MLE Ex 4

Task 1

① We have:

$$\mathcal{L} = \mathcal{L}_\theta(x,y) = -\sum_{i=1}^{c} y_i \ln \hat{y}_i$$

so

$$\delta_j^{(L)} = \frac{\partial \mathcal{L}}{\partial \tilde{z}_j^{(L)}}$$

$$= \frac{\partial}{\partial \tilde{z}_j^{(L)}} - \sum_{i=1}^{c} y_i \ln \hat{y}_i$$

$$= -\sum_{i=1}^{c} y_i \frac{\partial}{\partial \tilde{z}_j^{(L)}} \ln \hat{y}_i \,. \qquad (*)$$

Furthermore

$$\ln \hat{y}_i = \ln\left(\frac{\exp \tilde{z}_i^{(L)}}{\sum_{m=1}^{c} \exp \tilde{z}_m^{(L)}}\right)$$

$$= \tilde{z}_i^{(L)} - \ln \sum_{m=1}^{c} \exp \tilde{z}_m^{(L)}$$

$$\Rightarrow \frac{\partial \ln \hat{y}_i}{\partial \tilde{z}_j^{(L)}} = \delta_{ij} - \frac{1}{\sum_{m=1}^{c} \exp \tilde{z}_m^{(L)}} \frac{\partial}{\partial \tilde{z}_j^{(L)}} \sum_{m=1}^{c} \exp \tilde{z}_m^{(L)}$$

$$= \delta_{ij} - \frac{1}{\sum_{m=1}^{c} \exp \tilde{z}_m^{(L)}} \exp \tilde{z}_j^{(L)}$$

$$= \delta_{ij} - \hat{y}_j$$

Plugging into $(*)$:

$$\delta_j^{(L)} = -\sum_{i=1}^{c} y_i \left(\delta_{ij} - \hat{y}_j \right)$$

$$= \left(\sum_{i=1}^{c} y_i \hat{y}_j\right) - y_j$$

$= \hat{y}_j - y_j \qquad (\sum y_i = 1 \text{ \& classification task})$

Thus, from $\delta_j^{(L)} = \hat{y}_j - y_j$, we have $\underline{\delta}^{(L)} = \hat{\underline{y}} - \underline{y}$, as desired.

② a) We have (for $i \in [d_L^{\text{(max)}}]$ "$= [C]$" "$= \{1, \dots, d\}$", $j \in [d_{L-1}]$)

$$(\nabla_{W^{(L-1)}} \mathcal{L})_{ij} = \sum_{k=1}^{C} \underbrace{\frac{\partial \mathcal{L}}{\partial \tilde{z}_k^{(L)}}}_{\delta_k^{(L)}} \frac{\partial \tilde{z}_k^{(L)}}{\partial W_{ij}^{(L-1)}} \qquad (*)$$

where

$$\tilde{z}_k^{(L)} = (W^{(L-1)} z^{(L-1)} + b^{(L-1)})_k$$

$$= \left(\sum_{m=1}^{d_L} W_{km}^{(L-1)} z_m^{(L-1)}\right) + b_k^{(L-1)}$$

$$\Rightarrow \frac{\partial \tilde{z}_k^{(L)}}{\partial W_{ij}^{(L-1)}} = \sum_{m=1}^{d_L} \delta_{ik} \delta_{jm} z_m^{(L-1)}$$

$$= \delta_{ik} z_j^{(L-1)}$$

Plugging into $(*)$:

$$(\nabla_{W^{(L-1)}} \mathcal{L})_{ij} = \sum_{k=1}^{C} \delta_k^{(L)} \delta_{ik} z_j^{(L-1)}$$

$$= \delta_i^{(L)} z_j^{(L-1)}$$

$$= (\underline{\delta}^{(L)} (\underline{z}^{(L-1)})^T)_{ij}$$

So

$$\nabla_{W^{(L-1)}} \mathcal{L} = \underline{\delta}^{(L)} (\underline{z}^{(L-1)})^T$$

As desired. (Can plug in $\underline{\delta}^{(L)} = \hat{\underline{y}} - \underline{y}$).

b) Proceed similarly: (for $i \in [d_L]$) $= [C]$

$$(\nabla_{\underline{b}^{(L-1)}} \mathcal{L})_i = \sum_{k=1}^{C} \underbrace{\frac{\partial \mathcal{L}}{\partial \tilde{z}_k^{(L)}}}_{\delta_k^{(L)}} \frac{\partial \tilde{z}_k^{(L)}}{\partial b_i^{(L-1)}} \quad (*)$$

where

$$\tilde{z}_k^{(L)} = \left(\sum_{m=1}^{d_L} W_{km}^{(L-1)} z_m^{(L-1)} \right) + b_k^{(L-1)}$$

$$\Rightarrow \frac{\partial \tilde{z}_k^{(L)}}{\partial b_i^{(L-1)}} = \delta_{ik}$$

Plug into $(*)$:

$$(\nabla_{\underline{b}^{(L-1)}} \mathcal{L})_i = \sum_{k=1}^{C} \delta_k^{(L)} \delta_{ik}$$

$$= \delta_i^{(L)}$$

$$\Rightarrow \nabla_{\underline{b}^{(L-1)}} \mathcal{L} = \underline{\delta}^{(L)}$$

As desired. (Can plug in $\underline{\delta}^{(L)} = \hat{y} - y$).

Dimensions of the gradients:

We computed the components ~~$\frac{\partial \tilde{z}_k^{(L)}}{\partial z_k}$~~ $(\nabla_{W^{(L-1)}} \mathcal{L})_{ij}$ for $i \in [d_{L-1}] = [C]$ and $j = [d_{L-1}]$, and $(\nabla_{b^{(L-1)}} \mathcal{L})_i$ for $i \in [d_L] = [C]$, so necessarily they're of dimension $C \times d_{L-1}$ & C respectively. This is because ~~$\nabla_a f$~~ $\nabla_a f$ is defined as a tensor with the same dimension as tensor a, for any a, f.

In this case, ~~dim~~ from the results, $\dim(\nabla_b^{(L-1)} \mathcal{L}) = \dim \delta^{(L)} = C$ & $\dim(\nabla_{W^{(L-1)}} \mathcal{L}) = \dim \delta^L (z^{(L-1)})^T = \dim \delta^L \times \dim z^{(L-1)} = C \times d_{L-1}$ so they match.

③ Since $\varphi(x) = \max(0, x)$, we have (applying the convention for $x=0$)

$$\varphi'(x) = \mathbb{1}_{\{x>0\}}$$

Thus

$$\delta_i^{(L-1)} = [(W^{(L-1)})^T(\hat{y} - y) \odot \varphi'(\tilde{z}^{(L-1)})]_i$$

$$= [(W^{(L-1)})^T(\hat{y} - y)]_i [\varphi'(\tilde{z}^{(L-1)})]_i \quad \text{(Hadamard prod.)}$$

$$= [(W^{(L-1)})^T(\hat{y} - y)]_i \varphi'(\tilde{z}_i^{(L-1)}) \quad \text{(elementwise application)}$$

$$= [(W^{(L-1)})^T(\hat{y} - y)]_i \mathbb{1}_{\{\tilde{z}_i^{(L-1)} > 0\}}$$

Thus, $\tilde{z}_j^{(L-1)} < 0 \Rightarrow \mathbb{1}_{\{\tilde{z}_j^{(L-1)} > 0\}} = 0 \Rightarrow \delta_j^{(L-1)} = 0$

φ isn't diff'ble at 0 (at 0, the left derivative is 0 while the right derivative is 1), so we apply the convention $\varphi'(0) = 0$. As a consequence, the $\tilde{z}_j^{(L-1)} = 0$ case is the same as the $\tilde{z}_j^{(L-1)} < 0$ case.

Recall from the hint that

$$\cancel{\nabla_{b^{(L)}} \mathcal{L} = \delta^{(L+1)}}, \quad \cancel{\nabla_{W^{(L)}} \mathcal{L} = \delta^{(L+1)} z^{(}}$$

$$\nabla_{b^{(L-1)}} \mathcal{L} = \delta^{(L)}, \quad \nabla_{W^{(L-1)}} \mathcal{L} = \delta^{(L-1)} (z^{(L)})^T$$

Thus $(\nabla_{b^{(L-1)}} \mathcal{L})_j = 0$ and $(\nabla_{W^{(L-1)}} \mathcal{L})_{j} = 0$ $\forall i \in [d_{L-1}]$ so all the weights & the bias ~~governing~~ for neuron j in layer $L-1$ will not update under a gradient step.

How this propagates backwards through the prior layers is more difficult to comment on:

$$\delta^{(L-2)} = (W^{(L-2)})^T \underset{\sim}{\delta}^{(L-1)} \odot \varphi'(\underset{\sim}{z}^{(L-2)})$$

is the decisive equation since the error signals are (very) closely related to the parameter derivatives. However, the effect of $\delta_j^{(L-1)} = 0$ on the product $(W^{(L-2)})^T \underset{\sim}{\delta}^{(L-1)}$ isn't enormously significant. Prior components can still have non-zero derivative (they affect $\mathcal{L}$, just not on a path through neuron j of layer $L-1$).