1)

$$V_t = \beta_1 V_{t-1} + (1-\beta_1)\frac{\partial J}{\partial w}$$

$$S_t = \beta_2 S_{t-1} + (1-\beta_2)\left(\frac{\partial J}{\partial w}\right)^2$$

$$V_{corr} = \frac{V_t}{1-\beta_1^t}$$

$$S_{corr} = \frac{S_t}{1-\beta_2^t}$$

$$W_t = W_{t-1} - \alpha \frac{V_{corr}}{\sqrt{S_{corr}+\varepsilon}}$$

Using given data and calculating using python we get:

$$V_2 = \begin{bmatrix} -0.165 & 0.245 & -0.059 \\ -0.196 & -0.032 & 0.209 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 0.011 & 0.070 & 0.110 \\ 0.141 & 0.170 & 0.021 \end{bmatrix}$$

$$W_2 = \begin{bmatrix} -0.182 & 0.308 & 0.674 \\ 0.502 & -0.858 & -0.934 \end{bmatrix}$$

② $h^{(1)} = \sigma(1.x) - 0.5$

$$h^{(2)} = \sigma(w.h^{(1)}) - 0.5$$

$$h^{(n)} = \sigma(w.h^{(n-1)}) - 0.5$$

$$y = 1.h^{(T)}$$

$$L = f(y)$$

$$\frac{\partial L}{\partial h^k} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial h^T} \cdot \frac{\partial h^T}{\partial h^{T-1}} \cdots\cdots \frac{\partial h^{k+1}}{\partial h^k}$$

$$\nabla h^{t+1} = \nabla\sigma(w.h^{(t)})$$
$$= \sigma(wh^{(t)}) \cdot (1 - \sigma(wh^{(t)})) \cdot w \cdot \nabla h^{(t)}$$

↳ using chain rule

if $x = 0$

$$\nabla h^{t+1} = 0.5(1 - 0.5) \cdot w \cdot \nabla h^{(t)}$$

$w > 4 \rightarrow$ gradient explodes

$w < 4 \rightarrow$ gradient vanishes

$\therefore \boxed{\alpha = 4}$

③ a) False

If $x=0$

$$f_t = \sigma(U_f h_{t-1} + b_f)$$

$$i_t = \sigma(U_i h_{t-1} + b_i)$$

$$\tilde{C}_t = \tanh(U_c h_{t-1} + b_c)$$

$$\Rightarrow C_t = \frac{1}{1+e^{-U_f h_{t-1} - b_f}} \cdot C_{t-1} + \frac{1}{1+e^{-U_i h_{t-1} - b_i}} \cdot \frac{e^{U_c h_{t-1} + b_c} - e^{-U_c h_{t-1} - b_c}}{e^{U_c h_{t-1} + b_c} + e^{-U_c h_{t-1} - b_c}}$$

For general activation functions & values of the cell state we cannot say $h_{t-1} = h_t$

b) True

The forget gate's activation function appear in the backpropagation equations. Thus the gate control's what the neural network "forgets". Hence if f_t is very small the gradient will become very small and the error will not backpropagate

c) True

The range of the sigmoid function is $(0, 1)$ as f_t, i_t & o_t are sigmoid activations they are non negative by extension

d) False

While all entries of f_t, i_t & o_t are constrained to be non negative & between 0 and 1 there is no constraint that they should sum up to 1 Hence it cannot be viewed as a probability distribution

(4)

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$

$$\tilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{C}_t$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$

$$h_t = o_t \odot \tanh(c_t)$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad ; \quad \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

a) Dimensions

f_t : 1×1

i_t : 1×1

o_t : 1×1

h_t : 1×1

b) Using the formulas:

$$f_1 = 0.7685$$
$$i_1 = 0.2497$$
$$\tilde{c}_1 = 0.9051$$
$$c_1 = f_1 \times 0 + i_1 \times \tilde{c}_1 = 0.2261$$
$$o_1 = 0.9781$$
$$h_1 = 0.9781 \times \tanh(0.2261) = \boxed{0.2174}$$

Using these we get

$$f_2 = 0.2330$$
$$i_2 = 0.4588$$
$$\tilde{c}_2 = -0.5875$$
$$c_2 = 0.2330 \times 0.2261 - 0.4588 \times 0.5875$$
$$= -0.2169$$
$$o_2 = 0.8892$$

$$\therefore h_2 = 0.8892 \times \tanh(-0.2169)$$
$$\boxed{= -0.1899}$$

c) $MSE_1 = (h_1 - y_1)^2 = (0.2174 - 0.5)^2 = \underline{0.0798}$

$MSE_2 = (h_2 - y_2)^2 = (-0.1899 - 0.8)^2 = \underline{0.98}$