ECE 411 Introduction to Machine Learning
Fall 2022 Exam 1
Instructor: Dr. Chau-Wai Wong

This is a closed-book exam. You may use a scientific calculator with cleared memory, but not a smart phone or computer. One-sided letter-sized handwritten cheatsheet is allowed. You should answer ***all four*** problems.

Problem 1 [Linear Statistical Model] (25 pts) An ECE student named Tom plans to test the fuel economy of his car in terms of how many gallons is needed for driving one mile. He will do four test drives of x_i miles each, $i = 1, \ldots, 4$, and will measure the corresponding gas consumption Y_i gallons, $i = 1, \ldots, 4$ using a meter connected to his car's microcontroller. Denote the ground-truth fuel economy as k gallon/mile.

(a) Tom believes that the readings of the gas consumption Y_i are inaccurate but unbiased, so he set up a linear model $Y_i = kx_i + e_i$, $i = 1, \ldots, 4$, where e_i are measurement noise with zero-mean and variance σ^2. Express this model in the matrix–vector form. Explicitly define $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\mathbf{e}$.

(b) Use the normal equation $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$ to directly obtain the analytic form of the least-squares estimator $\hat{k}$ for the fuel economy, and simplify $\hat{\boldsymbol{\beta}}$ up to a point that it cannot be further simplified.

(c) Mathematically show that $\hat{k}$ is unbiased. (Hint: x_i's are constants whereas Y_i's are random variables.)

(d) Tom's friend proposed another way to estimate the fuel economy: $\tilde{k} = \frac{1}{4}\sum_{i=1}^{4} \frac{Y_i}{x_i}$. Is $\tilde{k}$ unbiased? If yes, prove it. If not, explain why.

(e) Tom's friend went back and figured out yet a third way to estimate the fuel economy: $\check{k} = \left(\sum_{i=1}^{4} Y_i\right) \Big/ \left(\sum_{i=1}^{4} x_i\right)$. Is $\check{k}$ unbiased? If yes, prove it. If not, explain why.

Problem 2 [Linear Independence, Basis, and Vector Space] (20 pts)

(a) Are vectors $[1 \;\; 2], [4 \;\; 5]$, and $[7 \;\; 8]$ linearly independent? Justify your answers.

(b) You are given a vector space $V = \text{span}\{[-1 \;\; 0 \;\; 0], [0 \;\; -1 \;\; 0]\}$.

(i) Express V in a set representation.

(ii) Can you find a basis for V?

(iii) Are $[5 \;\; 8 \;\; 0]$, $[8 \;\; 0 \;\; 5]$, and $[0 \;\; 5 \;\; 8]$ in vector space V? Is yes, what are the coefficient for each vector of the basis you found in (ii)?

(c) Let

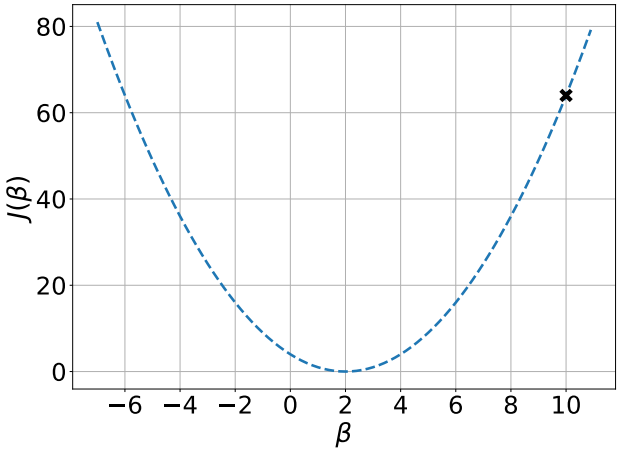
$$\mathbf{A} = \begin{bmatrix} 2 & 2 & -1 & 0 \\ -1 & -1 & 2 & -3 \\ 1 & 1 & -2 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

What is the dimension of the column vector space of $\mathbf{A}$?

Problem 3 [Deep Learning Concepts] (25 pts)

(a) A convolution neural network takes as input RGB pictures of size $512 \times 512 \times 3$. After passing through one convolutional layer, the output data is of size $512 \times 512 \times 10$. Explain what happened to the third dimension using no more than one complete sentence.

(b) Explain how transformer neural networks contextualize/"pay attention to" the embeddings of the sentence "walk by river bank" based on the initial embeddings that are not aware of the context. Keep your explanation simple by using no more than two sentences and ignore the projections to the "key," "value," and "query" semantic subspaces.



(c) Suppose we have a cost function/loss defined as $J(\beta) = (\beta - 2)^2$. Our goal is to use the gradient descent algorithm to find the best guess for β. We use an initial guess $\beta^{(0)} = 10$ and update $\beta^{(n)}$ via the update rule $\beta^{(n+1)} = \beta^{(n)} - \eta \cdot \frac{\partial J(\beta^{(n)})}{\partial \beta^{(n)}}$, where $\eta = 1$ is the learning rate/step size. Please (i) manually perform the first three steps of the gradient descent algorithm using the provided update rule, and (ii) mark the intermediate results $\left(\beta^{(1)}, J(\beta^{(1)})\right)$, $\left(\beta^{(2)}, J(\beta^{(2)})\right)$, and $\left(\beta^{(3)}, J(\beta^{(3)})\right)$ on the curve of the provided graph.

Problem 4 [Softmax in Neural Network Training] (25 pts) A softmax function $f : \mathbb{R}^K \to \mathbb{R}^K$, $K \geq 2$ is defined as follows:

$$f_i(\mathbf{x}) = \frac{\exp(\beta x_i)}{\sum_{k=1}^{K} \exp(\beta x_k)}, \quad i = 1, \ldots, K,$$

where x_i is the ith component of vector $\mathbf{x} \in \mathbb{R}^K$.

(a) Explain in your own words what the softmax function can do in machine learning. Please limit your explanation to at most two complete sentences.

(b) Derive the following two expressions:

$$\frac{\partial f_2(\mathbf{x})}{\partial x_2} = \beta f_2(\mathbf{x}) \left[1 - f_2(\mathbf{x})\right], \tag{1a}$$

$$\frac{\partial f_2(\mathbf{x})}{\partial x_j} = -\beta f_2(\mathbf{x}) f_j(\mathbf{x}), \text{ for } j \neq 2. \tag{1b}$$

[Recall that $(\frac{u}{v})' = \frac{u'v - uv'}{v^2}$.]

(c) When training a neural network for a classification task, the gradient update term will always contain at least one partial derivative of softmax as a multiplicative term. Based on the theoretical result in (b), could you explain what will happen to the training process if β is set to a number that is either too small or too large?

[This Page Intentionally Left Blank]