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 36 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.

Question 1. [1 MARK]

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

$$\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} g(x, y),$$

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

A. 17

B. 27

C. 37

D. 11

Solution 1. **Correct answer: C.**

To compute the double sum $\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} g(x, y)$, we evaluate $g(x, y) = x + y^2$ for each pair (x, y) where $x \in \{1, 2\}$ and $y \in \{1, 2, 3\}$.
Calculating each term:

$$\begin{aligned} g(1,1) &= 1 + 1^2 = 2, \\ g(1,2) &= 1 + 2^2 = 5, \\ g(1,3) &= 1 + 3^2 = 10, \\ g(2,1) &= 2 + 1^2 = 3, \\ g(2,2) &= 2 + 2^2 = 6, \\ g(2,3) &= 2 + 3^2 = 11. \end{aligned}$$

Summing these values:

$$2 + 5 + 10 + 3 + 6 + 11 = 37.$$

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. $\mathcal{X}$

B. $\mathcal{X}^3$

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

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

Solution 2. **Correct answer: B., C.**

Since each r.v. $X_1, X_2, X_3 \in \mathcal{X}$, the outcome space is $\mathcal{X}^3$, which is by definition option C.

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

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

Solution 3. **Correct answer: B.**

Starting from the definition of the weighted loss:

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

Taking the gradient with respect to $\mathbf{w}$:

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

Since $\nabla \ell(\mathbf{w}) = (\sigma(\mathbf{x}_i^\top \mathbf{w}) - y_i)\mathbf{x}_i$, we have:

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

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?

A. The expected value of X is 5.

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

C. $p(10) = 1/10$

D. The probability that X is between 4 and 10 is $2/3$.

Solution 4. **Correct answer: D.**

A. **False.** $\int_1^{10} x \cdot (1/9) dx = 5.5$.

B. **False.** $\mathbb{P}(X = 10) = 0$ for continuous variables.

C. **False.** $p(10) = 1/(10 - 1) = 1/9$ by definition.

D. **True.** The probability is calculated as:

$$\mathbb{P}(4 \leq X \leq 10) = \frac{10 - 4}{9} = \frac{2}{3}.$$

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:

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

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

A. 2.0

B. 2.2

C. 1.8

D. 2.4

Solution 5. **Correct answer: B.**

Find $p(X = 1)$:

$$p(X=1) = p(X=1, N=1) + p(X=1, N=2) + p(X=1, N=3) = \frac{4}{25} + \frac{8}{25} + \frac{8}{25} = \frac{20}{25} = 0.8.$$

Compute the conditional pmf $p_{N|X}(n|1)$: For each $n \in \{1, 2, 3\}$,

$$p_{N|X}(n|1) = \frac{p(X=1, N=n)}{p(X=1)}.$$

Thus,

$$p_{N|X}(1|1) = \frac{4/25}{20/25} = \frac{1}{5}, \quad p_{N|X}(2|1) = \frac{8/25}{20/25} = \frac{2}{5}, \quad p_{N|X}(3|1) = \frac{8/25}{20/25} = \frac{2}{5}.$$

Compute $\mathbb{E}[N|X = 1]$:

$$\mathbb{E}[N|X=1] = \sum_{n=1}^{3} n \cdot p_{N|X}(n|1) = (1)\left(\frac{1}{5}\right) + (2)\left(\frac{2}{5}\right) + (3)\left(\frac{2}{5}\right).$$

$$= \frac{1}{5} + \frac{4}{5} + \frac{6}{5} = \frac{11}{5} = 2.2.$$

Correct Answer: B.

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?

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

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

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

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

Solution 6. **Correct Answers: A., C., D.**

A. **True.**

$$\mathrm{Var}(\bar{Z}) = \mathrm{Var}\left(\frac{1}{4}(Z_1 + Z_2 + Z_3 + Z_4)\right)$$
$$= \left(\frac{1}{4}\right)^2 (\mathrm{Var}(Z_1) + \mathrm{Var}(Z_2) + \mathrm{Var}(Z_3) + \mathrm{Var}(Z_4)) = \frac{1}{16}(9+9+9+9) = \frac{36}{16} = \frac{9}{4}$$

B. **False.**

$$\mathbb{E}[\bar{Z}] = \mathbb{E}\left(\frac{1}{4}(Z_1 + Z_2 + Z_3 + Z_4)\right) = \frac{1}{4}\left(\mathbb{E}[Z_1] + \mathbb{E}[Z_2] + \mathbb{E}[Z_3] + \mathbb{E}[Z_4]\right) = \frac{1}{4}(8+8+8+8) = 8 \neq 2$$

C. **True.** Each Z_i has variance $\sigma^2 = 9$.

D. **True.**

$$\mathbb{E}[\bar{Z}] = \mathbb{E}\left(\frac{1}{4}(Z_1 + Z_2 + Z_3 + Z_4)\right) = \frac{1}{4}(8 + 8 + 8 + 8) = 8$$

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?

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

Solution 7. **Correct Answer: A.**

For $A = 1$:

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

For $A = 2$:

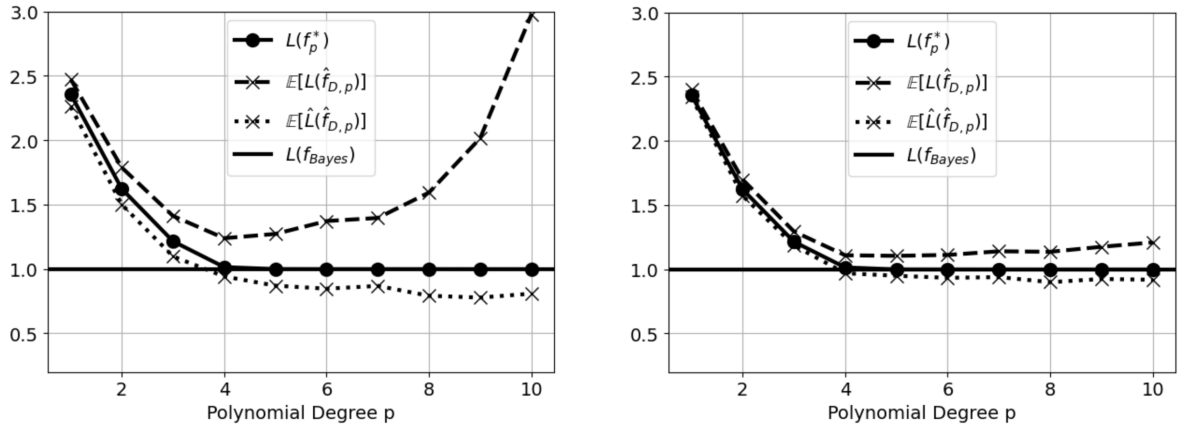
$$p_A(2) = p(2,0) + p(2,1) = \frac{1}{5} + \frac{1}{10} = \frac{2}{10} + \frac{1}{10} = \frac{3}{10}.$$

For $A = 3$:

$$p_A(3) = p(3,0) + p(3,1) = \frac{1}{10} + \frac{3}{10} = \frac{4}{10} = \frac{2}{5}.$$

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?



(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.

Solution 8. **Correct answer: A., B., C., D.**

A. **True.** In Fig 1, a low degree $p = 1$ leads to underfitting, and a high degree $p = 10$ results in overfitting.

B. **True.** Visually, the estimation error $\mathbb{E}[L(\hat{f}_{D,p})] - L(f_p^*)$ is smaller for $p = 2$ than for $p = 10$ in Fig 1.

C. **True.** Irreducible error is inherent to the data distribution and cannot be reduced by altering n or p.

D. **True.** The estimation error is $\mathbb{E}[L(\hat{f}_{D,p})] - L(f_p^*)$, which visually can be checked to be smaller in Fig 2 than in Fig 1 for all values of p.

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:

$$\hat{L}(f) = \frac{1}{4}\sum_{i=1}^{4} \ell(f(x_i), y_i).$$

A. 0.25

B. 0.3

C. 0.125

D. 0.5

Solution 9. **Correct Answer: B.**

First, compute $f(x_i)$ for each x_i:

$$f(1) = 5 + 0.5 \cdot 1 = 5.5$$
$$f(2) = 5 + 0.5 \cdot 2 = 6.0$$
$$f(3) = 5 + 0.5 \cdot 3 = 6.5$$
$$f(4) = 5 + 0.5 \cdot 4 = 7.0$$

Now compute the absolute losses:

$$\ell(f(1), y_1) = |5.5 - 5.5| = |0| = 0$$
$$\ell(f(2), y_2) = |6.0 - 5.0| = |1.0| = 1.0$$
$$\ell(f(3), y_3) = |6.5 - 6.5| = |0| = 0$$
$$\ell(f(4), y_4) = |7.0 - 7.2| = |-0.2| = 0.2$$

Summing these losses:

$$\sum_{i=1}^{4} \ell(f(x_i), y_i) = 0 + 1.0 + 0 + 0.2 = 1.2.$$

Taking the average:

$$\hat{L}(f) = \frac{1.2}{4} = 0.3.$$

Question 10. [1 MARK]

Consider the convex function

$$g(w_1, w_2) = w_1^2 + w_2^2 + w_1 w_2 + 10w_1,$$

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 $\min_{w_1,w_2\in\mathbb{R}} g(w_1, w_2)$.

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

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

Solution 10. **Correct answer: B.**

Step-by-step Solution:
We have:

$$g(w_1, w_2) = w_1^2 + w_2^2 + w_1 w_2 + 10w_1.$$

First, find the partial derivatives:

$$\frac{\partial g}{\partial w_1} = 2w_1 + w_2 + 10, \quad \frac{\partial g}{\partial w_2} = 2w_2 + w_1.$$

Set these equal to zero to find the stationary point:

$$2w_1 + w_2 + 10 = 0, \quad w_1 + 2w_2 = 0.$$

From the second equation:

$$w_1 + 2w_2 = 0 \implies w_1 = -2w_2.$$

Substitute $w_1 = -2w_2$ into the first equation:

$$2(-2w_2) + w_2 + 10 = 0 \implies -4w_2 + w_2 + 10 = 0 \implies -3w_2 + 10 = 0.$$

Solve for w_2:

$$-3w_2 = -10 \implies w_2 = \frac{10}{3}.$$

Given $w_2 = \frac{10}{3}$:

$$w_1 = -2\left(\frac{10}{3}\right) = -\frac{20}{3}.$$

Thus, the stationary point is:

$$(w_1^*, w_2^*) = \left(-\frac{20}{3}, \frac{10}{3}\right).$$

Now, evaluate g at this point:

$$g\left(-\frac{20}{3}, \frac{10}{3}\right) = \left(-\frac{20}{3}\right)^2 + \left(\frac{10}{3}\right)^2 + \left(-\frac{20}{3}\right)\left(\frac{10}{3}\right) + 10\left(-\frac{20}{3}\right).$$

Compute step-by-step:

$$\left(-\frac{20}{3}\right)^2 = \frac{400}{9}, \quad \left(\frac{10}{3}\right)^2 = \frac{100}{9}.$$

So,

$$g = \frac{400}{9} + \frac{100}{9} + \left(-\frac{200}{9}\right) + \left(-\frac{200}{3}\right).$$

Combine the fractions with denominator 9 first:

$$\frac{400}{9} + \frac{100}{9} - \frac{200}{9} = \frac{400 + 100 - 200}{9} = \frac{300}{9} = \frac{100}{3}.$$

Now:

$$g = \frac{100}{3} + \left(-\frac{200}{3}\right) = \frac{100 - 200}{3} = -\frac{100}{3}.$$

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?

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

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

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

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

Solution 11. **Correct answer: C.**

Step-by-step Solution:
Given:

$$g(w_1, w_2) = w_1^2 + w_2^2 + w_1 w_2 + 10 w_1.$$

First, we compute the gradient:

$$\nabla g(w_1, w_2) = \left(\frac{\partial g}{\partial w_1}, \frac{\partial g}{\partial w_2}\right).$$

Take partial derivatives:

$$\frac{\partial g}{\partial w_1} = 2w_1 + w_2 + 10, \quad \frac{\partial g}{\partial w_2} = 2w_2 + w_1.$$

Thus:

$$\nabla g(w_1, w_2) = (2w_1 + w_2 + 10,\ w_1 + 2w_2).$$

We use gradient descent with $\eta = 1$:

$$(w_1^{(t+1)}, w_2^{(t+1)}) = (w_1^{(t)}, w_2^{(t)}) - \eta \nabla g(w_1^{(t)}, w_2^{(t)}).$$

Iteration 0 (initialization):

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

Evaluate the gradient at $(0, 0)$:

$$\nabla g(0, 0) = (2(0) + 0 + 10,\ 0 + 2(0)) = (10, 0).$$

Update weights:

$$(w_1^{(1)}, w_2^{(1)}) = (0, 0) - 1 \cdot (10, 0) = (-10, 0).$$

Iteration 1: Evaluate the gradient at $(-10, 0)$:

$$\nabla g(-10, 0) = (2(-10) + 0 + 10,\ (-10) + 2(0)) = (-20 + 10,\ -10) = (-10, -10).$$

Update weights:

$$(w_1^{(2)}, w_2^{(2)}) = (-10, 0) - 1 \cdot (-10, -10) = (-10, 0) + (10, 10) = (0, 10).$$

After two iterations:

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

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 $1/\sqrt{2\pi}$.

C. The variance of Y is 2.

D. Y is a continuous random variable.

Solution 12. **Correct answer: A., C., D.**

A. **True.** The expected value of Y is 1 by definition.

B. **False.** The probability that $Y = 10$ is 0 for continuous variables.

C. **True.** The variance of Y is 2 by definition.

D. **True.** Y is a continuous random variable since it is normally distributed.

Question 13. [1 MARK]

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:

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

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

$$w^{(t+1)} = w^{(t)} - \frac{2}{n} \eta^{(t)} \left[\sum\nolimits_{i=1}^{n} c_i x_i (x_i w^{(t)} - y_i) + \lambda w^{(t)} \right].$$

C. The first-order gradient update rule is

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

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

Solution 13. Correct answer(s): A, C, D

Explanation:

A. **True.** Gradient descent, with a properly chosen step size, is designed to iteratively move towards a local minimum of the loss function. As the number of epochs increases, the algorithm should converge closer to the optimal $\hat{w}$. A good step size ensures convergence without overshooting the minimum.

B. **False.** This update rule is incorrect. It seems to be derived from a squared loss, not a fourth-power loss.

C. **True.** The gradient of the regularized loss function $\hat{L}_\lambda(w)$ with respect to w is:

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

The gradient descent update rule is:

$$w^{(t+1)} = w^{(t)} - \eta^{(t)} \nabla \hat{L}_\lambda(w^{(t)})$$

Substituting the gradient:

$$w^{(t+1)} = w^{(t)} - \frac{4}{n} \eta^{(t)} \left[\sum_{i=1}^{n} c_i x_i (x_i w^{(t)} - y_i)^3 + \lambda (w^{(t)})^3 \right]$$

D. **True.** A large value of λ heavily penalizes large values of w in the loss function. To minimize the loss, the algorithm will tend to find a $\hat{w}$ that is closer to 0, effectively shrinking the magnitude of the weight.

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 $\eta^{(t)} = \frac{1}{\hat{L}''_\lambda(w^{(t)})}$.

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

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

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

Solution 14. Correct answer(s): A, C

Explanation:

A. **True.** The second-order gradient descent (Newton's method) update rule is given by:

$$w^{(t+1)} = w^{(t)} - \left[\hat{L}''_\lambda(w^{(t)})\right]^{-1} \hat{L}'_\lambda(w^{(t)})$$

If we express this in a form similar to the first-order update rule:

$$w^{(t+1)} = w^{(t)} - \eta^{(t)} \hat{L}'_\lambda(w^{(t)})$$

then it's clear that setting $\eta^{(t)} = \frac{1}{\hat{L}''_\lambda(w^{(t)})}$ makes the second-order update equivalent to the first-order update with this specific step size.

B. **False.** This is an incorrect calculation of the second derivative.

C. **True.** To compute the second derivative, we first recall the first derivative from the previous question's solution:

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

Now, we differentiate this expression with respect to w:

$$\begin{aligned}\hat{L}''_\lambda(w) &= \frac{4}{n}\left[\sum_{i=1}^n 3 c_i x_i (x_i w - y_i)^2 x_i + 3\lambda w^2\right] \\ &= \frac{12}{n}\left[\sum_{i=1}^n c_i x_i^2 (x_i w - y_i)^2 + \lambda w^2\right]\end{aligned}$$

Therefore, the given statement is correct.

D. **False.** This expression does not correspond to the correct second derivative. It omits the term $(x_i w - y_i)^2$ within the summation, and also misses the factor of 12. It would be correct if the loss was quadratic.

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.

C. The estimated loss of a predictor f based on the M-th mini-batch is

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

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.

Solution 15. Correct answer(s): A, C.

A. **True.** We always form the first $M - 1$ mini-batches from b points each, and then we consider the last portion of the data. There are $n - (M - 1)b$ data points left. Since $M = \lfloor n/b \rfloor$, it follows that $n - (M - 1)b \geq b$, and at most it is less than $2b$. Thus, the last portion of the data (the M-th mini-batch under the new scheme) will contain at least b points and potentially more, but we never form an additional batch beyond these M batches: Therefore, there are always M mini-batches, no matter whether n is divisible by b or not.

B. **False.** If n is divisible by b, then $M = n/b$ and there are exactly M mini-batches of size b. There is no need to add a $(M + 1)$-th mini-batch. Thus, the claim in (2) that there are $M + 1$ mini-batches if n is divisible by b is incorrect.

C. **True.** Consider the M-th mini-batch. If n is divisible by b, this last mini-batch has exactly b points from indices $(M - 1)b + 1$ to $(M - 1)b + b = Mb$. In that case, $n = Mb$ and the normalization factor $n + b(1 - M)$ reduces to:

$$n + b(1 - M) = Mb + b(1 - M) = Mb + b - Mb = b.$$

So the formula matches the standard averaging over b points.

If n is not divisible by b, we have $n - Mb$ extra points appended to the M-th mini-batch, making it contain $b + (n - Mb) = n - (M - 1)b$ points. Substituting into the proposed formula:

$$n + b(1 - M) = n + b - bM = n - b(M - 1).$$

This is exactly the size of the M-th mini-batch in the new scheme. Therefore, the given expression correctly averages over all data points in the M-th mini-batch.

D. **False.** The variance of the sample mean of k i.i.d. datapoints is $\text{Var}(X)/k$, where X is the random variable representing the loss of a single data point. The variance of the estimated loss based on the M-th mini-batch is the variance of the sample mean of $b+n-Mb$ i.i.d. data points. This variance is $\text{Var}(X)/(b+n-Mb)$. The variance of the estimated loss based on any of the other mini-batches is the variance of the sample mean of b i.i.d. data points, which is $\text{Var}(X)/b$. Since $b+n-Mb > b$, the variance of the estimated loss based on the M-th mini-batch is smaller than the variance of the estimated loss based on any of the other mini-batches.

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**
    $\nabla \hat{L}(\mathbf{w}) \leftarrow \frac{2}{b} \sum_{i=(m-1)b+1}^{mb} (\mathbf{x}_i^\top \mathbf{w} - y_i)\mathbf{x}_i$
    $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \hat{L}(\mathbf{w})$
  $\nabla \hat{L}(\mathbf{w}) \leftarrow \frac{2}{n-Mb} \sum_{i=Mb+1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)\mathbf{x}_i$
  $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \hat{L}(\mathbf{w})$
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.

Solution 16. **Correct answer: C.**

A. **False.** The pseudocode is incorrect as described in C.

B. **False.** Since a constant step size is used, there is no need to update the step size at each epoch.

C. **True.** The gradient should be

$$\nabla \hat{L}(\mathbf{w}) = \frac{2}{n + b(1-M)} \sum_{i=(M-1)b+1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)\mathbf{x}_i \quad \text{for } m = M.$$

D. **False.** The gradient update only needs to occur M times, this is not the issue.

Question 17. [1 MARK]

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 $w \in (0, \infty)$?

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

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.

Solution 17. **Correct answer: B.**

Given the function $g(w) = e^{3w} - 2\log(w^2)$, where $w \in (0, \infty)$. Since $w > 0$, we can rewrite the function as $g(w) = e^{3w} - 4\log(w)$.
First, we find the first derivative $g'(w)$:

$$g'(w) = \frac{d}{dw}(e^{3w} - 4\log(w)) = 3e^{3w} - \frac{4}{w}.$$

Next, we find the second derivative $g''(w)$:

$$g''(w) = \frac{d}{dw}(3e^{3w} - \frac{4}{w}) = 9e^{3w} + \frac{4}{w^2}.$$

Now, let's analyze the convexity. Since $w \in (0, \infty)$, we have $w^2 > 0$, and thus $\frac{4}{w^2} > 0$. Also, $e^{3w} > 0$ for all w, so $9e^{3w} > 0$. Therefore, $g''(w) = 9e^{3w} + \frac{4}{w^2} > 0$ for all $w \in (0, \infty)$.
Since the second derivative $g''(w)$ is strictly positive for all w in the given domain, the function $g(w)$ is strictly convex on $(0, \infty)$.

Question 18. [1 MARK]

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 $-\log(p(\mathcal{D}|\alpha)$?
Hint: Recall the logarithm properties: $\log(a^b) = b\log(a)$ and $\log(ab) = \log(a) + \log(b)$.

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

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

Solution 18. Correct Answer: A., B.

The probability of each flip z_i is $p(z_i|\alpha) = \alpha^{z_i}(1 - \alpha)^{1-z_i}$. The likelihood is:

$$p(\mathcal{D}|\alpha) = \prod_{i=1}^{n} \alpha^{z_i}(1 - \alpha)^{1-z_i}$$

The negative log-likelihood is:

$$\begin{aligned} -\log p(\mathcal{D}|\alpha) &= -\sum_{i=1}^{n} \log\left(\alpha^{z_i}(1 - \alpha)^{1-z_i}\right) \\ &= \sum_{i=1}^{n} (-z_i \log \alpha - (1 - z_i) \log(1 - \alpha)) \\ &= -\left(\sum_{i=1}^{n} z_i\right) \log \alpha - \left(n - \sum_{i=1}^{n} z_i\right) \log(1 - \alpha) \end{aligned}$$

Question 19. [1 MARK]

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

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

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

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

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

Solution 19. Correct Answer: D.

Differentiating and setting $\frac{d}{d\alpha}(-\log p(\mathcal{D}|\alpha)) = 0$, we find:

$$\begin{aligned}
&\frac{d}{d\alpha}(-\log p(\mathcal{D}|\alpha)) = -\frac{\sum_{i=1}^{n} z_i}{\alpha} + \frac{n - \sum_{i=1}^{n} z_i}{1-\alpha} = 0 \\
&\implies \frac{\sum_{i=1}^{n} z_i}{\alpha} = \frac{n - \sum_{i=1}^{n} z_i}{1-\alpha} \\
&\implies (1-\alpha) \sum_{i=1}^{n} z_i = \alpha(n - \sum_{i=1}^{n} z_i) \\
&\implies \sum_{i=1}^{n} z_i - \alpha \sum_{i=1}^{n} z_i = \alpha n - \alpha \sum_{i=1}^{n} z_i \\
&\implies \sum_{i=1}^{n} z_i = \alpha n \\
&\implies \alpha = \frac{1}{n} \sum_{i=1}^{n} z_i = \alpha_{\text{MLE}}
\end{aligned}$$

Question 20. [1 MARK]

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 $-\log(p(\mathcal{D}|\mu) \cdot p(\mu))$?

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

Solution 20. **Correct Answer: A.**

$$-\log(p(\mathcal{D}|\mu)\cdot p(\mu)) = -\log(p(\mathcal{D}|\mu)) - \log(p(\mu))$$

We can expand and simplify each term seperately. First, the likelihood term is

$$\begin{aligned}
\log(p(\mathcal{D}|\mu)) &= \log\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(z_i-\mu)^2}{2}\right)\right)\\
&= \sum_{i=1}^{n}\log\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(z_i-\mu)^2}{2}\right)\right)\\
&= \sum_{i=1}^{n}\left(\log\left(\frac{1}{\sqrt{2\pi}}\right)-\frac{(z_i-\mu)^2}{2}\right)\\
&= n\log\left(\frac{1}{\sqrt{2\pi}}\right)-\sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2}.
\end{aligned}$$

Next, the prior term $\log(p(\mu))$ becomes

$$\begin{aligned}
\log(p(\mu)) &= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\mu-50)^2}{2\sigma^2}\right)\right)\\
&= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)-\frac{(\mu-50)^2}{2\sigma^2}.
\end{aligned}$$

Combining the two terms, we get

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

Question 21. [1 MARK]

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

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

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

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

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

Solution 21. **Correct Answer: B.**

The MAP solution is given by

$$\begin{aligned}\mu_{\text{MAP}} &= \arg\max_{\mu\in\mathbb{R}} p(\mu|\mathcal{D}) \\ &= \arg\min_{\mu\in\mathbb{R}} -\log(p(\mu|\mathcal{D})) \\ &= \arg\min_{\mu\in\mathbb{R}} -\log(p(\mathcal{D}|\mu)\cdot p(\mu)).\end{aligned}$$

Plugging our result from the previous question into the above expression, we get

$$\begin{aligned}\mu_{\text{MAP}} &= \arg\min_{\mu\in\mathbb{R}} \left[-n\log\left(\frac{1}{\sqrt{2\pi}}\right) + \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2} - \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{(\mu-50)^2}{2\sigma^2}\right] \\ &= \arg\min_{\mu\in\mathbb{R}} \left[\sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2} + \frac{(\mu-50)^2}{2\sigma^2}\right].\end{aligned}$$

The first equality holds since constants do not affect the minimization problem. To solve the optimization problem we can take the derivative of the last expression with respect to μ and set it to zero.

$$\begin{aligned}&\frac{d}{d\mu}\left[\sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2} + \frac{(\mu-50)^2}{2\sigma^2}\right] = 0 \\ \implies &-\sum_{i=1}^{n}(z_i-\mu) + \frac{(\mu-50)}{\sigma^2} = 0 \\ \implies &n\mu - \sum_{i=1}^{n} z_i + \frac{\mu}{\sigma^2} - \frac{50}{\sigma^2} = 0 \\ \implies &\mu_{\text{MAP}} = \frac{\sum_{i=1}^{n} z_i + \frac{50}{\sigma^2}}{n + \frac{1}{\sigma^2}}.\end{aligned}$$

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.

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

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

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

Solution 22. **Correct Answer: B., C., D.**

Recall the MAP estimate derived previously:

$$\mu_{\text{MAP}} = \frac{\sum_{i=1}^{n} z_i + \frac{50}{\sigma^2}}{n + \frac{1}{\sigma^2}}.$$

We analyze the behavior in different regimes:

A. **False. If σ^2 is large:** When $\sigma^2 \to \infty$, the terms $\frac{50}{\sigma^2} \to 0$ and $\frac{1}{\sigma^2} \to 0$. Therefore,

$$\mu_{\text{MAP}} \approx \frac{\sum_{i=1}^{n} z_i}{n}.$$

This means that if the prior variance is very large, the prior becomes less informative, and the MAP estimate approaches the sample mean. Thus, statement (A) is false because it incorrectly states that μ_{MAP} approaches 50 when σ^2 is large.

B. **True. If σ^2 is small:** When $\sigma^2 \to 0$, the term $\frac{50}{\sigma^2}$ dominates both the numerator and the denominator. Specifically,

$$\mu_{\text{MAP}} \approx \frac{50/\sigma^2}{1/\sigma^2} = 50.$$

Thus, if the prior is very confident (small variance), the MAP estimate is strongly pulled toward the prior mean of 50. This makes statement (B) true.

C. **True. If n is large:** As $n \to \infty$,

$$\mu_{\text{MAP}} = \frac{\sum_{i=1}^{n} z_i + \frac{50}{\sigma^2}}{n + \frac{1}{\sigma^2}}.$$

Because n grows large, the terms $\frac{50}{\sigma^2}$ and $\frac{1}{\sigma^2}$ become negligible:

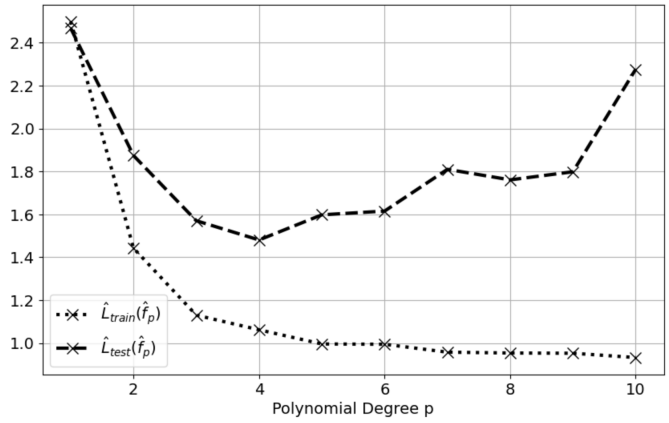
$$\mu_{\text{MAP}} \approx \frac{\sum_{i=1}^{n} z_i}{n}.$$

Thus, for large n, the MAP estimate is approximately the sample mean, making statement (C) true.

D. **True. If σ^2 is large:** As established in point (1), when σ^2 is large, the MAP estimate approaches the sample mean:

$$\mu_{\text{MAP}} \approx \frac{\sum_{i=1}^{n} z_i}{n}.$$

Therefore, statement (D) is true.



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?

A. The training 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 $p = 1$.

Solution 23. **Correct answer: B., C.**

A. **False.** Test loss is a better estimate of $L(\hat{f}_p)$ since $\hat{f}_p$ is chosen independent of the test dataset. $\hat{f}_p$ depends on the training dataset, so the training loss is not a good estimate of $L(\hat{f}_p)$.

B. **True.** From the plot for $p = 1$ the AE is likely high and the EE is likely low, indicating underfitting. For $p = 10$ the AE is likely low and the EE is likely high, indicating overfitting.

C. **True.** A higher p increases the function class $\mathcal{F}_p$, allowing lower training loss.

D. **False.** The AE is likely smallest for $p = 10$ since a higher p means a larger function class $\mathcal{F}_p$. The EE is likely smallest for $p = 1$ since a lower p means a smaller function class $\mathcal{F}_p$.

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

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

Which of the following statements are true?

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 $f \notin \mathcal{F}_{10}$.

C. If $d = 3$ then $\phi_2(\mathbf{x}) = (1, x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2x_3) \in \mathbb{R}^8$.

D. $\bar{p}$ increases as p increases.

Solution 24. **Correct answer: A., D.**

Explanation:

A. **True.** Since $\mathcal{F}_1 \subset \mathcal{F}_{10}$.

B. **False.** Since $\mathcal{F}_5 \subset \mathcal{F}_{10}$, every function in $\mathcal{F}_5$ is also in $\mathcal{F}_{10}$.

C. **False.** For $d = 3$, the polynomial feature map $\phi_2(\mathbf{x})$ of degree 2 includes all monomials up to degree 2 in variables x_1, x_2 and x_3. The total number of such monomials is:

Degree 0: 1 term (constant term)

Degree 1: 3 terms (x_1, x_2, x_3)

Degree 2: 6 terms (x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3)

Total terms: $1 + 2 + 3 + 4 = 10$ terms.

Therefore, the $\phi_2(\mathbf{x})$ given is incorrect since it only has 8 terms, and $x_1x_2x_3$ shouldn't even be part of it.

D. **True.** The number of terms in the polynomial feature map $\phi_p(\mathbf{x})$ of degree p is $\binom{d+p}{p}$, which increases as p increases. Alternatively this has to be the case since we know $\mathcal{F}_p \subset \mathcal{F}_{p+1}$.

Question 25. [1 MARK]

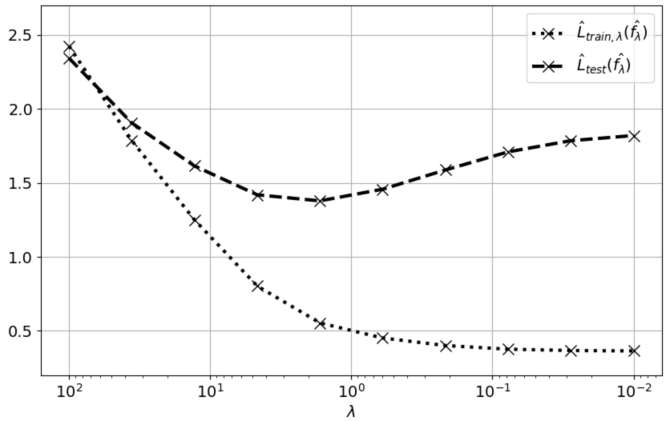
You are trying to decide which regularization parameter value λ to use for a closed-form polynomial regression learner with degree $p = 10$. You have a dataset of size n which you split into a training set and a test set as follows:

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

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?



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 $\lambda = 0.01$.

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

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

Solution 25. **Correct answer: A., D.**

A. **True.** A large λ causes for the predictor to likely be simpler and increasing bias.

B. **False.** Since the function class is fixed to $\mathcal{F}_{10}$ the approximation error is the same for all λ. However, the bias changes.

C. **False.** When λ is large, the predictor is simpler and the variance is likely lower not higher. The opposite is true when λ is small.

D. **True.** $\lambda \approx 2$ is the best choice since the test loss is the smallest there.

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

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

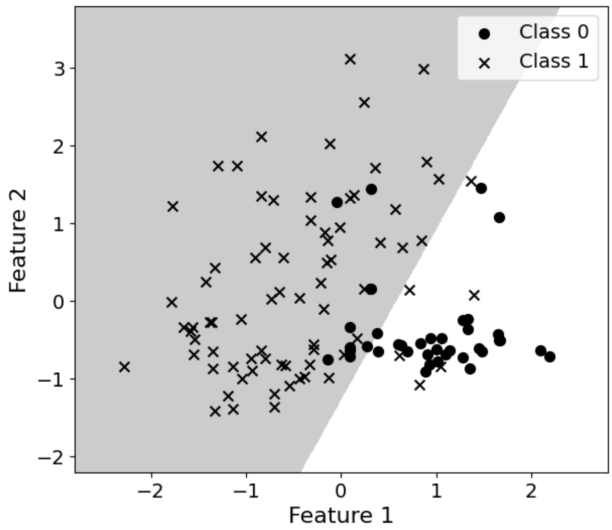
and the binary cross-entropy loss function

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

Suppose you use ERM and get the exact solution

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

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?



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 or equal to the area of the grey region in the figure.

Solution 26. **Correct answer: A., C.**

A. **True.** Since $\sigma(0) = 0.5 = c$.

B. **False.** See answer C.

C. **True.** Since there are much more Xs in the grey region than in the white region f_{ERM} would have been larger than 0.5 in the grey region, which implies $\mathbf{x}^\top \hat{\mathbf{w}}_{\text{ERM}} \geq 0$.

D. **False.** Since the grey region represents when $f_{\text{ERM}}(\mathbf{x}) \geq 0.5$, as discussed in the previous answer, the grey region would smaller or equal to the area of the grey region in the figure. This is because any features $\mathbf{x}_i$ that for which $0.5 \leq f_{\text{ERM}}(\mathbf{x}_i) < 0.7$ would now be classified as 0.

Question 27. [1 MARK]

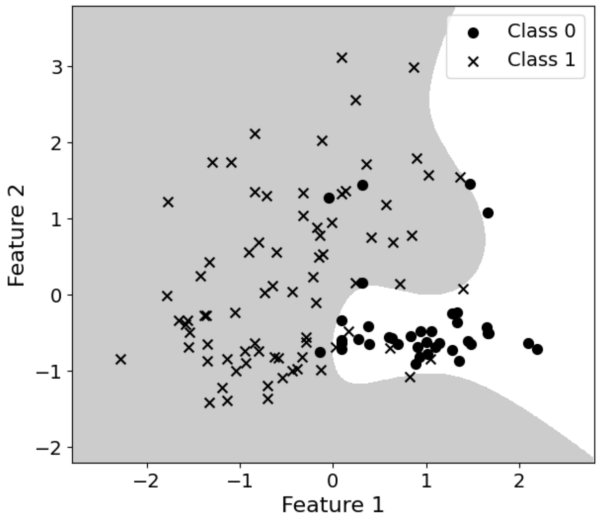
Let everything be as defined in the previous question. Suppose you decide to change to the polynomial function class to

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

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?

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 $f_{\text{Bin}}(\mathbf{x})$.

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 $\phi_p(\mathbf{x})^\top \hat{\mathbf{w}}_{\text{ERM},p} = 0$.

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

Solution 27. Correct answer: A., B., D.

A. **True.** Since $\mathcal{F}_1 = \mathcal{F}$.

B. **True.** See the explanation for D. in the previous question.

C. **False.** Since $\sigma(0) = 0.5 \neq 0.7 = c$.

D. **True.** Since $\sigma(0) = 0.5 = c$.

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?

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

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

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 the data than the classifier $f_{\text{Bin},10}$.

Solution 28. Correct answer: A., B., C.

A. **True.** By the discussion on option B. it means that $\hat{L}(f_{\text{Bin}}) > \hat{L}(f_{\text{Bin},p})$ is equivalent to $m_{\text{Bin}} > m_{\text{Bin},p}$, which we know is true.

B. **True.** The zero-one loss is 1 if $f_{\text{Bin}}(\mathbf{x}_i) \neq y_i$ and 0 otherwise. Thus, $\hat{L}(f_{\text{Bin}}) = m_{\text{Bin}}/n$.

C. **True.** Since the grey region represents when $f_{\text{Bin}}(\mathbf{x}) \geq 0.5$, and the white region represents when $f_{\text{Bin}}(\mathbf{x}) < 0.5$, the number of circles in the grey region and the number of Xs in the white region would be m_{Bin}.

D. **False.** The classifier $f_{\text{Bin},10}$ is more likely to overfit the data than the classifier f_{Bin}, since as p increases the estimation increase and the approximation error will likely decrease.

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

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

Which of the following are true?

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

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

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

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

Solution 29. **Correct answer: A., B.**

A. **True.** Every NN $f \in \mathcal{F}_{\text{NN}}^{(1)}$ has the form $f(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x})$, which is the same as the function class $\mathcal{F}$.

B. **True.** If $d = 2$, take for instance the function $f(\mathbf{x}) = \sigma(\mathbf{x}^\top \mathbf{w})$ where $\mathbf{w} = (1, 0, 0)^\top$. This function is in $\mathcal{F}$ but not in $\mathcal{F}_{\text{NN}}^{(2)}$.

C. **False.** Since A. holds.

D. **False.** Since A. and B. hold there is a function in $\mathcal{F}_{\text{NN}}^{(1)}$ that is not in $\mathcal{F}_{\text{NN}}^{(2)}$.

Question 30. [1 MARK]

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

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

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

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

You define your multiclass classifier as

$$f_{\text{Mul}}(\mathbf{x}) = \arg\max_{k \in \{0,\ldots,K-1\}} \hat{y}_k \quad \text{where} \quad \hat{\mathbf{y}} = \hat{f}_{\text{ERM}}(\mathbf{x}).$$

Which of the following are true?

A. $\sum_{q=0}^{K-1} \hat{y}_q = 1$.

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 $q \in \{0, \ldots, K-1\}$.

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

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

Solution 30. **Correct answer: A., B., C.**

A. **True.** Since softmax is used in the definition of $\hat{f}_{\text{ERM}}$.

B. **True.** By the definition of f_{Mul}.

C. **True.** If $\sigma_q(\mathbf{x}^\top \hat{\mathbf{w}}_0, \ldots, \mathbf{x}^\top \hat{\mathbf{w}}_{K-1}) > 0.5$ then $f_{\text{Mul}}(\mathbf{x}) = q$, thus it also holds for 0.7, which is a stricter condition.

D. **False.** Suppose that $\mathbf{x} = (1,1)^\top$ and $\hat{\mathbf{w}}_0 = (1,0)^\top, \hat{\mathbf{w}}_1 = (0,1)^\top$. Then $\sigma_0(\mathbf{x}^\top \hat{\mathbf{w}}_0, \mathbf{x}^\top \hat{\mathbf{w}}_1) = \sigma_1(\mathbf{x}^\top \hat{\mathbf{w}}_0, \mathbf{x}^\top \hat{\mathbf{w}}_1)$, but $\hat{\mathbf{w}}_0 \neq \hat{\mathbf{w}}_1$.

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. $\sigma(z_1, z_2) = \sigma(z_1 - z_2)$.

B. $\sigma_1(z_1, z_2) = \sigma(z_1)$.

C. $\sigma_1(z_1, z_2) = \sigma(z_1 - z_2)$.

D. $\sigma_2(z_1, z_2) = \sigma(z_2 - z_1)$.

Solution 31. Correct answer: C., D.

A. **False.** $\sigma(z_1, z_2) \in \mathbb{R}^2$, while $\sigma(z_1 - z_2) \in \mathbb{R}$.

B. **False.** Since C. is True.

C. **True.** Since

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

D. **True.** By same logic as 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

C. 1264

D. 260

Solution 32. Correct answer: C.

Each layer has a bias neuron except the output layer. Thus there are

$$1001 + 101 + 101 + 51 + 10 = 1264$$

neurons in the network.

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

C. 115760

D. 115510

Solution 33. **Correct answer: C.**

First layer: $100 \cdot 1001 = 100100$ weights (100 neurons with 1000 input features and 1 bias term).
Second layer: $100 \cdot 101 = 10100$ weights (100 neurons with 100 input features and 1 bias term).
Third layer: $50 \cdot 101 = 5050$ weights (50 neurons with 100 input features and 1 bias term).
Fourth layer: $10 \cdot 51 = 510$ weights (10 neurons with 50 input features and 1 bias term).
Total number of weights: $100100 + 10100 + 5050 + 510 = 115760$.

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

B. $(3, 0)^\top$

C. 3

D. 0

Solution 34. **Correct answer: B.**

The activations for the first layer are:

$$a_1^{(1)} = \text{ReLU}(\mathbf{x}^\top \mathbf{w}_1^{(1)}) = \text{ReLU}(1 \cdot (-1) + 1 \cdot (-1) + 1 \cdot (-1)) = 0,$$
$$a_2^{(1)} = \text{ReLU}(\mathbf{x}^\top \mathbf{w}_2^{(1)}) = \text{ReLU}(1 \cdot 0 + 1 \cdot 1 + 1 \cdot 1) = 2.$$

The activations for the second layer are:

$$a_1^{(2)} = \text{ReLU}(\mathbf{a}^{(1)\top} \mathbf{w}_1^{(2)}) = \text{ReLU}(1 \cdot 1 + 0 \cdot 1 + 2 \cdot 1) = 3,$$
$$a_2^{(2)} = \text{ReLU}(\mathbf{a}^{(1)\top} \mathbf{w}_2^{(2)}) = \text{ReLU}(1 \cdot 0 + 0 \cdot 1 + 2 \cdot 0) = 0.$$

Thus, $f(\mathbf{x}) = (3, 0)^\top$.

Question 35. [1 MARK]

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

A. $B = 1$, $d^{(1)} = 1$, $h^{(1)}(z) = z$, $\mathbf{w}_1^{(1)} = (1, 1, -1, 1)^\top$.

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

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

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

Solution 35. Correct answer: B., C., D.

A. **False.** $1 + x_1 \cdot 1 + x_2 \cdot (-1) + x_3 \cdot 1 \neq x_1 - x_2 + x_3$.

B. **True.** $0 + x_1 \cdot 1 + x_2 \cdot (-1) + x_3 \cdot 1 = x_1 - x_2 + x_3$.

C. **True.**

$$
\begin{aligned}
a_1^{(1)} &= 0 + x_1 \cdot 1 + x_2 \cdot (-1) + x_3 \cdot 0 = x_1 - x_2 \\
a_2^{(1)} &= 0 + x_1 \cdot 0 + x_2 \cdot 0 + x_3 \cdot 1 = x_3 \\
a_1^{(2)} &= 0 + (x_1 - x_2) \cdot 1 + x_3 \cdot 1 = x_1 - x_2 + x_3.
\end{aligned}
$$

D. **True.**

$$
\begin{aligned}
a_1^{(1)} &= 0 + x_1 \cdot 1 + x_2 \cdot 1 + x_3 \cdot 1 = x_1 + x_2 + x_3 \\
a_2^{(1)} &= 0 + x_1 \cdot 0 + x_2 \cdot (-2) + x_3 \cdot 0 = -2x_2 \\
a_1^{(2)} &= 0 + (x_1 + x_2 + x_3) \cdot 1 + (-2x_2) \cdot 1 = x_1 - x_2 + x_3.
\end{aligned}
$$

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 $\log(ab) = \log(a) + \log(b)$.

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

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

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

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

Solution 36. **Correct answer: A., B.**

A. **True.**

$$a_1^{(1)} = \log(0 + x_1 \cdot 1) = \log(x_1)$$
$$a_1^{(2)} = \exp(0 + \log(x_1) \cdot 2) = \exp(\log(x_1^2)) = x_1^2.$$

B. **True.**

$$a_1^{(1)} = \log(0 + x_1 \cdot 1) = \log(x_1)$$
$$a_2^{(1)} = \log(0 + x_1 \cdot 1) = \log(x_1)$$
$$a_1^{(2)} = \exp(0 + \log(x_1) \cdot 1 + \log(x_1) \cdot 1) = \exp(\log(x_1) + \log(x_1)) = x_1^2.$$

C. **False.** $\log(0 + x_1 \cdot 2) = \log(2x_1) \neq \log(x_1^2)$.

D. **False.** $f(\mathbf{x}) = (a_1^{(1)}, a_2^{(1)})^\top \in \mathbb{R}^2$ which is not the right dimension.

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.

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.

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.

Solution 37. **Correct answer: B., C.**

A. **False.** For small numbers of epochs the neurons in the network will likely not represent any meaningful features since the weights are initialized randomly.

B. **True.** After a large number of epochs the NN will be close to the ERM solution which we know will likely learn to represent more complex features than the pixel values in the input layer.

C. **True.** See A.

D. **False.** The neurons in the first layer will likely learn to represent simpler features than the neurons in the second layer, since after a large number of epochs the NN is close to the ERM solution.

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 $\mathcal{Y} = \{\texttt{once}, \texttt{upon}, \texttt{a}, \texttt{time}, \texttt{.}, \texttt{<EOS>}, \texttt{<PAD>}\}$. 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 $s_3 = (\texttt{<PAD>}, \texttt{once}, \texttt{upon}, \texttt{a})$ and $y_3 = \texttt{time}$. 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`.

Solution 38. Correct answer: C., D.

A. **False.** The context length is $c = 4$ since s_3 contains 4 tokens.

B. **False.** The one-hot vector label $\mathbf{y}_3$ is $(0, 0, 1, 0, 0, 0, 0)^\top$ since `time` is represented by the integer 3.

C. **True.** The vocabulary contains the 7 tokens listed in the question.

D. **True.** By definition, if we have a sequence of tokens $s = (v_1, v_2, \ldots, v_n)$, then if $i \in \{1, \ldots, c-1\}$ we create the input-output pairs (s_i, y_i) where $s_i = (\texttt{<PAD>}, \ldots, \texttt{<PAD>}, v_1, \ldots, v_i) \in \mathcal{Y}^c$ and $y_i = v_{i+1}$. In this case since $s_3 = (\texttt{<PAD>}, \texttt{once}, \texttt{upon}, \texttt{a})$, that means i is 3 and $v_1 = \texttt{once}$, $v_2 = \texttt{upon}$, $v_3 = \texttt{a}$, and $v_4 = \texttt{time}$.

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.

Solution 39. Correct answer: A., D.

A. **True.** The feature vector $\mathbf{x}_3 = \bar{E}(s_3) \in \mathbb{R}^{d+1}$ where $d = cd'$. In this case $c = 4$ and $d' = 3$, so $d = 12$.

B. **False.** By the definition of how the dataset is created it can be checked that the number of feature-label pairs $(\mathbf{x}_i, \mathbf{y}_i)$ in the dataset is $a - 1$ since there always needs to be one token for the output.

C. **False.** Since $n = a - 1 = 9 \neq 4 = c$ this is not true.

D. **True.** Since there are 7 tokens in the vocabulary, the output vector would be 7 dimensional.

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 $E(\texttt{exams})$ you could calculate $E(\texttt{exam}) + E(\texttt{dog}) - E(\texttt{dogs})$.

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

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`.

Solution 40. Correct answer: B., C.

A. **False.** The direction of the vector $E(\texttt{dog}) - E(\texttt{dogs})$ represents something like the concept of making a word not plural, but we want to make `exam` plural.

B. **True.** Since $E(\texttt{dogs}) - E(\texttt{dog})$ likely represents the concept of making a word plural, we can add this to the embedding of `exam` to get an estimate of the embedding of `exams`.

C. **True.** Since the words `dog` and `wolf` are both animals, it is likely that their embeddings are closer to each other than to the embedding of `exam`.

D. **False.** Since the words `exam` and `wolf` are not related in meaning, it is likely that their embeddings are not close to each other.