1 Linear Module

1.a

$$\left[\frac{\partial L}{\partial \mathbf{W}}\right]_{ij} = \frac{\partial L}{\partial W_{ij}} = \sum_{s,n} \frac{\partial L}{\partial Y_{sn}} \frac{\partial Y_{sn}}{\partial W_{ij}} \tag{1}$$

$$\frac{\partial Y_{sn}}{\partial W_{ij}} = \frac{\partial}{\partial W_{ij}} \left(\sum_m [\mathbf{X}]_{sm} [\mathbf{W}^\top]_{mn} + [\mathbf{B}]_{sn} \right) = \sum_m X_{sm} \frac{\partial W_{nm}}{\partial W_{ij}} + \frac{\partial B_{sn}}{\partial W_{ij}} \tag{2}$$

$$= \sum_m X_{sm} \delta_{ni} \delta_{mj} + 0 = X_{sj} \delta_{ni} \tag{3}$$

$$\sum_{s,n} \frac{\partial L}{\partial Y_{sn}} \frac{\partial Y_{sn}}{\partial W_{ij}} = \sum_{s,n} \frac{\partial L}{\partial Y_{sn}} X_{sj} \delta_{ni} = \sum_s \frac{\partial L}{\partial Y_{si}} X_{sj} \tag{4}$$

$$\therefore \frac{\partial L}{\partial \mathbf{W}} = \left(\frac{\partial L}{\partial \mathbf{Y}} \right)^\top \mathbf{X} \quad \in \mathbb{R}^{N \times M} \tag{5}$$

1.b

$$\left[\frac{\partial L}{\partial \mathbf{b}}\right]_{j} = \frac{\partial L}{\partial b_{j}} = \sum_{s,n} \frac{\partial L}{\partial Y_{sn}} \frac{\partial Y_{sn}}{\partial b_{j}} \tag{6}$$

$$\frac{\partial Y_{sn}}{\partial b_{j}} = \frac{\partial}{\partial b_{j}} \left(\sum_m [\mathbf{X}]_{sm} [\mathbf{W}^\top]_{mn} + [\mathbf{B}]_{sn} \right) = \sum_m \frac{\partial X_{sm} W_{nm}}{\partial b_j} + \frac{\partial B_{sn}}{\partial b_j} \tag{7}$$

$$= 0 + \delta_{nj} = \delta_{nj} \tag{8}$$

$$\sum_{s,n} \frac{\partial L}{\partial Y_{sn}} \frac{\partial Y_{sn}}{\partial b_{j}} = \sum_{s,n} \frac{\partial L}{\partial Y_{sn}} \delta_{nj} = \sum_s \frac{\partial L}{\partial Y_{sj}} \tag{9}$$

$$\therefore \frac{\partial L}{\partial \mathbf{b}} = \sum_s \frac{\partial L}{\partial \mathbf{Y}_s} \quad \in \mathbb{R}^{1 \times N} \tag{10}$$

For clarification, in equation (10) we sum over the rows $s \in S$ of $\mathbf{Y}$, so the j-th element of $\mathbf{b}$ is the sum of all elements in position j of the rows of $\mathbf{Y}$.

1.c

$$\left[\frac{\partial L}{\partial \mathbf{X}}\right]_{ij} = \frac{\partial L}{\partial X_{ij}} = \sum_{s,n} \frac{\partial L}{\partial Y_{sn}} \frac{\partial Y_{sn}}{\partial X_{ij}} \tag{11}$$

$$\frac{\partial Y_{sn}}{\partial X_{ij}} = \frac{\partial}{\partial X_{ij}} \left(\sum_m [\mathbf{X}]_{sm} [\mathbf{W}^\top]_{mn} + [\mathbf{B}]_{sn} \right) = \sum_m \frac{\partial X_{sm}}{\partial X_{ij}} W_{nm} + \frac{\partial B_{sn}}{\partial X_{ij}} \tag{12}$$

$$= \sum_m \delta_{si} \delta_{mj} W_{nm} + 0 = \delta_{si} W_{nj} \tag{13}$$

$$\sum_{s,n} \frac{\partial L}{\partial Y_{sn}} \frac{\partial Y_{sn}}{\partial X_{ij}} = \sum_{s,n} \frac{\partial L}{\partial Y_{sn}} \delta_{si} W_{nj} = \sum_n \frac{\partial L}{\partial Y_{in}} W_{nj} \tag{14}$$

$$\therefore \frac{\partial L}{\partial \mathbf{X}} = \frac{\partial L}{\partial \mathbf{Y}} \mathbf{W} \quad \in \mathbb{R}^{S \times M} \tag{15}$$

1.d

$$\left[\frac{\partial L}{\partial \mathbf{X}}\right]_{ij} = \frac{\partial L}{\partial X_{ij}} = \sum_{s,m} \frac{\partial L}{\partial Y_{sm}} \frac{\partial Y_{sm}}{\partial X_{ij}} \tag{16}$$

$$= \sum_{s,m} \frac{\partial L}{\partial Y_{sm}} \frac{\partial h(X_{sm})}{\partial X_{ij}} \tag{17}$$

$$= \sum_{s,m} \frac{\partial L}{\partial Y_{sm}} \delta_{si} \delta_{mj} = \frac{\partial L}{\partial Y_{ij}} \tag{18}$$

$$\therefore \frac{\partial L}{\partial \mathbf{X}} = \frac{\partial L}{\partial \mathbf{Y}} \circ h'(\mathbf{X}) \quad \in \mathbb{R}^{S \times M} \tag{19}$$

From equation (17) to (18) we get two Kronecker deltas because the derivative of the element-wise activation function is zero for all elements except the one we are taking the derivative of. This makes intuitive sense, given we are differentiating an element-wise function. The final derivative of the loss w.r.t. the input $\mathbf{X}$ is the Hadamard product of the derivative of the loss w.r.t. the output $\mathbf{Y}$ and the element-wise derivative of the activation function h w.r.t. the input $\mathbf{X}$. We can assume the shapes of $\mathbf{X}$ and $\mathbf{Y}$ are compatible, since $\mathbf{Y} = h(\mathbf{X})$