PURDUE

BME 64600 – 001 and ECE 60146 – 001

Midterm #1, Spring 2025

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.
- There are 32 sub-problems each worth 5pts, for a total score of 160.

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: Key

Problem 1. (35pt) Probability and Random Variables

Let X, Y, and Z be random variables such that $E[|X|] = E[|Y|] = E[|Z|] < \infty$ on the probability space $(\Omega, \mathcal{B}, P)$. Also, let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function.

a) Is X random? Is X a variable? What is X?

Solution: No, it is not random, and it is not a variable. X is a function, $X : \Omega \to \mathbb{R}$

---

b) Is $E[X]$ a random variable or number?

Solution: It is a number.

---

c) Is $E[X|Y]$ a random variable or number?

Solution: It is a random variable. In particular, it is a random variable with the form $Z = g(Y)$ for some measurable function g.

---

d) Is $f(X)$ a random variable or number?

Solution: Yes, $Z = f(X) = f(X(\omega))$, so it is a function of ω. So Z must be a random variable.

---

e) What is $E[X|X]$?

Solution: $E[X|X] = X$.

---

f) What is $E[E[Y|X]]$?

Solution: $E[E[Y|X]] = E[Y]$

---

g) Does $E[f(X)]$ always exist? Justify your answer.

Solution: No. For example, let $Z \sim p(z)$ where $p(z) = \frac{1}{(1+|z|)^3}$, and let $f(x) = x^3$, and define $W^+ = \max\{0, f(Z)\}$ and $W^- = \min\{0, f(Z)\}$. Then we have that

$$\begin{aligned} E[f(X)] &= E[W^+ + W^-] \\ &= E[W^+] + E[W^-] \\ &= \infty - \infty = \text{not defined} . \end{aligned}$$

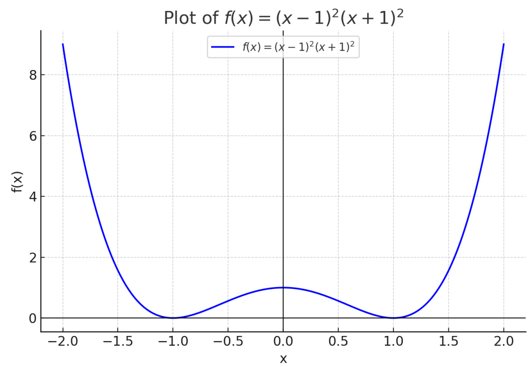
Name/PUID: ______________________________

Problem 2. (30pt) Convexity and Optimization

Consider the function $f : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = (x-1)^2(x+1)^2 \ .$$

a) Sketch the function $f(x)$ for $x \in [-2, 2]$.



Solution:

---

b) Is f convex? Justify your answer.

Solution: No. Take $a = 1$, $b = -1$, and $\lambda = 0.5$. Then we have that

$$\lambda f(a) + (1-\lambda) f(b) = 0 \leq f(\lambda a + (1-\lambda) b) = f(0) = 1 \ .$$

So f is not convex.

---

c) Does f have local minimum? If so, what are they?

Solution: Yes, it has local minimum at $x = 1$ and $x = -1$. This is true since $f(x) \geq 0$ and $f(1) = f(-1) = 0$. So $x = 1$ and $x = -1$ must be local minima.

---

d) Does f have a global minimum? Justify your answer.

Solution: Yes, since $f(x) \geq 0$ and $f(1) = f(-1) = 0$, then $x = 1$ and $x = -1$ must be global minima.

---

e) Does f have a **unique** global minimum? Justify your answer.

Solution: No, since both $x = 1$ and $x = -1$ are global minima, then the global minimum is not unique.

f) Does f have a saddle point? Justify your answer.

Solution: No, $x = 0$ is not a saddle point because it is a local maximum. So see this, notice that $\left.\frac{df(x)}{dx}\right|_{x=0} = 0$ and $\left.\frac{d^2 f(x)}{dx^2}\right|_{x=0} = -4$ and $f(0) = 1$.

****

However, the definition of saddle point in the class notes was wrong, so the following incorrect answer is also accepted.

Yes, $x = 0$ is a saddle point because its gradient is zero and it is not a local minimum.

Name/PUID: ______________________________

Problem 3. (35pt) Gradient Descent and Preconditioning

Consider the function

$$f(\theta) = \frac{1}{2}\theta^t H \theta \ ,$$

where $H = A^t A$ and $\theta \in \mathbb{R}^p$.
Our goals is to minimize this function using gradient descent optimization starting at $\theta_0 = \mathbf{1}$, where $\mathbf{1}$ denotes a vector of $1's$.

a) Calculate $\nabla f(\theta)$, the gradient of f at θ.

Solution:

$$\nabla f(\theta) = \theta^t H$$

---

b) Calculate $\nabla\nabla f(\theta)$, the Hessian of f at θ.

Solution:

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

---

c) Is f a convex function? Justify your answer.

Solution: Notice that $\forall\theta$, we have that

$$\theta^t H \theta = \theta^t A^t A \theta = \|A\theta\|^2 \geq 0 \ .$$

So therefore, H is non-negative definite, and f is convex.

---

d) Write the gradient descent update algorithm with a step size of $\alpha > 0$.

Solution:

$$\theta \leftarrow \theta - \alpha\left[\nabla f(\theta)\right] = \theta - \alpha H \theta$$

Important: For the remaining parts of the problem, assume that $A = diag\{a_0, \ldots, a_{p-1}\}$ such that $a_i^2 > a_{i+1}^2$.

e) Determine the largest value of α_{max} so that for all $0 < \alpha < \alpha_{max}$ gradient descent has guaranteed convergence.

Solution:

$$1 - \alpha_{max} a_0^2 \geq -1$$

So we have that

$$\alpha_{max} = \frac{2}{a_0^2}$$

---

f) If $a_0^2 >> a_{p-1}^2$, then will gradient descent have fast convergence? Justify your answer.

Solution: No, the convergence for the small eigenvalues will be slow. This is because

$$\begin{aligned}\theta_{p-1} &\leftarrow \theta_{p-1} - \alpha_{max} a_{p-1}^2 \theta_{p-1} \\ &\leftarrow \theta_{p-1}\left(1 - \frac{2a_{p-1}^2}{a_0^2}\right) \\ &\leftarrow \theta_{p-1}(1-\beta) \ ,\end{aligned}$$

where $\beta = \frac{2a_{p-1}^2}{a_0^2} << 1$.

---

g) What modification of gradient descent will have faster convergence? Be specific, and justify your answer.

Solution: You can speed convergence of gradient descent by using a preconditioner to adapt the step size for different components of θ. So this is

$$\theta \leftarrow \theta - \alpha M H \theta \ ,$$

where we choose $M = \text{diag}\{1/a_0^2, \ldots, 1/a_{p-1}^2\}$.

Name/PUID: ______________________________

Problem 4. (35pt) Forward and Backward Propagation Complexity

Our goal is to evaluate an expression for the vectors g and h given by

$$g = a^t BC$$
$$h = DEf \ ,$$

where $B, C, D, E \in \mathbb{R}^{p \times p}$ are matrices; $a, f \in \mathbb{R}^{p \times 1}$ are vectors, and $p >> 1$.
We call forward evaluation of these functions

$$g = a^t(BC))$$
$$h = D(Ef) \ ,$$

and we call backward evaluation of these functions

$$g = (a^t B)C$$
$$h = (DE)f \ .$$

a) Does forward and backward evaluation generate the same result for g and h? Justify your answer.

Solution: Yes, because multiplication is associative.

---

b) Give an expression for $\mathcal{FM}_g$ the number of multiples required for **forward** evaluation of g.
(Hint: Assume straight forward evaluation of the matrix vector products.)

Solution: Forward evaluation of g requires $\mathcal{FM}_g = p^3 + p^2$ multiplies.

---

c) Give an expression for $\mathcal{FM}_h$ the number of multiples required for **forward** evaluation of h.
(Hint: Assume straight forward evaluation of the matrix vector products.)

Solution: Forward evaluation of h requires $\mathcal{FM}_h = 2p^2$ multiplies.

---

d) Give expressions for $\mathcal{BM}_g$ and $\mathcal{BM}_h$, the number of multiples required for **backward** evaluation of g and h, respectively.

Solution: Backward evaluation of g requires $\mathcal{BM}_g = 2p^2$ multiplies. Backward evaluation of h requires $\mathcal{BM}_h = p^3 + p^2$ multiplies.

---

e) Which evaluation approach is best for g? Which evaluation approach is best for h?

Solution: Backward evaluation is best for g and forward evaluation is best for h.

---

f) Is there a general rule you can give for when to use forward evaluation and when to use backward evaluation?

Solution: Yes, use forward evaluation when the input dimension (on the right) is smaller than the output dimension (on the left). Use backward evaluation when the output dimension (on the left) is smaller than he input dimension (on the right).

---

g) Why is back propagation commonly used in optimization of deep neural networks?

Solution: Back propagation is commonly used because the output dimension is 1 for a loss function.

Name/PUID: ____________________________________

Problem 5. (25pt) Maximum Likelihood and Loss Functions

Define the open simplex, $\mathcal{S}$, as the following:

$$\mathcal{S} = \left\{ x \in \mathbb{R}^M : \forall i \in [0, \ldots, M-1], x_i > 0, \text{ and } \sum_{i=0}^{M-1} x_i = 1 \right\}$$

Then define the ground truth data $X_{k,i}$ for $0 \leq k < K$ and $0 \leq i < M$ to be one-hot encoded random variables were k denotes the training pair and i denotes the class.
Also define the cross-entropy loss function as

$$L(\theta; x) = \frac{1}{K} \sum_{k=0}^{K-1} \sum_{i=0}^{M-1} -x_{k,i} \log \theta_i \ ,$$

where $\theta \in \mathcal{S}$ is a parameter vector and x is a realization of X, i.e., x is not random.

a) Prove that the simplex is a convex set.

Solution: Let $a, b \in \mathcal{S}$, then $\forall i$, $a_i > 0$ and $b_i > 0$, and $\sum_i a_i = \sum_i b_i = 1$. So if we define

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

then we have that $c_i > 0$, and $\sum_i c_i = 1$. Q.E.D.

---

b) Is L is a convex function of θ on $\mathcal{S}$? Justify your answer.

Solution: Yes. To prove this, we will show that each term of the sum is convex. If we define, $f(z) = -x \log z$, then we have that

$$\frac{df(z)}{dz} = -x/z$$
$$\frac{d^2 f(z)}{dz^2} = x/z^2 \geq 0 \ .$$

So since $f(z)$ has positive second derivative for $z > 0$, it must be convex.

Important: For the remaining parts of the problem, assume that $X_{k,:}$ are independent and identically distributed (i.i.d.) for different values of k with

$$P_\theta\{X_{k,i} = 1\} = \theta_i$$

c) Calculate an expression for $l(\theta) = -\log P\{X = x\}$.

Solution:

$$
\begin{aligned}
l(\theta) &= -\log P\{X = x\} \\
&= \sum_{k=0}^{K-1} \sum_{i=0}^{M-1} -x_{k,i} \log \theta_i \\
&= K L(\theta; x)
\end{aligned}
$$

---

d) What is the relationship between minimizing the cross-entropy loss and the maximum likelihood estimate? Justify your answer.

Solution: Since $L(\theta; x) = \frac{1}{K} l(\theta)$, computing the arg min for L and l is the same. So the minimization of the cross-entropy loss results in the maximum likelihood estimate.

---

e) Calculate a closed from expression for the ML estimate of θ given X.[1]

Solution:

$$
\begin{aligned}
\hat{\theta} &= \arg\min_{\theta \in \mathcal{S}} l(\theta) \\
&= \arg\min_{\theta \in \mathcal{S}} \sum_{k=0}^{K-1} \sum_{i=0}^{M-1} -x_{k,i} \log \theta_i \ ,
\end{aligned}
$$

This results in

$$
\hat{\theta}_i = \frac{\sum_{k=0}^{K-1} x_{k,i}}{K} \ .
$$

---

[1] Assume that $\forall i \sum_k X_{k,i} > 0$ so that you don't have problems with the log.