1 Preprocess

Alphabet Size: $|\mathcal{A}| = 4$
Character to One-hot Encoding:
Given a vocabulary $\mathcal{A} = \{h, e, l, o\}$, the one-hot encoding of a character $c \in \mathcal{A}$ is defined as:

$$\text{one\_hot}(c)_i = \begin{cases} 1 & \text{if } \mathcal{A}_i = c, \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} h &: [1\ 0\ 0\ 0] \\ e &: [0\ 1\ 0\ 0] \\ l &: [0\ 0\ 1\ 0] \\ o &: [0\ 0\ 0\ 1] \end{aligned}$$

Input and Target Sequences:

$$\begin{aligned} x_1 &= h & y_1 &= e \\ x_2 &= e & y_2 &= l \\ x_3 &= l & y_3 &= l \\ x_4 &= l & y_4 &= o \end{aligned}$$

Replacing characters with one-hot encoding:

$$\begin{aligned} x_1^o &= [1\ 0\ 0\ 0] & y_1 &= [0\ 1\ 0\ 0] \\ x_2^o &= [0\ 1\ 0\ 0] & y_2 &= [0\ 0\ 1\ 0] \\ x_3^o &= [0\ 0\ 1\ 0] & y_3 &= [0\ 0\ 1\ 0] \\ x_4^o &= [0\ 0\ 1\ 0] & y_4 &= [0\ 0\ 0\ 1] \end{aligned}$$

2 Forward Pass Formulas

For a sequence of characters $x_1, x_2, \ldots, x_T$, the network computes: 1. Embeding Layer for computing $x_1, x_2, \ldots, x_T$

$$x_t = x_t^o E$$

2. Hidden State at time t, h_t:

$$h_t = \tanh(Ux_t + Wh_{t-1} + b)$$

3. Calculating Attention Score, $\mathcal{A}$:

$$\mathcal{A}_{t,i} = h_i \cdot h_t \quad for \quad i = 1, \ldots t$$

4. Calculating the Weights of the Attention Scores for each Hidden State:

$$\text{softmax}(\mathcal{A}_{t,i}) = \frac{e^{\mathcal{A}_{t,i}}}{\sum_{k=0}^{t} e^{\mathcal{A}_{t,k}}} \quad for \quad i = 1, \dots t$$

5. Context Vector, Sum of Weighted Attention Score

$$\mathcal{Z}_t = \sum_{k=0}^{t} \text{softmax}(\mathcal{A}_{t,k}) \cdot h_k$$

6. Output before softmax, Ω_t:

$$\Omega_t = V\mathcal{Z}_t + c$$

7. Softmax Output, $\hat{y}_t$, for each character:

$$\text{softmax}(\Omega_t) = \hat{y}_t = \frac{e^{\Omega_t}}{\sum_{k=1}^{|\mathcal{A}|} e^{\Omega_{t,k}}} \text{ for } t = 1, \dots, |\mathcal{A}|$$

8. Cross-Entropy Loss for the correct character y_t:

$$L_t = -\ y_t\ ln(\hat{y}_t)$$

3 Backpropagation Through Time Formulas

Gradients of the loss L with respect to the parameters U, W, V, b, c, E are computed as follows:

1. Gradient of Loss w.r.t. Output (Softmax Gradient):

$$\frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} = (\hat{y}_t - y_t)$$

2. Updates for V and c:

$$\begin{aligned}
\frac{\partial L}{\partial V} &= \sum_{t=1}^{S} \frac{\partial L_t}{\partial V} \\
&= \sum_{t=1}^{S} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial V} \\
&= \sum_{t=1}^{S} (\hat{y}_t - y_t) \cdot h_t^\top \\
\frac{\partial L}{\partial c} &= \sum_{t=1}^{T} \frac{\partial L_t}{\partial c} \\
&= \sum_{t=1}^{T} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial c} \\
&= \sum_{t=1}^{T} (\hat{y}_t - y_t)
\end{aligned}$$

3. Updates for U, W, b and E:

$$\begin{aligned}
\frac{\partial L}{\partial W} &= \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial \mathcal{Z}_t} \Big(\sum_{m=1}^{t} \sum_{k=1}^{m} \frac{\partial \mathcal{Z}_t}{\partial h_m} \prod_{j=k}^{m-1} \frac{\partial h_{j+1}}{\partial h_j} \Big(\frac{\partial h_k}{\partial W} \Big) \Big) \\
&= \sum_{t=1}^{\mathcal{S}} (y_t - \hat{y}_t)\ V^\top \Big(\sum_{m=1}^{t} \sum_{k=1}^{m} \mathcal{A}_{t,m} \prod_{j=k}^{m-1} W^\top\ (1 - h_{j+1}^2) \Big((1 - h_k^2)\ h_{k-1} \Big) \Big) \\
\frac{\partial L}{\partial U} &= \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial U} = \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial \mathcal{Z}_t} \Big(\sum_{m=1}^{t} \sum_{k=1}^{m} \frac{\partial \mathcal{Z}_t}{\partial h_m} \prod_{j=k}^{m-1} \frac{\partial h_{j+1}}{\partial h_j} \Big(\frac{\partial h_k}{\partial U} \Big) \Big) \\
&= \sum_{t=1}^{\mathcal{S}} (y_t - \hat{y}_t)\ V^\top \Big(\sum_{m=1}^{t} \sum_{k=1}^{m} \mathcal{A}_{t,m} \prod_{j=k}^{m-1} W^\top\ (1 - h_{j+1}^2) \Big((1 - h_k^2)\ x_k \Big) \Big) \\
\frac{\partial L}{\partial b} &= \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial \mathcal{Z}_t} \Big(\sum_{m=1}^{t} \sum_{k=1}^{m} \frac{\partial \mathcal{Z}_t}{\partial h_m} \prod_{j=k}^{m-1} \frac{\partial h_{j+1}}{\partial h_j} \Big(\frac{\partial h_k}{\partial b} \Big) \Big) \\
&= \sum_{t=1}^{\mathcal{S}} (y_t - \hat{y}_t)\ V^\top \Big(\sum_{m=1}^{t} \sum_{k=1}^{m} \mathcal{A}_{t,m} \prod_{j=k}^{m-1} W^\top\ (1 - h_{j+1}^2) \Big((1 - h_k^2) \Big) \Big) \\
\frac{\partial L}{\partial E} &= \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial E} = \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial \mathcal{Z}_t} \Big(\sum_{m=1}^{t} \sum_{k=1}^{m} \frac{\partial \mathcal{Z}_t}{\partial h_m} \prod_{j=k}^{m-1} \frac{\partial h_{j+1}}{\partial h_j} \Big(\frac{\partial h_k}{\partial x_k} \frac{\partial x_k}{\partial E} \Big) \Big) \\
&= \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial E} = \sum_{t=1}^{\mathcal{S}} (y_t - \hat{y}_t)\ V^\top \Big(\sum_{m=1}^{t} \sum_{k=1}^{m} \mathcal{A}_{t,m} \prod_{j=k}^{m-1} W^\top\ (1 - h_{j+1}^2) \Big((1 - h_k^2)\ U\ x_k^o \Big) \Big)
\end{aligned}$$

Parameter Updates

The parameters are updated by subtracting the gradient scaled by a learning rate η:

$$\begin{aligned}
V &= V - \eta \frac{\partial L}{\partial V} \\
c &= c - \eta \frac{\partial L}{\partial c} \\
W &= W - \eta \frac{\partial L}{\partial W} \\
U &= U - \eta \frac{\partial L}{\partial U} \\
b &= b - \eta \frac{\partial L}{\partial b} \\
E &= E - \eta \frac{\partial L}{\partial E}
\end{aligned}$$

4 Parameters Initialized

The network parameters are initialized as follows:

$$E = \begin{bmatrix} 0.8175 & 0.4613 \\ -0.3599 & -0.4824 \\ 0.1268 & 0.1540 \\ -0.0358 & -0.5677 \end{bmatrix}$$

$$U = \begin{bmatrix} 0.1442 & -0.2315 \end{bmatrix}$$

$$W = \begin{bmatrix} -0.5352 \end{bmatrix}$$

$$b = \begin{bmatrix} 0 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.6951 \\ -0.4402 \\ -0.2246 \\ -0.3053 \end{bmatrix}$$

$$c = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

5 Forward Pass

5.1 Hidden State at $t = 1$:

$$
\begin{aligned}
x_1 &= x_1^o \cdot E \\
&= [1\ 0\ 0\ 0] \cdot \begin{bmatrix} 0.8175 & 0.4613 \\ -0.3599 & -0.4824 \\ 0.1268 & 0.1540 \\ -0.0358 & -0.5677 \end{bmatrix} \\
x_1 &= \begin{bmatrix} 0.8175 & 0.4613 \end{bmatrix} \\
h_1 &= \tanh(U \cdot x_1 + W \cdot h_0 + b) \\
&= \tanh\left(\begin{bmatrix} 0.1442 & -0.2315 \end{bmatrix} \cdot \begin{bmatrix} 0.8175 \\ 0.4613 \end{bmatrix} + \begin{bmatrix} -0.5352 \end{bmatrix} \begin{bmatrix} 0 \end{bmatrix} + \begin{bmatrix} 0 \end{bmatrix} \right) \\
h_1 &= \begin{bmatrix} 0.0111 \end{bmatrix}
\end{aligned}
$$

5.2 Hidden State at $t = 2$:

$$
\begin{aligned}
x_2 &= x_2^o \cdot E \\
&= [0\ 1\ 0\ 0] \cdot \begin{bmatrix} 0.8175 & 0.4613 \\ -0.3599 & -0.4824 \\ 0.1268 & 0.1540 \\ -0.0358 & -0.5677 \end{bmatrix} \\
x_2 &= \begin{bmatrix} -0.3599 & -0.4824 \end{bmatrix} \\
h_2 &= \tanh(U \cdot x_2 + W \cdot h_1 + b) \\
&= \tanh\left(\begin{bmatrix} 0.1442 & -0.2315 \end{bmatrix} \begin{bmatrix} -0.3599 \\ -0.4824 \end{bmatrix} + \begin{bmatrix} -0.5352 \end{bmatrix} \begin{bmatrix} 0.0111 \end{bmatrix} + \begin{bmatrix} 0 \end{bmatrix} \right) \\
h_2 &= \begin{bmatrix} 0.0537 \end{bmatrix}
\end{aligned}
$$

5.3 Hidden State at $t = 3$:

$$
\begin{aligned}
x_3 &= x_3^o \cdot E \\
&= [0\ 0\ 1\ 0] \cdot \begin{bmatrix} 0.8175 & 0.4613 \\ -0.3599 & -0.4824 \\ 0.1268 & 0.1540 \\ -0.0358 & -0.5677 \end{bmatrix} \\
x_3 &= \begin{bmatrix} 0.1268 & 0.1540 \end{bmatrix} \\
h_3 &= \tanh(U \cdot x_3 + W \cdot h_2 + b) \\
&= \tanh\left(\begin{bmatrix} 0.1442 & -0.2315 \end{bmatrix} \begin{bmatrix} 0.1268 \\ 0.1540 \end{bmatrix} + \begin{bmatrix} -0.5352 \end{bmatrix} \begin{bmatrix} 0.0537 \end{bmatrix} + \begin{bmatrix} 0 \end{bmatrix} \right) \\
h_3 &= \begin{bmatrix} -0.0461 \end{bmatrix}
\end{aligned}
$$

5.4 Hidden State at $t = 4$:

$$
\begin{aligned}
x_4 &= x_4^o \cdot E \\
&= [0\ 0\ 1\ 0] \cdot \begin{bmatrix} 0.8175 & 0.4613 \\ -0.3599 & -0.4824 \\ 0.1268 & 0.1540 \\ -0.0358 & -0.5677 \end{bmatrix} \\
x_4 &= \begin{bmatrix} 0.1268 & 0.1540 \end{bmatrix} \\
h_4 &= \tanh(U \cdot x_4 + W \cdot h_3 + b) \\
&= \tanh\left(\begin{bmatrix} 0.1442 & -0.2315 \end{bmatrix} \begin{bmatrix} 0.1268 \\ 0.1540 \end{bmatrix} + \begin{bmatrix} -0.5352 \end{bmatrix} \begin{bmatrix} -0.0461 \end{bmatrix} + \begin{bmatrix} 0 \end{bmatrix} \right) \\
h_4 &= \begin{bmatrix} 0.0073 \end{bmatrix}
\end{aligned}
$$

Output Stage at $t = 1$

$$
\begin{aligned}
A_{1,1}, &= h_1 \cdot h_1 \\
A_{1,1} &= [0.0111] \cdot [0.0111] \\
\mathcal{Z}_1 &= \text{softmax}(\mathcal{A}_{1,1}) \cdot h_1 \\
\mathcal{Z}_1 &= [0.0111] \\
\Omega_1 &= V \cdot \mathcal{Z}_1 + c \\
&= \begin{bmatrix} 0.6951 \\ -0.4402 \\ -0.2246 \\ -0.3053 \end{bmatrix} \cdot [0.0111] + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \\
\Omega_1 &= \begin{bmatrix} 0.0077 \\ -0.0049 \\ -0.0025 \\ -0.0034 \end{bmatrix} \\
\hat{y}_1 &= \text{softmax}(\Omega_1) \\
\hat{y}_1 &= \begin{bmatrix} 0.2521 \\ 0.2490 \\ 0.2496 \\ 0.2493 \end{bmatrix}
\end{aligned}
$$

Based on the maximum value of the softmax function, our model predicts it as 'h', but we actually want it to predict 'e'.

$$
\begin{aligned}
\mathrm{L}_1 &= -y_1 \cdot \ln(\hat{y}_1) \\
&= -\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \cdot \ln \begin{bmatrix} 0.2521 \\ 0.2490 \\ 0.2496 \\ 0.2493 \end{bmatrix} \\
&= 1.3904
\end{aligned}
$$

Output Stage at $t = 2$

$$
\begin{aligned}
A_{2,1} &= h_2 \cdot h_1 \\
A_{2,1} &= [0.0537] \cdot [0.0111] \\
A_{2,2} &= h_2 \cdot h_2 \\
A_{2,2} &= [0.0537] \cdot [0.0537] \\
Z_2 &= \sum_{k=1}^{2} \mathrm{softmax}(A_{2,k}) \cdot h_k \\
Z_2 &= [0.0537] \cdot [0.0111] + [0.0537] \cdot [0.0537] \\
Z_2 &= [0.0325] \\
\Omega_2 &= V \cdot Z_2 + c \\
&= \begin{bmatrix} 0.6951 \\ -0.4402 \\ -0.2246 \\ -0.3053 \end{bmatrix} \cdot [0.0325] + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \\
\Omega_2 &= \begin{bmatrix} 0.0226 \\ -0.0143 \\ -0.0073 \\ -0.0099 \end{bmatrix} \\
\hat{y}_2 &= \mathrm{softmax}(\Omega_2) \\
\hat{y}_2 &= \begin{bmatrix} 0.2563 \\ 0.2470 \\ 0.2487 \\ 0.2481 \end{bmatrix}
\end{aligned}
$$

Based on the maximum value of the softmax function, our model predicts it as 'h', but we actually want it to predict 'l'.

$$
\begin{aligned}
\mathrm{L}_2 &= -y_2 \cdot \ln(\hat{y}_2) \\
&= -\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \cdot \ln \begin{bmatrix} 0.2563 \\ 0.2470 \\ 0.2487 \\ 0.2481 \end{bmatrix} \\
&= 1.3915
\end{aligned}
$$

Output Stage at $t = 3$

$$
\begin{aligned}
A_{3,1} &= h_3 \cdot h_1 \\
A_{3,1} &= [-0.0461] \cdot [0.0111] \\
A_{3,2} &= h_3 \cdot h_2 \\
A_{3,2} &= [-0.0461] \cdot [0.0537] \\
A_{3,3} &= h_3 \cdot h_3 \\
A_{3,3} &= [-0.0461] \cdot [-0.0461] \\
Z_3 &= \sum_{k=1}^{3} \mathrm{softmax}(A_{3,k}) \cdot h_k \\
Z_3 &= [-0.0461] \cdot [0.0111] + [-0.0461] \cdot [0.0537] + [-0.0461] \cdot [-0.0461] \\
Z_3 &= [0.0062] \\
\Omega_3 &= V \cdot Z_3 + c \\
&= \begin{bmatrix} 0.6951 \\ -0.4402 \\ -0.2246 \\ -0.3053 \end{bmatrix} \cdot [0.0062] + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \\
\Omega_3 &= \begin{bmatrix} 0.0043 \\ -0.0027 \\ -0.0014 \\ -0.0019 \end{bmatrix} \\
\hat{y}_3 &= \mathrm{softmax}(\Omega_3) \\
\hat{y}_3 &= \begin{bmatrix} 0.2512 \\ 0.2494 \\ 0.2498 \\ 0.2496 \end{bmatrix}
\end{aligned}
$$

Based on the maximum value of the softmax function, our model predicts it as 'h', but we actually want it to predict 'l'.

Output Stage at $t = 3$

$$
\begin{aligned}
A_{4,1} &= h_4 \cdot h_1 \\
A_{4,1} &= [0.0073] \cdot [0.0111] \\
A_{4,2} &= h_4 \cdot h_2 \\
A_{4,2} &= [0.0073] \cdot [0.0537] \\
A_{4,3} &= h_4 \cdot h_3 \\
A_{4,3} &= [0.0073] \cdot [-0.0461] \\
A_{4,3} &= h_4 \cdot h_4 \\
A_{4,3} &= [0.0073] \cdot [0.0073] \\
Z_4 &= \sum_{k=1}^{4} \text{softmax}(A_{4,k}) \cdot h_k \\
Z_4 &= [0.0073] \cdot [0.0111] + [0.0073] \cdot [0.0537] + [0.0073] \cdot [-0.0461] + [0.0073] \cdot [0.0073] \\
Z_4 &= [0.0066] \\
\Omega_4 &= V \cdot Z_4 + c \\
&= \begin{bmatrix} 0.6951 \\ -0.4402 \\ -0.2246 \\ -0.3053 \end{bmatrix} \cdot [-0.2694] + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \\
\Omega_4 &= \begin{bmatrix} 0.0045 \\ -0.0029 \\ -0.0015 \\ -0.0020 \end{bmatrix} \\
\hat{y}_4 &= \text{softmax}(\Omega_4) \\
\hat{y}_4 &= \begin{bmatrix} 0.2512 \\ 0.2494 \\ 0.2497 \\ 0.2496 \end{bmatrix}
\end{aligned}
$$

Based on the maximum value of the softmax function, our model predicts it as 'h', but we actually want it to predict 'o'.

$$
\begin{aligned}
\mathrm{L}_4 &= -y_4 \cdot \ln(\hat{y}_4) \\
&= -\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \cdot \ln \begin{bmatrix} 0.2512 \\ 0.2494 \\ 0.2497 \\ 0.2496 \end{bmatrix} \\
&= 1.3878
\end{aligned}
$$

$$
\sum_{t=1}^{k} L_t = 1.3904 + 1.3915 + 1.3873 + 1.3878 = 5.557
$$

6 Backpropagation Through Time

6.1 Gradient of L w.r.t. Output

$$\frac{\partial L_1}{\partial \hat{y}_1}\frac{\partial \hat{y}_1}{\partial \Omega_1} = (\hat{y}_1 - y_1)$$
$$= \begin{bmatrix}0.2521\\0.2490\\0.2496\\0.2493\end{bmatrix} - \begin{bmatrix}0\\1\\0\\0\end{bmatrix}$$
$$= \begin{bmatrix}0.2521\\-0.7510\\0.2496\\0.2493\end{bmatrix}$$
$$\frac{\partial L_2}{\partial \hat{y}_2}\frac{\partial \hat{y}_2}{\partial \Omega_2} = (\hat{y}_2 - y_2)$$
$$= \begin{bmatrix}0.2563\\0.2470\\0.2487\\0.2481\end{bmatrix} - \begin{bmatrix}0\\0\\1\\0\end{bmatrix}$$
$$= \begin{bmatrix}0.2563\\-0.7530\\0.2487\\0.2481\end{bmatrix}$$
$$\frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial \Omega_3} = (\hat{y}_3 - y_3)$$
$$= \begin{bmatrix}0.2512\\0.2494\\0.2498\\0.2496\end{bmatrix} - \begin{bmatrix}0\\0\\1\\0\end{bmatrix}$$
$$= \begin{bmatrix}0.2512\\0.2494\\-0.7502\\0.2496\end{bmatrix}$$
$$\frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial \Omega_4} = (\hat{y}_4 - y_4)$$
$$= \begin{bmatrix}0.2512\\0.2494\\0.2497\\0.2496\end{bmatrix} - \begin{bmatrix}0\\0\\0\\1\end{bmatrix}$$
$$= \begin{bmatrix}0.2512\\0.2494\\0.2497\\-0.7504\end{bmatrix}$$

6.2 Update V

$$
\begin{aligned}
\frac{\partial L}{\partial V} &= \sum_{t=1}^{S} \frac{\partial L_t}{\partial V} \\
&= \sum_{t=1}^{S} (\hat{y}_t - y_t) \cdot h_t^\top \\
&= \begin{bmatrix} -0.2677 \\ -0.4165 \\ 0.5373 \\ 0.1470 \end{bmatrix} \\
\eta &= 0.1 \\
V_{new} &= V - \eta \frac{\partial L}{\partial V} \\
&= \begin{bmatrix} 0.6951 \\ -0.4402 \\ -0.2246 \\ -0.3053 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} 0.0068 \\ -0.0048 \\ -0.0012 \\ -0.0009 \end{bmatrix} \\
V_{new} &= \begin{bmatrix} 0.6944 \\ -0.4397 \\ -0.2244 \\ -0.3052 \end{bmatrix}
\end{aligned}
$$

6.3 Update c:

$$\begin{aligned}
\frac{\partial L}{\partial c} &= \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial c} \\
&= \sum_{t=1}^{\mathcal{S}} (\hat{y}_t - y_t) \\
&= \begin{bmatrix} 1.011 \\ -0.005 \\ -1.002 \\ -0.003 \end{bmatrix} \\
\eta &= 0.1 \\
c_{new} &= c - \eta \frac{\partial L}{\partial c} \\
&= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - 0.1 \begin{bmatrix} 1.011 \\ -0.005 \\ -1.002 \\ -0.003 \end{bmatrix} \\
c_{new} &= \begin{bmatrix} -0.1011 \\ 0.0005 \\ 0.1002 \\ 0.0003 \end{bmatrix}
\end{aligned}$$

6.4 Update W:

$$\begin{aligned}
\frac{\partial L}{\partial W} &= \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial \mathcal{Z}_t} \Big(\sum_{m=1}^{t} \sum_{k=1}^{m} \frac{\partial \mathcal{Z}_t}{\partial h_m} \prod_{j=k}^{m-1} \frac{\partial h_{j+1}}{\partial h_j} \Big(\frac{\partial h_k}{\partial W} \Big) \Big) \\
&= \sum_{t=1}^{\mathcal{S}} (y_t - \hat{y}_t)\ V^\top \Big(\sum_{m=1}^{t} \sum_{k=1}^{m} \mathcal{A}_{t,m} \prod_{j=k}^{m-1} W^\top\ (1 - h_{j+1}^2) \Big((1 - h_k^2)\ h_{k-1} \Big) \Big) \\
&= \begin{bmatrix} 0.0032 \end{bmatrix} \\
\eta &= 0.1 \\
W_{new} &= W - \eta \frac{\partial L}{\partial W} \\
&= \begin{bmatrix} -0.5352 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} 0.0032 \end{bmatrix} \\
W_{new} &= \begin{bmatrix} -0.5355 \end{bmatrix}
\end{aligned}$$

6.5 Update U:

$$\begin{aligned}
\frac{\partial L}{\partial U} &= \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial \mathcal{Z}_t} \Bigg(\sum_{m=1}^{t} \sum_{k=1}^{m} \frac{\partial \mathcal{Z}_t}{\partial h_m} \prod_{j=k}^{m-1} \frac{\partial h_{j+1}}{\partial h_j} \left(\frac{\partial h_k}{\partial U} \right) \Bigg) \\
&= \sum_{t=1}^{\mathcal{S}} (y_t - \hat{y}_t)\ V^\top \Bigg(\sum_{m=1}^{t} \sum_{k=1}^{m} \mathcal{A}_{t,m} \prod_{j=k}^{m-1} W^\top\ (1 - h_{j+1}^2) \Big((1 - h_k^2)\ x_k \Big) \Bigg) \\
&= \begin{bmatrix} 0.3614 & 0.1736 \end{bmatrix} \\
\eta &= 0.1 \\
U_{new} &= U - \eta \frac{\partial L}{\partial U} \\
&= \begin{bmatrix} 0.1442 & -0.2315 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} 0.3614 & 0.1736 \end{bmatrix} \\
U_{new} &= \begin{bmatrix} 0.1081 - 0.2488 \end{bmatrix}
\end{aligned}$$

6.6 Update b:

$$\begin{aligned}
\frac{\partial L}{\partial b} &= \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial \mathcal{Z}_t} \Bigg(\sum_{m=1}^{t} \sum_{k=1}^{m} \frac{\partial \mathcal{Z}_t}{\partial h_m} \prod_{j=k}^{m-1} \frac{\partial h_{j+1}}{\partial h_j} \left(\frac{\partial h_k}{\partial b} \right) \Bigg) \\
&= \sum_{t=1}^{\mathcal{S}} (y_t - \hat{y}_t)\ V^\top \Bigg(\sum_{m=1}^{t} \sum_{k=1}^{m} \mathcal{A}_{t,m} \prod_{j=k}^{m-1} W^\top\ (1 - h_{j+1}^2) \Big((1 - h_k^2) \Big) \Bigg) \\
&= \begin{bmatrix} 0.7756 \end{bmatrix} \\
\eta &= 0.1 \\
b_{new} &= b - \eta \frac{\partial L}{\partial b} \\
&= \begin{bmatrix} 0 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} 0.7756 \end{bmatrix} \\
b_{new} &= \begin{bmatrix} -0.0775 \end{bmatrix}
\end{aligned}$$

6.7 Update E:

$$\frac{\partial L}{\partial E} = \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial E} = \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial \mathcal{Z}_t} \left(\sum_{m=1}^{t} \sum_{k=1}^{m} \frac{\partial \mathcal{Z}_t}{\partial h_m} \prod_{j=k}^{m-1} \frac{\partial h_{j+1}}{\partial h_j} \left(\frac{\partial h_k}{\partial x_k} \frac{\partial x_k}{\partial E} \right) \right)$$

$$= \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial E} = \sum_{t=1}^{\mathcal{S}} (y_t - \hat{y}_t)\ V^\top \left(\sum_{m=1}^{t} \sum_{k=1}^{m} \mathcal{A}_{t,m} \prod_{j=k}^{m-1} W^\top\ (1 - h_{j+1}^2) \left((1 - h_k^2)\ U\ x_k^o \right) \right)$$

$$= \begin{bmatrix} 0.0539 & -0.0865 \\ 0.0109 & -0.0174 \\ 0.0375 & -0.0602 \\ 0.0000 & 0.0000 \end{bmatrix}$$

$$\eta = 0.1$$

$$b_{new} = b - \eta \frac{\partial L}{\partial b}$$

$$= \begin{bmatrix} 0.8175 & 0.4613 \\ -0.3599 & -0.4824 \\ 0.1268 & 0.1540 \\ -0.0358 & -0.5677 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} 0.0539 & -0.0865 \\ 0.0109 & -0.0174 \\ 0.0375 & -0.0602 \\ 0.0000 & 0.0000 \end{bmatrix}$$

$$b_{new} = \begin{bmatrix} 0.8121 & 0.4699 \\ -0.3610 & -0.4807 \\ 0.1230 & 0.1600 \\ -0.0358 & -0.5677 \end{bmatrix}$$

6.8 Following Epochs Loss Values

$$L_1 = -y_1 \cdot \ln(\hat{y}_1)$$

$$= -\begin{bmatrix} 0. \\ 1. \\ 0. \\ 0. \end{bmatrix} \cdot \ln \left(\begin{bmatrix} 0.2071 \\ 0.2586 \\ 0.2793 \\ 0.2549 \end{bmatrix} \right)$$

$$L_1 = 1.3523$$

$$L_2 = -y_2 \cdot \ln(\hat{y}_2)$$

$$= -\begin{bmatrix} 0. \\ 0. \\ 1. \\ 0. \end{bmatrix} \cdot \ln \left(\begin{bmatrix} 0.2214 \\ 0.2515 \\ 0.2765 \\ 0.2507 \end{bmatrix} \right)$$

$$L_2 = 1.2856$$

$$
\begin{aligned}
L_3 &= -y_3 \cdot \ln(\hat{y}_3) \\
&= -\begin{bmatrix} 0. \\ 0. \\ 1. \\ 0. \end{bmatrix} \cdot \ln\left(\begin{bmatrix} 0.2147 \\ 0.2548 \\ 0.2778 \\ 0.2527 \end{bmatrix}\right) \\
L_3 &= 1.2807
\end{aligned}
$$

$$
\begin{aligned}
L_4 &= -y_4 \cdot \ln(\hat{y}_4) \\
&= -\begin{bmatrix} 0. \\ 0. \\ 0. \\ 1. \end{bmatrix} \cdot \ln\left(\begin{bmatrix} 0.2161 \\ 0.2541 \\ 0.2776 \\ 0.2523 \end{bmatrix}\right) \\
L_4 &= 1.3773
\end{aligned}
$$

$$
\sum_{t=1}^{k} L_t = 1.3523 + 1.2856 + 1.2807 + 1.3773 = 5.2959
$$

Up to Epoch 50 the minimum Total Loss will be 2.8254
Up to Epoch 100 the minimum Total Loss will be 2.0888
Up to Epoch 500 the minimum Total Loss will be 1.9524
Up to Epoch 1000 the minimum Total Loss will be 1.9261.