Question 3: Baby Convolutional Neural Network Solution

(a) Compute gradients $\frac{\partial h}{\partial \theta_j}$ and $\frac{\partial h}{\partial w_j}$ [6 marks]

Given:

$$u_1 = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_4 + \theta_4 x_5 \tag{1}$$
$$u_2 = \theta_1 x_2 + \theta_2 x_3 + \theta_3 x_5 + \theta_4 x_6 \tag{2}$$
$$u_3 = \theta_1 x_4 + \theta_2 x_5 + \theta_3 x_7 + \theta_4 x_8 \tag{3}$$
$$u_4 = \theta_1 x_5 + \theta_2 x_6 + \theta_3 x_8 + \theta_4 x_9 \tag{4}$$
$$h = \frac{1}{1 + e^{-(w_1 u_1 + w_2 u_2 + w_3 u_3 + w_4 u_4)}} \tag{5}$$

Let $z = w_1 u_1 + w_2 u_2 + w_3 u_3 + w_4 u_4$, so $h = \frac{1}{1+e^{-z}} = \sigma(z)$.
Using the chain rule:

$$\frac{\partial h}{\partial \theta_j} = \frac{\partial h}{\partial z} \cdot \frac{\partial z}{\partial \theta_j} \tag{6}$$
$$\frac{\partial h}{\partial w_j} = \frac{\partial h}{\partial z} \cdot \frac{\partial z}{\partial w_j} \tag{7}$$

First, compute $\frac{\partial h}{\partial z}$:

$$\frac{\partial h}{\partial z} = \frac{\partial}{\partial z}\left(\frac{1}{1+e^{-z}}\right) = \frac{e^{-z}}{(1+e^{-z})^2} = h(1-h) \tag{8}$$

Next, compute $\frac{\partial z}{\partial \theta_j}$:

$$\frac{\partial z}{\partial \theta_1} = w_1 x_1 + w_2 x_2 + w_3 x_4 + w_4 x_5 \tag{9}$$
$$\frac{\partial z}{\partial \theta_2} = w_1 x_2 + w_2 x_3 + w_3 x_5 + w_4 x_6 \tag{10}$$
$$\frac{\partial z}{\partial \theta_3} = w_1 x_4 + w_2 x_5 + w_3 x_7 + w_4 x_8 \tag{11}$$
$$\frac{\partial z}{\partial \theta_4} = w_1 x_5 + w_2 x_6 + w_3 x_8 + w_4 x_9 \tag{12}$$

And $\frac{\partial z}{\partial w_j}$:

$$\frac{\partial z}{\partial w_1} = u_1, \quad \frac{\partial z}{\partial w_2} = u_2, \quad \frac{\partial z}{\partial w_3} = u_3, \quad \frac{\partial z}{\partial w_4} = u_4 \tag{13}$$

Therefore:

$$\frac{\partial h}{\partial \theta_1} = h(1-h)(w_1x_1 + w_2x_2 + w_3x_4 + w_4x_5) \tag{14}$$

$$\frac{\partial h}{\partial \theta_2} = h(1-h)(w_1x_2 + w_2x_3 + w_3x_5 + w_4x_6) \tag{15}$$

$$\frac{\partial h}{\partial \theta_3} = h(1-h)(w_1x_4 + w_2x_5 + w_3x_7 + w_4x_8) \tag{16}$$

$$\frac{\partial h}{\partial \theta_4} = h(1-h)(w_1x_5 + w_2x_6 + w_3x_8 + w_4x_9) \tag{17}$$

$$\frac{\partial h}{\partial w_1} = h(1-h)u_1, \quad \frac{\partial h}{\partial w_2} = h(1-h)u_2, \quad \frac{\partial h}{\partial w_3} = h(1-h)u_3, \quad \frac{\partial h}{\partial w_4} = h(1-h)u_4 \tag{18}$$

(b) Derive the negative log-likelihood function [3 marks]

For binary classification with training dataset $\{(x_i, y_i)\}_{i=1}^N$ where $x_i \in \mathbb{R}^9$ and $y_i \in \{0, 1\}$:

The likelihood for each data point is:

$$P(y_i|x_i) = h_i^{y_i}(1-h_i)^{1-y_i}$$

where $h_i = h(x_i, \theta, w)$.

The total likelihood is:

$$L(\theta, w) = \prod_{i=1}^{N} P(y_i|x_i) = \prod_{i=1}^{N} h_i^{y_i}(1-h_i)^{1-y_i}$$

The log-likelihood is:

$$\ell(\theta, w) = \sum_{i=1}^{N} [y_i \log h_i + (1-y_i)\log(1-h_i)]$$

Therefore, the negative log-likelihood function to minimize is:

$$\boxed{\text{NLL}(\theta, w) = -\sum_{i=1}^{N} [y_i \log h_i + (1-y_i)\log(1-h_i)]}$$

This is the binary cross-entropy loss function.

(c) SGD Algorithm with Mini-batch [8 marks]

Mini-batch SGD proceeds by first sampling a batch of datapoints $\mathcal{B} = \{j_1, j_2, \ldots, j_{32}\}$, and then perform the updates

$$\theta^{t+1} = \theta^t - \alpha \cdot h_\theta(\theta^t, w^t; \mathcal{B}) \tag{19}$$

$$w^{t+1} = w^t - \alpha \cdot h_w(\theta^t, w^t; \mathcal{B}) \tag{20}$$

where $h_\theta(\theta^t, w^t; \mathcal{B})$ and $h_w(\theta^t, w^t; \mathcal{B})$ are computed as the average gradients over the batch $\mathcal{B}$:

$$h_\theta(\theta^t, w^t; \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} h_\theta(\theta^t, w^t; j) \tag{21}$$

$$h_w(\theta^t, w^t; \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} h_w(\theta^t, w^t; j) \tag{22}$$

where for each individual sample j in the batch:

- If $y_j = 1$, by the chain rule,

$$h_\theta(\theta^t, w^t; j) = \frac{\partial(-\ln h_j)}{\partial \theta} = -\frac{1}{h_j} \cdot \frac{\partial h_j}{\partial \theta} \tag{23}$$

$$= -\frac{1}{h_j} \cdot h_j(1 - h_j) \cdot \frac{\partial z_j}{\partial \theta} \tag{24}$$

$$= -(1 - h_j) \cdot \frac{\partial z_j}{\partial \theta} \tag{25}$$

$$h_w(\theta^t, w^t; j) = \frac{\partial(-\ln h_j)}{\partial w} = -\frac{1}{h_j} \cdot \frac{\partial h_j}{\partial w} \tag{26}$$

$$= -\frac{1}{h_j} \cdot h_j(1 - h_j) \cdot \frac{\partial z_j}{\partial w} \tag{27}$$

$$= -(1 - h_j) \cdot u_j \tag{28}$$

- If $y_j = 0$, by the chain rule,

$$h_\theta(\theta^t, w^t; j) = \frac{\partial(-\ln(1 - h_j))}{\partial \theta} = -\frac{1}{1 - h_j} \cdot \frac{\partial(1 - h_j)}{\partial \theta} \tag{29}$$

$$= \frac{1}{1 - h_j} \cdot h_j(1 - h_j) \cdot \frac{\partial z_j}{\partial \theta} \tag{30}$$

$$= h_j \cdot \frac{\partial z_j}{\partial \theta} \tag{31}$$

$$h_w(\theta^t, w^t; j) = \frac{\partial(-\ln(1 - h_j))}{\partial w} = -\frac{1}{1 - h_j} \cdot \frac{\partial(1 - h_j)}{\partial w} \tag{32}$$

$$= \frac{1}{1 - h_j} \cdot h_j(1 - h_j) \cdot \frac{\partial z_j}{\partial w} \tag{33}$$

$$= h_j \cdot u_j \tag{34}$$

where $h_j = \frac{1}{1+e^{-z_j}}$ is the sigmoid output, $z_j = w^T u_j$ is the linear combination, and u_j are the filter outputs for sample j.

Combining both cases: For any sample j, we can write:

$$h_\theta(\theta^t, w^t; j) = (h_j - y_j) \cdot \frac{\partial z_j}{\partial \theta} \tag{35}$$

$$h_w(\theta^t, w^t; j) = (h_j - y_j) \cdot u_j \tag{36}$$