ECE C147/C247, Winter 2025 **Homework #1**
Neural Networks & Deep Learning Prof. J.C. Kao
UCLA ECE TAs: B. Qu, K. Pang, S. Dong, S. Rajesh, T. Monsoor, X. Yan

Due Monday, 20 Jan 2025, by 11:59pm to Gradescope.
Covers material up to Introduction to machine learning refresher 1.
100 points total.

1. (25 points) **Linear algebra refresher.**

   (a) (12 points) Let $\mathbf{Q}$ be a real orthogonal matrix.

      i. (3 points) Show that $\mathbf{Q^T}$ and $\mathbf{Q^{-1}}$ are also orthogonal.
      ii. (3 points) Show that $\mathbf{Q}$ has eigenvalues with norm 1.
      iii. (3 points) Show that the determinant of $\mathbf{Q}$ is either +1 or -1.
      iv. (3 points) Show that $\mathbf{Q}$ defines a length preserving transformation.

   (b) (8 points) Let $\mathbf{A}$ be a matrix.

      i. (4 points) What is the relationship between the singular vectors of $\mathbf{A}$ and the eigenvectors of $\mathbf{A}\mathbf{A}^T$? What about $\mathbf{A}^T\mathbf{A}$?
      ii. (4 points) What is the relationship between the singular values of $\mathbf{A}$ and the eigenvalues of $\mathbf{A}\mathbf{A}^T$? What about $\mathbf{A}^T\mathbf{A}$?

   (c) (5 points) True or False. Partial credit on an incorrect solution may be awarded if you justify your answer.

      i. Every linear operator in an n-dimensional vector space has n distinct eigenvalues.
      ii. A non-zero sum of two eigenvectors of a matrix $\mathbf{A}$ is an eigenvector.
      iii. If a matrix $\mathbf{A}$ has the positive semidefinite property, i.e., $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x}$, then its eigenvalues must be non-negative.
      iv. The rank of a matrix can exceed the number of distinct non-zero eigenvalues.
      v. A non-zero sum of two eigenvectors of a matrix $\mathbf{A}$ corresponding to the same eigenvalue λ is always an eigenvector.

2. (25 points) **Probability refresher.**

   (a) (10 points) A and B are involved in a duel. The rules of the duel are that they are to pick up their guns and shoot at each other simultaneously. If one or both are hit, then the duel is over. If both shots miss, then they repeat the process. Suppose that the results of the shots are independent and that each shot of A will hit B with probability p_A and each shot of B will hit A with probability p_B. What is:

      i. (2 points) the probability that A is not hit?
      ii. (2 points) the probability that both duelists are hit?
      iii. (2 points) the probability that the duel ends after the n^{th} round of shots?

iv. (2 points) the conditional probability that the duel ends after the n^{th} round of shots given that A is not hit?

v. (2 points) the conditional probability that the duel ends after the n^{th} round of shots given that both duelists are hit?

(b) (5 points) Let X be a binary signal, such that $P(X = +1) = P(X = -1) = 0.5$. Suppose X is sent across a noisy channel, with noise N modeled by a zero-mean Gaussian distribution with variance σ^2, where the noise is independent of the signal that was sent. The received signal is $Y = X + N$.

i. (2 points) Find the conditional PDFs of Y given both $\{X = +1\}$ and $\{X = -1\}$.

ii. (2 points) Suppose a detector compares the received signal Y to a fixed threshold γ to decide which signal was sent. Specifically, the detector decides that $X = +1$ was sent if $Y \geq \gamma$ and that $X = -1$ was sent otherwise. For a given threshold γ, express the probability of error in terms of the Φ function, γ, and σ. Recall, if $Z \sim \mathcal{N}(0, 1)$, then

$$P(Z \leq z) = \Phi(z)$$

iii. (1 point) What is the optimal value of γ to minimize the probability of error? For that optimal value of γ, what is the probability of error? Leave your answer for the optimal value of γ as a real number and the probability of error in terms of the Φ function and σ.

(c) (5 points) There is a screening test for lung cancer that looks at the level of LSA (lung specific antigen) in the blood. There are a number of reasons besides lung cancer that a man can have elevated LSA levels. In addition, many types of lung cancer develop so slowly that they are never a problem. Unfortunately, there is currently no test to distinguish the different types and using the test is controversial because it's hard to quantify the accuracy rates and the harm done by false positives. For this problem, we will call a positive test a true positive if it catches a dangerous type of lung cancer. Also, we will assume the following numbers:

- Rate of dangerous type of lung cancer among men over 30 = 0.0005
- True positive rate for the test = 0.9
- False positive rate for the test = 0.01

Suppose you randomly select a man over 30 and perform a screening test.

i. (3 points) What is the probability that the man has a dangerous type of the disease given that he had a positive test?

ii. (2 points) What is the probability that the man has a dangerous type of the disease given that he had a negative test?

(d) (5 points) A family with three daughters and three sons needs to go to the grocery store. Besides the father, who is driving the car, exactly three of the children can come along to the grocery store with him. Suppose that the three children to join the father are chosen randomly, and all such choices are equally likely. Let X denote the number of daughters who accompany the father to the grocery. Let $X_i = 1$ if the i^{th} child who joins the father is a girl, and $X_i = 0$ otherwise.

i. (1 point) Find $Cov(X_i, X_i) = E(X_i X_i) - E(X_i)E(X_i)$, for $1 \leq i \leq 3$. Leave your answer as a fraction.

ii. (2 points) Find $Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$, for $1 \leq i, j \leq 3$ and $i \neq j$. Leave your answer as a fraction.

iii. (2 points) Find $Var(X)$. Leave your answer as a fraction.

3. (10 points) **Multivariate derivatives.**

(a) (1 points) Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$, and $\mathbf{A} \in \mathbb{R}^{n \times m}$. What is $\nabla_{\mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{y}$?

(b) (1 points) Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$, and $\mathbf{A} \in \mathbb{R}^{n \times m}$. What is $\nabla_{\mathbf{y}} \mathbf{x}^T \mathbf{A} \mathbf{y}$?

(c) (1 points) Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$, and $\mathbf{A} \in \mathbb{R}^{n \times m}$. What is $\nabla_{\mathbf{A}} \mathbf{x}^T \mathbf{A} \mathbf{y}$?

(d) (1 points) Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times n}$, and let $f = \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x}$. What is $\nabla_{\mathbf{x}} f$?

(e) (1 points) Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times n}$ and $f = \text{tr}(\mathbf{A}\mathbf{B})$. What is $\nabla_{\mathbf{A}} f$?

(f) (2 points) Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times n}$ and $f = \text{tr}(\mathbf{B}\mathbf{A} + \mathbf{A}^T \mathbf{B} + \mathbf{A}^2 \mathbf{B})$. What is $\nabla_{\mathbf{A}} f$?

(g) (3 points) Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times n}$ and $f = \|\mathbf{A} + \lambda \mathbf{B}\|_F^2$. What is $\nabla_{\mathbf{A}} f$?

4. (10 points) **Deriving least-squares with matrix derivatives.**
In least-squares, we seek to estimate some multivariate output $\mathbf{y}$ via the model

$$\hat{\mathbf{y}} = \mathbf{W}\mathbf{x}$$

In the training set we're given paired data examples $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ from $i = 1, \ldots, n$. Least-squares is the following quadratic optimization problem:

$$\min_{\mathbf{W}} \quad \frac{1}{2} \sum_{i=1}^{n} \left\| \mathbf{y}^{(i)} - \mathbf{W}\mathbf{x}^{(i)} \right\|^2$$

Derive the optimal $\mathbf{W}$.

Where $\mathbf{W}$ is a matrix, and for each example in the training set, both $\mathbf{x}^{(i)}$ and $\mathbf{y}^{(i)}$ $\forall i = 1, \ldots, n$ are vectors.
Hint: you may find the following derivatives useful:

$$\begin{aligned} \frac{\partial \text{tr}(\mathbf{W}\mathbf{A})}{\partial \mathbf{W}} &= \mathbf{A}^T \\ \frac{\partial \text{tr}(\mathbf{W}\mathbf{A}\mathbf{W}^T)}{\partial \mathbf{W}} &= \mathbf{W}\mathbf{A}^T + \mathbf{W}\mathbf{A} \end{aligned}$$

5. (10 points) **Regularized least squares**

In lecture, we worked through the following least squares problem

$$\arg\min_{\theta} \ \frac{1}{2} \sum_{i=1}^{N} (y^{(i)} - \theta^T \hat{\mathbf{x}}^{(i)})^2$$

However, the least squares has a tendency to overfit the training data. One common technique used to address the overfitting problem is regularization. In this problem, we work through one of the regularization techniques namely ridge regularization which is also known as the regularized least squares problem. In the regularized least squares we solve the following optimization problem

$$\arg\min_{\theta} \ \frac{1}{2}\sum_{i=1}^{N}(y^{(i)} - \theta^T \hat{\mathbf{x}}^{(i)})^2 + \frac{\lambda}{2}\|\theta\|_2^2$$

where λ is a tunable regularization parameter. From the above cost function it can be observed that we are seeking least squares solution with a smaller 2-norm. Derive the solution to the regularized least squares problem, i.e Find θ^*.

6. (20 points) **Linear regression.**
Complete the Jupyter notebook `linear_regression.ipynb`. Print out the Jupyter notebook as a PDF and submit it to Gradescope.