***Solution* 1 :**

1. Forward Pass:

$$z_H = W_H \cdot x + b_H \tag{1}$$

$$z_H = \begin{bmatrix} 2 & 4 \\ 3 & -5 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} -2 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ -2 \end{bmatrix} + \begin{bmatrix} -2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ -1 \end{bmatrix}$$

$$u_H = ReLU(z_H) = \begin{bmatrix} 4 \\ 0 \end{bmatrix} \tag{2}$$

$$z_O = W_O \cdot u_H + b_O \tag{3}$$

$$z_O = \begin{bmatrix} 4 & 6 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 0 \end{bmatrix} + \begin{bmatrix} -3.5 \end{bmatrix} = \begin{bmatrix} 16 - 3.5 \end{bmatrix} = \begin{bmatrix} 12.5 \end{bmatrix}$$

$$\hat{y} = sigmoid(z_O) = 1/(1 + e^{-z_O}) = 1/(1 + e^{-12.5}) = 0.9999962734$$

$$l(\hat{y}, y) = -log(\hat{y}) = 1.6185 \times 10^{-6}$$

---

2. Backward Pass:

(*a*) *Find* $\boldsymbol{\frac{\partial l}{\partial W_O}}$

$$\frac{\partial l}{\partial W_O} = \begin{bmatrix} \frac{\partial l}{\partial W_{O1}} & \frac{\partial l}{\partial W_{O2}} \end{bmatrix} \tag{4}$$

Applying chain rule with respect to children of previous nodes,

$$\frac{\partial l}{\partial W_O} = \frac{\partial l(\hat{y}, y)}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_O} \times \begin{bmatrix} \frac{\partial z_O}{\partial W_{O1}} & \frac{\partial z_O}{\partial W_{O2}} \end{bmatrix} \tag{5}$$

$$\frac{\partial l(\hat{y}, y)}{\partial \hat{y}} = \frac{\partial(-y.log(\hat{y}) - (1 - y).log(1 - \hat{y}))}{\partial \hat{y}} \tag{6}$$

At y=1,

$$\frac{\partial l(\hat{y}, y)}{\partial \hat{y}} = \frac{-1}{\hat{y}} \tag{7}$$

$$\frac{\partial \hat{y}}{\partial z_O} = \frac{\partial(1/(1 + e^{-z_O}))}{\partial z_O} \tag{8}$$

Calculating the partial derivative and re framing the expression in terms of $\hat{y}$

$$\frac{\partial \hat{y}}{\partial z_O} = \hat{y}.(1 - \hat{y}) \tag{9}$$

$$\frac{\partial z_O}{\partial W_{O1}} = \frac{\partial (W_0 \cdot u_H + b_O)}{\partial W_{O1}} \tag{10}$$

After expressing $W_0 \cdot u_H + b_O = (W_{01} \cdot u_{H1}) + (W_{02} \cdot u_{H2}) + b_O$ and taking partial derivative,

$$\frac{\partial z_O}{\partial W_{O1}} = u_{H1} \tag{11}$$

Similarly,

$$\frac{\partial z_O}{\partial W_{O2}} = u_{H2} \tag{12}$$

Combining all the terms, $\frac{\partial l}{\partial W_O} = \frac{-1}{\hat{y}} \times \hat{y}.(1 - \hat{y}) \times \begin{bmatrix} u_{H1} & u_{H2} \end{bmatrix}$

$$\frac{\partial l}{\partial W_O} = (0.9999962734 - 1) \times \begin{bmatrix} 4 & 0 \end{bmatrix}$$

$$\frac{\partial l}{\partial W_O} = \begin{bmatrix} -1.49064 \times 10^{-5} & 0 \end{bmatrix}$$

(b) *Find* $\frac{\partial l}{\partial b_O}$

Applying chain rule with respect to children of previous nodes,

$$\frac{\partial l}{\partial b_O} = \frac{\partial l(\hat{y}, y)}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_O} \times \left[\frac{\partial z_O}{\partial b_0} \right] \tag{13}$$

The derivation is similar to the previous one till (9)

$$\frac{\partial z_O}{\partial b_0} = \frac{\partial (W_0 \cdot u_H + b_O)}{\partial b_0} = 1 \tag{14}$$

Combining all the terms, $\frac{\partial l}{\partial b_O} = \frac{-1}{\hat{y}} \times \hat{y}.(1 - \hat{y}) \times \begin{bmatrix} b_0 \end{bmatrix}$

$$\frac{\partial l}{\partial b_O} = (0.9999962734 - 1) \times \begin{bmatrix} 1 \end{bmatrix}$$

$$\frac{\partial l}{\partial b_O} = \begin{bmatrix} -3.7266 \times 10^{-6} \end{bmatrix}$$

(c) *Find* $\frac{\partial l}{\partial W_H}$

$$\frac{\partial l}{\partial W_H} = \begin{bmatrix} \frac{\partial l}{\partial W_{H11}} & \frac{\partial l}{\partial W_{H12}} \\ \frac{\partial l}{\partial W_{H21}} & \frac{\partial l}{\partial W_{H22}} \end{bmatrix} \tag{15}$$

Applying chain rule with respect to children of previous nodes,

$$\frac{\partial l}{\partial W_H} = \frac{\partial l(\hat{y}, y)}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_O} \times \frac{\partial z_O}{\partial u_H} \times \frac{\partial u_H}{\partial z_H} \times \frac{\partial z_H}{\partial W_H} \tag{16}$$

$$\frac{\partial l}{\partial W_{H11}} = \frac{\partial l(\hat{y}, y)}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_O} \times \frac{\partial z_O}{\partial u_{H1}} \times \frac{\partial u_{H1}}{\partial z_{H1}} \times \frac{\partial z_{H1}}{\partial W_{H11}} \tag{17}$$

$$\frac{\partial l}{\partial W_{H12}} = \frac{\partial l(\hat{y}, y)}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_O} \times \frac{\partial z_O}{\partial u_{H1}} \times \frac{\partial u_{H1}}{\partial z_{H1}} \times \frac{\partial z_{H1}}{\partial W_{H12}} \tag{18}$$

$$\frac{\partial l}{\partial W_{H21}} = \frac{\partial l(\hat{y}, y)}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_O} \times \frac{\partial z_O}{\partial u_{H2}} \times \frac{\partial u_{H2}}{\partial z_{H2}} \times \frac{\partial z_{H2}}{\partial W_{H21}} \tag{19}$$

$$\frac{\partial l}{\partial W_{H22}} = \frac{\partial l(\hat{y}, y)}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_O} \times \frac{\partial z_O}{\partial u_{H2}} \times \frac{\partial u_{H2}}{\partial z_{H2}} \times \frac{\partial z_{H2}}{\partial W_{H22}} \tag{20}$$

Using (7) and (9),

$$\frac{\partial l(\hat{y}, y)}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_O} = \hat{y} - 1 \tag{21}$$

Using (3),
After expressing $W_0 \cdot u_H + b_O = (W_{01} \cdot u_{H1}) + (W_{02} \cdot u_{H2}) + b_O$ and taking partial derivative,

$$\frac{\partial z_O}{\partial u_{H1}} = W_{01} \; and \; \frac{\partial z_O}{\partial u_{H2}} = W_{02} \tag{22}$$

ReLU(x)= x when x is positive and 0 when x is negative. Hence, the differential of ReLU(x) with respect to x would be 1 if the input function is positive and 0 if negative. By this property,

$$\frac{\partial u_{H1}}{\partial z_{H1}} = 1 \; and \; \frac{\partial u_{H2}}{\partial z_{H2}} = 0 \tag{23}$$

After expressing $z_H = W_H \cdot x + b_H$, we can split the matrices into element wise operations for ease of calculation

$$z_{H1} = (W_{H11} \cdot x_1) + (W_{H12} \cdot x_2) + b_1 \tag{24}$$

$$z_{H2} = (W_{H21} \cdot x_1) + (W_{H22} \cdot x_2) + b_2 \tag{25}$$

Taking partial derivatives of above,

$$\frac{\partial z_{H1}}{\partial W_{H11}} = x_1 \; and \; \frac{\partial z_{H1}}{\partial W_{H12}} = x_2 \qquad \frac{\partial z_{H2}}{\partial W_{H21}} = x_1 \; and \; \frac{\partial z_{H2}}{\partial W_{H22}} = x_2 \tag{26}$$

Substituting (21), (22), (23) and (26) in (17),(18),(19),(20)

$$\frac{\partial l}{\partial W_{H11}} = (\hat{y} - 1) \times W_{01} \times 1 \times x_1 = -1.49064 \times 10^{-5}$$
$$\frac{\partial l}{\partial W_{H12}} = (\hat{y} - 1) \times W_{01} \times 1 \times x_2 = -1.49064 \times 10^{-5}$$
$$\frac{\partial l}{\partial W_{H21}} = (\hat{y} - 1) \times W_{02} \times 0 \times x_2 = 0$$
$$\frac{\partial l}{\partial W_{H22}} = (\hat{y} - 1) \times W_{02} \times 0 \times x_2 = 0$$

Hence, $\frac{\partial l}{\partial W_H} = \begin{bmatrix} -1.49064 \times 10^{-5} & -1.49064 \times 10^{-5} \\ 0 & 0 \end{bmatrix}$

(d) *Find* $\boldsymbol{\frac{\partial l}{\partial b_H}}$

$$\frac{\partial l}{\partial b_H} = \begin{bmatrix} \frac{\partial l}{\partial b_{H1}} \\ \frac{\partial l}{\partial b_{H2}} \end{bmatrix} \tag{27}$$

Applying chain rule with respect to children of previous nodes,

$$\frac{\partial l}{\partial b_H} = \frac{\partial l(\hat{y}, y)}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_O} \times \frac{\partial z_O}{\partial u_H} \times \frac{\partial u_H}{\partial z_H} \times \frac{\partial z_H}{\partial b_H} \tag{28}$$

$$\frac{\partial l}{\partial b_{H1}} = \frac{\partial l(\hat{y}, y)}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_O} \times \frac{\partial z_O}{\partial u_{H1}} \times \frac{\partial u_{H1}}{\partial z_{H1}} \times \frac{\partial z_{H1}}{\partial b_{H1}} \tag{29}$$

$$\frac{\partial l}{\partial b_{H2}} = \frac{\partial l(\hat{y}, y)}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z_O} \times \frac{\partial z_O}{\partial u_{H2}} \times \frac{\partial u_{H2}}{\partial z_{H2}} \times \frac{\partial z_{H2}}{\partial b_{H2}} \tag{30}$$

Using (24) and (25),

$$\frac{\partial z_{H1}}{\partial b_{H1}} = 1 \; and \; \frac{\partial z_{H2}}{\partial b_{H2}} = 1 \tag{31}$$

Substituting (21), (22), (23) and (31) in (29), (30)

$$\frac{\partial l}{\partial b_{H1}} = (\hat{y} - 1) \times W_{01} \times 1 \times 1 = -1.49064 \times 10^{-5}$$
$$\frac{\partial l}{\partial b_{H2}} = (\hat{y} - 1) \times W_{01} \times 0 \times 1 = 0$$

Hence,

$$\frac{\partial l}{\partial b_H} = \begin{bmatrix} -1.49064 \times 10^{-5} \\ 0 \end{bmatrix}$$

---

3. Yes, some derivatives were zero. These values of zero were propagated back due to ReLU function. ReLU(x)= x when x is positive and 0 when x is negative. Hence, the differential of ReLU(x) with respect to x would be 1 if the input function is positive and 0 if negative.

The zeros obtained in our case were due to z_{H2} being negative (-1) as observed by the value of z_H upon solving (1). These values of z_{H2} is back propagated as observed in (19), (20) and (30)

***Solution* 2 :**

1. Role of bias correction in ADAM optimizer -

Upon running the ADAM optimizer without the bias correction (like in 3^{rd} part of Q2) it was observed that the initial iterations provide a pretty bad average when compared to further iterations. This is because the moving averages m_t and v_t are initialized as (vectors) of 0's and hence leading to estimates that are heavily biased towards 0. Moreover, in the initial stages, we do not have large amount of values to average over and the moving averages are highly sensitive to β_1 and β_2. Thus, the bias correction helps in providing better averages for such cases.

Reference - Section 2 of ADAM Optimizer paper

---

2. Based on (4) of Section 3 of ADAM Optimizer paper -

$E[v_t] = E[{g_t}^2].(1 - {\beta_2}^t) + \zeta$

Using the bias correction formula from the algorithm, $E[\hat{v}_t] = \dfrac{E[{g_t}^2].(1 - {\beta_2}^t) + \zeta}{(1 - {\beta_2}^t)}$

We can use the similar derivation as given in paper to obtain the below equation

$$E[m_t] = E[g_t].(1 - {\beta_1}^t) + \zeta \tag{32}$$

$$E[\hat{m}_t] = \frac{E[g_t].(1 - {\beta_1}^t) + \zeta}{(1 - {\beta_1}^t)} \tag{33}$$

The paper also says that the value of $\zeta = 0$ if the moment is stationary (which is given condition in the question). Thus (33) can be written as

$E[\hat{m}_t] = E[g_t]$ when bias correction is applied.
$E[g_t] = E[g] = G = 2$ as provided in question

Hence, $E[\hat{m}_t] = 2$ is independent of t and hence the same value for t=2,10,100

---

3. Without using bias-correction,

$E[\hat{m}_t] = E[m_t] = E[g_t].(1 - {\beta_1}^t)$ (From (32) and $\zeta = 0$)
$E[\hat{m}_t] = 2 \times (1 - 0.9^t)$

At t = 2, $E[\hat{m}_t] = 2 \times (1 - 0.9^2) = 0.38$
At t = 10, $E[\hat{m}_t] = 2 \times (1 - 0.9^{10}) = 1.30264312$
At t = 100, $E[\hat{m}_t] = 2 \times (1 - 0.9^{100}) = 1.999946877$

***Solution* 3 :**

In attached ipynb file