CPSC 488/588 AI Foundation Models
Fall 2023
HW 1

Instructions: copy this project at `https://www.overleaf.com/read/gjvcgrzjyxfm`, complete the solutions, and return your solutions in pdf format.

Full Name: Your Name
Netid: yyyy8888

1. **Q1**

   Warmup. this part is about refreshing your calculus on calculating derivatives of functions.

   (a) Given the function: $f(x) = sin(3x^2 + 4cos(x))$, find $\partial f/\partial x$ using the chain rule.

   (b) Given the function: $f(x) = e^{\tanh(2x^3 - 5x^2 + x)}$, find $\partial f/\partial x$ using the chain rule.
   Recall: $\frac{\partial \tanh(x)}{\partial(x)} = 1 - \tanh^2(x) = \text{sech}^2(x)$

   (c) Consider the function: $f(x, y, z) = x^2 sin(yz) + e^{yz} - z^3 y$, find partial derivatives of the function with respect to each variable separately.

2. **Q2 Matrix calculus.** Recall that matrices are a way of organizing data into rows and columns. Answer the following questions:

   (a) Consider a function given by:

   $$f(\mathbf{x}) = \mathbf{c}^T \mathbf{A}\mathbf{x} \tag{1}$$

   where

   $$\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix}$$

is a column vector of variables,

$$\mathbf{c} = \begin{bmatrix} 2 & 3 & 1 \end{bmatrix}$$

is a constant vector, and

$$\mathbf{A} = \begin{bmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \\ a_{20} & a_{21} & a_{22} \end{bmatrix}$$

is a integer valued matrix,
Derive the gradient of f with respect to $\mathbf{x}$.

(b) Given the function

$$f = \mathbf{x}^\top \cdot \mathbf{A} \cdot \mathbf{x} + c \cdot \sin(\mathbf{y})^\top \cdot \mathbf{x}$$

where

- A is a symmetric matrix
- c is a scalar
- x is a vector
- y is a vector

Derive the gradients with respect to $\mathbf{x}$ and $\mathbf{y}$.

(c) Given the function

$$g = \mathbf{x}^\top \mathbf{B} \mathbf{y} + d \tanh(\mathbf{z})^\top \mathbf{x}$$

where

- $\mathbf{B}$ is an arbitrary matrix
- d is a scalar
- $\mathbf{x}$ is a vector
- $\mathbf{y}$ is a vector of the same dimension as $\mathbf{x}$
- $\mathbf{z}$ is a vector

Derive the gradients with respect to $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$.
Hint: recall $\frac{\partial \tanh(x)}{\partial(x)} = 1 - \tanh^2(x) = \text{sech}^2(x)$

3. **Q3 Automatic differentiation**
   (a) Consider the following function: $f(x_1, x_2, x_3, x_4) = \frac{1}{2}\exp(x_1 + x_2^2) - (x_3 * x_4^2)$
      i. draw the computation graph corresponding to this function and fill in the gradient values for all of the intermediate nodes and the leaves in the computation graph. Assume the values for x_1, x_2, x_3, x_4 are $-1, 2, 4, -3$ respectively.

      ii. Derive the gradients of f with respect to its inputs ∇f using symbolic differentiation and chain rule.

4. **Q4** Explain the reason behind using negative sampling in the SkipGram word embeddings model.

5. **Q5 transformer**

   (a) In the multi-head self-attention operation, what is the cost of computation (in terms of number of number of FLoating point OPerations)? Assume b is batch size, m is sequence length, d is the model dimension, and h is the number of attention heads. Assume the dimensionality of keys and queries are $d/2h$.

   Note: You need to derive the answer, just providing the final answer is not sufficient.

   (b) What is the cost of computation in terms of number of FLoating point Operations for multi-head attention in the backward pass, when the model is being trained? (use the same assumptions as the above questions).

   (c) What is the cost of computation in terms of number of FLoating point Operations for **Grouped Query Attention** where $G = k/4$ is the number of groups?