1. Read Deep Learning: An Introduction for Applied Mathematicians. Consider a network as defined in (3.1) and (3.2). Assume that $n_L = 1$, find an algorithm to calculate $\nabla a^{[L]}(x)$.

$$a^{[1]} = x \in \mathbb{R}^{n_1}, \tag{3.1}$$

$$a^{[l]} = \sigma\left(W^{[l]} a^{[l-1]} + b^{[l]}\right) \in \mathbb{R}^{n_l} \quad \text{for } l = 2, 3, \ldots, L. \tag{3.2}$$

Define: ① Input: $a^{[1]} = x \in \mathbb{R}^{n_1}$

② Hidden layer: $z^{[l]} = W^{[l]} a^{[l-1]} + b^{[l]} \in \mathbb{R}^{n_l}$, $a^{[l]} = \sigma(z^{[l]}) \in \mathbb{R}^{n_l}$. $l = 2, \ldots, L-1$

③ Ouput: $z^{[L]} = W^{[L]} a^{[L-1]} + b^{[L]} \in \mathbb{R}^{n_L}$, $a^{[L]} = \sigma(z^{[L]}) \in \mathbb{R}^{n_L}$

Use backpropagation,

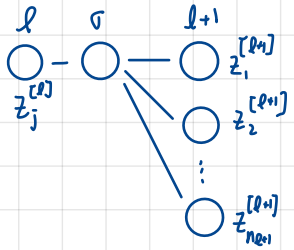
Define loss function $C = \frac{1}{2}\|y - a^{[L]}\|_2^2$.

Let $\delta^{[l]} = \frac{\partial C}{\partial z^{[l]}} = \begin{bmatrix} \frac{\partial C}{\partial z_1^{[l]}} \\ \vdots \\ \frac{\partial C}{\partial z_{n_l}^{[l]}} \end{bmatrix}$, $\frac{\partial a_j^{[L]}}{\partial z_j^{[L]}} = \sigma'(z_j^{[L]})$ for $j = 1, 2, \ldots, n_L$

Also, $\frac{\partial C}{\partial a_j^{[L]}} = \frac{\partial}{\partial a_j^{[L]}} \frac{1}{2} \sum_{k=1}^{n_L} \|y - a_k^{[L]}\|^2 = -(y - a_j^{[L]}) = a_j^{[L]} - y$

So, $\delta_j^{[L]} = \frac{\partial C}{\partial z_j^{[L]}} = \frac{\partial C}{\partial a_j^{[L]}} \frac{\partial a_j^{[L]}}{\partial z_j^{[L]}} = (a_j^{[L]} - y)\,\sigma'(z_j^{[L]}) \Rightarrow \underline{\delta^{[L]} = \sigma'(z^{[L]}) \circ (a^{[L]} - y)}$

$$\delta_j^{[l]} = \frac{\partial C}{\partial z_j^{[l]}} = \sum_{k=1}^{n_{l+1}} \frac{\partial C}{\partial z_k^{[l+1]}} \frac{\partial z_k^{[l+1]}}{\partial z_j^{[l]}} = \sum_{k=1}^{n_{l+1}} \delta_k^{[l+1]} \frac{\partial z_k^{[l+1]}}{\partial z_j^{[l]}}$$



Because $z_k^{[l+1]} = \sum_{s=1}^{n_l} w_{ks}^{[l+1]} a_s^{[l]} + b_k^{[l+1]}$,

$$\frac{\partial z_k^{[l+1]}}{\partial z_j^{[l]}} = \frac{\partial}{\partial z_j^{[l]}} \left(\sum_{s=1}^{n_l} w_{ks}^{[l+1]} a_s^{[l]} + b_k^{[l+1]} \right) = w_{kj}^{[l+1]} \sigma'(z_j^{[l]})$$

Hence, $\delta_j^{[l]} = \sum_{k=1}^{n_{l+1}} \delta_k^{[l+1]} w_{kj}^{[l+1]} \sigma'(z_j^{[l]}) = \sigma'(z_j^{[l]}) \sum_{k=1}^{n_{l+1}} \delta_k^{[l+1]} w_{kj}^{[l+1]} = \sigma'(z_j^{[l]}) \left((W^{[l+1]})^T \delta^{[l+1]}\right)_j$

$$\Rightarrow \underline{\delta^{[l]} = \sigma'(z^{[l]}) \circ \left((W^{[l+1]})^T \delta^{[l+1]}\right)}$$

Take $n_L = 1$, $\nabla_x a^{[L]}(x) = \begin{bmatrix} \frac{\partial a^{[L]}}{\partial x_1} \\ \frac{\partial a^{[L]}}{\partial x_2} \\ \vdots \\ \frac{\partial a^{[L]}}{\partial x_k} \end{bmatrix}$, where $\frac{\partial a^{[L]}}{\partial x_k} = \sum_{j=1}^{n_l} \frac{\partial a^{[L]}}{\partial z_j^{[l]}} \frac{\partial z_j^{[l]}}{\partial x_k}$ for $l = 2, \ldots, L-1$

$$\Rightarrow \frac{\partial a^{[L]}}{\partial x_k} = \sum_{j=1}^{n_2} \frac{\partial a^{[L]}}{\partial z_j^{[2]}} \frac{\partial z_j^{[2]}}{\partial x_k}$$

$\delta_j^{[2]} = \frac{\partial C}{\partial z_j^{[2]}} = \frac{\partial C}{\partial a^{[L]}} \frac{\partial a^{[L]}}{\partial z_j^{[2]}} = (a^{[L]} - y) \frac{\partial a^{[L]}}{\partial z_j^{[2]}} \Rightarrow \frac{\partial a^{[L]}}{\partial z_j^{[2]}} = \frac{\delta_j^{[2]}}{a^{[L]} - y}$

$\because z_j^{[2]} = \sum_{i=1}^{n_1} w_{ji}^{[2]} a_i^{[1]} + b_j^{[2]} = \sum_{i=1}^{n_1} w_{ji}^{[2]} x_i + b_j^{[2]} \Rightarrow \frac{\partial z_j^{[2]}}{\partial x_k} = w_{jk}^{[2]}$

Therefore, $\nabla_x a^{[L]}(x) = (w^{[2]})^T \frac{\delta^{[2]}}{a^{[L]} - y} = \frac{1}{a^{[L]} - y} (w^{[2]})^T \delta^{[2]}$ //

2. There are unanswered questions during the lecture, and there are likely more questions we haven't covered. Take a moment to think about them and write them down here.

In our class, when deriving the relationship between MSE and MLE, we assumed that the error terms follow an i.i.d. Gaussian distribution. However, in real-world data, these assumptions often do not hold.

If the error distribution is skewed (non-symmetric) or heavy-tailed, how would this affect the asymptotic properties of the regression parameter estimators (such as consistency and asymptotic normality)?

In such cases, are there alternative methods (such as robust regression) that can provide better estimates?