Name: **Student ID**:

Quiz 4

1. Suppose we have two matrices $\mathbf{X} \in \mathbb{R}^{N \times D}$, $\mathbf{W} \in \mathbb{R}^{D \times H}$, and $\mathbf{Y} = \mathbf{X}\mathbf{W} \in \mathbb{R}^{N \times H}$. In addition, we have a a function $L : \mathbb{R}^{N \times H} \to \mathbb{R}$ that maps a matrix to a scalar. Given $\nabla \mathbf{L}_y = \partial L / \partial \mathbf{Y} \in \mathbb{R}^{N \times H}$, show that

$$\frac{\partial L}{\partial X_{i,j}} = \frac{\partial \mathbf{Y}}{\partial X_{i,j}} \frac{\partial L}{\partial \mathbf{Y}} = [\nabla \mathbf{L}_y \mathbf{W}^T]_{i,j}, \tag{1}$$

where $[\cdot]_{i,j}$ represents the (i, j)th entry of the argument. Collecting the result of (1), we obtain

$$\frac{\partial L}{\partial \mathbf{X}} = \nabla \mathbf{L}_y \mathbf{W}^T. \tag{2}$$

Hints: Please refer to the slides of Lecture 3 (pp.187-204) by using the Jacobean formulation.

2. Although we have derive the gradient in Problem 1 using the Jacobean formulation, it is more convenient to follow the convention "the shape of the gradient equals the shape of the parameter" (the shape convention).

   Applying the shape convention, you are required to compute the gradients of a two-layer neural network trained with softmax loss. where $\mathbf{y} = [y_i]$ are labels of each inputs. Given ($\mathbf{X} \in \mathbb{R}^{N\times D}, \mathbf{y} \in \mathbb{R}^{N\times 1}$), the forward pass of the model is as follows:

$$\mathbf{X} = \text{input} \in \mathbb{R}^{N\times D} \tag{3}$$
$$\mathbf{Z} = \mathbf{X}\mathbf{W}_1 + \mathbf{1}_N\mathbf{b}_1 \in \mathbb{R}^{N\times H} \tag{4}$$
$$\mathbf{H} = \text{ReLu}(\mathbf{Z}) \in \mathbb{R}^{N\times H} \tag{5}$$
$$\mathbf{S} = \mathbf{H}\mathbf{W}_2 + \mathbf{1}_N\mathbf{b}_1 \in \mathbb{R}^{N\times C} \tag{6}$$
$$\mathbf{P} = \text{softmax}(\mathbf{S}) \in \mathbb{R}^{N\times C} \tag{7}$$
$$L = \text{cross\_entropy}(\mathbf{P}, \mathbf{y}) \in \mathbb{R} \tag{8}$$

   The dimensions of the model's parameters are

$$\mathbf{1}_N \in \mathbb{R}^{N\times 1}, \mathbf{W}_1 \in \mathbb{R}^{D\times H},\ \mathbf{W}_2 \in \mathbb{R}^{H\times C},\ \mathbf{b}_1 \in \mathbb{R}^{1\times H},\ \mathbf{b}_2 \in \mathbb{R}^{1\times C},$$

   Show that

$$\frac{\partial L}{\partial \mathbf{S}} = \begin{cases} (P_{i,j} - 1), & j = y_i, \\ P_{i,j}, & j \neq y_i \end{cases} \in \mathbb{R}^{N\times C} \tag{9}$$
$$\frac{\partial L}{\partial \mathbf{W}_2} = \mathbf{H}^T\frac{\partial L}{\partial \mathbf{S}} \tag{10}$$
$$\frac{\partial L}{\partial \mathbf{b}_2} = \mathbf{1}_N^T\frac{\partial L}{\partial \mathbf{S}} \tag{11}$$
$$\frac{\partial L}{\partial \mathbf{H}} = \frac{\partial L}{\partial \mathbf{S}}\mathbf{W}_2^T \tag{12}$$
$$\frac{\partial L}{\partial \mathbf{Z}} = \begin{cases} [\frac{\partial L}{\partial \mathbf{H}}]_{i,j}, & Z_{i,j} > 0, \\ 0, & \text{otherwise} \end{cases} \in \mathbb{R}^{N\times H} \tag{13}$$
$$\frac{\partial L}{\partial \mathbf{W}_2} = \mathbf{X}^T\frac{\partial L}{\partial \mathbf{Z}} \tag{14}$$
$$\frac{\partial L}{\partial \mathbf{b}_2} = \mathbf{1}_N^T\frac{\partial L}{\partial \mathbf{Z}}. \tag{15}$$

Hint: To better understand the shape convention principle, we take an example. First, we consider Problem 1 as a scalar case: suppose X, W, $Y = XW$, and $L(Y)$ are all scalars. Then, it is more easily to obtain $\frac{\partial L}{\partial X} = W\nabla L_y$. Finally, we extend the gradient to the matrix. You should try to put the relevant matrix $\mathbf{W} \in \mathbb{R}^{D\times H}$ in front/back of $\nabla\mathbf{L}_y \in \mathbb{R}^{N\times H}$ or take the transpose to make sure that $\mathbf{W}\nabla\mathbf{L}_y$ has the same shape of $\mathbf{X}$.