

RNN

HSK

April 2024

1 Introduce

$$h_t = \Theta_h(W h_{t-1} + U x_t + b) \quad (1)$$

$$\mathcal{A}_{t,i} = h_i \cdot h_t \text{ for } i = 1, \dots, t \quad (2)$$

$$\text{softmax}(\mathcal{A}_{t,i}) = \frac{e^{\mathcal{A}_{t,i}}}{\sum_{k=0}^t e^{\mathcal{A}_{t,k}}} \text{ for } i = 1, \dots, t \quad (3)$$

$$\mathcal{Z}_t = \sum_{k=0}^t \text{softmax}(\mathcal{A}_{t,k}) \cdot h_k \quad (4)$$

$$\Omega_t = V \cdot \mathcal{Z}_t + c \quad (5)$$

$$\hat{y}_t = \Theta_y(\Omega_t) \quad (6)$$

$$L_t = - y_t \ln(\hat{y}_t) \quad (7)$$

Activation functions are Θ_h which represents *tanh* in (1), and Θ_y *softmax* in (6).

2 Through Time for Recurrent Neural Network

2.1 Softmax

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

Let us compute $\frac{\partial}{\partial \Omega_{t,j}}(\hat{y}_i)$ for some arbitrary i and j :

$$\frac{\partial \hat{y}_i}{\partial \Omega_{t,j}} = \frac{\partial}{\partial \Omega_{t,j}} \left(\frac{e^{\Omega_{t,i}}}{\sum_k e^{\Omega_{t,k}}} \right)$$

Since $\frac{\partial}{\partial \Omega_{t,j}} e^{\Omega_{t,k}} = 0$ for $k \neq j$, we have:

$$\frac{\partial}{\partial \Omega_{t,j}} \left(\sum e^{\Omega_{t,k}} \right) = \sum \left(\frac{\partial e^{\Omega_{t,k}}}{\partial \Omega_{t,j}} \right) = e^{\Omega_{t,j}}$$

Only meaningful derivatives is obtained for $i = j$ case in the above equation for our example presented in this chapter. Recall that in our example only one of values is a one

$$\begin{aligned}\frac{\partial}{\partial \Omega_{t,j}} \left(\frac{e^{\Omega_{t,i}}}{\sum e^{\Omega_{t,k}}} \right) &= \frac{e^{\Omega_{t,i}} \sum e^{\Omega_{t,k}} - e^{\Omega_{t,j}} e^{\Omega_{t,i}}}{(\sum e^{\Omega_{t,k}})^2} \\ &= \frac{e^{\Omega_{t,i}} (\sum e^{\Omega_{t,k}} - e^{\Omega_{t,j}})}{(\sum e^{\Omega_{t,k}})^2} \\ &= \frac{e^{\Omega_{t,i}}}{\sum e^{\Omega_{t,k}}} \cdot \left(\frac{\sum e^{\Omega_{t,k}}}{\sum e^{\Omega_{t,k}}} - \frac{e^{\Omega_{t,j}}}{\sum e^{\Omega_{t,k}}} \right)\end{aligned}$$

$$\hat{y}_{t,i} (1 - \hat{y}_{t,j}) \tag{8}$$

2.2 Derivative of Loss Function w.r.t. Ω_t

Recall that cross-entropy loss is defined as:

$$L = - \sum_{t=1}^S y_t \ln(\hat{y}_t)$$

Let us compute the partial derivative of L_t with respect to Ω_t at step t :

$$\begin{aligned}\frac{\partial L_t}{\partial \Omega} &= - \frac{\partial}{\partial \Omega} (y_t \ln \hat{y}_t) \\ &= -y_t \cdot \frac{\hat{y}_{t,i} (1 - \hat{y}_{t,j})}{\hat{y}_t} \\ &= -\hat{y}_{t,i} (1 - \hat{y}_{t,j}) \cdot \frac{y_t}{\hat{y}_t} \\ &= -(1 - \hat{y}_{t,j}) y_t \\ &= -(y_t - \hat{y}_{t,j} \hat{y}_t) \\ &= \hat{y}_{t,j} \hat{y}_t - y_t \\ &= \hat{y}_{t,j} \cdot \hat{y}_t (?) - y_t \\ &= (\hat{y}_t - y_t)\end{aligned}$$

$$\frac{\partial L_t}{\partial \hat{\Omega}_t} = (\hat{y}_t - y_t) \tag{9}$$

2.3 Derivative of V

The weight V is consistent across the entire time sequence, allowing us to perform differentiation at each time step and then aggregate the results.

$$\begin{aligned}
\frac{\partial L}{\partial V} &= \sum_{t=1}^S \frac{\partial L_t}{\partial V} \\
&= \sum_{t=1}^S \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial V}
\end{aligned}$$

We know that this formula $\frac{\partial \hat{y}_t}{\partial \Omega_t}$ from (9) and no other function exists between Omega and V, so simply taking the derivative coefficient of V yields h, thus the answer is h.

$$= \sum_{t=1}^S (\hat{y}_t - y_t) \cdot h_t^\top \quad (10)$$

2.4 Derivative of c

Similar to V, but its derivative is easier to calculate since it stands alone in the function.

$$\begin{aligned}
\frac{\partial L}{\partial c} &= \sum_{t=1}^T \frac{\partial L_t}{\partial c} \\
&= \sum_{t=1}^T \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial c}
\end{aligned}$$

In this case, The Analytical Derivatives of c becomes:

$$= \sum_{t=1}^T (\hat{y}_t - y_t) \quad (11)$$

2.5 Derivative of W

This function employs recursion, therefore, computing its derivative may take some time. Recall our forward pass formulas

$$\begin{aligned}
 h_t &= \tanh(W \cdot h_{t-1} + U \cdot x_t + b) \\
 \mathcal{A}_{t,i} &= h_i \cdot h_t \text{ for } i = 1, \dots, t \\
 \text{softmax}(\mathcal{A}_{t,i}) &= \frac{e^{\mathcal{A}_{t,i}}}{\sum_{k=0}^t e^{\mathcal{A}_{t,k}}} \text{ for } i = 1, \dots, t \text{ (for the } i^{th} \text{ element of softmax)} \\
 Z_t &= \sum_{k=0}^t \text{softmax}(\mathcal{A}_{t,k}) \cdot h_k \\
 \Omega_t &= V \cdot Z_t + c \\
 \hat{y}_t &= \text{softmax}(\Omega_t) \\
 L_t &= -y_t \ln(\hat{y}_t)
 \end{aligned}$$

For $t = 1$, we get:

$$\begin{aligned}
h_1 &= \tanh(Wh_0 + Ux_1 + b) \\
\mathcal{A}_{1,1} &= h_1 \cdot h_1 \quad (\text{since there is only one element in } h_1) \\
\text{softmax}(\mathcal{A}_{1,1}) &= \frac{e^{h_1 \cdot h_1}}{e^{h_1 \cdot h_1}} = 1 \\
\mathcal{Z}_1 &= 1 \cdot h_1 \\
\Omega_1 &= V \cdot \mathcal{Z}_1 + c \\
\hat{y}_1 &= \text{softmax}(\Omega_1) \\
L_1 &= -y_1 \ln(\hat{y}_1)
\end{aligned}$$

We start from derivation of L_1 with respect to W at $t = 1$:

$$\frac{\partial L_1}{\partial W} = \frac{\partial L_1}{\partial \hat{y}_1} \frac{\partial \hat{y}_1}{\partial \Omega_1} \frac{\partial \Omega_1}{\partial \mathcal{Z}_1} \frac{\partial \mathcal{Z}_1}{\partial h_1} \frac{\partial h_1}{\partial W}$$

For $t = 2$, we get:

$$\begin{aligned}
h_1 &= \tanh(Wh_0 + Ux_1 + b) \\
h_2 &= \tanh(W \tanh(Wh_0 + Ux_1 + b) + Ux_2 + b) \\
\mathcal{A}_{2,1} &= h_1 \cdot h_2 \\
\mathcal{A}_{2,2} &= h_2 \cdot h_2 \\
\text{softmax}(\mathcal{A}_{2,1}) &= \frac{e^{h_1 \cdot h_2}}{e^{h_1 \cdot h_2} + e^{h_2 \cdot h_2}} \\
\text{softmax}(\mathcal{A}_{2,2}) &= \frac{e^{h_2 \cdot h_2}}{e^{h_1 \cdot h_2} + e^{h_2 \cdot h_2}} \\
\mathcal{Z}_2 &= \frac{e^{h_1 \cdot h_2}}{e^{h_1 \cdot h_2} + e^{h_2 \cdot h_2}} \cdot h_1 + \frac{e^{h_2 \cdot h_2}}{e^{h_1 \cdot h_2} + e^{h_2 \cdot h_2}} \cdot h_2 \\
\Omega_2 &= V \cdot \left(\frac{e^{h_1 \cdot h_2}}{e^{h_1 \cdot h_2} + e^{h_2 \cdot h_2}} \cdot h_1 + \frac{e^{h_2 \cdot h_2}}{e^{h_1 \cdot h_2} + e^{h_2 \cdot h_2}} \cdot h_2 \right) + c \\
\hat{y}_2 &= \text{softmax}(\Omega_2) \\
L_2 &= -y_2 \ln(\hat{y}_2)
\end{aligned}$$

We now get:

$$\begin{aligned}
\frac{\partial L_2}{\partial W} &= \left(\frac{\partial L_2}{\partial \hat{y}_2} \frac{\partial \hat{y}_2}{\partial \Omega_2} \frac{\partial \Omega_2}{\partial \mathcal{Z}_2} \frac{\partial \mathcal{Z}_2}{\partial h_1} \frac{\partial h_1}{\partial W} \right) \\
&\quad + \left(\frac{\partial L_2}{\partial \hat{y}_2} \frac{\partial \hat{y}_2}{\partial \Omega_2} \frac{\partial \Omega_2}{\partial \mathcal{Z}_2} \frac{\partial \mathcal{Z}_2}{\partial h_2} \frac{\partial h_2}{\partial W} + \frac{\partial L_2}{\partial \hat{y}_2} \frac{\partial \hat{y}_2}{\partial \Omega_2} \frac{\partial \Omega_2}{\partial \mathcal{Z}_2} \frac{\partial \mathcal{Z}_2}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W} \right)
\end{aligned}$$

Let us now include $t = 3$ terms as follows: For $t = 2$, we get:

$$\begin{aligned}
h_1 &= \tanh(Wh_0 + Ux_1 + b) \\
h_2 &= \tanh(W \tanh(Wh_0 + Ux_1 + b) + Ux_2 + b) \\
h_3 &= \tanh(W \tanh(W \tanh(Wh_0 + Ux_1 + b) + Ux_2 + b) + Ux_3 + b) \\
\mathcal{A}_{3,1} &= h_1 \cdot h_3 \\
\mathcal{A}_{3,2} &= h_2 \cdot h_3 \\
\mathcal{A}_{3,3} &= h_3 \cdot h_3 \\
\text{softmax}(\mathcal{A}_{3,1}) &= \frac{e^{h_1 \cdot h_3}}{e^{h_1 \cdot h_3} + e^{h_2 \cdot h_3} + e^{h_3 \cdot h_3}} = \frac{e^{\mathcal{A}_{3,1}}}{e^{\mathcal{A}_{3,1}} + e^{\mathcal{A}_{3,2}} + e^{\mathcal{A}_{3,3}}} \\
\text{softmax}(\mathcal{A}_{3,2}) &= \frac{e^{h_2 \cdot h_3}}{e^{h_1 \cdot h_3} + e^{h_2 \cdot h_3} + e^{h_3 \cdot h_3}} \\
\text{softmax}(\mathcal{A}_{3,3}) &= \frac{e^{h_3 \cdot h_3}}{e^{h_1 \cdot h_3} + e^{h_2 \cdot h_3} + e^{h_3 \cdot h_3}} \\
Z_3 &= \frac{e^{h_1 \cdot h_3}}{e^{h_1 \cdot h_3} + e^{h_2 \cdot h_3} + e^{h_3 \cdot h_3}} \cdot h_1 + \frac{e^{h_2 \cdot h_3}}{e^{h_1 \cdot h_3} + e^{h_2 \cdot h_3} + e^{h_3 \cdot h_3}} \cdot h_2 + \\
&\quad \frac{e^{h_3 \cdot h_3}}{e^{h_1 \cdot h_3} + e^{h_2 \cdot h_3} + e^{h_3 \cdot h_3}} \cdot h_3 \\
\Omega_3 &= V \cdot Z_3 + c \\
\hat{y}_3 &= \text{softmax}(\Omega_3) \\
L_3 &= -y_3 \ln(\hat{y}_3)
\end{aligned}$$

For $t = 3$, we now have:

$$\begin{aligned}
\frac{\partial L_3}{\partial W} &= \left(\frac{\partial L_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial \Omega_3} \frac{\partial \Omega_3}{\partial Z_3} \frac{\partial Z_3}{\partial h_1} \frac{\partial h_1}{\partial W} \right) + \\
&\quad \left(\frac{\partial L_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial \Omega_3} \frac{\partial \Omega_3}{\partial Z_3} \frac{\partial Z_3}{\partial h_2} \frac{\partial h_2}{\partial W} + \frac{\partial L_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial \Omega_3} \frac{\partial \Omega_3}{\partial Z_3} \frac{\partial Z_3}{\partial h_2} \frac{\partial h_2}{\partial W} \right) + \\
&\quad \left(\frac{\partial L_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial \Omega_3} \frac{\partial \Omega_3}{\partial Z_3} \frac{\partial Z_3}{\partial h_3} \frac{\partial h_3}{\partial W} + \frac{\partial L_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial \Omega_3} \frac{\partial \Omega_3}{\partial Z_3} \frac{\partial Z_3}{\partial h_3} \frac{\partial h_3}{\partial W} + \right. \\
&\quad \left. \frac{\partial L_3}{\partial \hat{y}_3} \frac{\partial \hat{y}_3}{\partial \Omega_3} \frac{\partial \Omega_3}{\partial Z_3} \frac{\partial Z_3}{\partial h_3} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W} \right)
\end{aligned}$$

For $t = 4$, we use the above formulas as follows:

$$\begin{aligned}
\frac{\partial L_4}{\partial W} = & \left(\frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial \Omega_4} \frac{\partial \Omega_4}{\partial \mathcal{Z}_4} \frac{\partial \mathcal{Z}_4}{\partial h_1} \frac{\partial h_1}{\partial W} \right) + \\
& \left(\frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial \Omega_4} \frac{\partial \Omega_4}{\partial \mathcal{Z}_4} \frac{\partial \mathcal{Z}_4}{\partial h_2} \frac{\partial h_2}{\partial W} + \frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial \Omega_4} \frac{\partial \Omega_4}{\partial \mathcal{Z}_4} \frac{\partial \mathcal{Z}_4}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W} \right) + \\
& \left(\frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial \Omega_4} \frac{\partial \Omega_4}{\partial \mathcal{Z}_4} \frac{\partial \mathcal{Z}_4}{\partial h_3} \frac{\partial h_3}{\partial W} + \frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial \Omega_4} \frac{\partial \Omega_4}{\partial \mathcal{Z}_4} \frac{\partial \mathcal{Z}_4}{\partial h_3} \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial W} + \right. \\
& \quad \left. \frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial \Omega_4} \frac{\partial \Omega_4}{\partial \mathcal{Z}_4} \frac{\partial \mathcal{Z}_4}{\partial h_3} \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W} \right) + \\
& \left(\frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial \Omega_4} \frac{\partial \Omega_4}{\partial \mathcal{Z}_4} \frac{\partial \mathcal{Z}_4}{\partial h_4} \frac{\partial h_4}{\partial W} + \frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial \Omega_4} \frac{\partial \Omega_4}{\partial \mathcal{Z}_4} \frac{\partial \mathcal{Z}_4}{\partial h_4} \frac{\partial h_4}{\partial h_3} \frac{\partial h_3}{\partial W} + \right. \\
& \quad \left. \frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial \Omega_4} \frac{\partial \Omega_4}{\partial \mathcal{Z}_4} \frac{\partial \mathcal{Z}_4}{\partial h_4} \frac{\partial h_4}{\partial h_3} \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial W} + \frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial \Omega_4} \frac{\partial \Omega_4}{\partial \mathcal{Z}_4} \frac{\partial \mathcal{Z}_4}{\partial h_4} \frac{\partial h_4}{\partial h_3} \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W} \right)
\end{aligned}$$

Let us now group the common terms:

$$\begin{aligned}
\frac{\partial L_4}{\partial W} = & \frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial \Omega_4} \frac{\partial \Omega_4}{\partial \mathcal{Z}_4} \left(\left(\frac{\partial \mathcal{Z}_4}{\partial h_1} \frac{\partial h_1}{\partial W} \right) + \left(\frac{\partial \mathcal{Z}_4}{\partial h_2} \frac{\partial h_2}{\partial W} + \frac{\partial \mathcal{Z}_4}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W} \right) + \right. \\
& \left(\frac{\partial \mathcal{Z}_4}{\partial h_3} \frac{\partial h_3}{\partial W} + \frac{\partial \mathcal{Z}_4}{\partial h_3} \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial W} + \frac{\partial \mathcal{Z}_4}{\partial h_3} \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W} \right) + \\
& \left. \left(\frac{\partial \mathcal{Z}_4}{\partial h_4} \frac{\partial h_4}{\partial W} + \frac{\partial \mathcal{Z}_4}{\partial h_4} \frac{\partial h_4}{\partial h_3} \frac{\partial h_3}{\partial W} + \frac{\partial \mathcal{Z}_4}{\partial h_4} \frac{\partial h_4}{\partial h_3} \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial W} + \frac{\partial \mathcal{Z}_4}{\partial h_4} \frac{\partial h_4}{\partial h_3} \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W} \right) \right)
\end{aligned}$$

Let us introduce summations and products into the formulation:

$$\frac{\partial L_4}{\partial W} = \frac{\partial L_4}{\partial \hat{y}_4} \frac{\partial \hat{y}_4}{\partial \Omega_4} \frac{\partial \Omega_4}{\partial \mathcal{Z}_4} \left(\sum_{m=1}^4 \sum_{k=1}^m \frac{\partial \mathcal{Z}_4}{\partial h_m} \prod_{j=k}^{m-1} \frac{\partial h_{j+1}}{\partial h_j} \left(\frac{\partial h_k}{\partial W} \right) \right)$$

Let us generalize for \mathcal{S} steps:

$$\frac{\partial L}{\partial W} = \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial W} = \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial \mathcal{Z}_t} \left(\sum_{m=1}^t \sum_{k=1}^m \frac{\partial \mathcal{Z}_t}{\partial h_m} \prod_{j=k}^{m-1} \frac{\partial h_{j+1}}{\partial h_j} \left(\frac{\partial h_k}{\partial W} \right) \right) \quad (12)$$

Finally, we insert the individual partial derivatives to calculate our final gradi-

ents of L with respect to W , where:

$$\begin{aligned}\frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} &= (y_t - \hat{y}_t) \\ \frac{\partial \Omega_t}{\partial Z_t} &= V^\top \\ \frac{\partial Z_t}{\partial h_m} &= \mathcal{A}_{t,m} \\ \frac{\partial h_{j+1}}{\partial h_j} &= W^\top (1 - h_{j+1}^2) \\ \frac{\partial h_k}{\partial W} &= (1 - h_k^2) h_{k-1}\end{aligned}$$

In this case, The Analytical Derivatives of Eq. (12) becomes:

$$\frac{\partial L}{\partial W} = \sum_{t=1}^S \frac{\partial L_t}{\partial W} = \sum_{t=1}^S (y_t - \hat{y}_t) V^\top \left(\sum_{m=1}^t \sum_{k=1}^m \mathcal{A}_{t,m} \prod_{j=k}^{m-1} W^\top (1 - h_{j+1}^2) \left((1 - h_k^2) h_{k-1} \right) \right) \quad (13)$$

2.6 Derivative of U

Now, let us compute the partial derivation of L with respect to U . Similar to the case of W in Eq. (12), for U we have:

$$\frac{\partial L}{\partial U} = \sum_{t=1}^S \frac{\partial L_t}{\partial U} = \sum_{t=1}^S \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial Z_t} \left(\sum_{m=1}^t \sum_{k=1}^m \frac{\partial Z_t}{\partial h_m} \prod_{j=k}^{m-1} \frac{\partial h_{j+1}}{\partial h_j} \left(\frac{\partial h_k}{\partial U} \right) \right) \quad (14)$$

We insert the individual partial derivatives into Eq. (14) as follows:

$$\begin{aligned}\frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} &= (y_t - \hat{y}_t) \\ \frac{\partial \Omega_t}{\partial Z_t} &= V^\top \\ \frac{\partial Z_t}{\partial h_m} &= \mathcal{A}_{t,m} \\ \frac{\partial h_{j+1}}{\partial h_j} &= W^\top (1 - h_{j+1}^2) \\ \frac{\partial h_k}{\partial U} &= (1 - h_k^2) x_k\end{aligned}$$

Inserting the above derivatives into Eq. (14), we have:

$$\frac{\partial L}{\partial U} = \sum_{t=1}^S \frac{\partial L_t}{\partial U} = \sum_{t=1}^S (y_t - \hat{y}_t) V^\top \left(\sum_{m=1}^t \sum_{k=1}^m \mathcal{A}_{t,m} \prod_{j=k}^{m-1} W^\top (1 - h_{j+1}^2) \left((1 - h_k^2) x_k \right) \right) \quad (15)$$

2.7 Derivative of b

In the same manner, gradient of L with respect to b is calculated similar to Eq. (14) as follows:

$$\frac{\partial L}{\partial b} = \sum_{t=1}^S \frac{\partial L_t}{\partial b} = \sum_{t=1}^S \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial \mathcal{Z}_t} \left(\sum_{m=1}^t \sum_{k=1}^m \frac{\partial \mathcal{Z}_t}{\partial h_m} \prod_{j=k}^{m-1} \frac{\partial h_{j+1}}{\partial h_j} \left(\frac{\partial h_k}{\partial b} \right) \right) \quad (16)$$

Recall that the derivatives used in Eq. (16) are:

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

In this case, Eq. (16) becomes:

$$\frac{\partial L}{\partial b} = \sum_{t=1}^S \frac{\partial L_t}{\partial b} = \sum_{t=1}^S (y_t - \hat{y}_t) V^\top \left(\sum_{m=1}^t \sum_{k=1}^m \mathcal{A}_{t,m} \prod_{j=k}^{m-1} W^\top (1 - h_{j+1}^2) \left((1 - h_k^2) \right) \right) \quad (17)$$