PURDUE

BME 64600 – 001 and ECE 60146 – 001

Midterm #1, Spring 2024

NAME ______________________________________

PUID ____

Exam instructions:

- You have 75 minutes to work the exam.
- This is a closed-book and closed-note exam. You may not use or have access to your book, notes, any supplementary reference, a calculator, or any communication device including a cell-phone or computer.
- You may not communicate with any person other than the official proctor during the exam.

To ensure Gradescope can read your exam:

- Write your full name and PUID above and on the top of every page.
- Answer all questions in the area designated for each problem.
- Write only on the front of the exam pages.
- DO NOT run over to the next question.

Name/PUID: ______________________________

Problem 1. (25pt) One Hot Encoding and the Simplex

Consider a problem in which X needs to represent the class of an image Y in which the three possible classes are $\{\text{chair}, \text{elephant}, \text{tree}\}$. You have the option of two possible encodings for the class.

Encoding A: $X \in \{0, 1, 2\}$ with $0 =$ chair, $1 =$ elephant, and $2 =$ tree.

Encoding B: $X \in \Re^3$ where $\sum_m X_m = 1$, and $X_0 = 1$, if chair; $X_1 = 1$, if elephant; and $X_2 = 1$, if tree.

a) What is Encoding B called?

Solution: One hot encoding

---

b) Give an advantage and a disadvantage of Encoding B over Encoding A.

Solution: The advantage of Encoding B is that it provides a better representation of each class since all classes are equally distant in this representation. The disadvantage of Encoding B is that it requires more store and memory since each value of X is a vector of dimension 3 rather than a scalar integer.

---

c) Give a mathematical explanation as to why Encoding B is better than Encoding A?

Solution: Let X^i and X^j be encodings of class i and j. For Encoding B, we have that

$$\|X^i - X^j\| = \delta(i - j) ,$$

but for Encoding A, we have that

$$\|X^i - X^j\| = |i - j| .$$

So in the second case, the difference depends on the specific classes.

---

d) For Encoding B, we say that $X \in \mathcal{S}$. State the name of the set $\mathcal{S}$, and give a precise mathematical definition for the set $\mathcal{S}$.

Solution: $\mathcal{S}$ is the Simplex, and it is defined by

$$\left\{ s \in \Re^P : \sum_{i=0}^{P-1} s_i = 1, \text{ and } \forall i, s_i \geq 0 \right\}$$

---

e) Prove that $\mathcal{S}$ is a convex set.

Solution: Let $a, b \in \mathcal{S}$, then select any $\lambda \in [0, 1]$. Then define

$$c = \lambda a + (1 - \lambda) b \ .$$

Then we need so show that c is also in the simplex. We can do this by showing

$$\sum_{i=0}^{P-1} c_i = \sum_{i=0}^{P-1} \{\lambda a_i + (1-\lambda) b_i\} = \lambda \sum_{i=0}^{P-1} a_i + (1-\lambda) \sum_{i=0}^{P-1} b_i = \lambda 1 + (1-\lambda) 1 = 1 \ ,$$

and

$$c_i = \lambda a_i + (1-\lambda) b_i \geq \lambda 0 + (1-\lambda) 0 = 0 \ .$$

Name/PUID: ______________________________

Problem 2. (25pt) Gradient of a Loss function

Consider a neural network with inference function $f_\theta(y)$ where $\theta \in \Re^p$ and $f_\theta : \Re^{N_y} \rightarrow \Re^{N_x}$, and loss function given by

$$L(\theta) = \frac{1}{K} \sum_{k=0}^{K-1} \|x_k - f_\theta(y_k)\|^2 \ ,$$

where $\{(x_k, y_k)\}_{k=0}^{K-1}$ are training pairs.

a) What is the shape of $A = \nabla_\theta f_\theta(y)$? What is the interpretation of the element $A_{i,j}$?

Solution: A is $N_x \times p$. The value $A_{i,j}$ has the interpretation of

$$A_{i,j} = \frac{\partial [f_\theta(y)]_i}{\partial \theta_j}$$

---

b) What is the shape of A^t? What is the interpretation of the element $[A^t]_{i,j}$?

Solution: A^t is $p \times N_x$. The value $A_{i,j}$ has the interpretation of

$$[A^t]_{i,j} = \frac{\partial [f_\theta(y)]_j}{\partial \theta_i}$$

---

c) Calculate an expression for $\nabla_\theta L(\theta)$.

Solution:

$$\nabla_\theta L(\theta) = -\frac{2}{K} \sum_{k=0}^{K-1} (x_k - f_\theta(y_k))^t \nabla_\theta f_\theta(y_k) = -\frac{2}{K} \sum_{k=0}^{K-1} (x_k - f_\theta(y_k))^t A$$

So therefore,

$$[\nabla_\theta L(\theta)]^t = -\frac{2}{K} \sum_{k=0}^{K-1} A^t (x_k - f_\theta(y_k))$$

---

d) For general A, how many multiplies are required to compute $\nabla_\theta L(\theta)$.

Solution: For each training sample indexed by k, the number of multiplications is $N_x \times P$. Then for K training samples, the number of multiplications is $K \times N_x \times P$. The final vector of shape $1 \times P$ is multiplied by $-2/K$, so the total number of multiplies is given by

$$\text{Total Multiplies} = K \times N_x \times P + P \ .$$

---

e) Consider the case when $A = \mathbf{1}\theta^t$, where $\mathbf{1} \in \Re^{N_x}$ is a column vector of 1's. Then how many multiplies are required to compute $\nabla_\theta L(\theta)$?

Solution: In this case, we have that

$$\begin{aligned}
\nabla_\theta L(\theta) &= -\frac{2}{K}\sum_{k=0}^{K-1}(x_k - f_\theta(y_k))^t A \\
&= -\frac{2}{K}\sum_{k=0}^{K-1}(x_k - f_\theta(y_k))^t \mathbf{1}\theta^t \\
&= -\frac{2}{K}\sum_{k=0}^{K-1}\left[(x_k - f_\theta(y_k))^t \mathbf{1}\right]\theta^t
\end{aligned}$$

Evaluation of each term in the sum requires P multiplications. (Here, multiplication by 1 is not counted as a multiplication.) Doing this for each of the K training samples requires KP multiplies. Finally, each of the P components of the resulting vector must be multiplied by $-2/K$. So the total number of multiplies is given by

$$\text{Total Multiplies} = (K+1)P \ .$$

Name/PUID: ______________________________

Problem 3. (25pt) Conditioning for Gradient Descent

Define the matrices

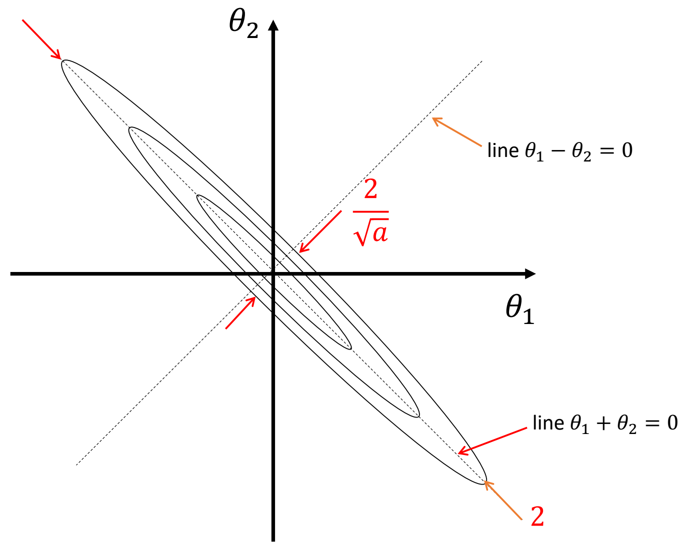
$$A = \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix}$$
$$\Sigma = \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix}$$
$$B = A^t \Sigma A \ .$$

where $a >> 1$ and $\phi = \pi/4$, (i.e. 45 deg). Then define the function $f(\theta) = \frac{1}{2}\theta^t B \theta$.
Also define the gradient descent algorithm as an iterative application of the following step:

$$\theta \leftarrow \theta + \alpha[-\nabla f(\theta)] \ .$$

a) Sketch the contours of the function $f(\theta)$. Label the key features of the plot.

Solution:



---

b) What is the condition number for this optimization problem?

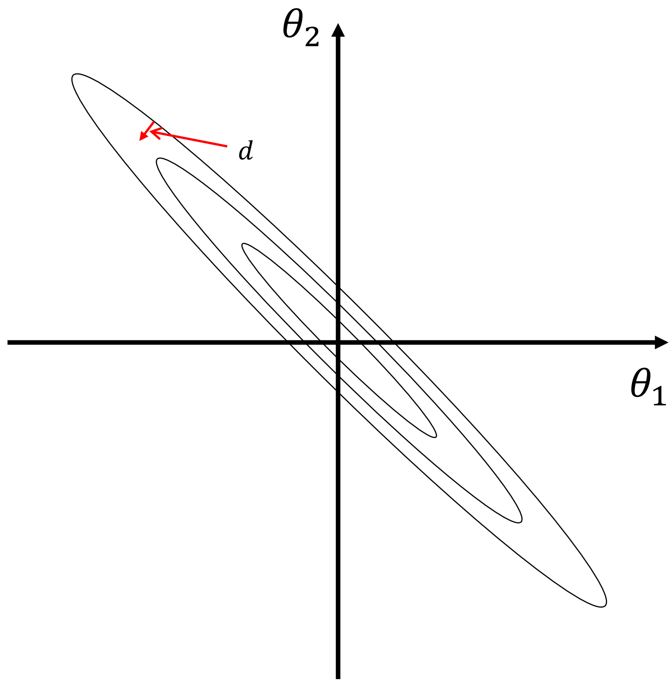
Solution: The condition number is a.

---

c) Calculate the negative gradient $d = -\nabla f(\theta)$. Draw another contour plot of f, and for a particular value of θ, draw the vector d on the plot.

Solution:

$$\nabla f(\theta) = B\theta$$



---

d) What is the largest value of α for which gradient descent is stable?

Solution: In order for gradient descent to be stable, we need that $\alpha < 2/a$. For values of $\alpha \geq 2/a$, the gradient descent algorithm will be unstable because along the 45 deg direction the solution will oscillate with increasing amplitude.

---

e) If $a = 10^6$ and you start gradient descent at $\theta = (1, 0)/\sqrt{2}$, what will happen?

Solution: In order to make convergence stable, the step size must be decreased so that $\alpha < 1/a$. However, this will make convergence very slow along the -45 deg axis. So each step will only move a small amount.

Name/PUID: ______________________________

Problem 4. (25pt) Convolution Blocks

A convolution block in a neural network can be represented by $x = f(y)$ where $y = [y_0, \cdots, y_N]$ is the input, $x = [0, \cdots, N-2]$ is the output for $N = 4$. Also it uses a 3-point convolution kernel of $w = [w_0, w_1, w_2]$ with the "valid" boundary condition and an offset of b. In this case, function can be written as

$$x = f(y) = y * w + b \ ,$$

where $*$ denotes conventional convolution. Also define the loss function

$$L(y) = \frac{1}{K} \sum_{k=0}^{K-1} \|x_k - f(y)\|^2 \ .$$

a) What is the shape of the gradient $A = \nabla_y f(y)$?

Solution: 3×5.

---

b) Write out an explicit expression for f in the form $f(y) = Ay + b$.

Solution:

$$f(y) = \begin{bmatrix} w_2 & w_1 & w_0 & 0 & 0 \\ 0 & w_2 & w_1 & w_0 & 0 \\ 0 & 0 & w_2 & w_1 & w_0 \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} + \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

---

c) Write out an explicit expression for the adjoint gradient, $A^t = [\nabla_y f(y)]^t$.

Solution:

$$A^t = \begin{bmatrix} w_2 & 0 & 0 \\ w_1 & w_2 & 0 \\ w_0 & w_1 & w_2 \\ 0 & w_0 & w_1 \\ 0 & 0 & w_0 \end{bmatrix}$$

---

d) Write out an explicit expression for the gradient of the loss function $\nabla_y L(y)$.

Solution:

$$\nabla_y L(y) = \frac{-2}{K} \sum_{k=0}^{K-1} (x_k - f(y))^t A \tag{1}$$

$$= \frac{-2}{K} \sum_{k=0}^{K-1} \begin{bmatrix} \epsilon_0 & \epsilon_1 & \epsilon_2 \end{bmatrix} \begin{bmatrix} w_2 & w_1 & w_0 & 0 & 0 \\ 0 & w_2 & w_1 & w_0 & 0 \\ 0 & 0 & w_2 & w_1 & w_0 \end{bmatrix} \tag{2}$$

$$[\nabla_y L(y)]^t = \frac{-2}{K} \sum_{k=0}^{K-1} A^t (x_k - f(y)) \tag{3}$$

$$= \begin{bmatrix} w_2 & 0 & 0 \\ w_1 & w_2 & 0 \\ w_0 & w_1 & w_2 \\ 0 & w_0 & w_1 \\ 0 & 0 & w_0 \end{bmatrix} \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \epsilon_2 \end{bmatrix} \tag{4}$$

$$\tag{5}$$

---

e) What is the interpretation of multiplication by A^t?

Solution: The interpretation is "same" boundary condition convolution with the time-reversed kernel, w_{2-n}, using an input that is padded with zeros at the first and last positions.

So in other words, it is convolution of $[0, \epsilon_0, \epsilon_1, \epsilon_2, 0]$ with the kernel $[w_2, w_1, w_0]$ using the "same" boundary condition.

Name/PUID: ______________________________

Problem 5. (25pt) Probability and Random Variables

a) Let X be a random variable. What precisely does $\{X \leq \lambda\}$ mean?

Solution: It means the event $A \subset \Omega$ defined by

$$A = \{\omega \in \Omega : X(\omega) \leq \lambda\}$$

---

b) Let X be a random variable. What precisely does $P\{X \leq \lambda\}$ mean?

Solution: It means

$$P(\{\omega \in \Omega : X(\omega) \leq \lambda\})$$

---

c) Let X, Y, Z be a random variables with Y and Z independent. Give a simplified expression for the following:

1. $E[Y|Z]$
2. $E[ZX|Z]$
3. $E[YZ]$

Solution:

$$E[Y|Z] = E[Y]$$
$$E[ZX|Z] = ZE[X|Z]$$
$$E[YZ] = E[Y]E[Z]$$

---

d) Consider that you use the cross entropy loss function to estimate θ in training the inference function $f_\theta(y)$ with training data $\{x_k, y_k\}_{k=0}^{K-1}$. What interpretation does minimization of the cross entropy loss function have in this case?

Solution: It is equivalent to maximum likelihood estimation of the parameter θ.

---

e) Let Y be a random variable with density $p_\theta(y)$ for $\theta \in \Re^P$, and let $\hat{\theta} = T(Y)$ be an estimator of θ. Define the bias of the estimator.

Solution:

$$\text{Bias} = E[\hat{\theta}|\theta] - \theta$$
