\alpha} = \mathbf{1}(\mathbf{1}^\top\mathbf{y}) \, .$$

Since we assumed the number of $+1$ labels equals the number of -1 labels, then $\mathbf{1}^\top\mathbf{y} = 0$. Hence $\boldsymbol{\alpha} = \mathbf{0}$ satisfies the normal equations. The resulting classifier is a constant function (depending on what the value of $\text{sign}(0)$ is).

(c) (4 points) Now we consider the regime when the bandwidth parameter is large but finite. Consider again the two sample dataset S with $(x_1, y_1) = (1, 1)$ and $(x_2, y_2) = (-1, -1)$. When $\sigma \gg 1$, we can approximate $k(x_1, x_2) \approx 1 + \frac{x_1 x_2}{2\sigma^2}$. **Show that the solution of the optimization problem** (1) **with the kernel** $k_a(x_1, x_2) = 1 + \frac{x_1 x_2}{2\sigma^2}$ **is given by** $\boldsymbol{\alpha} = (\sigma^2, -\sigma^2)$**. What is the resulting classifier** $\widehat{f}(x)$**?**

Hint: The inverse of a 2×2 *matrix is given by the formula* $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.

Solution: The kernel matrix $\mathbf{K} = \begin{bmatrix} 1 + \frac{1}{2\sigma^2} & 1 - \frac{1}{2\sigma^2} \\ 1 - \frac{1}{2\sigma^2} & 1 + \frac{1}{2\sigma^2} \end{bmatrix}$. The inverse matrix is:

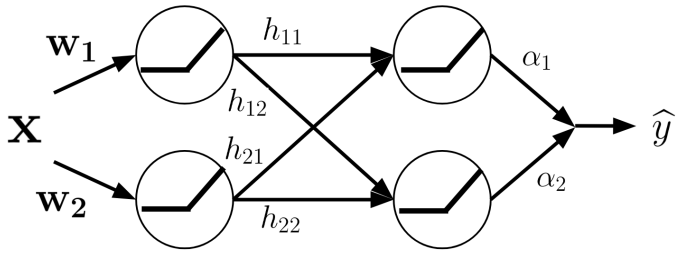
$$\mathbf{K}^{-1} = \frac{1}{4}\begin{bmatrix} 1 + 2\sigma^2 & 1 - 2\sigma^2 \\ 1 - 2\sigma^2 & 1 + 2\sigma^2 \end{bmatrix} \, .$$

Hence the optimal $\boldsymbol{\alpha} = (\sigma^2, -\sigma^2)$. Plugging this α into $\widehat{f}(x)$ we see that $\widehat{f}(x) = \text{sign}(x)$ (once again by symmetry).

3 Neural Networks (3 parts, 12 points)

We consider the following neural network structure, where:

$$\mathbf{x},\ \mathbf{w}_1,\ \mathbf{w}_2, \in \mathbb{R}^7, \quad \text{and}$$
$$\alpha_1,\ \alpha_2,\ h_{11},\ h_{12},\ h_{21},\ h_{22},\ \widehat{y} \in \mathbb{R}\,.$$



The output of this network $\widehat{y}(\mathbf{x})$ is given by $\widehat{y} = \alpha_1 v_1 + \alpha_2 v_2$, where

$$\begin{aligned} u_1 &= \max\{0, \mathbf{w}_1^\top \mathbf{x}\} \\ u_2 &= \max\{0, \mathbf{w}_2^\top \mathbf{x}\} \\ v_1 &= \max\{0, h_{11}u_1 + h_{21}u_2\} \\ v_2 &= \max\{0, h_{12}u_1 + h_{22}u_2\}\,. \end{aligned}$$

For a single data pair $(\mathbf{x}, y)$ we consider the loss function $\ell(\mathbf{x}, y) = \frac{1}{2}(\widehat{y}(\mathbf{x}) - y)^2$.

(a) (4 points) The average loss over an arbitrary set of labeled points $\{\mathbf{x}_i, y_i\}$ is given by $\frac{1}{n}\sum_{i=1}^n \ell(\mathbf{x}_i, y_i)$. In general, this function does not have a unique minimizer. **Give a specific reason why this is true.**

Solution: The vertical symmetry in the network architecture means that any parameter configuration leading to a local min can be permuted to achieve the same loss value.

(b) (4 points) **Compute the partial derivative $\frac{\partial}{\partial h_{ij}}\ell(\mathbf{x}, y)$. Your answer may be in terms of intermediate variables but must not contain any partial derivatives.** Please give an answer in terms of a general $\frac{\partial}{\partial h_{ij}}$ and do not simply give an explicit enumeration of the four different $\frac{\partial}{\partial h_{ij}}$ partial derivatives.

Note: You may take the derivative of a ReLu at zero to be zero.

Solution: By the chain rule, we have that

$$\frac{\partial}{\partial h_{ij}}\ell(\mathbf{x}, y) = (\widehat{y}(\mathbf{x}) - y)\left(\alpha_1 \frac{\partial}{\partial h_{ij}} v_1 + \alpha_2 \frac{\partial}{\partial h_{ij}} v_2\right) .$$

By viewing the graph, we note that only one of these partial derivatives will be nonzero—the one corresponding to α_j. Using the derivative rule for the ReLu, we arrive at

$$\begin{aligned}(\widehat{y}(\mathbf{x}) - y)\alpha_j \frac{\partial}{\partial h_{ij}} \max\{0, h_{1j}u_1 + h_{2j}u_2\}) \\ = (\widehat{y}(\mathbf{x}) - y)\alpha_j u_i \mathbf{1}_{\{h_{1j}u_1 + h_{2j}u_2 > 0\}} .\end{aligned}$$

(c) (4 points) Assuming $n > 2$, suppose we train this network with stochastic gradient descent using minibatches of size 2 and step size η. **Write down a procedure to update h_{ij} at time t.** (You may write it in terms of an unexpanded $\frac{\partial \ell}{\partial h_{ij}}$, i.e. you do not need to rely on a correct part (b).)

Solution:

One such answer:

- Choose z_1, z_2 uniformly at random from $[n]$.
- Perform the update

$$h_{ij}^{(t)} \leftarrow h_{ij}^{(t-1)} - \eta \sum_{k=1}^{2} \frac{\partial}{\partial h_{ij}} \ell(\mathbf{x}_{z_k}, y_{z_k})$$

4 Demographic Criteria (3 parts, 12 points)

In this problem, we analyze the recommendations made by a hypothetical resume screening tool. This tool automatically scans applicants' resumes and recommends a subset of applicants to be considered more closely by hiring managers. The table below shows counts of individuals, organized by the screening tool's recommendation ($\widehat{Y}$ which indicates whether a candidate was recommended or not) and by the actual hiring outcome (Y which indicates whether an individual was ultimately hired). The data is split by age: candidates below the age of 35 and candidates above the age of 35 (A indicates one of these two age groups). We have data from 400 candidates under 35 and 600 candidates over 35.

	Age < 35			Age≥ 35		
	$Y=0$	$Y=1$		$Y=0$	$Y=1$	
$\widehat{Y}=0$	150	50	200	200	100	300
$\widehat{Y}=1$	50	150	200	50	250	300
	200	200		250	350	

A perfect resume screener would be able to recommend only candidates who would later be hired, i.e. $\widehat{Y} = Y$. For this problem, we will recall the definitions of some demographic criteria.

- **Independence** is satisfied if $P(\widehat{Y} = i \mid A = a) = P(\widehat{Y} = i \mid A = a')$ for all i and any pairs a, a'.
- **Separation** is satisfied if $P(\widehat{Y} = i \mid A = a,\ Y = j) = P(\widehat{Y} = i \mid A = a',\ Y = j)$ for all i, j and any pairs a, a'.

(a) (4 points) First, we consider the accuracy of the tool. **What percent of the resume screening tool's recommendations are correct for:**

- the population as a whole?
- applicants under 35?
- applicants over 35?

Solution: All are 75%.

(b) (4 points) Next, we consider whether this tool satisfies demographic criteria. Let the sensitive characteristic A denote age less than 35 ($A = 0$) or greater than 35 ($A = 1$). **Does this tool satisfy independence or separation? Justify your answer mathematically.**

Solution: This tool satisfies independence, since $P(\widehat{Y} = 1|A = 1) = P(\widehat{Y} = 1|A = 0) = 0.5$. However, it does not satisfy separation, because $TPR_{A=0} = 0.75$ while $TPR_{A=1} = \frac{5}{7} \approx 0.71 < 0.75$.

(c) (4 points) Suppose you can change the recommendation $\widehat{Y}$ for specific applicants. **Describe changes in recommendations for applicants over 35 that would cause this tool to satisfy the separation criteria.** You do not need to compute exact numbers. **How would these changes affect the satisfaction of the independence criteria?**

Solution: To satisfy separation, we need to make true and false positive rates equal. Since we have $TPR_{A=1} < TPR_{A=0}$, we should accept more applicants over the age of 35 who are qualified ($Y = 1$). We also have that $FPR_{A=1} < FPR_{A=0}$, so we should accept more applicants over the age of 35 who are unqualified ($Y = 0$). Making this change would cause the overall acceptance rates to become unequal: we would have that $P(\widehat{Y} = 1|A = 1) > P(\widehat{Y} = 1|A = 0)$ so the classifier would no longer satisfy independence.

5 Population and Sample Risk (4 parts, 16 points)

Let $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be a set of n points sampled i.i.d. from a distribution $\mathcal{D}$. This is the training set with $\mathbf{x}_i \in \mathbb{R}^d$ being the features and $y_i \in \{-1, 1\}$ being the labels.

Recall that the 0-1 loss function is defined as

$$\text{loss}_{0/1}(\hat{y}, y) := \begin{cases} 0 & \text{if } \text{sign}(\hat{y}) = y \\ 1 & \text{otherwise} \end{cases}$$

For any d-dimensional weight vector $\mathbf{w}$, the resulting classifier we will consider is

$$h_{\mathbf{w}}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{w}^\top \mathbf{x} > 0 \\ -1 & \text{otherwise} \end{cases}.$$

We then define the population (distribution) *risk* of $\mathbf{w}$ as

$$R_{\mathcal{D}}[\mathbf{w}] = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\text{loss}_{0/1}(h_{\mathbf{w}}(\mathbf{x}), y)]\,,$$

and the *empirical (sample) risk* of $\mathbf{w}$ as

$$R_S[\mathbf{w}] = \frac{1}{n}\sum_{i=1}^{n} \text{loss}_{0/1}(h_{\mathbf{w}}(\mathbf{x}_i), y_i)\,.$$

(a) (4 points) **Prove that $\mathbb{E}_S[R_S[\mathbf{w}]] = R_{\mathcal{D}}[\mathbf{w}]$ for a fixed vector $\mathbf{w} \in \mathbb{R}^d$.**
Solution:

$$\mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^{n} \text{loss}_{0/1}(h_w(x_i), y_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_S\left[\text{loss}_{0/1}(h_w(x_i), y_i)\right] = \frac{1}{n}\sum_{i=1}^{n} R_{\mathcal{D}}[w] = R_{\mathcal{D}}[w]$$

(b) (4 points) **Prove that $\text{Var}(R_S[\mathbf{w}]) \leq \frac{1}{n}$ for a fixed vector $\mathbf{w} \in \mathbb{R}^d$.**
Solution:

$$
\begin{aligned}
\text{Var}(R_S[w]) &= \mathbb{E}_S\left[(R_S[w] - R_{\mathcal{D}}[w])^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}_S\left[(\text{loss}(h_w(x_i), y_i) - R_{\mathcal{D}}[w])(\text{loss}(h_w(x_j), y_j) - R_{\mathcal{D}}[w])\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}_S\left[(\text{loss}(h_w(x_i), y_i) - R_{\mathcal{D}}[w])^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[(\text{loss}(h_w(x_i), y) - R_{\mathcal{D}}[w])^2\right] \\
&\leq \frac{1}{n}
\end{aligned}
$$

Here, the first line is the definition of variance, the second line expands the square, the third line follows because (x_i, y_i) and (x_j, y_j) are independent. The fourth line follows because the (x_i, y_i) are identically distributed. The last line follows because the 0-1 loss loss is nonnegative and bounded above by 1.

(c) (4 points) Now let $\widehat{\mathbf{w}}_S$ be any minimizer of the empirical risk $\frac{1}{n}\sum_{(\mathbf{x}_i,y_i)\in S}\text{loss}_{0/1}(h_{\mathbf{w}}(\mathbf{x}_i),y_i)$. Suppose that a dataset S of n samples drawn i.i.d. from the distribution $\mathcal{D}$ is linearly separable with probability 1. Show that in this setting,

$$\mathbb{E}_S[R_S[\widehat{\mathbf{w}}_S]] \leq \mathbb{E}_S[R_{\mathcal{D}}[\widehat{\mathbf{w}}_S]] \,.$$

Solution: $\mathbb{E}_S[R_S[\hat{w}_S]] = 0$ and $\mathbb{E}_S[R_{\mathcal{D}}[\hat{w}_S]] \geq 0$.

(d) (4 points) Give an example of a distribution $\mathcal{D}$ and a sample from that distribution such that the sample risk is **strictly** less than the expected population risk (i.e. $R_S[\widehat{\mathbf{w}}_S] < \mathbb{E}_S[R_{\mathcal{D}}[\widehat{\mathbf{w}}_S]]$ for some minimizing $\widehat{\mathbf{w}}_S$). You may define the setting mathematically or you may draw a picture, but make sure it is clear what you mean.

Solution: Any picture where $\hat{w}_S$ separates the sample points but not everything from the distribution.

6 PCA and Least Squares (3 parts, 12 points)

Consider the ridge regression estimator,

$$\widehat{\mathbf{w}}_{\text{ridge}} := \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2,$$

where $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{y} \in \mathbb{R}^n$. Suppose that $\mathbf{X}$ has singular value decomposition $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \sum_{i=1}^d \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, where $\mathbf{U} \in \mathbb{R}^{n\times d}$, $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$, and $\mathbf{V} \in \mathbb{R}^{d\times d}$, and where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0$ are the diagonal elements of $\mathbf{\Sigma}$. Recall that

$$\widehat{\mathbf{w}}_{\text{ridge}} = \sum_{i=1}^d \rho_\lambda(\sigma_i)\mathbf{u}_i\mathbf{v}_i^\top\mathbf{y}, \text{ where } \rho_\lambda(\sigma) := \frac{\sigma}{\lambda+\sigma^2}.$$

We previously found a similar expression for $\widehat{\mathbf{w}}_{\text{PCA}} := \arg\min_{\mathbf{w}} \|\mathbf{X}\mathbf{V}_k\mathbf{w} - \mathbf{y}\|_2^2$, where $\mathbf{V}_k = [\mathbf{v}_1|\mathbf{v}_2|\ldots|\mathbf{v}_k]$. In this problem, we will consider a slightly different way of performing PCA, using the matrix $\mathbf{P}_k := \mathbf{V}_k\mathbf{V}_k^\top$. Throughout, you may assume $\sigma_k > \sigma_{k+1}$.

(a) (4 points) **Show that you can express $\mathbf{X}\mathbf{P}_k = \mathbf{U}\mathbf{\Sigma}_k\mathbf{V}^\top$, where $\mathbf{\Sigma}_k \in \mathbb{R}^{d\times d}$ is a diagonal matrix for you to define. What is $\mathbf{X}\mathbf{P}_k$ if $k \geq \text{rank}(\mathbf{X})$?**

Solution:

$$\begin{aligned}
\mathbf{X}\mathbf{P}_k &= \left(\sum_{i=1}^d \sigma_i\mathbf{u}_i\mathbf{v}_i^\top\right)\sum_{j=1}^k \mathbf{v}_j\mathbf{v}_j^\top \\
&= \sum_{i=1}^d\sum_{j=1}^k \sigma_i\mathbf{u}_i\mathbf{v}_j^\top\langle\mathbf{v}_i,\mathbf{v}_j\rangle \\
&= \sum_{i=1}^d\sum_{j=1}^k \sigma_i\mathbf{u}_i\mathbf{v}_j^\top\mathbf{I}(i=j) \\
&= \sum_{i}^k \sigma_i\mathbf{u}_i\mathbf{v}_i^\top,
\end{aligned}$$

so we see that $\mathbf{\Sigma}_k$ is the matrix with $\sigma_1, \ldots, \sigma_k$ on the first k diagonals, and 0 on the remaining ones. If $k \geq \text{rank}(x)$, thne $\mathbf{\Sigma}_k = \mathbf{\Sigma}$ and so $\mathbf{X}\mathbf{P}_k = \mathbf{X}$.

(b) (4 points) Consider the ridge-PCA estimator

$$\widehat{\mathbf{w}}_{\mathrm{RP}} := \arg\min_{\mathbf{w}} \|\mathbf{X}\mathbf{P}_k\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2.$$

Show that the predictor $\mathbf{X}\widehat{\mathbf{w}}_{\mathrm{RP}}$ can be written as

$$\mathbf{X}\widehat{\mathbf{w}}_{\mathrm{RP}} = \sum_{i=1}^{d} \rho_{k,\lambda}(\sigma_i)\mathbf{u}_i\mathbf{u}_i^\top\mathbf{y},$$

where $\rho_{k,\lambda}$ is a function for you to define.

Hint: The instructions ask you to find a formula for the predictor $\mathbf{X}\widehat{\mathbf{w}}_{\mathrm{RP}}$*, not for the parameter* $\widehat{\mathbf{w}}$.

Solution: Let $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{P}_k$. Then,

$$\mathbf{X}\widehat{\mathbf{w}}_{\mathrm{RP}} = \mathbf{X}(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} + \lambda I)^{-1}\widetilde{\mathbf{X}}^\top\mathbf{y}.$$

By the previous problem, we can write $\widetilde{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}_k\mathbf{V}^\top$, and thus

$$\begin{aligned}
\mathbf{X}(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} + \lambda I)^{-1}\widetilde{\mathbf{X}}^\top &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top(\mathbf{V}\mathbf{\Sigma}_k\mathbf{U}^\top\mathbf{U}\mathbf{\Sigma}_k\mathbf{V}^\top + \lambda I)^{-1}\mathbf{V}\mathbf{\Sigma}_k\mathbf{U}^\top \\
&= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top(\mathbf{V}\mathbf{\Sigma}_k\mathbf{\Sigma}_k\mathbf{V}^\top + \lambda I)^{-1}\mathbf{V}\mathbf{\Sigma}_k\mathbf{U}^\top \\
&= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top(\mathbf{V}(\mathbf{\Sigma}_k\mathbf{\Sigma}_k + \lambda I)\mathbf{V})^{-1}\mathbf{V}\mathbf{\Sigma}_k\mathbf{U}^\top \\
&= \mathbf{U}\mathbf{\Sigma}(\mathbf{\Sigma}_k\mathbf{\Sigma}_k + \lambda I)^{-1}\mathbf{\Sigma}_k\mathbf{U}^\top.
\end{aligned}$$

We see that Σ is diagonal with σ_i on the diagonals, $(\mathbf{\Sigma}_k\mathbf{\Sigma}_k+\lambda I)^{-1}$ is diagonal with $\frac{1}{\sigma_i^2\mathbf{I}(\sigma_i\geq\sigma_k)+\lambda}$ on the diagonals, and $\mathbf{\Sigma}_k$ has $\sigma_i\mathbf{I}(\sigma_i \geq \sigma_k)$ on the diagonals. Thus, $\Sigma(\mathbf{\Sigma}_k\mathbf{\Sigma}_k + \lambda I)^{-1}\mathbf{\Sigma}_k$ has diagonals $\frac{\sigma_i^2\mathbf{I}(\sigma_i\geq\sigma_k)}{\lambda+\sigma_i^2\mathbf{I}(\sigma_i\geq\sigma_k} = \frac{\sigma_i^2}{\lambda+\sigma_i^2}\mathbf{I}(\sigma_i \geq \sigma_k) := \rho_{k,\lambda}(\sigma_i)$ on the diagonals. Concluding, we have

$$\begin{aligned}
\mathbf{X}(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} + \lambda I)^{-1}\widetilde{\mathbf{X}}^\top\mathbf{y} &= \mathbf{U}\mathrm{Diag}(\rho_{k,\lambda}(\sigma_i))\mathbf{U}^\top\mathbf{y} \\
&= \sum_{i=1}^{d} \rho_{k,\lambda}(\sigma_i)\mathbf{u}_i\mathbf{u}_i^\top\mathbf{y}.
\end{aligned}$$

(c) (4 points) Suppose that $\mathbf{y} = \mathbf{X}\mathbf{w}_*$, where $\|\mathbf{w}_*\|_2 = 1$. Consider the unregularized ridge-PCA estimator with $\lambda = 0$ and $k \leq \text{rank}(\mathbf{X})$. **Show that the prediction error satisfies**

$$\|\mathbf{X}\widehat{\mathbf{w}}_{\text{RP}} - \mathbf{y}\|_2^2 \leq \sigma_{k+1}^2.$$

Hint: You might want to use the inequality that for any real numbers $\alpha_{k+1}, \ldots, \alpha_n$ and any $\sigma_{k+1} \geq \sigma_{k+2} \geq \cdots \geq \sigma_d \geq 0$, $\sum_{j=k+1}^d \sigma_j^2 \alpha_j^2 \leq \sigma_{k+1}^2 \sum_{j=k+1}^d \alpha_j^2$. **Solution:** Define $\alpha_i = \langle \mathbf{v}_i, \mathbf{y} \rangle$ to denote the projection of y onto the i-th singular vector. If $\mathbf{y} = \mathbf{X}\mathbf{w}_*$, then $\mathbf{y} = \sum_{i=1}^d \mathbf{u}_i \sigma_i \langle \mathbf{v}_i, \mathbf{w}_* \rangle = \sum_{i=1}^d \alpha_i \sigma_i \mathbf{u}_i$. Moreover, we have

$$\begin{aligned}
\mathbf{X}\widehat{\mathbf{w}}_{\text{RP}} &= \sum_{i=1}^d \rho_{k,0}(\sigma_i) \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} \\
&= \sum_{i=1}^d \rho_{k,0}(\sigma_i) \mathbf{u}_i \mathbf{u}_i^\top (\sum_{j=1}^d \alpha_i \sigma_i \mathbf{u}_j) \\
&= \sum_{i=1}^d \alpha_i \sigma_i \rho_{k,0}(\sigma_i) \mathbf{u}_i,
\end{aligned}$$

since $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = \mathbf{I}(i = j)$. Hence,

$$\begin{aligned}
\|\mathbf{X}\widehat{\mathbf{w}}_{\text{RP}} - \mathbf{y}\|_2^2 &= \| \sum_{i=1}^d \alpha_i \sigma_i \left(\rho_{k,0}(\sigma_i) - 1 \right) \mathbf{u}_i \|_2^2 \\
&= \| \sum_{i=1}^d \alpha_i \sigma_i \left(\mathbf{I}(\sigma_i \geq \sigma_k) - 1 \right) \mathbf{u}_i \|_2^2. \\
&= \| \sum_{i=k+1}^d \alpha_i \sigma_i \mathbf{u}_i \|_2^2.
\end{aligned}$$

Since $\mathbf{u}_1, \ldots, \mathbf{u}_d$ are an orthogonal , we have that

$$\| \sum_{i=k+1}^d \alpha_i \sigma_i \mathbf{u}_i \|_2^2 = \sum_{i=1}^d \|\mathbf{u}_i\|_2^2 (\alpha_i \sigma_i^2) = \sum_{i=k+1}^d \alpha_i^2 \sigma_i^2 \leq \sigma_{k+1}^2 \sum_{i=k+1}^d \alpha_i^2.$$

To conclude, we show that $\sum_{i=k+1}^d \alpha_i^2 = \sum_{i=k+1}^d \langle \mathbf{v}_i, \mathbf{w}_* \rangle^2 \|\mathbf{w}_*\|_2^2 = 1$, where we use that $\mathbf{v}_i$ are orthonormal.

7 Stability and Generalization (3 parts, 12 points; +3pt. extra credit part)

Let $S = \{x_1, \ldots, x_n\}$ be a dataset where each sample $x_i \in \mathbb{R}$ is a scalar value drawn independently from the same distribution $\mathcal{D}$. Furthermore, assume that the distribution $\mathcal{D}$ satisfies $\mathbb{E}_{x\sim\mathcal{D}}[x] = 0$ and $\mathbb{E}_{x\sim\mathcal{D}}[x^2] = \sigma^2$. In this problem, we will see that the sample mean $\frac{1}{n}\sum_{i=1}^n x_i$ is a stable estimator of the population mean under reasonable assumptions.

Assume we are trying to estimate some property of the distribution $\mathcal{D}$. In this setting, recall that the *risk* of an estimator w under a given loss function is given by

$$R_{\mathcal{D}}[w] = \mathbb{E}_{x\sim\mathcal{D}}[\text{loss}(w, x)]\,.$$

Furthermore, the *empirical risk* of an estimator w under a given loss function is given by

$$R_S[w] = \frac{1}{n}\sum_{i=1}^n \text{loss}(w, x_i)\,.$$

(a) (4 points) Consider the square loss defined as $\text{loss}(w, x) = \frac{1}{2}(w - x)^2$. Under the square loss function, we will minimize the empirical risk to get the estimator $\widehat{w}_S = \arg\min_{w\in\mathbb{R}} R_S[w]$. **Show that $\widehat{w}_S$ is equal to the sample mean.**

Solution: Since the objective is convex, we can minimize it by setting the gradient to zero. Thus,

$$\nabla_w f = \frac{1}{n}\sum_{i=1}^n (w - x_i) = 0$$
$$\implies w = \frac{1}{n}\sum_{i=1}^n x_i\,.$$

(b) (4 points) Define the *generalization gap* of an estimator w by

$$\epsilon_{\text{gen}}(w, S) := R_{\mathcal{D}}[w] - R_S[w] \, .$$

Recall that the *average stability* of an estimator $\widetilde{w}_S$ (an estimator which, in general, depends on the sample S) can be computed as the average of the generalization gap over the randomness in the sample, i.e. it is equal to

$$\mathbb{E}_S[\epsilon_{\text{gen}}(\widetilde{w}_S, S)] \, .$$

Now, assume that $\widetilde{w}_S$ is a symmetric function of the sample S—that is, re-arranging the ordering of the samples in S does not affect the value of $\widetilde{w}_S$. **Show that the average stability for $\widetilde{w}_S$ (under the given loss function and sample distribution) is given by $\mathbb{E}_S[\widetilde{w}_S x_1]$.**

Solution: Expanding the term for average stability, we see that

$$\begin{aligned}
\mathbb{E}_S[\epsilon_{\text{gen}}(\widetilde{w}_S, S)] &= \mathbb{E}_{S,x}[\frac{1}{2}(\widetilde{w}_S - x)^2] - \mathbb{E}_S[\frac{1}{2n}\sum_{i=1}^{n}(\widetilde{w}_S - x_i)^2] \\
&= \frac{1}{2}\mathbb{E}_{S,x}[\widetilde{w}_S^2 - 2\widetilde{w}_S x + x^2] - \frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}_S[\widetilde{w}_S^2 - 2\widetilde{w}_S x_i + x_i^2] \\
&= \frac{1}{2}\mathbb{E}_S[\widetilde{w}_S^2] - \mathbb{E}_S[\widetilde{w}_S]\mathbb{E}_x[x] + \frac{1}{2}\mathbb{E}_x[x^2] - \frac{1}{2}\mathbb{E}_S[\widetilde{w}_S^2] + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_S[\widetilde{w}_S x_i] - \frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}_S[x_i^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_S[\widetilde{w}_S x_i] \, ,
\end{aligned}$$

as most of the terms cancel. Then, by symmetry, this is equal to $\mathbb{E}_S[\widetilde{w}_S x_1]$.

(c) (4 points) **Calculate the average stability $\mathbb{E}_S[\epsilon_{\text{gen}}(\widehat{w}_S, S)]$ for the sample mean $\widehat{w}_S$, under the square loss function and sample distribution $\mathcal{D}$.**

Solution: Since the sample mean is symmetric, we can use the previous part. Thus,

$$\mathbb{E}_S[\widetilde{w}_S x_1] = \mathbb{E}_S[\frac{1}{n}\sum_{i=1}^{n} x_i x_1] = \frac{\sigma^2}{n}\,.$$

(d) (*extra credit, 3 points*) Recall from lecture that the average stability of a generic estimator $\widetilde{w}_S$ under a given loss ℓ is often difficult to compute, and can instead be upper bounded by the *uniform stability*, which, for a symmetric estimator, is given by

$$\max_{S,x_1'} |\ell(\widetilde{w}_S, x_1) - \ell(\widetilde{w}_{S'}, x_1)| \,,$$

where the maxima is taken over all possible samples. Here, we introduce an auxilliary set $S' = \{x_1', x_2, x_3, ..., x_n\}$ where the first sample point from S is replaced by an independent and identically drawn sample point x_1' from $\mathcal{D}$, and the rest are taken from S. This quantity represents how much the loss function can change if we perturb the estimator at one sample point (because we assume the estimator is symmetric in its arguments, it is without loss of generality to consider only perturbing the first sample point).

Now, let us make the following two assumptions on both the loss function ℓ and the distribution $\mathcal{D}$:

(i) For a fixed second argument, the loss function ℓ is L-Lipschitz in its first argument. Recall that a function $\phi(u) : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz if for all $u,\ v \in \mathbb{R}$, the following inequality holds:

$$|\phi(u) - \phi(v)| \leq L|u - v| \,.$$

(ii) For any x drawn from $\mathcal{D}$, we have that $|x| \leq b$ holds.

Derive an upper bound on the uniform stability of the sample mean. This bound should hold for all loss functions ℓ and distributions $\mathcal{D}$ satisfying assumptions (i) and (ii) above, and should decay to zero as $n \to \infty$.

Solution: Since the loss is $L-$Lipschitz, we can upper bound the uniform stability by

$$\begin{aligned} L \max_{S,x_1'} |\widehat{w}_S - \widehat{w}_{S'}| &= \frac{L}{n} \max_{S,x_1'} |x_1 - x_1'| \\ &= \frac{L}{n} \max_{x_1,x_1'} |x_1 - x_1'| \\ &\leq \frac{2Lb}{n} \,. \end{aligned}$$

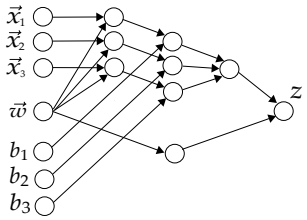
8 Multiple Choice (5 parts, 10 points)

For the following multiple choice questions, select all that apply.

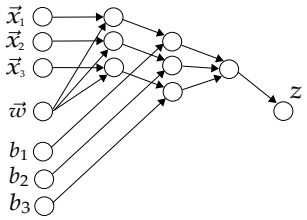
(a) (2 points) **Which of the following computation graphs describes the function**

$$z = \sum_{i=1}^{n} \phi(\mathbf{x}_i^\top \mathbf{w} + b_i) + \|\mathbf{w}\|_2^2$$

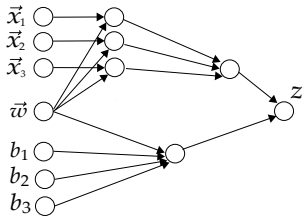
where $z \in \mathbb{R}$, $\mathbf{w}, \mathbf{x}_i \in \mathbb{R}^d$, and $b_i \in \mathbb{R}$?



(A)



(B)



(B)

(a) ◯ computation graph (A)

(b) ◯ computation graph (B)

(c) ◯ computation graph (C)

Solution: (A)

(b) (2 points) **Select all the parts that are always true.** A *saddle point*:

(a) Is a stationary point.

(b) Must always exist for a differentiable function.

(c) Cannot be a local minima.

(d) Never exists for a convex function.

Solution: (a), (c), and (d) are true.

(c) (2 points) Consider a distribution $\mathcal{D}$ on labelled pairs $(x, y) \in \{\pm 1\} \times \{\pm 1\}$ where the conditional distribution of y given x is:

$$y = \begin{cases} x & \text{with probability } \alpha \\ -x & \text{with probability } 1 - \alpha \end{cases}.$$

The $0/1$ population risk of a classifier the form $\text{sign}(w \cdot x)$, $w \in \mathbb{R}$ is defined as:

$$R_{\mathcal{D}}[w] = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{I}(\text{sign}(w \cdot x) \neq y)\right].$$

What is $\min_{w\in\mathbb{R}} R_{\mathcal{D}}[w]$**?**

(a) $1/2$.

(b) $\max\{\alpha, 1 - \alpha\}$.

(c) $\min\{\alpha, 1 - \alpha\}$.

(d) Depends on the marginal distribution of x.

Solution: (c).

(d) (2 points) Using a fully connected neural net with one hidden layer and ReLU activations, **what is the minimum number of hidden nodes needed to achieve perfect classification, with respect to the hinge loss, of the data shown in Figure 3** (the plus signs denote data points labelled $+1$ while the minus signs denote data points labelled -1)? Mathematically, the neural network we are considering has the form

$$f(\mathbf{x}) = \beta + \sum_{j=1}^{N} \alpha_j \max\{\langle \mathbf{w}_j, \mathbf{x}\rangle + b_j, 0\},$$

where β, α_j, $\mathbf{w}_j$, and b_j are the parameters of the network.

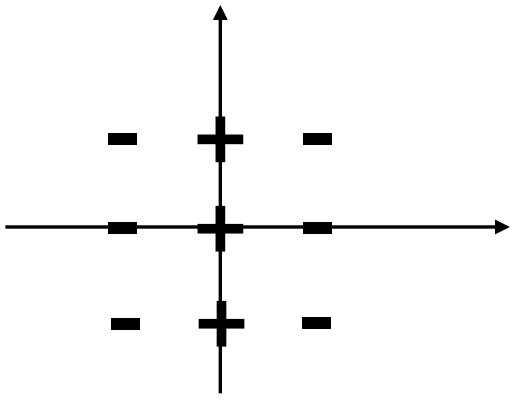


Figure 3: Data set for part (d).

(a) $N = 1$

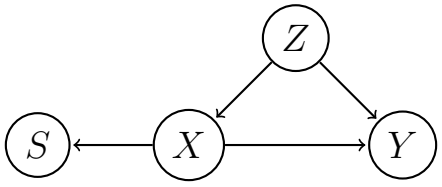
(b) $N = 2$

(c) $N = 3$

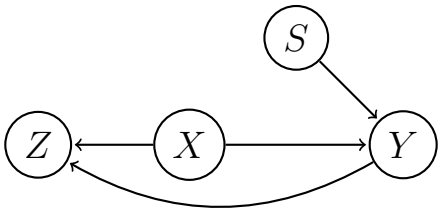
(d) $N = 4$

Solution: (b) $N = 2$. Choose $\beta = 1$. If we assume that the scale of the picture is so that the points have coordinates in $\{-1, 0, 1\}$ then we can choose $\mathbf{w}_1 = [-1, 0]$ and $\mathbf{w}_2 = [1, 0]$ and $b_1 = b_2 = -1/2$. Finally, $\alpha_1 = \alpha_2 = -2$.

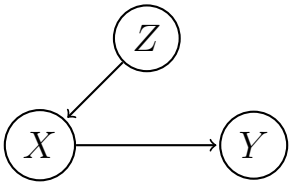
(e) (2 points) When trying to determine the causal effect of variable X on variable Y, we must account for confounders. **In which of the following structural equation models (represented by their respective causal DAGs) are variables X and Y confounded by variable Z?**



(a)



(b)



(c)

Solution: (a). In (a), Z has directed edges going into both X and Y. One can check by definition that there is in fact no confounding in (b) and (c).