UNIVERSITY OF ALBERTA
CMPUT 267 Fall 2024

Final Exam

Do Not Distribute

Duration: 3 Hours

Last Name: ______________________________

First Name: ______________________________

Carefully read all of the instructions and questions. Good luck!

---

1. **Do not turn this page** until instructed to begin.
2. Verify that your exam package includes 18 pages, along with a formula sheet and a blank page at the end.
3. **Only the scantron will be marked**. All of your answers must be clearly marked on the scantron.
4. Use **pencil only** to fill out the scantron (preferably an HB or #2 pencil).
5. **Erase mistakes completely** on the scantron to avoid misreading by the scanner.
6. **Mark answers firmly and darkly**, filling in the bubbles completely.
7. This exam consists of **40 questions**. Each question is worth **1 mark**. The exam is worth a total of **40 marks**.
8. Some questions may have **multiple correct answers**. To receive **full marks**, you must select **all correct answers**. If you select only **some** of the correct answers, you will receive **partial marks**. Selecting an incorrect option will cancel out a correct one. For example, if you select two answers—one correct and one incorrect—you will receive zero points for that question. If the number of incorrect answers exceeds the correct ones, your score for that question will be zero. **No negative marks** will be given.

C **Question 1.** [1 MARK]

Let $g(x, y) = x + y^2$ where $x, y \in \mathbb{R}$. What is

$$\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} g(x, y), = \sum_{x=\{1,2\}}\sum_{y=\{1,2,3\}} x+y^2 = (1+1)+(1+4)+(1+9)+(2+1)+(2+4)+(2+9)$$

$$= 2+5+10+3+6+11$$

$$= 17+20$$

$$= 37$$

where $\mathcal{Y} = \{1, 2, 3\}$ and $\mathcal{X} = \{1, 2\}$?

A. 17

B. 27

$\boxed{\text{C.}}$ 37

D. 11

BC **Question 2.** [1 MARK]

Suppose you roll three fair six-sided dice. Let the dice be represented by the random variables $X_1, X_2, X_3 \in \mathcal{X}$ where $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$. Which of the following sets represents the outcome space of the random variable $X = (X_1, X_2, X_3)$?

A. $\mathcal{X}$ X

$\boxed{\text{B}}$ $\mathcal{X}^3$

$\boxed{\text{C.}}$ $\{(x_1, x_2, x_3) \mid x_1, x_2, x_3 \in \mathcal{X}\}$

D. $\{(x, x, x) \mid x \in \mathcal{X}\}$

$X = \{1,2,3,4,5,6\} \Rightarrow X^2 = \{(1,1),(1,2),\cdots,(1,6),\cdots,(2,1),\cdots,(2,6),\cdots,(6,6)\}$

$X^3 = \{(X_1,X_2,X_3) \mid X_1,X_2,X_3 \in X\}$

B **Question 3.** [1 MARK]

When we did logistic regression, we minimized the estimated loss

$$\hat{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\sigma(\mathbf{x}_i^\top\mathbf{w}), y_i) \quad \text{where} \quad \ell(\sigma(\mathbf{x}_i^\top\mathbf{w}), y_i) = -y_i \ln(\sigma(\mathbf{x}_i^\top\mathbf{w})) - (1 - y_i)\ln(1 - \sigma(\mathbf{x}_i^\top\mathbf{w})).$$

Imagine we decided some samples are more important to get right than other samples. To do this we introduce a scalar importance-weight $a_i > 0$ on each sample, and get the following weighted estimated loss

$$\hat{L}_{\text{weight}}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} a_i \ell(\sigma(\mathbf{x}_i^\top\mathbf{w}), y_i).$$

If $a_i > a_k$ then that means we care more about reducing the loss on sample $(\mathbf{x}_i, y_i)$ than on $(\mathbf{x}_k, y_k)$. Recall that the gradient of $\ell(\sigma(\mathbf{x}_i^\top\mathbf{w}), y_i)$ with respect to $\mathbf{w}$ is $\nabla\ell(\mathbf{w}) = (\sigma(\mathbf{x}_i^\top\mathbf{w}) - y_i)\mathbf{x}_i$. What is the gradient of $\hat{L}_{\text{weight}}(\mathbf{w})$?

A. $\nabla\hat{L}_{\text{weight}}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\sigma(\mathbf{x}_i^\top\mathbf{w}) - y_i)\mathbf{x}_i$

$\boxed{\text{B.}}$ $\nabla\hat{L}_{\text{weight}}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} a_i(\sigma(\mathbf{x}_i^\top\mathbf{w}) - y_i)\mathbf{x}_i$

C. $\nabla\hat{L}_{\text{weight}}(\mathbf{w}) = \sum_{i=1}^{n} a_i(\sigma(\mathbf{x}_i^\top\mathbf{w}) - y_i)$

D. $\nabla\hat{L}_{\text{weight}}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\sigma(\mathbf{x}_i^\top\mathbf{w}) - y_i)a_i$

$$\nabla\hat{L}_{weight}(\vec{w}) = \frac{\partial}{\partial\vec{w}}\frac{1}{n}\sum_{i=1}^{n} a_i \ell(\sigma(\vec{x}_i\vec{w}), y_i)$$

$$= \frac{1}{n}\frac{\partial}{\partial\vec{w}}\sum_{i=1}^{n} a_i \ell(\sigma(\vec{x}_i\vec{w}), y_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} a_i \frac{\partial}{\partial\vec{w}}\ell(\sigma(\vec{x}_i\vec{w}), y_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} a_i(\sigma(\vec{x}^T_i\vec{w}) - y_i)\vec{x}_i$$

D

Question 4. [1 MARK]

Suppose you have a random variable X representing the time (in minutes) it takes for the bus to arrive at the bus stop. You know X is distributed according to the continuous uniform distribution over the interval $\mathcal{X} = [1, 10]$. Let p be the pdf of X. Which of the following statements are true?

$X \sim U(1,10) \Rightarrow p(X) = \frac{1}{9}$ $\quad E(X) = \int_1^{10} x \cdot \frac{1}{9}\, dx = \frac{1}{18}[x^2]_1^{10} = \frac{99}{18} = \frac{33}{6} = \frac{11}{2} = 5.5$

A. The expected value of X is ~~5~~.5 5.5

B. The probability that $X = 10$ is ~~1/9~~. 0 X

C. $p(10) = \cancel{1/10}$ 1/9

$\boxed{\text{D.}}$ The probability that X is between 4 and 10 is 2/3. $P(4 \le X \le 10) = \int_4^{10} \frac{dx}{9} = \frac{1}{9} \cdot 6 = \frac{2}{3}$

B

Question 5. [1 MARK]

Let X and N be two random variables where $X \in \{0, 1\}$ and $N \in \{1, 2, 3\}$. Their joint probability mass function is:

$$p_{X,N}(x, n) = \begin{cases} \frac{4}{25}, & (x, n) = (1, 1), \\ \frac{8}{25}, & (x, n) \in \{(1, 2), (1, 3)\}, \\ \frac{1}{25}, & (x, n) = (0, 1), \\ \frac{2}{25}, & (x, n) \in \{(0, 2), (0, 3)\}, \\ 0, & \text{otherwise.} \end{cases}$$

$p(n|X=1) = \frac{p(n, X=1)}{p(X=1)} = \frac{p(n,X=1)}{4/5} = \frac{5}{4} p(n, X=1)$

$P(X=1) = \frac{4}{25} + \frac{16}{25} = \frac{4}{5}$

$P(1|X=1) = \frac{5}{4} \cdot \frac{4}{25} = \frac{1}{5}$ $\quad P(2|X=1) = \frac{5}{4} \cdot \frac{8}{25} = \frac{2}{5}$

$P(3|X=1) = \frac{2}{5}$

What is $\mathbb{E}[N \mid X = 1]$?

$E[N|X=1] = \sum n \cdot P(n|X=1) = 1 \cdot \frac{1}{5} + \frac{2}{5}(2+3) = 2.2$

A. 2.0

$\boxed{\text{B.}}$ 2.2

C. 1.8

D. 2.4

ACD

Question 6. [1 MARK]

Suppose Z_1, Z_2, Z_3, Z_4 are independent random variables, each with $Z_i \sim \mathcal{N}(8, 9)$. Let $\bar{Z} = \frac{1}{4}(Z_1 + Z_2 + Z_3 + Z_4)$. Which of the following statements are true?

$Var(\bar{z}) = Var\left[\frac{1}{4}(z_1+z_2+z_3+z_4)\right] = \frac{1}{16}[Var(z_1)+\cdots+Var(z_4)] = \frac{1}{16} \cdot 36 = \frac{9}{4}$

$E(\bar{z}) = \frac{1}{4}[E(z_1)+\cdots+E(z_4)] = \frac{1}{4} \cdot 4 \cdot 8 = 8$

$\boxed{\text{A.}}$ The variance $\text{Var}(\bar{Z}) = 9/4$.

B. The expected value $\mathbb{E}[\bar{Z}] = \cancel{2}.8$ 8

$\boxed{\text{C.}}$ The variance $\text{Var}(Z_1) = 9$.

$\boxed{\text{D.}}$ The expected value $\mathbb{E}[\bar{Z}] = 8$.

A

Question 7. [1 MARK]

Suppose you have two discrete random variables $A \in \{1,2,3\}$ and $B \in \{0,1\}$. The joint probability mass function (pmf) of A and B is given by the following values:

$$p(1,0) = \frac{1}{10}, \qquad p(2,0) = \frac{1}{5}, \qquad p(3,0) = \frac{1}{10},$$
$$p(1,1) = \frac{1}{5}, \qquad p(2,1) = \frac{1}{10}, \qquad p(3,1) = \frac{3}{10}.$$

Which of the following are the marginal pmf of A?

$p_A(1) = p(1,0) + p(1,1) = \frac{3}{10}$

$p_A(2) = p(2,0) + p(2,1) = \frac{3}{10}$

$p_A(3) = p(3,0) + p(3,1) = \frac{2}{5}$

A. $p_A(1) = \frac{3}{10}, \quad p_A(2) = \frac{3}{10}, \quad p_A(3) = \frac{2}{5}$

B. $p_A(1) = \frac{1}{2}, \quad p_A(2) = \frac{2}{5}, \quad p_A(3) = \frac{1}{10}$

C. $p_A(1) = \frac{2}{10}, \quad p_A(2) = \frac{3}{10}, \quad p_A(3) = \frac{5}{10}$

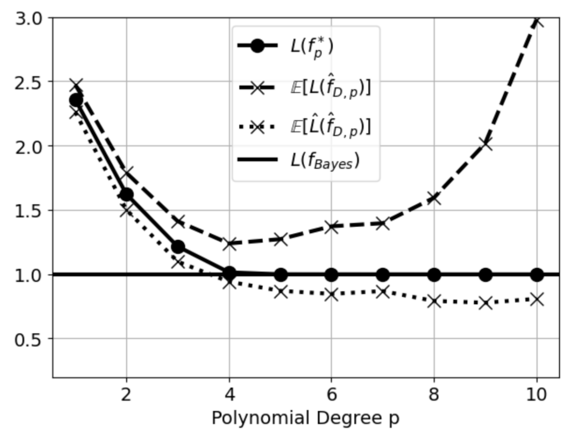
D. $p_A(1) = \frac{2}{5}, \quad p_A(2) = \frac{2}{5}, \quad p_A(3) = \frac{1}{5}$

ABCD

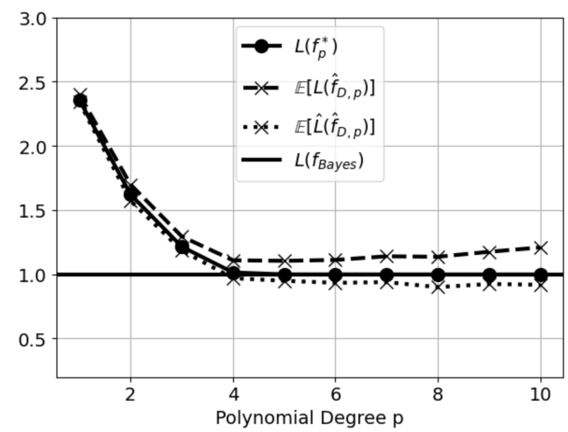
Question 8. [1 MARK]

You have access to the true feature-label distribution $\mathbb{P}_{\mathbf{X},Y}$. You are interested in studying the estimation, approximation, and irreducible errors as a function of polynomial degree p and dataset size n. To do this, you plot the following figures. Note that $L(f_p^*), L(f_{\text{Bayes}})$ are identical in both Fig 1 and Fig 2. Which of the following statements are true?

$EE = \mathbb{E}[L(\hat{f}_{D,p})] - L(f_p^*) \quad IE = L(f_{Bayes})$



(a) Fig 1



(b) Fig 2

A. In Fig 1 the predictor $\hat{f}_{D,p}$ is underfitting for $p = 1$ and overfitting for $p = 10$.

B. In Fig 1 the estimation error is smaller for $p = 2$ than for $p = 10$.

C. It is impossible to make the irreducible error smaller by changing n or p.

D. In Fig 2 the estimation error is smaller than in Fig 1 for all values of p.

B

Question 9. [1 MARK]

Consider a linear predictor that estimates an employee's salary (in tens of thousands of dollars) based on their years of experience $f(x) = 5 + 0.5x$. You have the following dataset of four observations:

$$\mathcal{D} = ((x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)) = ((1, 5.5),\ (2, 5.0),\ (3, 6.5),\ (4, 7.2)).$$

You decide to use the absolute loss $\ell(f(x), y) = |f(x) - y|$. What is the estimated loss:

$$\hat{L}(f) = \frac{1}{4}\sum_{i=1}^{4} \ell(f(x_i), y_i). = \frac{1}{4}\left(|5.5-5.5| + |6-5| + |6.5-6.5| + |7-7.2|\right)$$

$$= \frac{1}{4}(1 + 0.2)$$

$$= 0.3$$

A. 0.25

$\boxed{\text{B.}}$ 0.3

C. 0.125

D. 0.5

B

Question 10. [1 MARK]

Consider the convex function

$$g(w_1, w_2) = w_1^2 + w_2^2 + w_1 w_2 + 10w_1, \Rightarrow g_{w_1} = 2w_1 + w_2 + 10 = 0 \quad g_{w_2} = 2w_2 + w_1 = 0$$

where $w_1, w_2 \in \mathbb{R}$. Find $(w_1^*, w_2^*) = \arg\min_{w_1, w_2 \in \mathbb{R}} g(w_1, w_2)$ and $\min_{w_1, w_2 \in \mathbb{R}} g(w_1, w_2)$.

$$\Rightarrow -4w_2 + w_2 + 10 = 0$$

$$w_2 = \frac{10}{3} \quad w_1 = -\frac{20}{3}$$

$$\Rightarrow \min_{w_1, w_2 \in \mathbb{R}} g = \frac{400}{9} + \frac{100}{9} - \frac{200}{9} - \frac{200}{3}$$

$$= \frac{300}{9} - \frac{600}{9} = -\frac{300}{9} = -\frac{100}{3}$$

A. $(w_1^*, w_2^*) = (0, 0), \quad \min_{w_1, w_2 \in \mathbb{R}} g(w_1, w_2) = 0$

$\boxed{\text{B.}}$ $(w_1^*, w_2^*) = (-20/3, 10/3), \quad \min_{w_1, w_2 \in \mathbb{R}} g(w_1, w_2) = -100/3$

C. $(w_1^*, w_2^*) = (-5, 0), \quad \min_{w_1, w_2 \in \mathbb{R}} g(w_1, w_2) = -50$

D. $(w_1^*, w_2^*) = (-10, 10), \quad \min_{w_1, w_2 \in \mathbb{R}} g(w_1, w_2) = -200$

C

Question 11. [1 MARK]

Let everything be defined as in the previous question. You decide to use gradient descent to approximate the minimum of $g(w_1, w_2)$. You choose a constant step size of $\eta^{(t)} = 1$ for all iterations. You initialize the weights as $(w_1^{(0)}, w_2^{(0)}) = (0, 0)$. What are the weights $(w_1^{(2)}, w_2^{(2)})$ after two iterations of gradient descent?

$$w_1^{(1)} = w_1^{(0)} - \eta^{(t)} g'_{w_1}(\vec{w}^{(0)}) = 0 - 10 = -10$$

$$w_2^{(1)} = w_2^{(0)} - \eta^{(t)} g'_{w_2}(\vec{w}^{(0)}) = 0 - 0 = 0$$

$$w_2^{(1)} = w_1^{(1)} - \eta^{(t)} g'_{w_1}(\vec{w}^{(1)}) = -10 + 10 = 0$$

$$w_2^{(2)} = w_2^{(1)} - \eta^{(t)} g'_{w_2}(\vec{w}^{(1)}) = 0 + 10 = 10$$

A. $(w_1^{(2)}, w_2^{(2)}) = (0, 0)$

B. $(w_1^{(2)}, w_2^{(2)}) = (-10, 0)$

$\boxed{\text{C.}}$ $(w_1^{(2)}, w_2^{(2)}) = (0, 10)$

D. $(w_1^{(2)}, w_2^{(2)}) = (-20, 30)$

ACD **Question 12.** [1 MARK]

Suppose that $Y \sim \mathcal{N}(1, 2)$. Which of the following statements are true?

A. The expected value of Y is 1.

B. The probability that $Y = 10$ is $1/\cancel{\sqrt{2\pi}}$. 0 X

C. The variance of Y is 2.

D. Y is a continuous random variable.

AC **Question 13.** [1 MARK] $\hat{L}'_\lambda(w) = \frac{1}{n}\sum_{i=1}^{n} 4c_i(x_iw-y_i)^3x_i + \frac{4\lambda}{n}w^3 = \frac{4}{n}\left[\sum_{i=1}^{n} c_ix_i(x_iw-y_i)^3+\lambda w^3\right]$

$-\frac{1}{3}$

Consider the predictor $f(x) = xw$, where $w \in \mathbb{R}$ is a one-dimensional parameter, and x represents the feature with no bias term. Suppose you are given a dataset of n data points $\mathcal{D} = ((x_1, y_1), \ldots, (x_n, y_n))$, where each y_i is the target variable corresponding to feature x_i. You define a new regularized estimated loss as follows:

$$\hat{L}_\lambda(w) = \frac{1}{n}\sum\nolimits_{i=1}^{n} c_i(x_iw - y_i)^4 + \frac{\lambda}{n}w^4$$

where $\lambda \geq 0$ is the regularization parameter and $c_i \geq 0$ weights the importance of each data point. In this question, we are interested in finding $\hat{w} = \arg\min_{w\in\mathbb{R}} \hat{L}_\lambda(w)$ using first-order gradient descent. Which of the following statements are true?

A. If we pick a good step size $\eta^{(t)}$ then $w^{(t)}$ will get closer to $\hat{w}$ as t increases.

B. The first-order gradient update rule is

$$w^{(t+1)} = w^{(t)} - \frac{2}{n}\eta^{(t)}\left[\sum\nolimits_{i=1}^{n} c_ix_i(x_iw^{(t)} - y_i) + \lambda w^{(t)}\right]^{3}. \quad X$$

C. The first-order gradient update rule is

$$w^{(t+1)} = w^{(t)} - \frac{4}{n}\eta^{(t)}\left[\sum\nolimits_{i=1}^{n} c_ix_i(x_iw^{(t)} - y_i)^3 + \lambda(w^{(t)})^3\right].$$

D. If λ is large, then $\hat{w}$ will likely be close to 0.

AC **Question 14.** [1 MARK]

Let everything be defined as in the previous question. Suppose that we are now interested in using second-order gradient descent to find $\hat{w}$. Which of the following statements are true?

A. The second-order gradient descent update rule is the same as the first-order gradient descent update rule if the step size is $\eta^{(t)} = \frac{1}{\hat{L}''_\lambda(w^{(t)})}$.

$\hat{L}'_\lambda(w) = \frac{4}{n}\left[\sum_{i=1}^{n} c_ix_i(x_iw-y_i)^3+\lambda w^3\right]$

B. $\hat{L}''_\lambda(w) = \frac{12}{n}\left[\sum_{i=1}^{n} c_ix_i(x_iw^{(t)} - y_i)^2 + \lambda(w^{(t)})^2\right]$.

$\hat{L}''_\lambda(w) = \frac{12}{n}\left[\sum_{i=1}^{n} c_ix_i^2(x_iw-y_i)^2+\lambda w^2\right]$

C. $\hat{L}''_\lambda(w) = \frac{12}{n}\left[\sum_{i=1}^{n} c_ix_i^2(x_iw^{(t)} - y_i)^2 + \lambda(w^{(t)})^2\right]$.

D. $\hat{L}''_\lambda(w) = \frac{2}{n}\left[\sum_{i=1}^{n} c_ix_i^2 + \lambda\right]$

AC **Question 15.** [1 MARK]

Let the dataset be $\mathcal{D} = ((x_1, y_1), \ldots, (x_n, y_n))$, the mini-batch size $b \in \mathbb{N}$, and $M = \lfloor n/b \rfloor$. In class we learned about mini-batch gradient descent. However, if the size of the dataset n was not divisible by the mini-batch size b, then we discarded the last batch of data. In this question, we are interested in developing a mini-batch gradient descent algorithm that uses *all* the data points. To achieve this, if n is not divisible by b, we will append the last $n - Mb$ data points to the M-th mini-batch, such that the M-th mini-batch will now contain $b + n - Mb$ data points. Which of the following statements are true?

A. There are always M mini-batches.

B. If n is divisible by b, then there are $M + 1$ mini-batches. X

C. The estimated loss of a predictor f based on the M-th mini-batch is $n - i + 1 = b + n - Mb$

$$\frac{1}{n + b(1 - M)} \sum_{i=(M-1)b+1}^{n} \ell(f(\mathbf{x}_i), y_i).$$

$\Rightarrow i = n + 1 - b - n + Mb = (M-1)b + 1$

D. If n is not divisible by b then the variance of the estimated loss based on the M-th mini-batch (of a predictor f that is chosen independent of the dataset) is larger than the variance of the estimated loss based on any of the other mini-batches. smaller X

C **Question 16.** [1 MARK]

Let everything be defined as in the previous question. Your friend is trying to implement the version of mini-batch gradient descent discussed in the previous question with a constant step size. They have written the following pseudocode and asked you to review it. Which of the following statements are true?

Algorithm 1: MBGD Linear Regression Learner (with a constant step size and using all data)

input: $\mathcal{D} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$, step size η, number of epochs T, mini-batch size b
$\mathbf{w} \leftarrow$ random vector in $\mathbb{R}^{d+1}$
$M \leftarrow \text{floor}\left(\frac{n}{b}\right)$
for $t = 1, \ldots, T$ **do**
  randomly shuffle $\mathcal{D}$
  for $m = 1, \ldots, M - 1$ **do**
    $\nabla \hat{L}(\mathbf{w}) \leftarrow \frac{2}{b} \sum_{i=(m-1)b+1}^{mb} (\mathbf{x}_i^\top \mathbf{w} - y_i)\mathbf{x}_i$
    $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \hat{L}(\mathbf{w})$
  $\nabla \hat{L}(\mathbf{w}) \leftarrow \frac{2}{n - Mb} \sum_{i=Mb+1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)\mathbf{x}_i$ X $n - Mb + b$
  $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \hat{L}(\mathbf{w})$
return $\hat{f}(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}$

A. The pseudocode is correct.

B. The pseudocode is incorrect because the step size should be updated at each epoch.

C. The pseudocode is incorrect because the gradient for the last mini-batch is incorrect.

D. The pseudocode is incorrect because the gradient update occurs only M times.

B

Question 17. [1 MARK]

$= e^{3w} - 2\ln(w^2) = e^{3w} - 4\ln(w)$

Consider the function $g(w) = e^{3w} - 2\log(w^2)$, where $w \in (0, \infty)$. What is the second derivative $g''(w)$, and is $g(w)$ convex for $w \in (0, \infty)$?

$f'(w) = 3e^{3w} - \frac{4}{w} \Rightarrow f''(w) = 9e^{3w} + \frac{4}{w^2} \geq 0$

A. $g''(w) = 9e^{3w} - 4/w^2$. Not convex.

$\boxed{\text{B.}}$ $g''(w) = 9e^{3w} + 4/w^2$. Convex.

C. $g''(w) = 3e^{3w} - 2/w$. Not Convex.

D. $g''(w) = 9e^{3w} - 2/w$. Convex.

B

Question 18. [1 MARK]

$-\frac{1}{2}$

Suppose you have a dataset $\mathcal{D} = (z_1, \ldots, z_n)$ containing n i.i.d. flips of a coin. Since the flips are i.i.d. you know they all follow the distribution Bernoulli (α^*). However, you do not know what α^* is so you would like to estimate it using MLE. Which of the following are equal to the negative log-likelihood function $-\log(p(\mathcal{D}|\alpha)$? $p(z|\alpha^*) = (\alpha^*)^z(1-\alpha^*)^{1-z}$

Hint: Recall the logarithm properties: $\log(a^b) = b\log(a)$ and $\log(ab) = \log(a) + \log(b)$.

$\boxed{\text{A.}}$ $\sum_{i=1}^n (-z_i \log(\alpha) - (1 - z_i)\log(1 - \alpha))$

$\boxed{\text{B.}}$ $-(\sum_{i=1}^n z_i)\log(\alpha) - (n - \sum_{i=1}^n z_i)\log(1 - \alpha)$

C. $\sum_{i=1}^n (z_i(1 - z_i)\log(\alpha)\log(1 - \alpha))$

D. $\sum_{i=1}^n (z_i \log(\alpha) + (1 - z_i)\log(1 - \alpha))$

$-\ln(p(D|\alpha)) = -\ln\left(\prod_{i=1}^n p(z_i|\alpha)\right) = -\ln\left[(\alpha)^{z_1+\cdots+z_n}(1-\alpha)^{n-(z_1+\cdots+z_n)}\right]$

$= -\left[\ln\left((\alpha)^{z_1+\cdots+z_n}\right) + \ln\left((1-\alpha)^{n-(z_1+\cdots+z_n)}\right)\right]$

$= -\left[\sum_{i=1}^n z_i \cdot \ln(\alpha) + (n - \sum_{i=1}^n z_i)\ln(1-\alpha)\right]$

$= -(\sum_{i=1}^n z_i)\ln(\alpha) - (n - \sum_{i=1}^n z_i)\ln(1-\alpha)$

d

Question 19. [1 MARK]

Let everything be defined as in the previous question. Which of the following are equal to the MLE solution $\alpha_{\text{MLE}} = \arg\max_{\alpha \in [0,1]} p(\mathcal{D}|\alpha)$?

$\alpha_{MLE} = \underset{\alpha\in[0,1]}{\operatorname{argmax}}\ p(D|\alpha) = \underset{\alpha\in[0,1]}{\operatorname{argmin}}\ (-p(D|\alpha))$

a. $\alpha_{\text{MLE}} = \sum_{i=1}^n z_i$

$= \underset{\alpha\in[0,1]}{\operatorname{argmin}} \left[-(\sum_{i=1}^n z_i)\ln(\alpha) - (n - \sum_{i=1}^n z_i)\ln(1-\alpha)\right]$

b. $\alpha_{\text{MLE}} = \frac{1}{n-1}\sum_{i=1}^n z_i$

c. $\alpha_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n-1} z_i$

$\boxed{\text{d.}}$ $\alpha_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^n z_i$

let $f(\alpha) = -\sum_{i=1}^n z_i \cdot \ln(\alpha) - (n - \sum_{i=1}^n z_i)\ln(1-\alpha)$

$\Rightarrow f'(\alpha) = \frac{-\sum_{i=1}^n z_i}{\alpha} + \frac{n - \sum_{i=1}^n z_i}{1-\alpha} = \frac{(\alpha-1)\sum_{i=1}^n z_i + \alpha n - \alpha\sum_{i=1}^n z_i}{\alpha(1-\alpha)} = 0$

$\Rightarrow (\alpha-1)\sum_{i=1}^n z_i + \alpha n - \alpha\sum_{i=1}^n z_i = 0$

$\Rightarrow \alpha n = \sum_{i=1}^n z_i \Rightarrow \alpha = \frac{\sum_{i=1}^n z_i}{n}$

$P(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mu-50)^2}{2\sigma^2}\right)$ $P(D|\mu) = \prod_{i=1}^{n} P(z_i|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z_i-\mu^*)^2}{2}\right)$

$\Rightarrow -\ln(P(D|\mu)P(\mu)) = -\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mu-50)^2}{2\sigma^2}\right) \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z_i-\mu^*)^2}{2}\right)\right) = -\left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mu-50)^2}{2\sigma^2}\right)\right) + \ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z_i-\mu)^2}{2}\right)\right)\right]$

A

Question 20. [1 MARK]

$P(z|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right)$

$= -\left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(\mu-50)^2}{2\sigma^2} + \ln\left(\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{(z_i-\mu)^2}{2}\right)\right)\right]$

Suppose you have a dataset $\mathcal{D} = (z_1, \ldots, z_n)$ containing n i.i.d. samples from a normal distribution $\mathcal{N}(\mu^*, 1)$ Each data point z_i represents the age of a person in years. We would like to estimate μ^* using MAP. Suppose we have some prior knowledge that the average age μ of a person is around 50. We decide that a normal distribution with mean 50 and variance σ^2 accurately represents our prior knowledge of μ. Which of the following are equal to $-\log(p(\mathcal{D}|\mu) \cdot p(\mu))$?

$N(50, \sigma^2)$

$= -\left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(\mu-50)^2}{2\sigma^2} + n\ln\left(\frac{1}{\sqrt{2\pi}}\right) - \sum_{i=1}^{n} \frac{(z_i-\mu)^2}{2}\right]$

$= -\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{(\mu-50)^2}{2\sigma^2} - n\ln\left(\frac{1}{\sqrt{2\pi}}\right) + \sum_{i=1}^{n} \frac{(z_i-\mu)^2}{2}$

$\boxed{\text{A.}}$ $-n\log\left(\frac{1}{\sqrt{2\pi}}\right) + \sum_{i=1}^{n} \frac{(z_i-\mu)^2}{2} - \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{(\mu-50)^2}{2\sigma^2}$

B. $\left[-\sum_{i=1}^{n} \frac{(z_i-\mu)^2}{2} - \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}}\right)\right] \cdot \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(\mu-50)^2}{2\sigma^2}\right]$

C. $\sum_{i=1}^{n} \frac{(z_i-\mu)^2}{2} - \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}}\right)$

D. $n\log\left(\frac{1}{\sqrt{2\pi}}\right) - \sum_{i=1}^{n} \frac{(z_i-\mu)^2}{2} + \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(\mu-50)^2}{2\sigma^2}$

B

Question 21. [1 MARK]

Let everything be defined as in the previous question. Which of the following are the MAP solution $\mu_{\text{MAP}} = \arg\max_{\mu\in\mathbb{R}} p(\mu|\mathcal{D})$?

$\mu_{MAP} = \arg\max_{\mu\in\mathbb{R}} P(\mu|D) = \arg\max_{\mu\in\mathbb{R}} P(D|\mu)P(\mu) = \arg\min_{\mu\in\mathbb{R}} (-P(D|\mu)P(\mu))$

$= \arg\min_{\mu\in\mathbb{R}} (-\ln(P(D|\mu)P(\mu))) = \arg\min_{\mu\in\mathbb{R}} \left[-n\ln\left(\frac{1}{\sqrt{2\pi}}\right) + \sum_{i=1}^{n} \frac{(z_i-\mu)^2}{2} - \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{(\mu-50)^2}{2\sigma^2}\right]$

$= \arg\min_{\mu\in\mathbb{R}} \left[\sum_{i=1}^{n} \frac{(z_i-\mu)^2}{2} + \frac{(\mu-50)^2}{2\sigma^2}\right]$

let $f(\mu) = \sum_{i=1}^{n} \frac{(z_i-\mu)^2}{2} + \frac{(\mu-50)^2}{2\sigma^2} \Rightarrow f'(\mu) = \sum_{i=1}^{n} \mu - z_i + \frac{\mu-50}{\sigma^2} = n\mu - \sum_{i=1}^{n} z_i + \frac{\mu-50}{\sigma^2}$

$= \left(n + \frac{1}{\sigma^2}\right)\mu - \sum_{i=1}^{n} z_i - \frac{50}{\sigma^2} = 0$

$\Rightarrow \mu = \frac{\sum_{i=1}^{n} z_i + \frac{50}{\sigma^2}}{n + \frac{1}{\sigma^2}}$

A. $\frac{\sum_{i=1}^{n} z_i}{n + 1/\sigma^2}$

$\boxed{\text{B.}}$ $\frac{\sum_{i=1}^{n} z_i + 50/\sigma^2}{n + 1/\sigma^2}$

C. $\frac{\sum_{i=1}^{n} z_i - 50/\sigma^2}{n - 1/\sigma^2}$

D. $\frac{\sum_{i=1}^{n} z_i}{n}$

BCD

Question 22. [1 MARK]

Let everything be defined as in the previous two questions. Which of the following are true.

A. If σ^2 is large, then μ_{MAP} is approximately 50. $\sum_{i=1}^{n} z_i / n$ X

$\lim_{\sigma^2\to\infty} \frac{\sum_{i=1}^{n} z_i + \frac{50}{\sigma^2}}{n + \frac{1}{\sigma^2}} = \frac{\sum_{i=1}^{n} z_i}{n}$

$\boxed{\text{B.}}$ If σ^2 is small, then μ_{MAP} is approximately 50.

$\boxed{\text{C.}}$ If n is large, then μ_{MAP} is approximately $\frac{1}{n}\sum_{i=1}^{n} z_i$.

$\lim_{\sigma^2\to 0} \frac{\sum_{i=1}^{n} z_i + \frac{50}{\sigma^2}}{n + \frac{1}{\sigma^2}} = \frac{\sigma^2 \sum_{i=1}^{n} z_i + 50}{\sigma^2 n + 1}$

$\boxed{\text{D.}}$ If σ^2 is large, then μ_{MAP} is approximately $\frac{1}{n}\sum_{i=1}^{n} z_i$.

$= 50$

BC **Question 23.** [1 MARK]

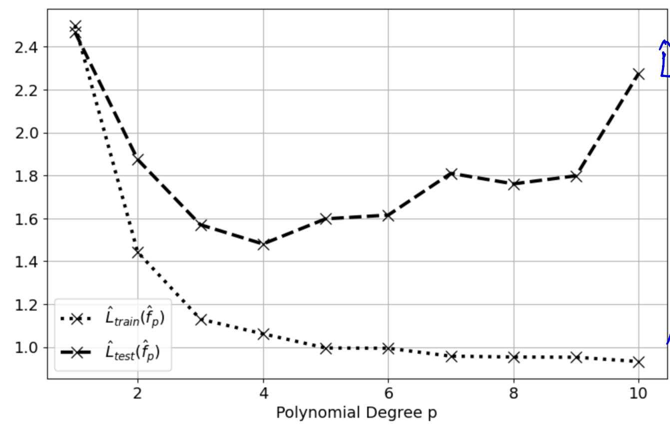
You are trying to decide which polynomial degree p to use for the function class $\mathcal{F}_p$ for a closed-form polynomial regression learner. You have a dataset of size n which you split into a training set and a test set as follows:

$$\mathcal{D}_{\text{train}} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n-m}, y_{n-m})), \quad \text{and} \quad \mathcal{D}_{\text{test}} = ((\mathbf{x}_{n-m+1}, y_{n-m+1}), \ldots, (\mathbf{x}_n, y_n)).$$

You train a polynomial regression learner for each p on $\mathcal{D}_{\text{train}}$, giving you a predictor $\hat{f}_p$ for each p. The training and test loss are defined as follows:

$$\hat{L}_{\text{train}}(f) = \frac{1}{n-m} \sum_{i=1}^{n-m} \ell(f(\mathbf{x}_i), y_i), \quad \hat{L}_{\text{test}}(f) = \frac{1}{m} \sum_{i=n-m+1}^{n} \ell(f(\mathbf{x}_i), y_i).$$

You plot the training loss $\hat{L}_{\text{train}}(\hat{f}_p)$ and the test loss $\hat{L}_{\text{test}}(\hat{f}_p)$ as a function of p, which is shown below. Which of the following statements are true?



$\hat{L}_{test}(\hat{f}_p) \approx L(\hat{f}_p)$

$\hat{L}_{train}(\hat{f}_p) \approx \hat{L}(\hat{f}_p)$

A. The ~~training~~ test loss loss is usually a better estimate of $L(\hat{f}_p)$ than the test loss.

B. The predictor $\hat{f}_1$ is likely underfitting, and the predictor $\hat{f}_{10}$ is likely overfitting.

C. The reason that the train loss decreases as p increases is because $\mathcal{F}_p$ becomes a larger function class as p increases.

D. The approximation error is likely the largest for $p = 10$ and the estimation error is likely the ~~largest~~ for $p = 1$. $AE = L(f^*) - L(f_{Bayes})$ $EE = E[L(\hat{f}_p)] - L(f^*)$

A

$-\frac{1}{2}$

Question 24. [1 MARK]

Let ϕ_p be the polynomial feature map of degree p. The function class containing all polynomials of degree p or less is

$$\mathcal{F}_p = \{f \mid f : \mathbb{R}^{d+1} \to \mathbb{R}, \text{ and } f(\mathbf{x}) = \phi_p(\mathbf{x})^\top \mathbf{w}, \text{ for some } \mathbf{w} \in \mathbb{R}^{\bar{p}}\}.$$

$d+p-(p-1)=d+1$

Which of the following statements are true?

$d+p+1-p=$

$\boxed{\text{A.}}$ $\min_{f\in\mathcal{F}_1} \hat{L}(f) \geq \min_{f\in\mathcal{F}_{10}} \hat{L}(f)$.

B. There exists as function $f \in \mathcal{F}_5$ such that $f \notin \mathcal{F}_{10}$. X

C. If $d = 3$ then $\phi_2(\mathbf{x}) = (1, x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2x_3) \in \mathbb{R}^8$. X $d=3, p=2$ $\bar{p} = C^p_{d+p} = C^d_{d+p} = C^2_5 = 10$

$\boxed{\text{D.}}$ $\bar{p}$ increases as p increases.? $\bar{p} = C^p_{d+p} = \frac{(d+p)(d+p-1)\cdots(d+1)}{p!}$ $\bar{p}' = C^{p+1}_{d+p+1} = \frac{(d+p+1)(d+p)\cdots(d+1)}{(p+1)!}$

$\Rightarrow \bar{p}'/\bar{p} = \frac{(d+p+1)(d+p)\cdots(d+1)}{p!\,(p+1)} \cdot \frac{p!}{(d+p)(d+p-1)\cdots(d+1)} = \frac{d+p+1}{p+1} = 1 + \frac{d}{p+1} > 1 \Rightarrow \bar{p}' > \bar{p}$

D

$-\frac{1}{2}$

Question 25. [1 MARK]

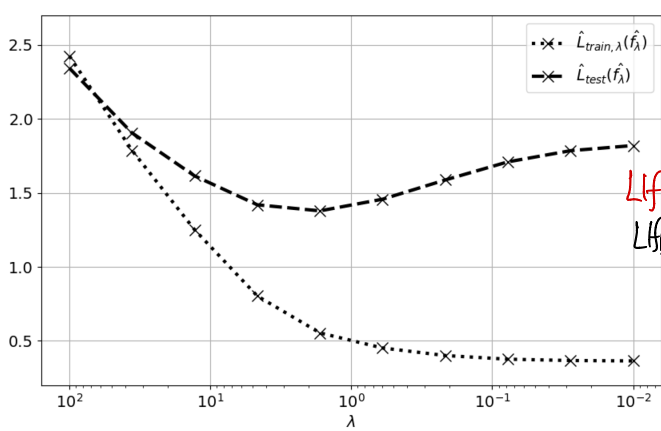
You are trying to decide which regularization parameter value λ to use for a closed-form polynomial regression learner with degree $p = 10$. You have a dataset of size n which you split into a training set and a test set as follows:

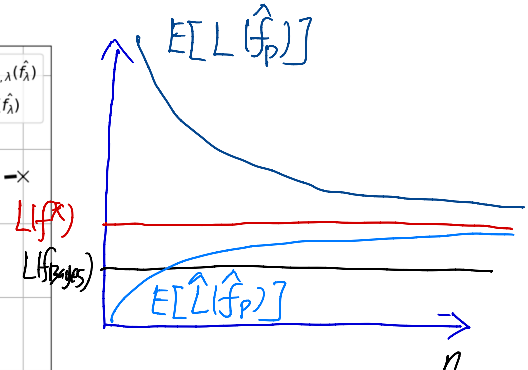
$$\mathcal{D}_{\text{train}} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n-m}, y_{n-m})), \quad \mathcal{D}_{\text{test}} = ((\mathbf{x}_{n-m+1}, y_{n-m+1}), \ldots, (\mathbf{x}_n, y_n)).$$

You train a polynomial regression learner for 10 different values of λ on $\mathcal{D}_{\text{train}}$, giving you a different predictor $\hat{f}_\lambda$ for each value of λ. The training and test loss are defined as follows:

$$\hat{L}_{\text{train},\lambda}(f) = \frac{1}{n-m}\sum_{i=1}^{n-m} \ell(f(\mathbf{x}_i), y_i) + \frac{\lambda}{n-m}\sum_{j=1}^{\bar{p}-1} w_j^2, \quad \hat{L}_{\text{test}}(f) = \frac{1}{m}\sum_{i=n-m+1}^{n} \ell(f(\mathbf{x}_i), y_i).$$

You plot the training loss $\hat{L}_{\text{train},\lambda}(\hat{f}_\lambda)$ and the test loss $\hat{L}_{\text{test}}(\hat{f}_\lambda)$ as a function of λ, which is shown below. Which of the following statements are true?





$\boxed{\text{A.}}$ Based on the plot, for large λ values, such as $\lambda = 100$, the bias is likely high.

B. The approximation error is likely ~~higher~~ for $\lambda = 100$ than for $\lambda = 0.01$. ~~lower~~ unchanged $AE = L(f^*) - L(f_{Bayes})$

C. The variance is likely higher for $\lambda = 100$ than for $\lambda = 0.01$. lower

$\boxed{\text{D.}}$ The best choice of λ based on the plot is $\lambda \approx 2$ since the test loss is the smallest there.

AC

Question 26. [1 MARK]

You are interested in getting a binary classifier. You have a dataset $\mathcal{D} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$ where $\mathbf{x}_i \in \mathbb{R}^{2+1}$ and $y_i \in \{0, 1\}$. You select the following function class

$$\mathcal{F} = \left\{ f | f : \mathbb{R}^{d+1} \to [0,1], \text{ where } f(\mathbf{x}) = \sigma(\mathbf{x}^\top \mathbf{w}), \text{ and } \mathbf{w} \in \mathbb{R}^{d+1} \right\},$$

and the binary cross-entropy loss function

$$\ell(\hat{y}, y) = -y \log(\hat{y}) - (1-y)\log(1-\hat{y}).$$

Suppose you use ERM and get the exact solution

$$\hat{f}_{\text{ERM}} = \arg\min_{f \in \mathcal{F}} \hat{L}(f).$$

$\hat{f}_{ERM}(\vec{x}) = \sigma(\vec{x}^\top \vec{\hat{w}}_{ERM}) = \frac{1}{1+\exp(-\vec{x}^\top \vec{\hat{w}}_{ERM})} = 0.5$

$\Rightarrow -\vec{x}^\top \vec{\hat{w}}_{ERM} = 0$

That is $\hat{f}_{\text{ERM}}(\mathbf{x}) = \sigma(\mathbf{x}^\top \hat{\mathbf{w}}_{\text{ERM}})$ for some $\hat{\mathbf{w}}_{\text{ERM}} \in \mathbb{R}^{d+1}$. You define your binary classifier $f_{\text{Bin}}(\mathbf{x}) = 1$ if $\hat{f}_{\text{ERM}}(\mathbf{x}) \geq c$ and $f_{\text{Bin}}(\mathbf{x}) = 0$ otherwise, where $c = 0.5$. The decision boundary of f_{Bin} is the set of points $\mathbf{x}$ where $\hat{f}_{\text{ERM}}(\mathbf{x}) = c$. You plot the decision boundary of f_{Bin} and the points in the dataset in the figure below. Which of the following are true?

$f_{Bin}(\vec{x}) = 1 \Rightarrow -\vec{x}^\top \vec{\hat{w}}_{ERM} \leq 0$

$\Rightarrow \vec{x}^\top \vec{\hat{w}}_{ERM} \geq 0$

$f_{Bin}(\vec{x}) = 0 \Rightarrow \vec{x}^\top \vec{\hat{w}}_{ERM} < 0$

$\vec{x}^\top \vec{\hat{w}}_{ERM} = x_0 \hat{w}_{ERM,0} + x_1 \hat{w}_{ERM,1} + x_2 \hat{w}_{ERM,2}$

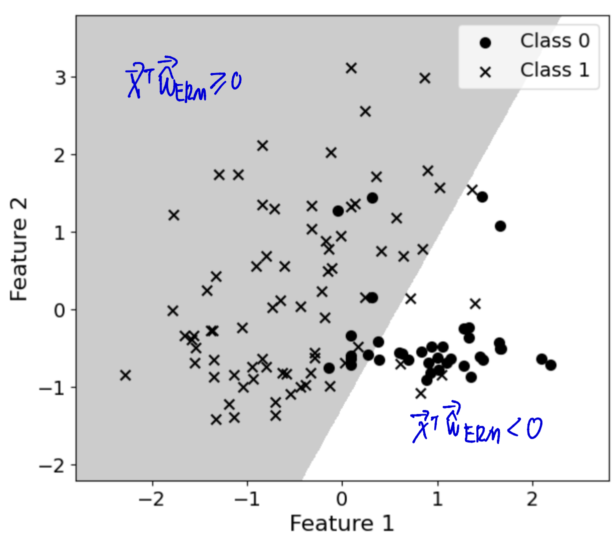
$= \hat{w}_{ERM,0} + x_1 \hat{w}_{ERM,1} + x_2 \hat{w}_{ERM,2} = 0$

$\Rightarrow x_2 \hat{w}_{ERM,2} = -x_1 \hat{w}_{ERM,1} - \hat{w}_{ERM,0}$

$\Rightarrow x_2 = -\frac{\hat{w}_{ERM,1}}{\hat{w}_{ERM,2}} x_1 - \frac{\hat{w}_{ERM,0}}{\hat{w}_{ERM,2}}$

$\Rightarrow \frac{\hat{w}_{ERM,1}}{\hat{w}_{ERM,2}} < 0 \quad \frac{\hat{w}_{ERM,0}}{\hat{w}_{ERM,2}} < 0$

$\Rightarrow \hat{w}_{ERM,0} \hat{w}_{ERM,1} > 0$



A. The decision boundary of $f_{\text{Bin}}(\mathbf{x})$ is represented by the line $\mathbf{x}^\top \hat{\mathbf{w}}_{\text{ERM}} = 0$.

B. The grey region represents the values of $\mathbf{x}$ where $\mathbf{x}^\top \hat{\mathbf{w}}_{\text{ERM}} < 0$.

C. The grey region represents the values of $\mathbf{x}$ where $\mathbf{x}^\top \hat{\mathbf{w}}_{\text{ERM}} \geq 0$.

D. If you changed the threshold c to 0.7, the grey region would cover an area that is ~~larger~~ smaller or equal to the area of the grey region in the figure.

$\hat{f}_{ERM}(\vec{x}) = \sigma(\vec{x}^\top \vec{\hat{w}}_{ERM}) = \frac{1}{1+\exp(-\vec{x}^\top \vec{\hat{w}}_{ERM})} = 0.7 \Rightarrow 1+\exp(-\vec{x}^\top \vec{\hat{w}}_{ERM}) = \frac{10}{7} \Rightarrow \exp(-\vec{x}^\top \vec{\hat{w}}_{ERM}) = \frac{3}{7}$

$-\vec{x}^\top \vec{\hat{w}}_{ERM} = \ln(\frac{3}{7})$

$\Rightarrow \vec{x}^\top \vec{\hat{w}}_{ERM} = \ln(\frac{7}{3})$

$\Rightarrow \hat{w}_{ERM,2} x_2 = -\hat{w}_{ERM,1} x_1 + \ln(\frac{7}{3}) - \hat{w}_{ERM,0}$

$\Rightarrow \hat{w}_{ERM,0} + \hat{w}_{ERM,1} x_1 + \hat{w}_{ERM,2} x_2 = \ln(\frac{7}{3})$

$\Rightarrow x_2 = -\frac{\hat{w}_{ERM,1}}{\hat{w}_{ERM,2}} x_1 + \frac{\ln(\frac{7}{3}) - \hat{w}_{ERM,0}}{\hat{w}_{ERM,2}}$

ABD

Question 27. [1 MARK]

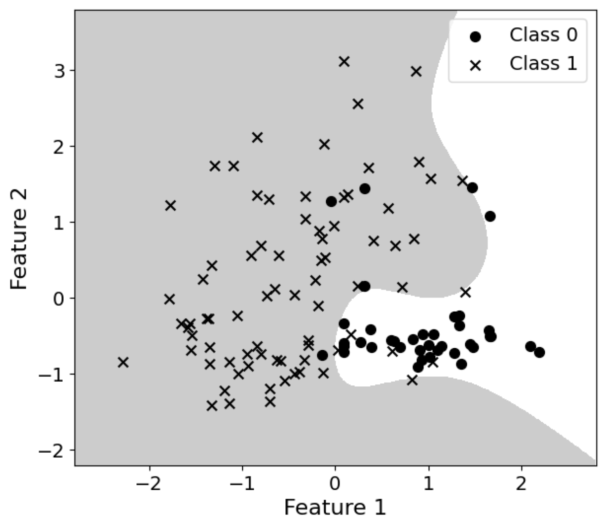
Let everything be as defined in the previous question. Suppose you decide to change to the polynomial function class to

$$\mathcal{F}_p = \left\{ f | f : \mathbb{R}^{d+1} \to [0,1], \text{ where } f(\mathbf{x}) = \sigma(\phi_p(\mathbf{x})^\top \mathbf{w}), \text{ and } \mathbf{w} \in \mathbb{R}^{\bar{p}} \right\},$$

for some $p > 1$. You go through the same process as before and get the exact solution

$$\hat{f}_{\text{ERM},p} = \arg\min_{f \in \mathcal{F}_p} \hat{L}(f) \quad \text{where} \quad \hat{f}_{\text{ERM},p}(\mathbf{x}) = \sigma(\phi_p(\mathbf{x})^\top \hat{\mathbf{w}}_{\text{ERM},p}) \quad \text{for some } \hat{\mathbf{w}}_{\text{ERM},p} \in \mathbb{R}^{\bar{p}}.$$

You define your polynomial binary classifier $f_{\text{Bin},p}(\mathbf{x}) = 1$ if $\hat{f}_{\text{ERM},p}(\mathbf{x}) \geq c$ and $f_{\text{Bin},p}(\mathbf{x}) = 0$ otherwise, where $c = 0.5$. The decision boundary of $f_{\text{Bin, p}}$ is the set of points $\mathbf{x}$ where $\hat{f}_{\text{ERM},p}(\mathbf{x}) = c$. You plot the decision boundary of $f_{\text{Bin},p}$ and the points in the dataset in the figure below. Which



of the following are true?

$\boxed{\text{A.}}$ If you set $p = 1$, the decision boundary of $f_{\text{Bin},p}(\mathbf{x})$ would be the same as the decision boundary of $f_{\text{Bin}}(\mathbf{x})$.

$\boxed{\text{B.}}$ If you changed the threshold c to 0.7, the grey region would cover an area that is smaller or equal to the area of the grey region in the figure.

C. If you changed the threshold c to 0.7, the decision boundary of $f_{\text{Bin},p}(\mathbf{x})$ can be represented by the curve $\phi_p(\mathbf{x})^\top \hat{\mathbf{w}}_{\text{ERM},p} = \cancel{0}.$ $\ln(\frac{7}{3})$

$\boxed{\text{D.}}$ The decision boundary of $f_{\text{Bin},p}(\mathbf{x})$ is represented by the curve $\phi_p(\mathbf{x})^\top \hat{\mathbf{w}}_{\text{ERM},p} = 0$.

ABC

Question 28. [1 MARK]

Let everything be as defined in the previous two questions. For your dataset $\mathcal{D}$ you count the number of datapoints $(\mathbf{x}_i, y_i)$ that were misclassified by f_{Bin} (i.e. $f_{\text{Bin}}(\mathbf{x}_i) \neq y_i$) and call this number m_{Bin}. You also count the number of datapoints $(\mathbf{x}_i, y_i)$ that where misclassified by $f_{\text{Bin},p}$ (i.e. $f_{\text{Bin},p}(\mathbf{x}_i) \neq y_i$) and call this number $m_{\text{Bin},p}$. You find that $m_{\text{Bin}} > m_{\text{Bin},p}$. Which of the following are true?

$\boxed{\text{A.}}$ If the zero-one loss function is used, then $\hat{L}(f_{\text{Bin}}) > \hat{L}(f_{\text{Bin},p})$.

$\boxed{\text{B.}}$ If the zero-one loss function is used, then $\hat{L}(f_{\text{Bin}}) = m_{\text{Bin}}/n$.

$\boxed{\text{C.}}$ If you count the number of circles in the grey region and the number of Xs in the white region in the figure for f_{Bin}, you would get m_{Bin}.

D. The classifier f_{Bin} is more likely to ~~overfit~~ underfit the data than the classifier $f_{\text{Bin},10}$.

AB

Question 29. [1 MARK]

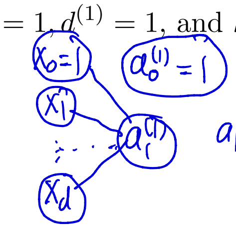
In class we used the following function class for logistic regression:

$$\mathcal{F} = \left\{ f \mid f : \mathbb{R}^{d+1} \to [0,1], \text{ where } f(\mathbf{x}) = \sigma(\mathbf{x}^\top \mathbf{w}), \text{ and } \mathbf{w} \in \mathbb{R}^{d+1} \right\}.$$

Suppose that we would like to use some different function classes that contain NNs with a fixed architecture. We define two new function classes as follows

$$\mathcal{F}_{\text{NN}}^{(1)} = \left\{ f \mid f : \mathbb{R}^{d+1} \to [0,1], \text{ where } f \text{ is a NN with } B = 1, d^{(1)} = 1, \text{ and } h^{(1)}(z) = \sigma(z) \right\},$$
$$\mathcal{F}_{\text{NN}}^{(2)} = \left\{ f \mid f : \mathbb{R}^{d+1} \to [0,1], \text{ where } f \text{ is a NN with } B = 1, d^{(1)} = 1, \text{ and } h^{(1)}(z) = z \right\}.$$

$\mathcal{F}_{NN}^{(1)}$:



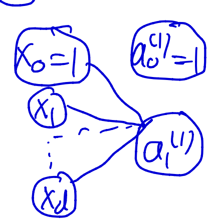
$a_1^{(1)} = h^{(1)}(\vec{x}^T\vec{w}) = \sigma(\vec{x}^T\vec{w}) = \frac{1}{1+e^{-(\vec{x}^T\vec{w})}}$

Which of the following are true?

$\boxed{\text{A.}}$ $\mathcal{F} = \mathcal{F}_{\text{NN}}^{(1)}$.

$\boxed{\text{B.}}$ There is a function $f \in \mathcal{F}$ that is not in $\mathcal{F}_{\text{NN}}^{(2)}$.

$\mathcal{F}_{NN}^{(2)}$



$a_1^{(1)} = h^{(1)}(\vec{x}^T\vec{w}) = \vec{x}^T\vec{w}$

C. There is a function $f \in \mathcal{F}_{\text{NN}}^{(1)}$ that is not in $\mathcal{F}$.

D. $\mathcal{F}_{\text{NN}}^{(1)} \subset \mathcal{F}_{\text{NN}}^{(2)}$.

AD

-1

Question 30. [1 MARK]

You are interested in getting a multiclass classifier. You have a dataset $\mathcal{D} = ((\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n))$ where $\mathbf{y}_i \in [0,1]^K$ is a one-hot vector with K elements. You select the following function class

$$\mathcal{F} = \left\{ f | f : \mathbb{R}^{d+1} \to [0,1]^K, \text{ where } f(\mathbf{x}) = \sigma(\mathbf{x}^\top \mathbf{w}_0, \dots, \mathbf{x}^\top \mathbf{w}_{K-1}), \text{ and } \mathbf{w}_0, \dots, \mathbf{w}_{K-1} \in \mathbb{R}^{d+1} \right\},$$

and the multiclass cross-entropy loss function

$$\ell(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_{q=0}^{K-1} y_q \log(\hat{y}_q) \quad \text{where} \quad \hat{\mathbf{y}}, \mathbf{y} \in [0,1]^K.$$

Suppose you use ERM and get the exact solution

$$\hat{f}_{\text{ERM}} = \arg\min_{f \in \mathcal{F}} \hat{L}(f) \text{ where } \hat{f}_{\text{ERM}}(\mathbf{x}) = \sigma(\mathbf{x}^\top \hat{\mathbf{w}}_0, \dots, \mathbf{x}^\top \hat{\mathbf{w}}_{K-1}) \text{ for some } \hat{\mathbf{w}}_0, \dots, \hat{\mathbf{w}}_{K-1} \in \mathbb{R}^{d+1}.$$

You define your multiclass classifier as $f_{Mul}(\vec{x}) = \arg\max_{k\in\{0,\dots,K-1\}} \hat{y}_k = \arg\max_{k\in\{0,\dots,K-1\}} \frac{\exp(\vec{x}^T\vec{\hat{w}}_k)}{\sum_{t=0}^{K-1}\exp(\vec{x}^T\vec{\hat{w}}_t)} = \arg\max_{k\in\{0,\dots,K-1\}} \exp(\vec{x}^T\vec{\hat{w}}_k)$

$$f_{\text{Mul}}(\mathbf{x}) = \arg\max_{k \in \{0,\dots,K-1\}} \hat{y}_k \quad \text{where} \quad \hat{\mathbf{y}} = \hat{f}_{\text{ERM}}(\mathbf{x}). = \sigma(\vec{x}^T\vec{\hat{w}}_0, \dots, \vec{x}^T\vec{\hat{w}}_{K-1})$$

$= \arg\max_{k\in\{0,\dots,K-1\}} (\vec{x}^T\vec{\hat{w}}_k)$

$\frac{5}{2} + \frac{1}{3} = \frac{15+2}{6} = \frac{17}{6}$

Which of the following are true?

$\boxed{\text{A.}}$ $\sum_{q=0}^{K-1} \hat{y}_q = 1.$ $\Rightarrow \hat{y}_q = \sigma_q(\vec{x}^T\vec{\hat{w}}_0, \dots, \vec{x}^T\vec{\hat{w}}_{K-1}) = \frac{\exp(\vec{x}^T\vec{\hat{w}}_q)}{\sum_{t=0}^{K-1}\exp(\vec{x}^T\vec{\hat{w}}_t)} \Rightarrow \sum_{q=0}^{K-1}\hat{y}_q = 1$ $\quad \frac{17}{6} \times \frac{\cancel{100}\ 50}{37} = \frac{850}{111}$

$\boxed{\text{B.}}$ If $\mathbf{x}^\top \hat{\mathbf{w}}_q > \mathbf{x}^\top \hat{\mathbf{w}}_k$ for all $k \neq q$, then $f_{\text{Mul}}(\mathbf{x}) = q$ where $q \in \{0, \dots, K-1\}$.

$\boxed{\text{C.}}$ If $\sigma_q(\mathbf{x}^\top \hat{\mathbf{w}}_0, \dots, \mathbf{x}^\top \hat{\mathbf{w}}_{K-1}) > 0.7$, then $f_{\text{Mul}}(\mathbf{x}) = q$ where $q \in \{0, \dots, K-1\}$. ≈ 7.65

$\boxed{\text{D.}}$ If $\sigma_q(\mathbf{x}^\top \hat{\mathbf{w}}_0, \dots, \mathbf{x}^\top \hat{\mathbf{w}}_{K-1}) = \sigma_k(\mathbf{x}^\top \hat{\mathbf{w}}_0, \dots, \mathbf{x}^\top \hat{\mathbf{w}}_{K-1})$ then $\hat{\mathbf{w}}_q = \hat{\mathbf{w}}_k$ for $q, k \in \{0, \dots, K-1\}$. 92.35

$\Rightarrow \vec{x}^T\vec{\hat{w}}_q = \vec{x}^T\vec{\hat{w}}_k$

CD

Question 31. [1 MARK]

You are interested in understanding the relationship between the two input softmax function $\sigma(z_1, z_2) = (\sigma_1(z_1, z_2), \sigma_2(z_1, z_2))^\top$ and the logistic function. Which of the following are true?

A. $\sigma(z_1, z_2) = \sigma(z_1 - z_2)$. X $\quad \sigma(z_1,z_2) = (\sigma_1(z_1,z_2), \sigma_2(z_1,z_2))^T = \left(\frac{e^{z_1}}{e^{z_1}+e^{z_2}}, \frac{e^{z_2}}{e^{z_1}+e^{z_2}}\right)^T$

B. $\sigma_1(z_1, z_2) = \sigma(z_1)$. X $\quad \sigma_1(z_1,z_2) = \frac{e^{z_1}}{e^{z_1}+e^{z_2}} \quad \sigma(z_1-z_2) = \frac{1}{1+e^{z_2-z_1}} = \frac{e^{z_1}}{e^{z_1}+e^{z_2}}$

$\boxed{\text{C.}}$ $\sigma_1(z_1, z_2) = \sigma(z_1 - z_2)$.

$\boxed{\text{D.}}$ $\sigma_2(z_1, z_2) = \sigma(z_2 - z_1)$. $\quad \sigma_2(z_1,z_2) = \frac{e^{z_2}}{e^{z_1}+e^{z_2}} \quad \sigma(z_2-z_1) = \frac{1}{1+e^{z_1-z_2}} = \frac{e^{z_2}}{e^{z_2}+e^{z_1}}$

C

Question 32. [1 MARK]

You are designing a neural network architecture for a multiclass classification problem. You decide to have $B = 4$ layers and $d^{(1)} = 100$, $d^{(2)} = 100$, $d^{(3)} = 50$, $d^{(4)} = 10$ neurons in each layer respectively. The input dimension is $d = 1000$. How many neurons are there in the network (you should include all the input neurons and bias neurons in your calculation)?

A. 1260

B. 263

$\boxed{\text{C.}}$ 1264

D. 260

$(1000+1)+(100+1)+(100+1)+(50+1)+10 = 1260+4=1264$

C

Question 33. [1 MARK]

Let everything be as defined in the previous question. If you sum up the dimension of all the weight vectors in the neural network you get the number of weights in the network. How many weights are there in the network?

A. 115500

B. 115771

$\boxed{\text{C.}}$ 115760

D. 115510

$1001 \cdot 100 + 101 \cdot 100 + 101 \cdot 50 + 51 \cdot 10 = 100100 + 10100 + 5050 + 510$

$= 110200 + 5560$

$= 115760$

B

Question 34. [1 MARK]

You have a neural network f with $B = 2$ layers and $d^{(1)} = 2$, $d^{(2)} = 2$ neurons in each layer respectively. The input dimension is $d = 2$. You choose to use the ReLU activation function for both the layers, defined as $\text{ReLU}(z) = \max(0, z)$, where $z \in \mathbb{R}$. The weight vectors have the following values:

$$\mathbf{w}_1^{(1)} = (-1,-1,-1)^\top \quad \mathbf{w}_2^{(1)} = (0,1,1)^\top \quad \mathbf{w}_1^{(2)} = (1,1,1)^\top \quad \mathbf{w}_2^{(2)} = (0,1,0)^\top$$

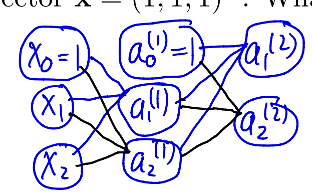
Suppose you get a feature vector $\mathbf{x} = (1,1,1)^\top$. What is $f(\mathbf{x})$?

A. $(0,-3)^\top$

$\boxed{\text{B.}}$ $(3,0)^\top$

C. 3

D. 0



$\vec{x} = (x_0, x_1, x_2)^T = (1,1,1)^T \Rightarrow \vec{x}^T \vec{w}_1^{(1)} = (1,1,1)\begin{pmatrix}-1\\-1\\-1\end{pmatrix} = -3$

$\Rightarrow a_1^{(1)} = h^{(1)}(\vec{x}^T \vec{w}_1^{(1)}) = 0$ $\quad \vec{x}^T \vec{w}_2^{(2)} = (1,1,1)\begin{pmatrix}0\\1\\1\end{pmatrix} = 2$

$\Rightarrow a_2^{(1)} = h^{(1)}(\vec{x}^T \vec{w}_2^{(2)}) = 2$ $\Rightarrow \vec{a}^{(1)} = (1,0,2)^T$

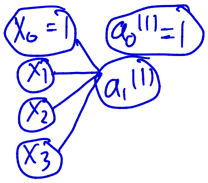
$\Rightarrow \vec{a}^{T(1)} \vec{w}_1^{(2)} = (1,0,2)\begin{pmatrix}1\\1\\1\end{pmatrix} = 3$

$\Rightarrow a_1^{(2)} = h^{(2)}(\vec{a}^{T(1)} \vec{w}_1^{(2)}) = 3$

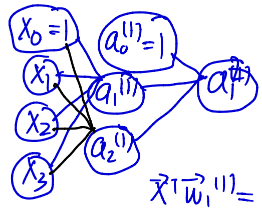
$\vec{a}^{T(1)} \vec{w}_2^{(2)} = (1,0,2)\begin{pmatrix}0\\1\\0\end{pmatrix} = 0$

$\Rightarrow a_2^{(2)} = h^{(2)}(\vec{a}^{T(1)} \vec{w}_1^{(2)}) = 0$

$\Rightarrow f(\vec{x}) = \vec{a}^{(2)} = (3\ 0)^T$



$\Rightarrow f(\vec{x}) = a_1^{(1)} = h^{(1)}(\vec{x}^T\vec{w}_1^{(1)}) = h^{(1)}\left((1,x_1,x_2,x_3)\begin{pmatrix}1\\1\\-1\\1\end{pmatrix}\right)$
$= h^{(1)}(1+x_1-x_2+x_3)$
$= 1+x_1-x_2+x_3$



$\vec{x}^T\vec{w}_1^{(1)} = (1\ x_1\ x_2\ x_3)\begin{pmatrix}0\\1\\-1\\0\end{pmatrix} = x_1-x_2$

$a_1^{(1)} = h^{(1)}(\vec{x}^T\vec{w}_1^{(1)}) = x_1-x_2$

BCD

Question 35. [1 MARK]

Let $d = 3$. Which of the following NNs f satisfy $f(\mathbf{x}) = x_1 - x_2 + x_3$.

A. $B = 1$, $d^{(1)} = 1$, $h^{(1)}(z) = z$, $\mathbf{w}_1^{(1)} = (1, 1, -1, 1)^\top$. X $f(x) = 1+x_1-x_2+x_3$

$\vec{x}^T\vec{w}_1^{(2)} = (1\ x_1\ x_2\ x_3)\begin{pmatrix}0\\0\\0\\1\end{pmatrix} = x_3$

$\boxed{\text{B.}}$ $B = 1$, $d^{(1)} = 1$, $h^{(1)}(z) = z$, $\mathbf{w}_1^{(1)} = (0, 1, -1, 1)^\top$. $f(x) = x_1-x_2+x_3$

$a_1^{(2)} = h^{(2)}(\vec{a}^{T(1)}\vec{w}_2^{(1)}) = h^{(2)}\left((1,x_1-x_2,x_3)\begin{pmatrix}0\\1\\1\end{pmatrix}\right)$
$= x_1-x_2+x_3$

$\boxed{\text{C.}}$ $B = 2$, $d^{(1)} = 2$, $d^{(2)} = 1$, $h^{(1)}(z) = h^{(2)}(z) = z$, $\mathbf{w}_1^{(1)} = (0, 1, -1, 0)^\top$, $\mathbf{w}_2^{(1)} = (0, 0, 0, 1)^\top$, $\mathbf{w}_1^{(2)} = (0, 1, 1)^\top$.

$\boxed{\text{D.}}$ $B = 2$, $d^{(1)} = 2$, $d^{(2)} = 1$, $h^{(1)}(z) = h^{(2)}(z) = z$, $\mathbf{w}_1^{(1)} = (0, 1, 1, 1)^\top$, $\mathbf{w}_2^{(1)} = (0, 0, -2, 0)^\top$, $\mathbf{w}_1^{(2)} = (0, 1, 1)^\top$.

$\vec{x}^T\vec{w}_1^{(1)} = (1\ x_1\ x_2\ x_3)\begin{pmatrix}0\\1\\1\\1\end{pmatrix} = x_1+x_2+x_3$

$a_1^{(1)} = h^{(1)}(\vec{x}^T\vec{w}_1^{(1)}) = x_1+x_2+x_3$

$\vec{x}^T\vec{w}_2^{(1)} = (1\ x_1\ x_2\ x_3)\begin{pmatrix}0\\0\\-2\\0\end{pmatrix} = -2x_2$

$a_2^{(1)} = h^{(1)}(\vec{x}^T\vec{w}_2^{(1)}) = -2x_2$

$\vec{a}^{(1)}\vec{w}_1^{(2)} = (1\ x_1+x_2+x_3\ -2x_2)\begin{pmatrix}0\\1\\1\end{pmatrix}$
$= x_1+x_2+x_3-2x_2$
$= x_1-x_2+x_3$

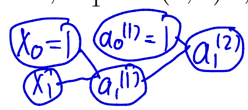
AB

Question 36. [1 MARK]

Let $d = 1$. Assume $x_1 > 0$. Which of the following NNs f satisfy $f(\mathbf{x}) = x_1^2$.
Hint: recall the logarithm property $\log(a^b) = b \log(a)$ and $\log(ab) = \log(a) + \log(b)$.

$\boxed{\text{A.}}$ $B = 2$, $d^{(1)} = 1$, $d^{(2)} = 1$, $h^{(1)}(z) = \log(z)$, $h^{(2)}(z) = e^z$, $\mathbf{w}_1^{(1)} = (0, 1)^\top$, $\mathbf{w}_1^{(2)} = (0, 2)^\top$.

$\boxed{\text{B.}}$ $B = 2$, $d^{(1)} = 2$, $d^{(2)} = 1$, $h^{(1)}(z) = \log(z)$, $h^{(2)}(z) = e^z$, $\mathbf{w}_1^{(1)} = (0, 1)^\top$, $\mathbf{w}_2^{(1)} = (0, 1)^\top$, $\mathbf{w}_1^{(2)} = (0, 1, 1)^\top$.



$f(\vec{x}) = a_1^{(2)} = h^{(2)}(\vec{a}^{T(1)}\vec{w}_1^{(2)}) = x_1^2$

C. $B = 1$, $d^{(1)} = 1$, $h^{(1)}(z) = \log(z)$, $\mathbf{w}_1^{(1)} = (0, 2)^\top$. X

D. $B = 1$, $d^{(1)} = 2$, $h^{(1)}(z) = \log(z)$, $\mathbf{w}_1^{(1)} = (0, 1, 1)^\top$ X

$\vec{x}^T\vec{w}_1^{(1)} = (1, x_1)\begin{pmatrix}0\\1\end{pmatrix} = x_1$

$a_1^{(1)} = h^{(1)}(\vec{x}^T\vec{w}_1^{(1)}) = \ln(x_1)$

$\vec{a}^{(1)}\vec{w}_1^{(2)} = (1, \ln(x_1))\begin{pmatrix}0\\2\end{pmatrix} = 2\ln(x_1)$

BC

Question 37. [1 MARK]

Suppose you want to use a NN function class with $B = 3$ layers and non-linear activation functions to predict the probability that a picture contains a dog from its pixel values. You decide to use ERM with the binary cross-entropy loss function. To solve the optimization you use gradient descent with a step size that you know will guarantee convergence to the minimum if you run for enough epochs T. You initialize the weights of the NN randomly. Which of the following are true?

A. For small numbers of epochs (such as $T = 1$), the neurons in the network will likely represent meaningful features. X

$\boxed{\text{B.}}$ After a large number of epochs the neurons in the first layer will likely learn a more complex feature representation than the pixel values in input layer. For example a neuron in the first layer might learn to detect an edge.

$\boxed{\text{C.}}$ For small numbers of epochs (such as $T = 1$), the neurons in the network will likely not represent any meaningful features.

D. After a large number of epochs the neurons in the first layer will likely learn to represent more complex features than the neurons in the second layer. For example a neuron in the first layer might learn to represent a dogs head, while a neuron in the second layer might learn to represent an edge.

Question 38. [1 MARK]

The only tokens you will encounter are `a`, `upon`, `time`, `once`, `.`, <EOS>, <PAD>. You represent each token by an integer as follows `a` $= 1$, `upon` $= 2$, `time` $= 3$, `once` $= 4$, `.` $= 5$, <EOS> $= 6$, <PAD> $= 7$. Thus, the vocabulary can be either $\mathcal{Y} = \{1, 2, 3, 4, 5, 6, 7\}$, or $\mathcal{Y} = \{$`once`, `upon`, `a`, `time`, `.`, <EOS>, <PAD>$\}$. You are given a sequence of tokens $s \in \mathcal{Y}^a$, where $a = 10$. Suppose that you are creating a dataset and the third input-output pair that you create is (s_3, y_3) where $s_3 = ($<PAD>, `once`, `upon`, `a`$)$ and $y_3 =$ `time`. Which of the following are true?

A. The context length is $c = 3$.

B. The one-hot vector label $\mathbf{y}_3$ is $(0, 0, 0, 1, 0, 0, 0)^\top$.

C. The vocabulary contains $|\mathcal{Y}| = K = 7$ tokens.

D. The sequence s must have started with the tokens `once`, `upon`, `a`, `time`.

Question 39. [1 MARK]

Let everything be as defined in the previous question. Suppose that you are using an embedding function $E : \mathcal{Y} \to \mathbb{R}^3$. Which of the following are true?

A. The feature vector based on s_3 is an element of $\mathbb{R}^{13}$.

B. Since the input sequence contains a tokens the number of feature-label pairs $(\mathbf{x}_i, \mathbf{y}_i)$ in the dataset is a.

C. The number of feature-label pairs $(\mathbf{x}_i, \mathbf{y}_i)$ in the dataset is the same as the context length c.

D. If you wanted to learn a model that outputs a vector containing the probability of each token in the vocabulary, you would have a 7 dimensional output vector.

Question 40. [1 MARK]

Suppose that you have the embeddings for the following words `dog`, `dogs`, `exam`, and `wolf`. Which of the following are true?

A. If you wanted to get an estimate of the embedding $E($`exams`$)$ you could calculate $E($`exam`$) + E($`dog`$) - E($`dogs`$)$.

B. If you wanted to get an estimate of the embedding $E($`exams`$)$ you could calculate $E($`exam`$) + E($`dogs`$) - E($`dog`$)$.

C. The embedding of `wolf` is likely closer to the embedding of `dog` (since they are similar animals) than the embedding of `exam`.

D. The embedding of `wolf` is likely closer to the embedding of `exam` than the embedding of `dog`.