HOMEWORK 4
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/dzbjyqhnncvy#352b6d

Question	Points
Scalar, Vector, & Matrix Derivatives	10
Partial Derivatives for "Neural Networks"	10
Using the Chain Rule on Cross-Entropy Loss	5
Total:	25

*Compiled on Monday 16th September, 2024 at 20:44

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	10-6067

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

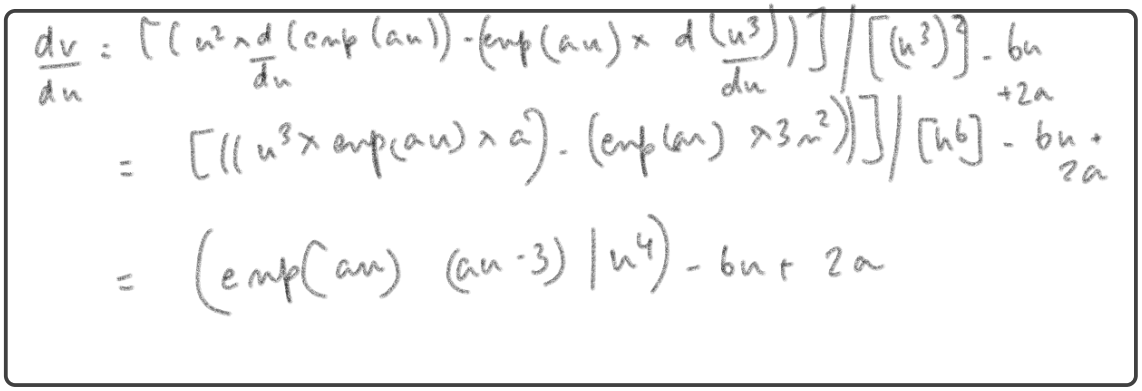
$$v = \frac{\exp(au)}{u^3} - 3u^2 + 2au + 5a - 10$$

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

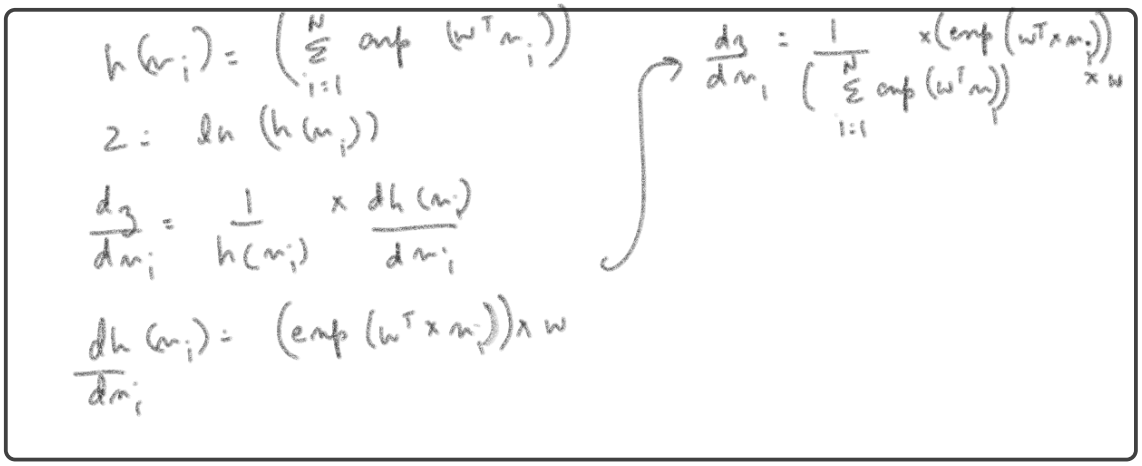
$$y = a \ln(\exp(\mathbf{b}^T \mathbf{X} \mathbf{c}) + \exp(\mathbf{d}^T \mathbf{X} \mathbf{e})) .$$

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 $\mathbf{X}$.

$u = \left(\exp(b^T X c) + \exp(d^T X e)\right)$ $\quad \frac{dy}{dX} = \frac{a}{u} \frac{du}{dX}$

$y = a \ln(u)$

$\frac{d}{dX}\left(\exp(b^T X c)\right) = \exp(b^T X c) \times \frac{d}{dX}(b^T X c)$

$= \exp(b^T X c) \cdot bc^T$

$\frac{d}{dX}\left(\exp(d^T X e)\right) = \exp(d^T X e) \cdot de^T$

$\frac{du}{dX} = \exp(b^T X c) \cdot bc^T + \exp(d^T X e) \cdot de^T$

$\frac{dy}{dX} = \frac{a}{\left(\exp(b^T X c) + \exp(d^T X e)\right)} \left(\exp(b^T X c) \cdot bc^T + \exp(d^T X e) \cdot de^T\right)$

2 Partial Derivatives for "Neural Networks" (10 points)

A neural network always includes a nonlinear function such as the sigmoid function. However, we could create one that has no nonlinear function as follows.

Suppose we have a prediction function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ defined as

$$\hat{y} = f(\mathbf{x}; \mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \mathbf{w}^{(2)}, b^{(2)}) = \mathbf{w}^{(2)}(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) + b^{(2)},$$

where the matrix $\mathbf{W}^{(1)}$, the column vector $\mathbf{b}^{(1)}$, scalar $b^{(2)}$ and the row vector $\mathbf{w}^{(2)}$ are parameters of the function f with approproate sizes.

1. (5 points) What is the partial derivative of $\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$ with respect to $\mathbf{b}^{(1)}$? Show your derivation in the first box and include the final result in the second box below.

Derivation

$$\frac{\partial L}{\partial \hat{y}} = -2(y - \hat{y})$$

$$\frac{\partial \hat{y}}{\partial b^{(1)}} = w^2$$

$$\frac{\partial L}{\partial b^1} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial b^1}$$

$$= -2(y - \hat{y}) \times w^2$$

Final result

$$\frac{\partial L}{\partial b^{(1)}} = -2(y - \hat{y}) \times w^{(2)}$$

2. (5 points) What is the partial derivative of $\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$ with respect to $\mathbf{W}^{(1)}$? Show your derivation in the first box and include the final result in the second box below.

Derivation

$\frac{\partial L}{\partial W^{(1)}} = ?$

$\frac{\partial L}{\partial W^1} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial W^1}$

$\frac{\partial L}{\partial \hat{y}} = -2(y - \hat{y})$

$= -2(y - \hat{y}) \times W^2 m^T$

$\frac{\partial \hat{y}}{\partial W^{(1)}} = ?$

taking $W^{(2)}(W^{(1)} m)$

$\frac{\partial (W^1 m)}{\partial W^1} : m^T$

$\frac{\partial \hat{y}}{\partial W^1} = W^2 m^T$ $\frac{\partial \hat{y}}{\partial W^1} = W^2 m^T$

Final result

$\frac{\partial L}{\partial W^{(1)}} = -2(y - \hat{y}) \times W^{(2)} m^T$

3 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

$g(w)^T x_i = \sum_{l=1}^{D} w_l^2 x_{il}$

$u_i =$

$h(w, x_i) = \frac{1}{1 + \exp(-g(w)^T x_i)}$

$= \frac{1}{1 + \exp(-u)}$ (quotient rule: $f(z) = \frac{g(z)}{h(z)}$, $f'(z) = \frac{g'(z)h(z) - g(z)h'(z)}{h(z)^2}$)

$\frac{dh(w, x_i)}{du_i} = \frac{0 \cdot (1 + \exp(-u)) - 1 \cdot (-\exp(-u))}{(1 + \exp(-u))^2}$

$= \exp(-u) / (1 + \exp(-u))^2$

$= h(w, x_i)(1 - h(w, x_i))$

Final result

$$\Delta_w L(w) = -\frac{2}{N} \sum_{i=1}^{N} (y_i - h(w, m_i)) \cdot w \odot m_i$$

↳ $u_i = \sum_{l=1}^{D} w_l^2 m_{il}$

$$\frac{\partial u_i}{\partial w_l} = 2 w_l m_{il} \quad \text{(Power Rule)}$$

$$\frac{\partial h(w, m_i)}{\partial w_l} = \frac{dh(w, m_i)}{du_i} \times \frac{\partial u_i}{\partial w_l}$$

$$= h(w, m_i)(1 - h(w, m_i)) \times 2 w_l m_{il}$$

$$\frac{\partial L(w)}{\partial w_l} = -\frac{1}{N} \sum_{i=1}^{N} \left[\frac{y_i}{h(w, m_i)} \cdot \frac{\partial h(w, m_i)}{\partial w_l} - \frac{1 - y_i}{1 - h(w, m_i)} \cdot \frac{\partial h(w, m_i)}{\partial w_l} \right]$$

↓ after factoring & substituting $\frac{\partial h}{\partial w_l}(w, m_i)$

$$\nabla_w L(w) = -\frac{2}{N} \sum_{i=1}^{N} (y_i - h(w, m_i)) \cdot w \odot m_i$$

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

1. No help whatsoever.
2. Didn't help anyone whatsoever.
3. Didn't come across code that's being implemented.

I only referred to the textbook, slides & the internet to help in some calculus equation formulas.