Name: Ryan Toh
Student ID: 1005129

(Note: Questions 1 and 3 are in the attached `.ipynb` file.)

Question 2: Neural Networks

2.1:

$$\hat{y} = \sigma(\mathbf{z}_2), \mathbf{z}_2 = \mathbf{w}_2\mathbf{a}_1 + b_2$$

Differentiating $\mathbf{z}_2$ with respect to $\mathbf{w}_2$:

$$\frac{\partial \mathbf{z}_2}{\partial \mathbf{w}_2} = \mathbf{a}_1 \tag{2.1.1}$$

Differentiating $\mathbf{z}_2$ with respect to b_2:

$$\frac{\partial \hat{y}}{\partial b_2} = 1 \tag{2.1.2}$$

$$\begin{aligned} \frac{\partial \hat{y}}{\partial \mathbf{w}_2} &= \frac{\partial \hat{y}}{\partial \mathbf{z}_2}\frac{\partial \mathbf{z}_2}{\partial \mathbf{w}_2} \\ &= \mathbf{a}_1 \frac{\partial \hat{y}}{\partial \mathbf{z}_2} \quad \text{using (2.1.1)} \end{aligned} \tag{2.1.3}$$

$$\begin{aligned} \frac{\partial \hat{y}}{\partial b_2} &= \frac{\partial \hat{y}}{\partial \mathbf{z}_2}\frac{\partial \mathbf{z}_2}{\partial b_2} \\ &= \frac{\partial \hat{y}}{\partial \mathbf{z}_2} \quad \text{using (2.1.2)} \end{aligned} \tag{2.1.4}$$

2.2:

Assume the activation function is the softplus function:

$$\begin{aligned} \hat{y} &= \sigma(\mathbf{z}_2) \\ &= \log(1 + e^{\mathbf{z}_2}) \end{aligned}$$

Differentiating $\hat{y}$ with respect to $\mathbf{z}_2$, we have:

$$\begin{aligned} \frac{\partial \hat{y}}{\partial \mathbf{z}_2} &= \frac{\partial}{\partial \mathbf{z}_2}\left(\log(1 + e^{\mathbf{z}_2})\right) \\ &= \frac{1}{1 + e^{\mathbf{z}_2}}(e^{\mathbf{z}_2}) \\ &= \frac{e^{\mathbf{z}_2}}{1 + e^{\mathbf{z}_2}} \\ &= \frac{1}{1 + e^{-\mathbf{z}_2}} \end{aligned} \tag{2.2.1}$$

$$
\begin{aligned}
\frac{\partial \hat{y}}{\partial \mathbf{w}_2} &= \mathbf{a}_1 \frac{\partial \hat{y}}{\partial \mathbf{z}_2} \quad \text{using (2.1.3)} \\
&= \frac{\mathbf{a}_1}{1+e^{-\mathbf{z}_2}} \quad \text{using (2.2.1)}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \hat{y}}{\partial b_2} &= \frac{\partial \hat{y}}{\partial \mathbf{z}_2} \quad \text{using (2.1.4)} \\
&= \frac{1}{1+e^{-\mathbf{z}_2}} \quad \text{using (2.2.1)}
\end{aligned}
$$

2.3:

No, $\frac{\partial \hat{y}}{\partial \mathbf{x}}$ does not change.

$$
\begin{aligned}
\hat{y} = \sigma(\mathbf{z}_2), \mathbf{z}_2 = \mathbf{w}_2\mathbf{a}_1 + b_2 \\
\mathbf{a}_1 = \sigma(\mathbf{z}_1), \mathbf{z}_1 = \mathbf{w}_1\mathbf{x} + b_1
\end{aligned}
$$

$$
\frac{\partial \hat{y}}{\partial \mathbf{x}} = \frac{\partial \hat{y}}{\partial \mathbf{z}_2}\frac{\partial \mathbf{z}_2}{\partial \mathbf{a}_1}\frac{\partial \mathbf{a}_1}{\partial \mathbf{z}_1}\frac{\partial \mathbf{z}_1}{\partial \mathbf{x}} \tag{2.3.1}
$$

Differentiate $\mathbf{z}_2$ with respect to $\mathbf{a}_1$:

$$
\frac{\partial \mathbf{z}_2}{\partial \mathbf{a}_1} = \mathbf{w}_2 \tag{2.3.2a}
$$

Differentiate $\mathbf{a}_1$ with respect to $\mathbf{z}_1$:

$$
\begin{aligned}
\mathbf{a}_1 &= \sigma(\mathbf{z}_1) \\
&= \log(1+e^{\mathbf{z}_1})
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \mathbf{a}_1}{\partial \mathbf{z}_1} &= \frac{\partial}{\partial \mathbf{z}_1}\Big(\log(1+e^{\mathbf{z}_1})\Big) \\
&= \frac{1}{1+e^{-\mathbf{z}_1}} \quad \text{similar to (2.2.1)}
\end{aligned} \tag{2.3.2b}
$$

Differentiate $\mathbf{z}_1$ with respect to $\mathbf{x}$:

$$
\frac{\partial \mathbf{z}_1}{\partial \mathbf{x}} = \mathbf{w}_1 \tag{2.3.2c}
$$

Using (2.2.1), (2.3.2a), (2.3.2b) and (2.3.2c) on (2.3.1):

$$
\begin{aligned}
\frac{\partial \hat{y}}{\partial \mathbf{x}} &= \frac{\partial \hat{y}}{\partial \mathbf{z}_2}\frac{\partial \mathbf{z}_2}{\partial \mathbf{a}_1}\frac{\partial \mathbf{a}_1}{\partial \mathbf{z}_1}\frac{\partial \mathbf{z}_1}{\partial \mathbf{x}} \\
&= \Big(\frac{1}{1+e^{-\mathbf{z}_2}}\Big)(\mathbf{w}_2)\Big(\frac{1}{1+e^{-\mathbf{z}_1}}\Big)(\mathbf{w}_1) \\
&= \frac{\mathbf{w}_1\mathbf{w}_2}{(1+e^{-\mathbf{z}_2})(1+e^{-\mathbf{z}_1})}
\end{aligned}
$$

Since there is no b_2 value in this expression, a change in bias Δb_2 will not change the overall value.

Hence, $\frac{\partial \hat{y}}{\partial \mathbf{x}}$ does not change.

2.4:

Assume activation function is the logistic function:

$$\begin{aligned}\hat{y} &= \sigma(\mathbf{z}_2) \\ &= \frac{1}{1+e^{-\mathbf{z}_2}}\end{aligned} \tag{2.4.1}$$

$$\begin{aligned}\hat{y} = \sigma(\mathbf{z}_2), \mathbf{z}_2 = \mathbf{w}_2\mathbf{a}_1 + b_2 \\ \mathbf{a}_1 = \sigma(\mathbf{z}_1), \mathbf{z}_1 = \mathbf{w}_1\mathbf{x} + b_1\end{aligned}$$

$$\frac{\partial \hat{y}}{\partial \mathbf{w}_1} = \frac{\partial \hat{y}}{\partial \mathbf{z}_2}\frac{\partial \mathbf{z}_2}{\partial \mathbf{a}_1}\frac{\partial \mathbf{a}_1}{\partial \mathbf{z}_1}\frac{\partial \mathbf{z}_1}{\partial \mathbf{w}_1} \tag{2.4.2}$$

From $2.4.1$, we differentiate $\hat{y}$ with respect to $\mathbf{z}_2$:

$$\begin{aligned}\frac{\partial \hat{y}}{\partial \mathbf{z}_2} &= \frac{\partial}{\partial \mathbf{z}_2}\left((1+e^{-\mathbf{z}_2})^{-1}\right) \\ &= -(1+e^{-\mathbf{z}_2})^{-2}e^{-\mathbf{z}_2}(-1) \\ &= \frac{e^{-\mathbf{z}_2}}{(1+e^{-\mathbf{z}_2})^2} \\ &= \frac{1}{1+e^{-\mathbf{z}_2}} \cdot \frac{e^{-\mathbf{z}_2}}{1+e^{-\mathbf{z}_2}} \\ &= \frac{1}{1+e^{-\mathbf{z}_2}} \cdot \left(1 - \frac{1}{1+e^{-\mathbf{z}_2}}\right) \\ &= \hat{y}(1-\hat{y})\end{aligned} \tag{2.4.3a}$$

Differentiate $\mathbf{z}_2$ with respect to $\mathbf{a}_1$:

$$\frac{\partial \mathbf{z}_2}{\partial \mathbf{a}_1} = \mathbf{w}_2 \tag{2.4.3b}$$

Differentiate $\mathbf{a}_1$ with respect to $\mathbf{z}_1$:

$$\begin{aligned}\frac{\partial \mathbf{a}_1}{\partial \mathbf{z}_1} &= \sigma(\mathbf{z}_1) \\ &= \mathbf{a}_1(1-\mathbf{a}_1) \quad \text{similar to (2.4.3a)}\end{aligned} \tag{2.4.3c}$$

Differentiate $\mathbf{z}_1$ with respect to $\mathbf{w}_1$:

$$\frac{\partial \mathbf{z}_1}{\partial \mathbf{w}_1} = \mathbf{x} \tag{2.4.3d}$$

Using $(2.4.3a)$, $(2.4.3b)$, $(2.4.3c)$ and $(2.4.3d)$ on $(2.4.2)$:

$$\begin{aligned}\frac{\partial \hat{y}}{\partial \mathbf{w}_1} &= \frac{\partial \hat{y}}{\partial \mathbf{z}_2}\frac{\partial \mathbf{z}_2}{\partial \mathbf{a}_1}\frac{\partial \mathbf{a}_1}{\partial \mathbf{z}_1}\frac{\partial \mathbf{z}_1}{\partial \mathbf{w}_1} \\ &= \big(\hat{y}(1-\hat{y})\big)(\mathbf{w}_2)\big(\mathbf{a}_1(1-\mathbf{a}_1)\big)(\mathbf{x}) \\ &= \mathbf{a}_1\mathbf{w}_2\mathbf{x}\hat{y}(1-\mathbf{a}_1)(1-\hat{y})\end{aligned}$$