CSC413 Math Assignment 1

Deadline: January 26, 2024 by 5pm

Submission: Compile and submit a PDF report containing your written solutions. You may also submit an image of your legible hand-written solutions. Submissions will be done on Markus.

Late Submission: Please see the syllabus for the late submission criteria.

Collaboration Policy: Please see the syllabus for the collaboration policy.

Question 1: Gradients [2 points]

Recall that the gradient of a scalar-valued function $f : \mathbb{R}^m \to \mathbb{R}$ is the $m \times 1$ dimensional vector of partial derivatives of f with respect to each of its inputs,

$$\nabla_{\mathbf{x}} f(x_1, \dots, x_m) = \frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f(x_1,\dots,x_m)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x_1,\dots,x_m)}{\partial x_m} \end{bmatrix}$$

Compute the gradients of the following function:

$$f(x, y, z) = xyz + 3x^2 + \log(y/2) - z^4$$

Question 2: Computation Graph [4 points]

Consider the following set of equations:

$$\begin{aligned} f(p, q) &= p + q^2 \\ p(x, y) &= xy \\ q(z) &= \sin(z) \\ x(t) &= t^2 \\ y(t) &= 2t + 1 \\ z(t) &= t^3 \end{aligned}$$

Draw the computation graph expressing the dependencies between the variables $\{f, p, q, x, y, z, t\}$.

Then, using the computation graph you found above and the chain rule, write the expression for the derivative $\frac{\partial f}{\partial t}$.

Question 3. MLP [12 points]

Consider a neural network that receives a 10×10 image of a digit and decides which of $\{0, 1, \dots, 9\}$ it is. Suppose we were to use a 2-layer multilayer Perceptron to solve this classification problem. We vectorize the input before passing it to the neural network.

$$\begin{aligned} \mathbf{x} &= 100 \text{ dimensional input} \\ \mathbf{m} &= \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)} \\ \mathbf{h} &= \text{ReLU}(\mathbf{m}) \\ \mathbf{z} &= \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)} \\ \mathbf{y} &= \text{softmax}(\mathbf{z}) \end{aligned}$$

Part (a) [2 points]

What is the shape of the output vector $\mathbf{y}$? Let k represent the size of the hidden layer. What are the dimension of $W^{(1)}$ and $W^{(2)}$?

Part (b) [2 points]

Draw a computation graph for $\mathbf{y}$. Your graph should include the quantities $\mathbf{W^{(1)}}$, $\mathbf{W^{(2)}}$, $\mathbf{b^{(1)}}$, $\mathbf{b^{(2)}}$, $\mathbf{x}$, $\mathbf{m}$, $\mathbf{h}$, $\mathbf{z}$ and $\mathbf{y}$.

Part (c) [4 points]

Derive the gradient descent update rule for $\mathbf{W}^{(2)}$. You should begin by deriving the update rule for $W_{ij}^{(2)}$, and then vectorize your answer. Assume that we will use the softmax activation and cross-entropy loss.

Note: If you use the derivative of the softmax activation and the cross-entropy loss, you **must** derive them.

Part (d) [4 points]

What would be the update rule for $W_{ij}^{(2)}$, if we use the square loss $\mathcal{L}_{SE}(\mathbf{y}, \mathbf{t}) = \mathbf{\frac{1}{2}}(\mathbf{y} - \mathbf{t})^{\mathbf{2}}$?

Show that we will not get good gradient signal to update $W_{ij}^{(2)}$ if we use this square loss.

Question 4. Diminishing Gradient [6 points]

Let's assume we have a deep neural network as follows:

$$
\begin{aligned}
h_1 &= \sigma(w_1 x + b_1) \\
h_2 &= \sigma(w_2 h_1 + b_2) \\
&\vdots
\end{aligned}
$$

For simplicity, assume that x, w_1, b_1, h_1, w_2, b_2, h_2, etc., are all scalars. This is a very narrow neural network.

Part (a) [2 points]

Show that

$$
\left| \frac{\partial h_1}{\partial x} \right| \leq \frac{1}{4}|w_1|
$$

In order to do so, you will need to first show that $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ (worth 1 point). Include a plot (or sketch) of the function $\sigma'(z)$ (worth 1 point).

Part (b) [2 points]

Continue from the previous question, show that for a deeper neural network.

$$
\left| \frac{\partial h_N}{\partial x} \right| \leq \frac{1}{4^N}|w_1||w_2| \cdots |w_N|
$$

What would be a problem with this result?

Part (c) [2 points]

Would we have the same issue as in Part (a) if we replaced the sigmoid activation with ReLU activations? Why or why not?

Question 5. Work Allocation [2 points]

This question is to make sure that if you are working with a partner, that you and your partner contributed equally to the assignment.

Please have each team member write down the times that you worked on the assignment, and your contribution to the assignment.

```
# Example answer:
# I worked on the assignment on Jan 20 afternoon, Jan 26th 12pm-2pm,
# and then Feb 4th in the evening. My partner and I had a meeting on
# Jan 20th to read the entire assignment, and we did Question 1 together
# while screensharing. I worked out the math for Q2, and checked my
# partner's implementation in Q3. I also wrote the Q3 helper functions,
# and Q4(b).
```