# Question 1

1a

$$\left[\frac{\partial L}{\partial \mathbf{W}}\right]_{ij} = \frac{\partial L}{\partial W_{ij}} = \sum_s^S \sum_n^N \frac{\partial L}{\partial Y_{sn}} \frac{\partial Y_{sn}}{\partial W_{ij}}$$

$$\frac{\partial Y_{sn}}{\partial W_{ij}} = \sum_m^M X_{sm} \frac{\partial W_{nm}}{\partial W_{ij}} + 0 = \sum_m^M X_{sm} \delta_{ni} \delta_{mj} = X_{sj} \delta_{ni}$$

$$\frac{\partial L}{\partial W_{ij}} = \sum_s^S \sum_n^N \frac{\partial L}{\partial Y_{sn}} X_{sj} \delta_{ni} = \sum_s^S \frac{\partial L}{\partial Y_{si}} X_{sj} \Leftrightarrow \frac{\partial L}{\partial \mathbf{W}} = \left(\frac{\partial L}{\partial \mathbf{Y}}\right)^T \mathbf{X}$$

1b

$$\left[\frac{\partial L}{\partial \mathbf{b}}\right]_j = \frac{\partial L}{\partial b_j} \Leftrightarrow \frac{\partial L}{\partial B_{ij}} = \sum_s^S \sum_n^N \frac{\partial L}{\partial Y_{sn}} \frac{\partial Y_{sn}}{\partial B_{ij}}$$

$$\frac{\partial Y_{sn}}{\partial B_{ij}} = \frac{\partial (B_{sn} + \sum_m^M X_{sm} W_{nm})}{\partial B_{ij}} = \frac{\partial B_{sn}}{\partial B_{ij}} \Leftrightarrow \frac{\partial b_n}{\partial b_j} = \delta_{nj}$$

$$\frac{\partial L}{\partial B_{ij}} = \sum_s^S \sum_n^N \frac{\partial L}{\partial Y_{sn}} \delta_{nj} = \sum_s^S \frac{\partial L}{\partial Y_{sj}}$$

This means that $\frac{\partial L}{\partial \mathbf{b}}$ is the row vector

$$\left[\sum_s^S \frac{\partial L}{\partial Y_{s1}}, \sum_s^S \frac{\partial L}{\partial Y_{s2}}, \ldots, \sum_s^S \frac{\partial L}{\partial Y_{sn}}\right] \in \mathbb{R}^{1 \times N}$$

which can be obtained with a dot product between a one's vector and $\frac{\partial L}{\partial \mathbf{Y}}$, giving

$$\frac{\partial L}{\partial \mathbf{b}} = \mathbf{1}^T \left(\frac{\partial L}{\partial \mathbf{Y}}\right)$$

1c

$$\left[\frac{\partial L}{\partial \mathbf{X}}\right]_{ij} = \frac{\partial L}{\partial X_{ij}} = \sum_s^S \sum_n^N \frac{\partial L}{\partial Y_{sn}} \frac{\partial Y_{sn}}{\partial X_{ij}}$$

$$\frac{\partial Y_{sn}}{\partial X_{ij}} = \sum_m^M \frac{\partial X_{sm}}{\partial X_{ij}} W_{nm} + 0 = \sum_m^M \delta_{si} \delta_{mj} W_{nm} = \delta_{si} W_{nj}$$

$$\frac{\partial L}{\partial X_{ij}} = \sum_s^S \sum_n^N \frac{\partial L}{\partial Y_{sn}} \delta_{si} W_{nj} = \sum_n^N \frac{\partial L}{\partial Y_{in}} W_{nj} \Leftrightarrow \frac{\partial L}{\partial \mathbf{X}} = \frac{\partial L}{\partial \mathbf{Y}} \mathbf{W}$$

1d Given $\mathbf{Y} = h(\mathbf{X}) \in \mathbb{R}^{S\times N}$ — an activation function applied element-wise to its input — we can find $\frac{\partial L}{\partial \mathbf{X}}$ by first differentiating w.r.t. $\mathbf{X}$ component-wise:

$$\left[\frac{\partial L}{\partial \mathbf{X}}\right]_{ij} = \frac{\partial L}{\partial X_{ij}} = \sum_s^S \sum_n^N \frac{\partial L}{\partial Y_{sn}} \frac{\partial Y_{sn}}{\partial X_{ij}}$$

$$\frac{\partial Y_{sn}}{\partial X_{ij}} = \frac{\partial h(X_{sn})}{\partial X_{ij}} = h'(X_{sn})\frac{\partial X_{sn}}{\partial X_{ij}} = h'(X_{sn})\delta_{si}\delta_{nj}$$

$$\frac{\partial L}{\partial X_{ij}} = \sum_s^S \sum_n^N \frac{\partial L}{\partial Y_{sn}} \frac{\partial Y_{sn}}{h'(X_{sn})\delta_{si}\delta_{nj}} = \frac{\partial Y_{ij}}{h'(X_{sn})} \Leftrightarrow \frac{\partial L}{\partial \mathbf{X}} = \frac{\partial L}{\partial \mathbf{Y}} \odot h'(\mathbf{X})$$

1e

$$\left[\frac{\partial L}{\partial \mathbf{X}}\right]_{ij} = \frac{\partial L}{\partial X_{ij}} = \sum_s^S \sum_c^C \frac{\partial L}{\partial Y_{sc}} \frac{\partial Y_{sc}}{\partial X_{ij}}$$

$$\frac{\partial Y_{sc}}{\partial X_{ij}} = \frac{\frac{\partial (e^{X_{sc}})}{\partial X_{ij}} \sum_k^C e^{X_{sk}} - \frac{\partial (\sum_k^C e^{X_{sk}})}{\partial X_{ij}} e^{X_{sc}}}{\left(\sum_k^C e^{X_{sk}}\right)^2}$$

$$\frac{\partial (e^{X_{sc}})}{\partial X_{ij}} = e^{X_{sc}} \frac{\partial X_{sc}}{\partial X_{ij}} = e^{X_{sc}} \delta_{si} \delta_{cj}$$

$$\frac{\partial (\sum_k^C e^{X_{sk}})}{\partial X_{ij}} = \sum_k^C e^{X_{sk}} \delta_{si} \delta_{kj} = e^{X_{sj}} \delta_{si}$$

$$\frac{\partial Y_{sc}}{\partial X_{ij}} = \frac{e^{X_{sc}} \delta_{si} \delta_{cj} \sum_k^C e^{X_{sk}} - e^{X_{sj}} \delta_{si} e^{X_{sc}}}{\left(\sum_k^C e^{X_{sk}}\right)^2}$$

$$\frac{e^{X_{sc}} \delta_{si} \delta_{cj} \sum_k^C e^{X_{sk}}}{\left(\sum_k^C e^{X_{sk}}\right)^2} = \frac{e^{X_{sc}} \delta_{si} \delta_{cj}}{\sum_k^C e^{X_{sk}}} = \delta_{si} \delta_{cj} Y_{sc}$$

$$\frac{e^{X_{sj}} \delta_{si} e^{X_{sc}}}{\left(\sum_k^C e^{X_{sk}}\right)^2} = \frac{e^{X_{sj}} \delta_{si}}{\sum_k^C e^{X_{sk}}} Y_{sc} = \delta_{si} Y_{sj} Y_{sc}$$

$$\frac{\partial Y_{sc}}{\partial X_{ij}} = \delta_{si} \delta_{cj} Y_{sc} - \delta_{si} Y_{sj} Y_{sc} = \delta_{si} Y_{sc} (\delta_{cj} - Y_{sj})$$

$$\frac{\partial L}{\partial X_{ij}} = \sum_s^S \sum_c^C \frac{\partial L}{\partial Y_{sc}} \delta_{si} Y_{sc} (\delta_{cj} - Y_{sj}) = \sum_c^C \frac{\partial L}{\partial Y_{ic}} Y_{ic} (\delta_{cj} - Y_{ij})$$

$$= \frac{\partial L}{\partial Y_{ij}} - \sum_c^C \frac{\partial L}{\partial Y_{ic} Y_{ic} Y_{ij}} = Y_{ij} \Big( \frac{\partial L}{\partial Y_{ij}} - \sum_c^C \frac{\partial L}{\partial Y_{ic}} Y_{ic} \Big)$$

Generalizing $\frac{\partial L}{\partial X_{ij}}$, we see that the full Jacobian matrix of the loss w.r.t. $\mathbf{X}$ is

$$\frac{\partial L}{\partial \mathbf{X}} = \begin{bmatrix} Y_{11} \Big( \frac{\partial L}{\partial Y_{11}} - \sum_c^C \frac{\partial L}{\partial Y_{1c}} Y_{1c} \Big) & \cdots & Y_{1C} \Big( \frac{\partial L}{\partial Y_{1C}} - \sum_c^C \frac{\partial L}{\partial Y_{1c}} Y_{1c} \Big) \\ \vdots & \ddots & \vdots \\ Y_{S1} \Big( \frac{\partial L}{\partial Y_{S1}} - \sum_c^C \frac{\partial L}{\partial Y_{Sc}} Y_{Sc} \Big) & \cdots & Y_{SC} \Big( \frac{\partial L}{\partial Y_{SC}} - \sum_c^C \frac{\partial L}{\partial Y_{Sc}} Y_{Sc} \Big) \end{bmatrix} \in \mathbb{R}^{S \times C}$$

This can be obtained by suitably multiplying the component-wise derivative with a one's matrix in $\mathbb{R}^{C\times C}$, giving the final answer

$$\frac{\partial L}{\partial \mathbf{X}} = \mathbf{Y} \odot \Big(\frac{\partial L}{\partial \mathbf{Y}} - \Big(\frac{\partial L}{\partial \mathbf{Y}} \odot \mathbf{Y}\Big)\mathbf{1}\mathbf{1}^T\Big).$$

1f

$$\begin{aligned}
\Big[\frac{\partial L}{\partial \mathbf{X}}\Big]_{ij} &= \frac{\partial(-\frac{1}{S}\sum_s^S \sum_c^C T_{sc} \ln X_{sc})}{\partial X_{ij}} \\
&= -\frac{1}{S}\sum_s^S \sum_c^C \frac{\partial (T_{sc} \ln X_{sc})}{\partial X_{ij}} = -\frac{1}{S}\sum_s^S \sum_c^C \frac{T_{sc}}{X_{sc}} \delta_{si}\delta_{cj} = -\frac{1}{S}\frac{T_{ij}}{X_{ij}} \\
&\Leftrightarrow \frac{\partial L}{\partial \mathbf{X}} = -\frac{1}{S}\mathbf{T} \oslash \mathbf{X}
\end{aligned}$$