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:

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

2 Forward Pass Formulas

For a sequence of characters $x_1, x_2, \ldots, x_T$, the network computes:
1. Hidden State at time t, h_t:

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

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

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

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

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

4. Context Vector, Sum of Weighted Attention Score

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

5. Output before softmax, Ω_t:

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

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

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

5. 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 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 and b:

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

Parameter Updates

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

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

4 Parameters Initialized

The network parameters are initialized as follows:

$$
\begin{aligned}
U &= \begin{bmatrix} 0.1442 & -0.2315 & -0.6690 & 1.1585 \end{bmatrix} \\
W &= \begin{bmatrix} -0.5870 \end{bmatrix} \\
b &= \begin{bmatrix} 0 \end{bmatrix} \\
V &= \begin{bmatrix} -0.2246 \\ -0.3053 \\ 0.4905 \\ 0.2768 \end{bmatrix} \\
c &= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
\end{aligned}
$$

5 Forward Pass

5.1 Hidden State at $t = 1$:

$$
\begin{aligned}
h_1 &= \tanh(U \cdot x_1 + W \cdot h_0 + b) \\
&= \tanh\left(\begin{bmatrix} 0.1442 & -0.2315 & -0.6690 & 1.1585 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -0.5869 \end{bmatrix} \begin{bmatrix} 0 \end{bmatrix} + \begin{bmatrix} 0 \end{bmatrix}\right) \\
h_1 &= \begin{bmatrix} 0.1432 \end{bmatrix}
\end{aligned}
$$

5.2 Hidden State at $t = 2$:

$$
\begin{aligned}
h_2 &= \tanh(U \cdot x_2 + W \cdot h_1 + b) \\
&= \tanh\left(\begin{bmatrix} 0.1442 & -0.2315 & -0.6690 & 1.1585 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -0.5869 \end{bmatrix} \begin{bmatrix} 0.1432 \end{bmatrix} + \begin{bmatrix} 0 \end{bmatrix}\right) \\
h_2 &= \begin{bmatrix} -0.3055 \end{bmatrix}
\end{aligned}
$$

5.3 Hidden State at $t = 3$:

$$
\begin{aligned}
h_3 &= \tanh(U \cdot x_3 + W \cdot h_2 + b) \\
&= \tanh\left(\begin{bmatrix} 0.1442 & -0.2315 & -0.6690 & 1.1585 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} -0.5869 \end{bmatrix} \begin{bmatrix} -0.3055 \end{bmatrix} + \begin{bmatrix} 0 \end{bmatrix}\right) \\
h_3 &= \begin{bmatrix} -0.4540 \end{bmatrix}
\end{aligned}
$$

5.4 Hidden State at $t = 4$:

$$
\begin{aligned}
h_4 &= \tanh(U \cdot x_4 + W \cdot h_3 + b) \\
&= \tanh\left(\begin{bmatrix} 0.1442 & -0.2315 & -0.6690 & 1.1585 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} -0.5869 \end{bmatrix} \begin{bmatrix} -0.4540 \end{bmatrix} + \begin{bmatrix} 0 \end{bmatrix}\right) \\
h_4 &= \begin{bmatrix} -0.3821 \end{bmatrix}
\end{aligned}
$$

Output Stage at $t = 1$

$$
\begin{aligned}
\mathcal{A}_{1,1}, &= h_1 \cdot h_1 \\
\mathcal{A}_{1,1} &= [0.1432] \cdot [0.1432] \\
\mathcal{Z}_1 &= \mathrm{softmax}(\mathcal{A}_{1,1}) \cdot h_1 \\
\mathcal{Z}_1 &= [0.1432] \\
\Omega_1 &= V \cdot \mathcal{Z}_1 + c \\
&= \begin{bmatrix} -0.2246 \\ -0.3053 \\ 0.4905 \\ 0.2768 \end{bmatrix} \cdot [0.1432] + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \\
\Omega_1 &= \begin{bmatrix} -0.0322 \\ -0.0437 \\ 0.0703 \\ 0.0396 \end{bmatrix} \\
\hat{y}_1 &= \mathrm{softmax}(\Omega_1) \\
\hat{y}_1 &= \begin{bmatrix} 0.2398 \\ 0.2370 \\ \color{red}{0.2656} \\ 0.2576 \end{bmatrix}
\end{aligned}
$$

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

$$
\begin{aligned}
\mathrm{L}_1 &= -y_1 \cdot \ln(\hat{y}_1) \\
&= -\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \cdot \ln \begin{bmatrix} 0.2398 \\ 0.2370 \\ \color{red}{0.2656} \\ 0.2576 \end{bmatrix} \\
&= 1.4397
\end{aligned}
$$

Output Stage at $t = 2$

$$
\begin{aligned}
\mathcal{A}_{2,1} &= h_2 \cdot h_1 \\
\mathcal{A}_{2,1} &= [-0.3055] \cdot [0.1432] \\
\mathcal{A}_{2,2} &= h_2 \cdot h_2 \\
\mathcal{A}_{2,2} &= [-0.3055] \cdot [-0.3055] \\
\mathcal{Z}_2 &= \sum_{k=1}^{2} \text{softmax}(\mathcal{A}_{2,k}) \cdot h_k \\
\mathcal{Z}_2 &= [0.4658] \cdot [0.1432] + [0.5342] \cdot [-0.3055] \\
\mathcal{Z}_2 &= [-0.0965] \\
\Omega_2 &= V \cdot \mathcal{Z}_2 + c \\
&= \begin{bmatrix} -0.2246 \\ -0.3053 \\ 0.4905 \\ 0.2768 \end{bmatrix} \cdot [-0.0965] + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \\
\Omega_2 &= \begin{bmatrix} 0.0217 \\ 0.0294 \\ -0.0473 \\ -0.0267 \end{bmatrix} \\
\hat{y}_2 &= \text{softmax}(\Omega_2) \\
\hat{y}_2 &= \begin{bmatrix} 0.2568 \\ 0.2588 \\ 0.2397 \\ 0.2447 \end{bmatrix}
\end{aligned}
$$

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

$$
\begin{aligned}
\mathrm{L}_2 &= -y_2 \cdot \ln(\hat{y}_2) \\
&= -\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \cdot \ln \begin{bmatrix} 0.2568 \\ 0.2588 \\ 0.2397 \\ 0.2447 \end{bmatrix} \\
&= 1.4284
\end{aligned}
$$

Output Stage at $t = 3$

$$
\begin{aligned}
\mathcal{A}_{3,1} &= h_3 \cdot h_1 \\
\mathcal{A}_{3,1} &= [-0.4540] \cdot [0.1432] \\
\mathcal{A}_{3,2} &= h_3 \cdot h_2 \\
\mathcal{A}_{3,2} &= [-0.4540] \cdot [-0.3055] \\
\mathcal{A}_{3,3} &= h_3 \cdot h_3 \\
\mathcal{A}_{3,3} &= [-0.4540] \cdot [-0.4540] \\
\mathcal{Z}_3 &= \sum_{k=1}^{3} \text{softmax}(\mathcal{A}_{3,k}) \cdot h_k \\
\mathcal{Z}_3 &= [0.2827] \cdot [0.1432] + [0.3466] \cdot [-0.3055] + [0.3707] \cdot [-0.4540] \\
\mathcal{Z}_3 &= [-0.2337] \\
\Omega_3 &= V \cdot \mathcal{Z}_3 + c \\
&= \begin{bmatrix} -0.2246 \\ -0.3053 \\ 0.4905 \\ 0.2768 \end{bmatrix} \cdot [-0.2337] + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \\
\Omega_3 &= \begin{bmatrix} 0.0525 \\ 0.0713 \\ -0.1146 \\ -0.0647 \end{bmatrix} \\
\hat{y}_3 &= \text{softmax}(\Omega_3) \\
\hat{y}_3 &= \begin{bmatrix} 0.2663 \\ \color{red}{0.2714} \\ 0.2254 \\ 0.2369 \end{bmatrix}
\end{aligned}
$$

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

$$
\begin{aligned}
\text{L}_3 &= -y_3 \cdot \ln(\hat{y}_3) \\
&= -\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \cdot \ln \begin{bmatrix} 0.2663 \\ \color{red}{0.2714} \\ 0.2254 \\ 0.2369 \end{bmatrix} \\
&= 1.4901
\end{aligned}
$$

Output Stage at $t = 3$

$$\mathcal{A}_{4,1} = h_4 \cdot h_1$$
$$\mathcal{A}_{4,1} = [-0.3821] \cdot [0.1432]$$
$$\mathcal{A}_{4,2} = h_4 \cdot h_2$$
$$\mathcal{A}_{4,2} = [-0.3821] \cdot [-0.3055]$$
$$\mathcal{A}_{4,3} = h_4 \cdot h_3$$
$$\mathcal{A}_{4,3} = [-0.3821] \cdot [-0.4540]$$
$$\mathcal{A}_{4,3} = h_4 \cdot h_4$$
$$\mathcal{A}_{4,3} = [-0.3821] \cdot [-0.3821]$$

$$\mathcal{Z}_4 = \sum_{k=1}^{4} \text{softmax}(\mathcal{A}_{4,k}) \cdot h_k$$
$$\mathcal{Z}_4 = [0.2143] \cdot [0.1432] + [0.2544] \cdot [-0.3055] + [0.2693] \cdot [-0.4540] + [0.2620] \cdot [-0.3821]$$
$$\mathcal{Z}_4 = [-0.2694]$$
$$\Omega_4 = V \cdot \mathcal{Z}_4 + c$$
$$= \begin{bmatrix} -0.2246 \\ -0.3053 \\ 0.4905 \\ 0.2768 \end{bmatrix} \cdot [-0.2694] + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$
$$\Omega_4 = \begin{bmatrix} 0.0605 \\ 0.0822 \\ -0.1321 \\ -0.0746 \end{bmatrix}$$
$$\hat{y}_4 = \text{softmax}(\Omega_4)$$
$$\hat{y}_4 = \begin{bmatrix} 0.2688 \\ 0.2747 \\ 0.2217 \\ 0.2348 \end{bmatrix}$$

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

$$\begin{aligned}
\mathrm{L}_4 &= -y_4 \cdot \ln(\hat{y}_4) \\
&= -\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \cdot \ln \begin{bmatrix} 0.2688 \\ 0.2747 \\ 0.2217 \\ 0.2348 \end{bmatrix} \\
&= 1.4489 \\
\sum_{t=1}^{k} L_t &= 1.4397 + 1.4284 + 1.4901 + 1.4489 = 5.8071
\end{aligned}$$

6 Backpropagation Through Time

6.1 Gradient of L w.r.t. Output

$$
\begin{aligned}
\frac{\partial L_1}{\partial \hat{y}_1}\frac{\partial \hat{y}_1}{\partial \Omega_1} &= (\hat{y}_1 - y_1) \\
&= \begin{bmatrix} 0.2398 \\ 0.2370 \\ 0.2656 \\ 0.2576 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 0.2398 \\ -0.7630 \\ 0.2656 \\ 0.2576 \end{bmatrix} \\
\frac{\partial L_2}{\partial \hat{y}_2}\frac{\partial \hat{y}_2}{\partial \Omega_2} &= (\hat{y}_2 - y_2) \\
&= \begin{bmatrix} 0.2568 \\ 0.2588 \\ 0.2397 \\ 0.2447 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 0.2568 \\ 0.2588 \\ -0.7603 \\ 0.2447 \end{bmatrix} \\
\frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial \Omega_3} &= (\hat{y}_3 - y_3) \\
&= \begin{bmatrix} 0.2663 \\ 0.2714 \\ 0.2254 \\ 0.2369 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 0.2663 \\ 0.2714 \\ -0.7746 \\ 0.2369 \end{bmatrix} \\
\frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial \Omega_4} &= (\hat{y}_4 - y_4) \\
&= \begin{bmatrix} 0.2688 \\ 0.2747 \\ 0.2217 \\ 0.2348 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} 0.2688 \\ 0.2747 \\ 0.2217 \\ -0.7652 \end{bmatrix}
\end{aligned}
$$

6.2 Update V

$$
\begin{aligned}
\frac{\partial L}{\partial V} &= \sum_{t=1}^{S} \frac{\partial L_t}{\partial V} \\
&= \sum_{t=1}^{S} (\hat{y}_t - y_t) \cdot h_t^\top \\
&= \begin{bmatrix} -0.2677 \\ -0.4165 \\ 0.5373 \\ 0.1470 \end{bmatrix} \\
\eta &= 0.1 \\
V_{new} &= V - \eta \frac{\partial L}{\partial V} \\
&= \begin{bmatrix} -0.2246 \\ -0.3053 \\ 0.4905 \\ 0.2768 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} -0.2677 \\ -0.4165 \\ 0.5373 \\ 0.1470 \end{bmatrix} \\
V_{new} &= \begin{bmatrix} -0.2219 \\ -0.3011 \\ 0.4851 \\ 0.2753 \end{bmatrix}
\end{aligned}
$$

6.3 Update c:

$$\begin{aligned}
\frac{\partial L}{\partial c} &= \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial c} \\
&= \sum_{t=1}^{\mathcal{S}} (\hat{y}_t - y_t) \\
&= \begin{bmatrix} 1.0317 \\ 0.0419 \\ -1.0476 \\ -0.0260 \end{bmatrix} \\
\eta &= 0.1 \\
c_{new} &= c - \eta \frac{\partial L}{\partial c} \\
&= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - 0.1 \begin{bmatrix} 1.0317 \\ 0.0419 \\ -1.0476 \\ -0.0260 \end{bmatrix} \\
c_{new} &= \begin{bmatrix} -0.0103 \\ -0.0004 \\ 0.0105 \\ 0.0003 \end{bmatrix}
\end{aligned}$$

6.4 Update W:

$$\begin{aligned}
\frac{\partial L}{\partial W} &= \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial \hat{y}_t} \; \frac{\partial \hat{y}_t}{\partial \Omega_t} \; \frac{\partial \Omega_t}{\partial \mathcal{Z}_t} \left(\sum_{m=1}^{t} \sum_{k=1}^{m} \frac{\partial \mathcal{Z}_t}{\partial h_m} \prod_{j=k}^{m-1} \frac{\partial h_{j+1}}{\partial h_j} \left(\frac{\partial h_k}{\partial W} \right) \right) \\
&= \sum_{t=1}^{\mathcal{S}} (y_t - \hat{y}_t) \; V^\top \left(\sum_{m=1}^{t} \sum_{k=1}^{m} \mathcal{A}_{t,m} \prod_{j=k}^{m-1} W^\top \; (1 - h_{j+1}^2) \Big((1 - h_k^2) \; h_{k-1} \Big) \right) \\
&= \begin{bmatrix} 0.0274 \end{bmatrix} \\
\eta &= 0.1 \\
W_{new} &= W - \eta \frac{\partial L}{\partial W} \\
&= \begin{bmatrix} -0.5870 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} 0.0274 \end{bmatrix} \\
W_{new} &= \begin{bmatrix} -0.5872 \end{bmatrix}
\end{aligned}$$

6.5 Update U:

$$\begin{aligned}
\frac{\partial L}{\partial U} &= \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial \mathcal{Z}_t} \left(\sum_{m=1}^{t} \sum_{k=1}^{m} \frac{\partial \mathcal{Z}_t}{\partial h_m} \prod_{j=k}^{m-1} \frac{\partial h_{j+1}}{\partial h_j} \left(\frac{\partial h_k}{\partial U} \right) \right) \\
&= \sum_{t=1}^{\mathcal{S}} (y_t - \hat{y}_t)\ V^\top \left(\sum_{m=1}^{t} \sum_{k=1}^{m} \mathcal{A}_{t,m} \prod_{j=k}^{m-1} W^\top\ (1 - h_{j+1}^2) \Big((1 - h_k^2)\ x_k \Big) \right) \\
&= \begin{bmatrix} 0.1817 & -0.3287 & -0.2169 & 0 \end{bmatrix} \\
\eta &= 0.1 \\
U_{new} &= U - \eta \frac{\partial L}{\partial U} \\
&= \begin{bmatrix} 0.1442 & -0.2315 & -0.6690 & 1.1585 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} 0.1817 & -0.3287 & -0.2169 & 0 \end{bmatrix} \\
U_{new} &= \begin{bmatrix} 0.1424 & -0.2282 & -0.6668 & 1.1585 \end{bmatrix}
\end{aligned}$$

6.6 Update b:

$$\begin{aligned}
\frac{\partial L}{\partial b} &= \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial \mathcal{Z}_t} \left(\sum_{m=1}^{t} \sum_{k=1}^{m} \frac{\partial \mathcal{Z}_t}{\partial h_m} \prod_{j=k}^{m-1} \frac{\partial h_{j+1}}{\partial h_j} \left(\frac{\partial h_k}{\partial b} \right) \right) \\
&= \sum_{t=1}^{\mathcal{S}} (y_t - \hat{y}_t)\ V^\top \left(\sum_{m=1}^{t} \sum_{k=1}^{m} \mathcal{A}_{t,m} \prod_{j=k}^{m-1} W^\top\ (1 - h_{j+1}^2) \Big((1 - h_k^2) \Big) \right) \\
&= \begin{bmatrix} -0.3639 \end{bmatrix} \\
\eta &= 0.1 \\
b_{new} &= b - \eta \frac{\partial L}{\partial b} \\
&= \begin{bmatrix} 0 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} -0.3639 \end{bmatrix} \\
b_{new} &= \begin{bmatrix} 0.0036 \end{bmatrix}
\end{aligned}$$

6.7 Following Epochs Loss Values

$$\begin{aligned}
L_1 &= -y_1 \cdot \ln(\hat{y}_1) \\
&= - \begin{bmatrix} 0. \\ 1. \\ 0. \\ 0. \end{bmatrix} \cdot \ln \left(\begin{bmatrix} 0.2372 \\ 0.2368 \\ \color{red}{0.2683} \\ 0.2576 \end{bmatrix} \right) \\
L_1 &= 1.4404
\end{aligned}$$

$$
\begin{aligned}
L_2 &= -y_2 \cdot \ln(\hat{y}_2) \\
&= -\begin{bmatrix} 0. \\ 0. \\ 1. \\ 0. \end{bmatrix} \cdot \ln\left(\begin{bmatrix} 0.2539 \\ 0.2583 \\ 0.2428 \\ 0.2450 \end{bmatrix}\right) \\
L_2 &= 1.4155
\end{aligned}
$$

$$
\begin{aligned}
L_3 &= -y_3 \cdot \ln(\hat{y}_3) \\
&= -\begin{bmatrix} 0. \\ 0. \\ 1. \\ 0. \end{bmatrix} \cdot \ln\left(\begin{bmatrix} 0.2633 \\ 0.2709 \\ 0.2285 \\ 0.2373 \end{bmatrix}\right) \\
L_3 &= 1.4764
\end{aligned}
$$

$$
\begin{aligned}
L_4 &= -y_4 \cdot \ln(\hat{y}_4) \\
&= -\begin{bmatrix} 0. \\ 0. \\ 0. \\ 1. \end{bmatrix} \cdot \ln\left(\begin{bmatrix} 0.2657 \\ 0.2741 \\ 0.2248 \\ 0.2353 \end{bmatrix}\right) \\
L_4 &= 1.4468
\end{aligned}
$$

$$
\sum_{t=1}^{k} L_t = 1.4404 + 1.4155 + 1.4764 + 1.4468 = 5.7791
$$

in Epoch #100 Total Loss Will be 4.6789
in Epoch #1000 Total Loss Will be 2.1040
in Epoch #10000 Total Loss Will be 0.3418