5. **[16 points] Neural networks with shortcut connections**

In this problem, we'll perform classification using a modified two-layer neural network. For any input vector $x \in \mathbb{R}^d$, our neural network outputs a probability distribution over 2 classes following the forward propagation rules:

$$\begin{aligned} z^{[1]} &= W^{[1]}x + b^{[1]} \\ a^{[1]} &= \mathrm{ReLU}(z^{[1]}) = \max(0, z^{[1]}) \\ z^{[2]} &= W^{[2]}a^{[1]} + b^{[2]} \\ \hat{y} = a^{[2]} &= \sigma(z^{[2]}) = \frac{1}{1+e^{-z^{[2]}}} \end{aligned}$$

where $W^{[1]} \in \mathbb{R}^{h\times d}$, $b^{[1]} \in \mathbb{R}^h$, $W^{[2]} \in \mathbb{R}^{1\times h}$, $b^{[2]} \in \mathbb{R}$. The first layer of the network is a fully-connected layer, followed by a Rectified Linear Unit (ReLU) activation function $\mathrm{ReLU}(z)$. The second layer of the network is a fully-connected layer, followed by a sigmoid activation function.

We evaluate our model using a mean squared loss. For a single example (x, y), the squared loss is:

$$\mathcal{L}(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2.$$

where $\hat{y} \in (0, 1)$ and $y \in \{0, 1\}$.

For n training examples, we average the mean squared loss over the n examples:

$$J(W^{[1]}, W^{[2]}, b^{[1]}, b^{[2]}) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\hat{y}^{(i)}, y^{(i)}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(\hat{y}^{(i)} - y^{(i)})^2.$$

We modify the described network by adding a "shortcut" connection between the input x and the second layer. The forward propagation equations then become:

$$\begin{aligned} z^{[1]} &= W^{[1]}x + b^{[1]} \\ a^{[1]} &= \mathrm{ReLU}(z^{[1]}) \\ z^{[2]} &= W^{[2]}a^{[1]} + b^{[2]} + Wx \\ \hat{y} = a^{[2]} &= \sigma(z^{[2]}) \end{aligned}$$

where $W \in \mathbb{R}^{1\times d}$, and $J(W^{[1]}, W^{[2]}, b^{[1]}, b^{[2]}, W)$ is defined as before.

Figure 1 (on the next page) shows the two-layer neural network, before and after adding the shortcut connection. In practice, it is often observed that shortcut connections improve the learning of neural networks.

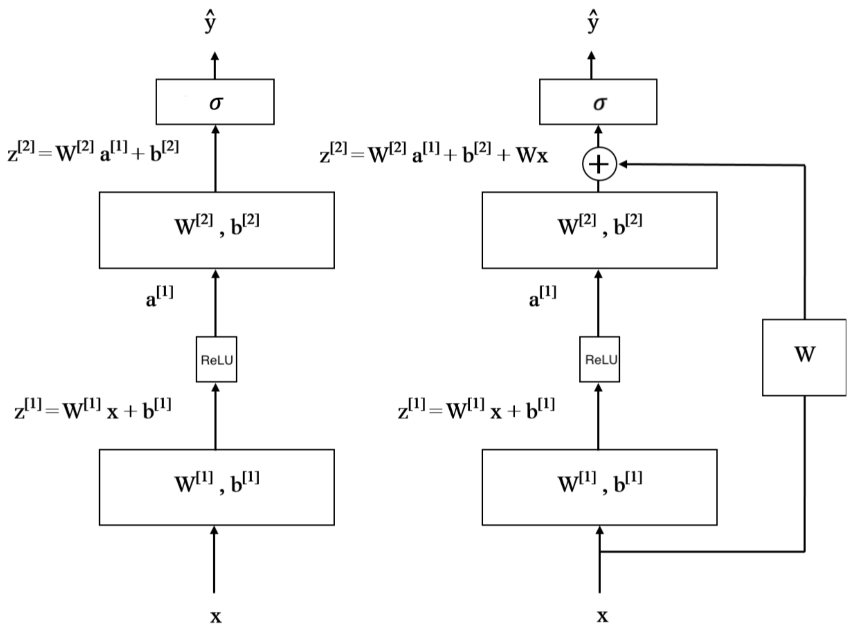

Figure 1: On the left, a two-layer neural network without shortcut connection. On the right, the same two-layer neural network with a shortcut connection.

(a) [4 points] How many parameters does the model including the shortcut connection have? Your answer should be expressed in terms of n, d, or h.

Answer: $W^{[1]}$: $h \times d$, $b^{[1]}$: h, $W^{[2]}$: $1 \times h$, $b^{[2]}$: 1, W: $1 \times d$.
Total number of parameters: $(d + 1)h + (h + 1) + d$.

(b) [12 points] Find the expressions for $\frac{\partial \mathcal{L}}{\partial W^{[1]}}$, $\frac{\partial \mathcal{L}}{\partial b^{[1]}}$, $\frac{\partial \mathcal{L}}{\partial W^{[2]}}$, $\frac{\partial \mathcal{L}}{\partial b^{[2]}}$, $\frac{\partial \mathcal{L}}{\partial W}$ of the neural network with a shortcut connection, given a **single** training example (x, y).
Hint: You may find the indicator function useful for the derivative of ReLU function.

Answer:

$$\frac{\partial \mathcal{L}}{\partial \hat{y}} = (\hat{y} - y) \qquad \in \mathbb{R}$$

$$\frac{\partial \mathcal{L}}{\partial b^{[2]}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial b^{[2]}} = (\hat{y} - y) \cdot (1 - \hat{y}) \cdot \hat{y} \qquad \in \mathbb{R}$$

$$\frac{\partial \mathcal{L}}{\partial W^{[2]}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial W^{[2]}} = (\hat{y} - y) \cdot (1 - \hat{y}) \cdot \hat{y} \cdot a^{[1]T} \qquad \in \mathbb{R}^{1 \times h}$$

$$\frac{\partial \mathcal{L}}{\partial W} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial W} = (\hat{y} - y) \cdot (1 - \hat{y}) \cdot \hat{y} \cdot x^T \qquad \in \mathbb{R}^{1 \times d}$$

$$\frac{\partial \mathcal{L}}{\partial a^{[1]}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial a^{[1]}} \cdot = (\hat{y} - y) \cdot (1 - \hat{y}) \cdot \hat{y} \cdot W^{[2]T} \qquad \in \mathbb{R}^{1 \times h}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial b^{[1]}} &= \frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial a^{[1]}} \cdot \frac{\partial a^{[1]}}{\partial z^{[1]}} \cdot \frac{\partial z^{[1]}}{\partial b^{[1]}} \\ &= (\hat{y} - y) \cdot (1 - \hat{y}) \cdot \hat{y} \cdot W^{[2]T} \cdot \mathbb{1}\{z^{[1]} > 0\} \qquad \in \mathbb{R}^{1 \times h} \\ \frac{\partial \mathcal{L}}{\partial W^{[1]}} &= \frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial a^{[1]}} \cdot \frac{\partial a^{[1]}}{\partial z^{[1]}} \cdot \frac{\partial z^{[1]}}{\partial W^{[1]}} \\ &= (\hat{y} - y) \cdot (1 - \hat{y}) \cdot \hat{y} \cdot W^{[2]T} \cdot \mathbb{1}\{z^{[1]} > 0\} \cdot x^T \qquad \in \mathbb{R}^{h \times d}\end{aligned}$$