Name: **Student ID**:

Quiz 5

1. Please sketch the computational graph of the two-layer neural network with the ReLu activation function and softmax loss. The dimensions of the model's parameters are

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

2. Please sketch the backward propagation of the above computational graph using the shape convention.

   Hints: Please refer to the slides of Lecture 4 (pp.127-133).

3. The exponential weighted average of the gradient is given by

$$m_t = \beta m_{t-1} + (1 - \beta)\nabla L_t.$$

If we initialize $m_0 = 0$, then we have a significant amount of bias initially towards smaller values. Show that to correct this bias, we have to introduce the bias correction term $1/(1 - \beta^t)$. The normalized gradient is given by

$$\hat{m}_t = \frac{m_t}{1 - \beta^t}.$$

Hints: For convenient, we denote ∇L_t as g_t.

$$\begin{aligned}
m_t &= \beta m_{t-1} + (1-\beta)g_t \\
m_0 &= 0 \\
m_1 &= \beta m_0 + (1-\beta)g_1 \\
&= (1-\beta)g_1 \\
m_2 &= \beta m_1 + (1-\beta)g_2 \\
&= \beta(1-\beta)g_1 + (1-\beta)g_2 \\
m_3 &= \beta m_2 + (1-\beta)g_3 \\
&= \beta^2(1-\beta)g_1 + \beta(1-\beta)g_2 + (1-\beta)g_3 \\
&= (1-\beta)\sum_{i=1}^{3}\beta^{3-i}g_i
\end{aligned}$$

So, we have

$$m_t = (1-\beta)\sum_{i=1}^{t}\beta^{t-i}g_i$$

Take expectation on the both sides

$$\mathsf{E}\{m_t\} = (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\mathsf{E}\{g_i\}$$

Assume $\mathsf{E}\{g_i\} = \mathsf{E}\{g\}\ \forall i$. We obtain

$$\mathsf{E}\{m_t\} = (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\mathsf{E}\{g\} = (1-\beta)\frac{1-\beta^t}{1-\beta}\mathsf{E}\{g\}.$$