HOMEWORK 5
MATRIX CALCULUS [*]

10-606 MATHEMATICAL FOUNDATIONS FOR MACHINE LEARNING

START HERE: Instructions

- **Collaboration Policy**: Please read the collaboration policy in the syllabus.
- **Late Submission Policy:** See the late submission policy in the syllabus.
- **Submitting your work:** You will use Gradescope to submit answers to all questions.
  - **Written:** For written problems such as short answer, multiple choice, derivations, proofs, or plots, please use the provided template. Submissions can be handwritten onto the template, but should be labeled and clearly legible. If your writing is not legible, you will not be awarded marks. Alternatively, submissions can be written in LaTeX. Each derivation/proof should be completed in the boxes provided. To receive full credit, you are responsible for ensuring that your submission contains exactly the same number of pages and the same alignment as our PDF template.
  - **Latex Template:** https://www.overleaf.com/read/qwvphdsnkstf#127b9f

Question	Points
Scalar, Vector, & Matrix Derivatives	10
Using the Chain Rule on Cross-Entropy Loss	5
Matrix Derivatives for Weighted Linear Regression	12
Total:	27

[*] Compiled on Monday 22nd September, 2025 at 20:53

Instructions for Specific Problem Types

For "Select One" questions, please fill in the appropriate bubble completely:

Select One: Who taught this course?

● Matt Gormley

○ Marie Curie

○ Noam Chomsky

If you need to change your answer, you may cross out the previous answer and bubble in the new answer:

Select One: Who taught this course?

● Henry Chai

○ Marie Curie

⊗ Noam Chomsky

For "Select all that apply" questions, please fill in all appropriate squares completely:

Select all that apply: Which are scientists?

■ Stephen Hawking

■ Albert Einstein

■ Isaac Newton

□ I don't know

Again, if you need to change your answer, you may cross out the previous answer(s) and bubble in the new answer(s):

Select all that apply: Which are scientists?

■ Stephen Hawking

■ Albert Einstein

■ Isaac Newton

□ I don't know

For questions where you must fill in a blank, please make sure your final answer is fully included in the given space. You may cross out answers or parts of answers, but the final answer must still be within the given space.

Fill in the blank: What is the course number?

10-606

10-60~~6~~7

1 Scalar, Vector, & Matrix Derivatives (10 points)

Consider a constant scalar a, constant vectors $\mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{e}, \mathbf{w}$, variable scalar u, variable vectors $\mathbf{x}_i, i = 1...N$, and a variable matrix $\mathbf{X}$. Suppose

$$v = \frac{\exp(au)}{u^3} - 3u^2 + 2au + 5a - 10$$

$$z = \ln\left(\sum_{i=1}^{N} \exp(\mathbf{w}^T \mathbf{x}_i)\right)$$

$$y = a \ln(\exp(\mathbf{b}^T \mathbf{X} \mathbf{c}) + \exp(\mathbf{d}^T \mathbf{X} \mathbf{e}))\,.$$

Note that the derivative of a scalar y with respect to a vector $\mathbf{x}$ is a vector $\nabla = \frac{dy}{d\mathbf{x}}$ whose individual elements are derivatives $\frac{dy}{d\mathbf{x}_i}$. Similarly, the derivative of a scalar y with respect to a matrix $\mathbf{X}$ is a matrix $\nabla' = \frac{dy}{d\mathbf{X}}$ whose individual elements are derivatives $\frac{dy}{d\mathbf{X}_{ij}}$.

1. (2 points) Derive an expression for the derivative of v with respect to u i.e. $\frac{dv}{du}$ in terms of a and u.

2. (4 points) Derive an expression for the derivative of z with respect to $\mathbf{x}_i$ i.e. $\frac{dz}{d\mathbf{x}_i}$ in terms of $\mathbf{w}$ and $\mathbf{x}_i, \forall i$.

3. (4 points) Derive an expression for the derivative of y with respect to $\mathbf{X}$ i.e. $\frac{dy}{d\mathbf{X}}$ in terms of $a, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{e}$ and $\mathbf{X}$.

2 Using the Chain Rule on Cross-Entropy Loss (5 points)

In this question, you will compute the derivative of some functions related to the cross entropy loss. Cross entropy loss is commonly used to train neural networks for classification tasks. More specifically, you are given N examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ where $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \{0, \ldots, K-1\}$. This is a multi-class setting: each point $\mathbf{x}_i$ belongs to one of the K classes.

When $K = 2$, this is binary classification, and the cross-entropy loss is:

$$L(\mathbf{w}) = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log h(\mathbf{w}, \mathbf{x}_i) + (1 - y_i) \log \left(1 - h(\mathbf{w}, \mathbf{x}_i)\right) \right] \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^D$ and $\hat{y}_i = h(\mathbf{w}, \mathbf{x}_i)$ is our predicted probability that y_i is 1. We can predict $h(\mathbf{w}, \mathbf{x}_i)$ using any function of $\mathbf{x}_i$ and $\mathbf{w}$; for this problem we will use the following form:

$$h(\mathbf{w}, \mathbf{x}_i) = \frac{1}{1 + \exp\left(-g(\mathbf{w})^T \mathbf{x}_i\right)} \tag{2}$$

where g is a function applied element-wise to $\mathbf{w}$. In other words, the l-th entry of $g(\mathbf{w})$ is defined as $g(\mathbf{w}_l)$ where $\mathbf{w}_l$ is the l-th entry of $\mathbf{w}$. For this question, g is defined as $g(w) = w^2$.

1. (5 points) Write down the gradient $\nabla_{\mathbf{w}} L(\mathbf{w})$, when $L(\mathbf{w})$ is defined above in (1). Show your derivation in the first box and include the final result in the second box below. **Hint:** it might be helpful to compute the derivatives $\nabla_{\mathbf{w}} \hat{y}_i = \nabla_{\mathbf{w}} h(\mathbf{w}, \mathbf{x}_i)$ first. You can also "shape check" your answer, by considering what shape (i.e., vector or matrix dimensions) any intermediate derivatives should have.

Derivation

Final result

3 Matrix Derivatives for Weighted Linear Regression (12 points)

In statistics, linear regression is an approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). Relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. In this problem, we will try to fit a weighted regression model to a training dataset $(\mathbf{X}, \mathbf{Y}, \mathbf{w})$ of n datapoints where the input features $\mathbf{X} \in \mathbb{R}^{n\times m}$, the real-valued output $\mathbf{Y} \in \mathbb{R}^{n\times 1}$, and the non-negative real-valued datapoint weights $\mathbf{w} \in \mathbb{R}^{n\times 1}$.

Now, given a training set, we would like to learn a parameter $\boldsymbol{\theta}$, such that the function $h_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^T\mathbf{x}$ is close to y for the training examples we have. To formalize this, we will define a function that measures, for each value of $\boldsymbol{\theta}$, how close the $h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})$'s are to the corresponding $y^{(i)}$'s. We define the objective function using provided weights on the datapoints as follows:

$$J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n} \mathbf{w}^{(i)}(h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - \mathbf{y}^{(i)})^2$$

1. (3 points) Show that $J(\boldsymbol{\theta}) = \frac{1}{2}(\mathbf{X}\boldsymbol{\theta}-\mathbf{Y})^T\mathbf{W}(\mathbf{X}\boldsymbol{\theta}-\mathbf{Y})$ where $\mathbf{W}$ is a diagonal matrix with the weights from $\mathbf{w}$ placed on the diagonal and zeros in all off-diagonal entries. Show your work in the box below. (**Hint:** The i^{th} row of $\mathbf{X}$ is $\mathbf{x}^{(i)}$ and the i^{th} element of $\mathbf{y}$ is $\mathbf{y}^{(i)}$. Consider how multiplying the rows of $\mathbf{X}$ to $\boldsymbol{\theta}$ relates to $h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) = \boldsymbol{\theta}^T\mathbf{x}^{(i)}$. Carefully expand and simplify the matrix-vector notation to derive the weighted sum of squared errors version.)

2. (3 points) To minimize J, we need to find its gradient with respect to $\boldsymbol{\theta}$. Derive $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$. (**Hint:** You can use the following gradient formulae $\nabla_{\mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{x} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$ and $\nabla_{\mathbf{x}} \mathbf{x}^T \mathbf{y} = \nabla_{\mathbf{x}} \mathbf{y}^T \mathbf{x} = \mathbf{y}$)

3. (3 points) To minimize J, set its gradient to zero, and solve for $\boldsymbol{\theta}$.

Now suppose we used an l_2 regularizer. Now, our objective function is,

$$J'(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n} \mathbf{w}^{(i)}(h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - y^{(i)})^2 + \frac{1}{2}\lambda\|\boldsymbol{\theta}\|^2$$

1. (3 points) To minimize J', we need to find its gradient with respect to $\boldsymbol{\theta}$. Derive $\triangledown_{\boldsymbol{\theta}} J'(\boldsymbol{\theta})$. (**Hint:** You can use the following gradient formulae $\nabla_{\mathbf{x}}\mathbf{x}^T\mathbf{A}\mathbf{x} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$ and $\nabla_{\mathbf{x}}\mathbf{x}^T\mathbf{y} = \nabla_{\mathbf{x}}\mathbf{y}^T\mathbf{x} = \mathbf{y}$. Also, $\|\boldsymbol{\theta}\|^2 = \boldsymbol{\theta}^T\boldsymbol{\theta}$)

4 Collaboration Questions

After you have completed all other components of this assignment, report your answers to these questions regarding the collaboration policy. Details of the policy can be found in the syllabus.

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details.
2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details.
3. Did you find or come across code that implements any part of this assignment? If so, include full details.

Your Answer