

## Long-short Term Memory (LSTM): back-propagation derivation

Harsha Vardhan

May 5, 2022

## Abstract

This document contains derivation of the gradients for a Long-short Term Memory Unit, using back-propagation (or reverse-mode differentiation). For the implementation of the neural-network see the accompanying notebooks.

## 1 Block Architecture

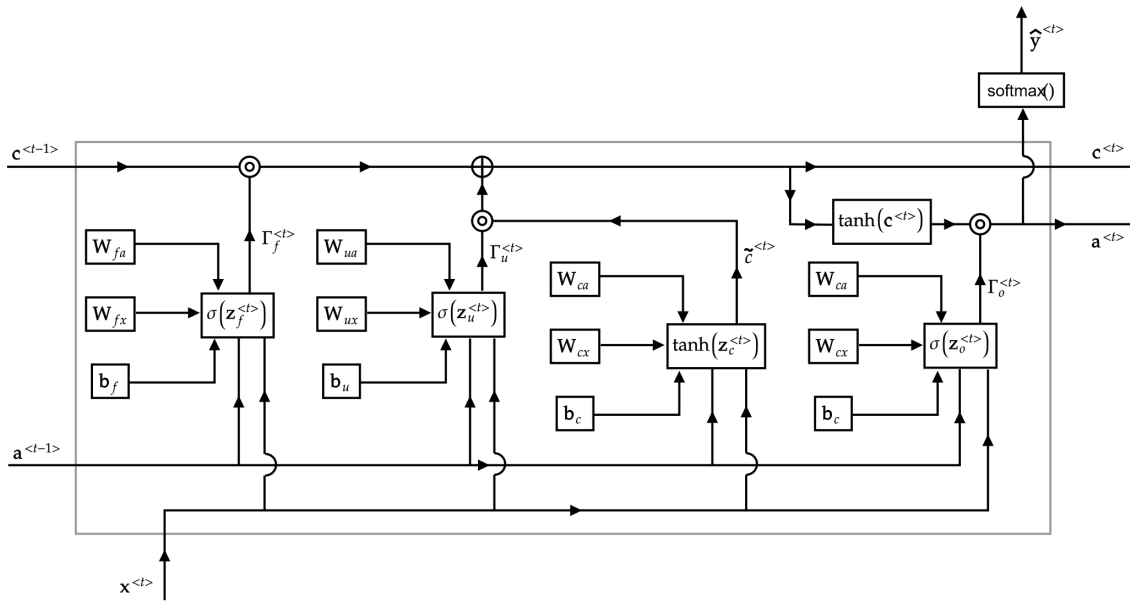


Figure 1: forward propagation diagram for a LSTM-block at time-step  $t$ . The concentric circles represent a hadamard-product (i.e.  $\mathbf{x} \circ \mathbf{y}$ ) of the input vectors.

## 2 Forward Propagation

Given,  $\mathbf{c}^{(t-1)}$  and  $\mathbf{a}^{(t-1)}$  from the  $(t-1)^{th}$  time-step, the equations for forward propagation through the  $t^{th}$  time-step, are as follows: (see figure-1)

$$\mathbf{z}_f^{(t)} = \mathbf{W}_{fa}^\top \mathbf{a}^{(t-1)} + \mathbf{W}_{fx}^\top \mathbf{x}^{(t)} + \mathbf{b}_f \quad \text{eq.1}$$

$$\Gamma_f^{(t)} = \sigma(\mathbf{z}_f^{(t)}) \quad \text{eq.2}$$

$$\mathbf{z}_u^{(t)} = \mathbf{W}_{ua}^\top \mathbf{a}^{(t-1)} + \mathbf{W}_{ux}^\top \mathbf{x}^{(t)} + \mathbf{b}_u \quad \text{eq.3}$$

$$\Gamma_u^{(t)} = \sigma(\mathbf{z}_u^{(t)}) \quad \text{eq.4}$$

$$\mathbf{z}_o^{(t)} = \mathbf{W}_{oa}^\top \mathbf{a}^{(t-1)} + \mathbf{W}_{ox}^\top \mathbf{x}^{(t)} + \mathbf{b}_o \quad \text{eq.5}$$

$$\Gamma_o^{(t)} = \sigma(\mathbf{z}_o^{(t)}) \quad \text{eq.6}$$

$$\mathbf{z}_c^{(t)} = \mathbf{W}_{ca}^\top \mathbf{a}^{(t-1)} + \mathbf{W}_{cx}^\top \mathbf{x}^{(t)} + \mathbf{b}_c \quad \text{eq.7}$$

$$\tilde{\mathbf{c}}^{(t)} = \tanh(\mathbf{z}_c^{(t)}) \quad \text{eq.8}$$

$$\mathbf{c}^{(t)} = \Gamma_u \circ \tilde{\mathbf{c}}^{(t)} + \Gamma_f \circ \mathbf{c}^{(t-1)} \quad \text{eq.9}$$

$$\mathbf{a}^{(t)} = \tanh(\mathbf{c}^{(t)}) \circ \Gamma_o \quad \text{eq.10}$$

$$\hat{\mathbf{y}}^{(t)} = \text{softmax}(\mathbf{a}^{(t)}) \quad \text{eq.11}$$

where,

$\mathbf{x}^{(t)}$  is a  $(n_x, 1)$ -dimensional vector

$\mathbf{c}^{(t)}$  &  $\mathbf{c}^{(t-1)}$  &  $\tilde{\mathbf{c}}^{(t)}$  are  $(n_a, 1)$ -dimensional vectors

$\mathbf{a}^{(t)}$  &  $\mathbf{a}^{(t-1)}$  are  $(n_a, 1)$ -dimensional vectors

$\mathbf{W}_{*a}, * \in \{f, u, o, c\}$  are  $(n_a, n_a)$ -dimensional parameter matrices

$\mathbf{W}_{*x}, * \in \{f, u, o, c\}$  are  $(n_x, n_a)$ -dimensional parameter matrices

$\mathbf{b}_*, * \in \{f, u, o, c\}$  are  $(n_a, 1)$ -dimensional bias-vectors

$\Gamma_*, * \in \{f, u, o\}$  are  $(n_a, 1)$ -dimensional vector, which represents a gate

$\mathbf{z}_*, * \in \{f, u, o, c\}$  are  $(n_a, 1)$ -dimensional vector, is input to a gates' activation

Also, the sub-scripts denote the following:  $_o$  - output-gate,  $_u$  - update-gate, and  $_f$  - forget-gate.

## 2.1 Backward Propagation

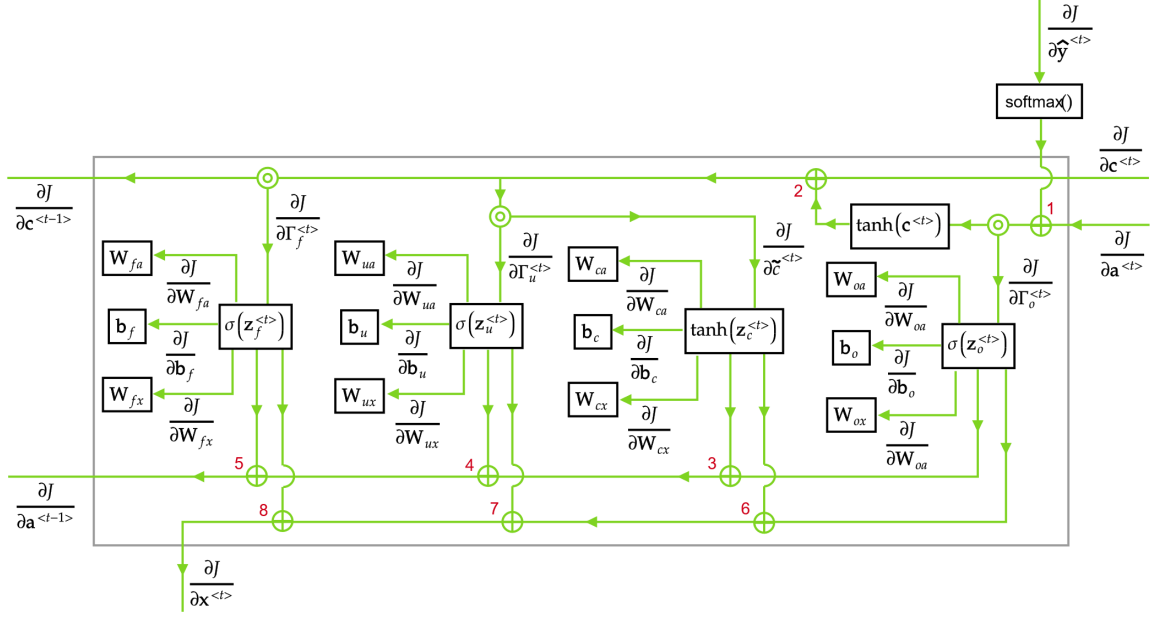


Figure 2: backward-propagation/gradient-flow diagram for a LSTM-block at time-step  $t$ . The concentric circles represent a hadamard-product (i.e.  $\mathbf{x} \circ \mathbf{y}$ ) of the input vectors.

### 2.1.1 Computing $\frac{\partial J}{\partial \hat{\mathbf{y}}^{(t)}}$

Since,  $\hat{\mathbf{y}}^{(t)}$  is computed using the  $\text{softmax}()$  activation-function, the loss  $J$  is computed using the cross-entropy loss. Also, let  $\mathbf{y}^{(t)}$  be the output-label corresponding to the  $t^{\text{th}}$  time-step. Then,

$$\begin{aligned} \frac{\partial J}{\partial \hat{\mathbf{y}}^{(t)}} &= \begin{bmatrix} \frac{\partial J}{\partial \hat{y}_1^{(t)}} & \frac{\partial J}{\partial \hat{y}_2^{(t)}} & \cdots & \frac{\partial J}{\partial \hat{y}_{n_y}^{(t)}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{y_1^{(t)}}{\hat{y}_1^{(t)}} & \frac{y_2^{(t)}}{\hat{y}_2^{(t)}} & \cdots & \frac{y_{n_y}^{(t)}}{\hat{y}_{n_y}^{(t)}} \end{bmatrix} \end{aligned}$$

### 2.1.2 Computing $\frac{\partial J}{\partial \mathbf{a}^{(t)}}$

This derivative has two components

**Comp-1** one flows-in from the  $(t+1)^{\text{th}}$  time-step, and

**Comp-2** one flows-in as the derivative from  $\hat{\mathbf{y}}^{(t)}$ . This derivative is computed as follows,

$$\begin{aligned} \frac{\partial J}{\partial \mathbf{a}^{(t)}} &= \frac{\partial J}{\partial \hat{\mathbf{y}}^{(t)}} \frac{\partial \hat{\mathbf{y}}^{(t)}}{\partial \mathbf{a}^{(t)}} \\ &= \frac{\partial J}{\partial \hat{\mathbf{y}}^{(t)}} (\text{diag}(\hat{\mathbf{y}}^{(t)}) - \hat{\mathbf{y}}^{(t)} (\hat{\mathbf{y}}^{(t)})^\top) \end{aligned}$$

Then, these two components are added to compute the true derivative (see the  $\oplus$  labeled as 1 in the figure-2), i.e.

$$\frac{\partial J}{\partial \mathbf{a}^{(t)}} = \left. \frac{\partial J}{\partial \mathbf{a}^{(t)}} \right|_{\text{Comp-1}} + \left. \frac{\partial J}{\partial \mathbf{a}^{(t)}} \right|_{\text{Comp-2}}$$

### 2.1.3 Computing $\frac{\partial J}{\partial \mathbf{c}^{(t)}}$ , and $\frac{\partial J}{\partial \Gamma_o^{(t)}}$

From *eq.10* (see section-2), we have

$$\begin{aligned} \frac{\partial J}{\partial \Gamma_o^{(t)}} &= \frac{\partial J}{\partial \mathbf{a}^{(t)}} \frac{\partial \mathbf{a}^{(t)}}{\partial \Gamma_o^{(t)}} \\ &= \frac{\partial J}{\partial \mathbf{a}^{(t)}} \text{diag}(\tanh(\mathbf{c}^{(t)})) \\ &= \frac{\partial J}{\partial \mathbf{a}^{(t)}} \circ (\tanh(\mathbf{c}^{(t)}))^{\top} \end{aligned}$$

**Note:** For more information about computing the derivative of a Hadamard-product, see section-4.

The derivative  $\frac{\partial J}{\partial \mathbf{c}^{(t)}}$  is made of two components, described as follows

**Comp-1** one flows-in from the  $(t+1)^{th}$  time-step, and

**Comp-2** one flows-in as the derivative from the  $\mathbf{a}^{(t)}$ . This derivative is computed as follows,

$$\begin{aligned} \frac{\partial J}{\partial \mathbf{c}^{(t)}} &= \frac{\partial J}{\partial \mathbf{a}^{(t)}} \frac{\partial \mathbf{a}^{(t)}}{\partial \mathbf{c}^{(t)}} \\ &= \frac{\partial J}{\partial \mathbf{a}^{(t)}} \frac{\partial \mathbf{a}^{(t)}}{\partial \tanh(\mathbf{c}^{(t)})} \frac{\partial \tanh(\mathbf{c}^{(t)})}{\partial \mathbf{c}^{(t)}} \\ &= \frac{\partial J}{\partial \mathbf{a}^{(t)}} (\mathbf{I}_{n_a \times n_a} \circ \Gamma_o^{(t)}) (\mathbf{I}_{n_a \times n_a} - \mathbf{I}_{n_a \times n_a} \circ \tanh^2(\mathbf{c}^{(t)})) \\ &= \frac{\partial J}{\partial \mathbf{a}^{(t)}} \circ [\Gamma_o^{(t)} \circ (\mathbf{1}_{(n_a,1)} - \tanh^2(\mathbf{c}^{(t)}))]^{\top} \end{aligned}$$

These two components are then added to compute the true derivative (see  $\oplus$  labeled as 2 in figure-2), i.e.

$$\frac{\partial J}{\partial \mathbf{c}^{(t)}} = \left. \frac{\partial J}{\partial \mathbf{c}^{(t)}} \right|_{\text{Comp-1}} + \left. \frac{\partial J}{\partial \mathbf{c}^{(t)}} \right|_{\text{Comp-2}}$$

### 2.1.4 Computing $\frac{\partial J}{\partial \mathbf{W}_{oa}}$ , $\frac{\partial J}{\partial \mathbf{W}_{ox}}$ , and $\frac{\partial J}{\partial \mathbf{b}_o}$

From *eq.6* (see section-2), we have

$$\begin{aligned} \frac{\partial J}{\partial \mathbf{z}_o^{(t)}} &= \frac{\partial J}{\partial \Gamma_o^{(t)}} \frac{\partial \Gamma_o^{(t)}}{\partial \mathbf{z}_o^{(t)}} \\ &= \frac{\partial J}{\partial \Gamma_o^{(t)}} \circ (\Gamma_o^{(t)} \circ (\mathbf{1}_{(n_a,1)} - \Gamma_o^{(t)}))^{\top} \end{aligned}$$

From *eq.5* (see section-2), using the trace-method, we have

$$\text{tr}(J) = \text{tr} \left( \frac{\partial J}{\partial \mathbf{z}_o^{(t)}} d\mathbf{z}_o^{(t)} \right)$$

where,  $d\mathbf{z}_o^{(t)}$  can be expanded as follows:

$$\begin{aligned} d\mathbf{z}_o^{(t)} &= d(\mathbf{W}_{oa}^\top \mathbf{a}^{(t-1)} + \mathbf{W}_{ox}^\top \mathbf{x}^{(t)} + \mathbf{b}_o) \\ &= d\mathbf{W}_{oa}^\top \mathbf{a}^{(t-1)} + \mathbf{W}_{oa}^\top d\mathbf{a}^{(t-1)} + d\mathbf{W}_{ox}^\top \mathbf{x}^{(t)} + \mathbf{W}_{ox}^\top d\mathbf{x}^{(t)} + d\mathbf{b}_o \end{aligned} \quad \text{eq.12-1}$$

when differentiating w.r.t.  $\mathbf{W}_{oa}$ , we have  $d\mathbf{W}_{ox} = 0$ ,  $d\mathbf{a}^{(t-1)} = 0$ ,  $d\mathbf{b}_o = 0$ , and  $d\mathbf{x}^{(t)} = 0$ . So,

$$\begin{aligned} dJ &= \text{tr} \left( \frac{\partial J}{\partial \mathbf{z}_o^{(t)}} d\mathbf{W}_{oa}^\top \mathbf{a}^{(t-1)} \right) \\ &= \text{tr} \left( (\mathbf{a}^{(t-1)})^\top d\mathbf{W}_{oa} \left( \frac{\partial J}{\partial \mathbf{z}_o^{(t)}} \right)^\top \right) \\ &= \text{tr} \left( \left( \frac{\partial J}{\partial \mathbf{z}_o^{(t)}} \right)^\top (\mathbf{a}^{(t-1)})^\top d\mathbf{W}_{oa} \right) \\ &\Rightarrow \frac{\partial J}{\partial \mathbf{W}_{oa}} = \left( \frac{\partial J}{\partial \mathbf{z}_o^{(t)}} \right)^\top (\mathbf{a}^{(t-1)})^\top \end{aligned}$$

when differentiating w.r.t.  $\mathbf{W}_{ox}$ , we have  $d\mathbf{W}_{oa} = 0$ ,  $d\mathbf{a}^{(t-1)} = 0$ ,  $d\mathbf{b}_o = 0$ , and  $d\mathbf{x}^{(t)} = 0$ . So,

$$\begin{aligned} dJ &= \text{tr} \left( \frac{\partial J}{\partial \mathbf{z}_o^{(t)}} d\mathbf{W}_{ox}^\top \mathbf{x}^{(t)} \right) \\ &= \text{tr} \left( (\mathbf{x}^{(t)})^\top d\mathbf{W}_{ox} \left( \frac{\partial J}{\partial \mathbf{z}_o^{(t)}} \right)^\top \right) \\ &= \text{tr} \left( \left( \frac{\partial J}{\partial \mathbf{z}_o^{(t)}} \right)^\top (\mathbf{x}^{(t)})^\top d\mathbf{W}_{ox} \right) \\ &\Rightarrow \frac{\partial J}{\partial \mathbf{W}_{ox}} = \left( \frac{\partial J}{\partial \mathbf{z}_o^{(t)}} \right)^\top (\mathbf{x}^{(t)})^\top \end{aligned}$$

when differentiating w.r.t.  $\mathbf{b}_o$ , we have  $d\mathbf{W}_{oa} = 0$ ,  $d\mathbf{a}^{(t-1)} = 0$ ,  $d\mathbf{W}_{ox} = 0$ , and  $d\mathbf{x}^{(t)} = 0$ . So,

$$\begin{aligned} dJ &= \text{tr} \left( \frac{\partial J}{\partial \mathbf{z}_o^{(t)}} d\mathbf{b}_o \right) \\ &\Rightarrow \frac{\partial J}{\partial \mathbf{b}_o} = \frac{\partial J}{\partial \mathbf{z}_o^{(t)}} \end{aligned}$$

### 2.1.5 Computing $\frac{\partial J}{\partial \Gamma_u^{(t)}}$ , $\frac{\partial J}{\partial \tilde{\mathbf{c}}^{(t)}}$ , $\frac{\partial J}{\partial \Gamma_f^{(t)}}$ , and $\frac{\partial J}{\partial \mathbf{c}^{(t-1)}}$

From *eq.9* (see section-2), we have

$$\begin{aligned}
\frac{\partial J}{\partial \Gamma_u^{(t)}} &= \frac{\partial J}{\partial \mathbf{c}^{(t)}} \frac{\partial \mathbf{c}^{(t)}}{\partial \Gamma_u^{(t)}} \\
&= \frac{\partial J}{\partial \mathbf{c}^{(t)}} \text{diag}(\tilde{\mathbf{c}}^{(t)}) \\
&= \frac{\partial J}{\partial \mathbf{c}^{(t)}} \circ (\tilde{\mathbf{c}}^{(t)})^\top \\
\frac{\partial J}{\partial \tilde{\mathbf{c}}^{(t)}} &= \frac{\partial J}{\partial \mathbf{c}^{(t)}} \frac{\partial \mathbf{c}^{(t)}}{\partial \tilde{\mathbf{c}}^{(t)}} \\
&= \frac{\partial J}{\partial \mathbf{c}^{(t)}} \text{diag}(\Gamma_u^{(t)}) \\
&= \frac{\partial J}{\partial \mathbf{c}^{(t)}} \circ (\Gamma_u^{(t)})^\top \\
\frac{\partial J}{\partial \Gamma_f^{(t)}} &= \frac{\partial J}{\partial \mathbf{c}^{(t)}} \frac{\partial \mathbf{c}^{(t)}}{\partial \Gamma_f^{(t)}} \\
&= \frac{\partial J}{\partial \mathbf{c}^{(t)}} \text{diag}(\mathbf{c}^{(t-1)}) \\
&= \frac{\partial J}{\partial \mathbf{c}^{(t)}} \circ (\mathbf{c}^{(t-1)})^\top \\
\frac{\partial J}{\partial \mathbf{c}^{(t-1)}} &= \frac{\partial J}{\partial \mathbf{c}^{(t)}} \frac{\partial \mathbf{c}^{(t)}}{\partial \mathbf{c}^{(t-1)}} \\
&= \frac{\partial J}{\partial \mathbf{c}^{(t)}} \text{diag}(\Gamma_f^{(t)}) \\
&= \frac{\partial J}{\partial \mathbf{c}^{(t)}} \circ (\Gamma_f^{(t)})^\top
\end{aligned}$$

### 2.1.6 Computing $\frac{\partial J}{\partial \mathbf{W}_{ua}}$ , $\frac{\partial J}{\partial \mathbf{W}_{ux}}$ , and $\frac{\partial J}{\partial \mathbf{b}_u}$

From *eq.4* (see section-2), we have

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{z}_u^{(t)}} &= \frac{\partial J}{\partial \Gamma_u^{(t)}} \frac{\partial \Gamma_u^{(t)}}{\partial \mathbf{z}_u^{(t)}} \\
&= \frac{\partial J}{\partial \Gamma_u^{(t)}} \circ (\Gamma_u^{(t)} \circ (\mathbf{1}_{(n_a,1)} - \Gamma_u^{(t)}))^\top
\end{aligned}$$

From *eq.3* (see section-2), using the trace-method, we have

$$\text{tr}(J) = \text{tr} \left( \frac{\partial J}{\partial \mathbf{z}_u^{(t)}} d\mathbf{z}_u^{(t)} \right)$$

where,  $d\mathbf{z}_u^{(t)}$  can be expanded as follows:

$$\begin{aligned}
d\mathbf{z}_u^{(t)} &= d(\mathbf{W}_{ua}^\top \mathbf{a}^{(t-1)} + \mathbf{W}_{ux}^\top \mathbf{x}^{(t)} + \mathbf{b}_u) \\
&= d\mathbf{W}_{ua}^\top \mathbf{a}^{(t-1)} + \mathbf{W}_{ua}^\top d\mathbf{a}^{(t-1)} + d\mathbf{W}_{ux}^\top \mathbf{x}^{(t)} + \mathbf{W}_{ux}^\top d\mathbf{x}^{(t)} + d\mathbf{b}_u \quad \text{eq.12-2}
\end{aligned}$$

when differentiating w.r.t.  $\mathbf{W}_{ua}$ , we have  $d\mathbf{W}_{ux} = 0$ ,  $d\mathbf{a}^{(t-1)} = 0$ ,  $d\mathbf{b}_u = 0$ , and  $d\mathbf{x}^{(t)} = 0$ . So,

$$\begin{aligned}
dJ &= \text{tr} \left( \frac{\partial J}{\partial \mathbf{z}_u^{(t)}} d\mathbf{W}_{ua}^\top \mathbf{a}^{(t-1)} \right) \\
&= \text{tr} \left( (\mathbf{a}^{(t-1)})^\top d\mathbf{W}_{ua} \left( \frac{\partial J}{\partial \mathbf{z}_u^{(t)}} \right)^\top \right) \\
&= \text{tr} \left( \left( \frac{\partial J}{\partial \mathbf{z}_u^{(t)}} \right)^\top (\mathbf{a}^{(t-1)})^\top d\mathbf{W}_{ua} \right) \\
&\Rightarrow \frac{\partial J}{\partial \mathbf{W}_{ua}} = \left( \frac{\partial J}{\partial \mathbf{z}_u^{(t)}} \right)^\top (\mathbf{a}^{(t-1)})^\top
\end{aligned}$$

when differentiating w.r.t.  $\mathbf{W}_{ux}$ , we have  $d\mathbf{W}_{ua} = 0$ ,  $d\mathbf{a}^{(t-1)} = 0$ ,  $d\mathbf{b}_u = 0$ , and  $d\mathbf{x}^{(t)} = 0$ . So,

$$\begin{aligned}
dJ &= \text{tr} \left( \frac{\partial J}{\partial \mathbf{z}_u^{(t)}} d\mathbf{W}_{ux}^\top \mathbf{x}^{(t)} \right) \\
&= \text{tr} \left( (\mathbf{x}^{(t)})^\top d\mathbf{W}_{ux} \left( \frac{\partial J}{\partial \mathbf{z}_u^{(t)}} \right)^\top \right) \\
&= \text{tr} \left( \left( \frac{\partial J}{\partial \mathbf{z}_u^{(t)}} \right)^\top (\mathbf{x}^{(t)})^\top d\mathbf{W}_{ux} \right) \\
&\Rightarrow \frac{\partial J}{\partial \mathbf{W}_{ux}} = \left( \frac{\partial J}{\partial \mathbf{z}_u^{(t)}} \right)^\top (\mathbf{x}^{(t)})^\top
\end{aligned}$$

when differentiating w.r.t.  $\mathbf{b}_u$ , we have  $d\mathbf{W}_{ua} = 0$ ,  $d\mathbf{a}^{(t-1)} = 0$ ,  $d\mathbf{W}_{ux} = 0$ , and  $d\mathbf{x}^{(t)} = 0$ . So,

$$\begin{aligned}
dJ &= \text{tr} \left( \frac{\partial J}{\partial \mathbf{z}_u^{(t)}} d\mathbf{b}_u \right) \\
&\Rightarrow \frac{\partial J}{\partial \mathbf{b}_u} = \frac{\partial J}{\partial \mathbf{z}_u^{(t)}}
\end{aligned}$$

### 2.1.7 Computing $\frac{\partial J}{\partial \mathbf{W}_{fa}}$ , $\frac{\partial J}{\partial \mathbf{W}_{fx}}$ , and $\frac{\partial J}{\partial \mathbf{b}_f}$

From *eq.2* (see section-2), we have

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{z}_f^{(t)}} &= \frac{\partial J}{\partial \Gamma_f^{(t)}} \frac{\partial \Gamma_f^{(t)}}{\partial \mathbf{z}_f^{(t)}} \\
&= \frac{\partial J}{\partial \Gamma_f^{(t)}} \circ \left( \Gamma_f^{(t)} \circ (\mathbf{1}_{(n_a,1)} - \Gamma_f^{(t)}) \right)^\top
\end{aligned}$$

From *eq.1* (see section-2), using the trace-method, we have

$$\text{tr}(J) = \text{tr} \left( \frac{\partial J}{\partial \mathbf{z}_f^{(t)}} d\mathbf{z}_f^{(t)} \right)$$

where,  $d\mathbf{z}_f^{(t)}$  can be expanded as follows:

$$\begin{aligned}
d\mathbf{z}_f^{(t)} &= d(\mathbf{W}_{fa}^\top \mathbf{a}^{(t-1)} + \mathbf{W}_{fx}^\top \mathbf{x}^{(t)} + \mathbf{b}_f) \\
&= d\mathbf{W}_{fa}^\top \mathbf{a}^{(t-1)} + \mathbf{W}_{fa}^\top d\mathbf{a}^{(t-1)} + d\mathbf{W}_{fx}^\top \mathbf{x}^{(t)} + \mathbf{W}_{fx}^\top d\mathbf{x}^{(t)} + d\mathbf{b}_f \quad \text{eq.12-3}
\end{aligned}$$

when differentiating w.r.t.  $\mathbf{W}_{fa}$ , we have  $d\mathbf{W}_{fx} = 0$ ,  $d\mathbf{a}^{(t-1)} = 0$ ,  $d\mathbf{b}_f = 0$ , and  $d\mathbf{x}^{(t)} = 0$ . So,

$$\begin{aligned} dJ &= \text{tr} \left( \frac{\partial J}{\partial \mathbf{z}_f^{(t)}} d\mathbf{W}_{fa}^\top \mathbf{a}^{(t-1)} \right) \\ &= \text{tr} \left( (\mathbf{a}^{(t-1)})^\top d\mathbf{W}_{fa} \left( \frac{\partial J}{\partial \mathbf{z}_f^{(t)}} \right)^\top \right) \\ &= \text{tr} \left( \left( \frac{\partial J}{\partial \mathbf{z}_f^{(t)}} \right)^\top (\mathbf{a}^{(t-1)})^\top d\mathbf{W}_{fa} \right) \\ &\Rightarrow \frac{\partial J}{\partial \mathbf{W}_{fa}} = \left( \frac{\partial J}{\partial \mathbf{z}_f^{(t)}} \right)^\top (\mathbf{a}^{(t-1)})^\top \end{aligned}$$

when differentiating w.r.t.  $\mathbf{W}_{fx}$ , we have  $d\mathbf{W}_{fa} = 0$ ,  $d\mathbf{a}^{(t-1)} = 0$ ,  $d\mathbf{b}_f = 0$ , and  $d\mathbf{x}^{(t)} = 0$ . So,

$$\begin{aligned} dJ &= \text{tr} \left( \frac{\partial J}{\partial \mathbf{z}_f^{(t)}} d\mathbf{W}_{fx}^\top \mathbf{x}^{(t)} \right) \\ &= \text{tr} \left( (\mathbf{x}^{(t)})^\top d\mathbf{W}_{fx} \left( \frac{\partial J}{\partial \mathbf{z}_f^{(t)}} \right)^\top \right) \\ &= \text{tr} \left( \left( \frac{\partial J}{\partial \mathbf{z}_f^{(t)}} \right)^\top (\mathbf{x}^{(t)})^\top d\mathbf{W}_{fx} \right) \\ &\Rightarrow \frac{\partial J}{\partial \mathbf{W}_{fx}} = \left( \frac{\partial J}{\partial \mathbf{z}_f^{(t)}} \right)^\top (\mathbf{x}^{(t)})^\top \end{aligned}$$

when differentiating w.r.t.  $\mathbf{b}_f$ , we have  $d\mathbf{W}_{fa} = 0$ ,  $d\mathbf{a}^{(t-1)} = 0$ ,  $d\mathbf{W}_{fx} = 0$ , and  $d\mathbf{x}^{(t)} = 0$ . So,

$$\begin{aligned} dJ &= \text{tr} \left( \frac{\partial J}{\partial \mathbf{z}_f^{(t)}} d\mathbf{b}_f \right) \\ &\Rightarrow \frac{\partial J}{\partial \mathbf{b}_f} = \frac{\partial J}{\partial \mathbf{z}_f^{(t)}} \end{aligned}$$

### 2.1.8 Computing $\frac{\partial J}{\partial \mathbf{W}_{ca}}$ , $\frac{\partial J}{\partial \mathbf{W}_{cx}}$ , and $\frac{\partial J}{\partial \mathbf{b}_c}$

From *eq.8* (see section-2), we have

$$\begin{aligned} \frac{\partial J}{\partial \mathbf{z}_c^{(t)}} &= \frac{\partial J}{\partial \tilde{\mathbf{c}}^{(t)}} \frac{\partial \tilde{\mathbf{c}}^{(t)}}{\partial \mathbf{z}_c^{(t)}} \\ &= \frac{\partial J}{\partial \tilde{\mathbf{c}}^{(t)}} \circ [\mathbf{1}_{(n_a, 1)} - \tanh^2(\mathbf{z}_c^{(t)})]^\top \end{aligned}$$

From *eq.7* (see section-2), using the trace-method, we have



$$\text{tr}(J) = \text{tr} \left( \frac{\partial J}{\partial \mathbf{z}_c^{(t)}} d\mathbf{z}_c^{(t)} \right)$$

where,  $d\mathbf{z}_c^{(t)}$  can be expanded as follows:

$$\begin{aligned} d\mathbf{z}_c^{(t)} &= d(\mathbf{W}_{ca}^\top \mathbf{a}^{(t-1)} + \mathbf{W}_{cx}^\top \mathbf{x}^{(t)} + \mathbf{b}_c) \\ &= d\mathbf{W}_{ca}^\top \mathbf{a}^{(t-1)} + \mathbf{W}_{ca}^\top d\mathbf{a}^{(t-1)} + d\mathbf{W}_{cx}^\top \mathbf{x}^{(t)} + \mathbf{W}_{cx}^\top d\mathbf{x}^{(t)} + d\mathbf{b}_c \end{aligned} \quad \text{eq.12-4}$$

when differentiating w.r.t.  $\mathbf{W}_{ca}$ , we have  $d\mathbf{W}_{cx} = 0$ ,  $d\mathbf{a}^{(t-1)} = 0$ ,  $d\mathbf{b}_c = 0$ , and  $d\mathbf{x}^{(t)} = 0$ . So,

$$\begin{aligned} dJ &= \text{tr} \left( \frac{\partial J}{\partial \mathbf{z}_c^{(t)}} d\mathbf{W}_{ca}^\top \mathbf{a}^{(t-1)} \right) \\ &= \text{tr} \left( (\mathbf{a}^{(t-1)})^\top d\mathbf{W}_{ca} \left( \frac{\partial J}{\partial \mathbf{z}_c^{(t)}} \right)^\top \right) \\ &= \text{tr} \left( \left( \frac{\partial J}{\partial \mathbf{z}_c^{(t)}} \right)^\top (\mathbf{a}^{(t-1)})^\top d\mathbf{W}_{ca} \right) \\ &\Rightarrow \frac{\partial J}{\partial \mathbf{W}_{ca}} = \left( \frac{\partial J}{\partial \mathbf{z}_c^{(t)}} \right)^\top (\mathbf{a}^{(t-1)})^\top \end{aligned}$$

when differentiating w.r.t.  $\mathbf{W}_{cx}$ , we have  $d\mathbf{W}_{ca} = 0$ ,  $d\mathbf{a}^{(t-1)} = 0$ ,  $d\mathbf{b}_c = 0$ , and  $d\mathbf{x}^{(t)} = 0$ . So,

$$\begin{aligned} dJ &= \text{tr} \left( \frac{\partial J}{\partial \mathbf{z}_c^{(t)}} d\mathbf{W}_{cx}^\top \mathbf{x}^{(t)} \right) \\ &= \text{tr} \left( (\mathbf{x}^{(t)})^\top d\mathbf{W}_{cx} \left( \frac{\partial J}{\partial \mathbf{z}_c^{(t)}} \right)^\top \right) \\ &= \text{tr} \left( \left( \frac{\partial J}{\partial \mathbf{z}_c^{(t)}} \right)^\top (\mathbf{x}^{(t)})^\top d\mathbf{W}_{cx} \right) \\ &\Rightarrow \frac{\partial J}{\partial \mathbf{W}_{cx}} = \left( \frac{\partial J}{\partial \mathbf{z}_c^{(t)}} \right)^\top (\mathbf{x}^{(t)})^\top \end{aligned}$$

when differentiating w.r.t.  $\mathbf{b}_c$ , we have  $d\mathbf{W}_{ca} = 0$ ,  $d\mathbf{a}^{(t-1)} = 0$ ,  $d\mathbf{W}_{cx} = 0$ , and  $d\mathbf{x}^{(t)} = 0$ . So,

$$\begin{aligned} dJ &= \text{tr} \left( \frac{\partial J}{\partial \mathbf{z}_c^{(t)}} d\mathbf{b}_c \right) \\ &\Rightarrow \frac{\partial J}{\partial \mathbf{b}_c} = \frac{\partial J}{\partial \mathbf{z}_c^{(t)}} \end{aligned}$$

### 2.1.9 Computing $\frac{\partial J}{\partial \mathbf{x}^{(t)}}$ , and $\frac{\partial J}{\partial \mathbf{a}^{(t-1)}}$

Each of these derivatives has four components, described as follows

**Comp-1** flows-in as derivative from  $\Gamma_o^{(t)}$

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{a}^{(t-1)}} &= \frac{\partial J}{\partial \mathbf{z}_o^{(t)}} \frac{\partial \mathbf{z}_o^{(t)}}{\partial \mathbf{a}^{(t-1)}} \\
&= \frac{\partial J}{\partial \mathbf{z}_o^{(t)}} \mathbf{W}_{oa}^\top \\
\frac{\partial J}{\partial \mathbf{x}^{(t)}} &= \frac{\partial J}{\partial \mathbf{z}_o^{(t)}} \frac{\partial \mathbf{z}_o^{(t)}}{\partial \mathbf{x}^{(t)}} \\
&= \frac{\partial J}{\partial \mathbf{z}_o^{(t)}} \mathbf{W}_{ox}^\top
\end{aligned}$$

**Comp-2** flows in as derivative from  $\tilde{\mathbf{c}}^{(t)}$

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{a}^{(t-1)}} &= \frac{\partial J}{\partial \mathbf{z}_c^{(t)}} \frac{\partial \mathbf{z}_c^{(t)}}{\partial \mathbf{a}^{(t-1)}} \\
&= \frac{\partial J}{\partial \mathbf{z}_c^{(t)}} \mathbf{W}_{ca}^\top \\
\frac{\partial J}{\partial \mathbf{x}^{(t)}} &= \frac{\partial J}{\partial \mathbf{z}_c^{(t)}} \frac{\partial \mathbf{z}_c^{(t)}}{\partial \mathbf{x}^{(t)}} \\
&= \frac{\partial J}{\partial \mathbf{z}_c^{(t)}} \mathbf{W}_{cx}^\top
\end{aligned}$$

**Comp-3** flows-in as derivative from  $\Gamma_u^{(t)}$

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{a}^{(t-1)}} &= \frac{\partial J}{\partial \mathbf{z}_u^{(t)}} \frac{\partial \mathbf{z}_u^{(t)}}{\partial \mathbf{a}^{(t-1)}} \\
&= \frac{\partial J}{\partial \mathbf{z}_u^{(t)}} \mathbf{W}_{ua}^\top \\
\frac{\partial J}{\partial \mathbf{x}^{(t)}} &= \frac{\partial J}{\partial \mathbf{z}_u^{(t)}} \frac{\partial \mathbf{z}_u^{(t)}}{\partial \mathbf{x}^{(t)}} \\
&= \frac{\partial J}{\partial \mathbf{z}_u^{(t)}} \mathbf{W}_{ux}^\top
\end{aligned}$$

**Comp-4** flows-in as derivative from  $\Gamma_f^{(t)}$

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{a}^{(t-1)}} &= \frac{\partial J}{\partial \mathbf{z}_f^{(t)}} \frac{\partial \mathbf{z}_f^{(t)}}{\partial \mathbf{a}^{(t-1)}} \\
&= \frac{\partial J}{\partial \mathbf{z}_f^{(t)}} \mathbf{W}_{fa}^\top \\
\frac{\partial J}{\partial \mathbf{x}^{(t)}} &= \frac{\partial J}{\partial \mathbf{z}_f^{(t)}} \frac{\partial \mathbf{z}_f^{(t)}}{\partial \mathbf{x}^{(t)}} \\
&= \frac{\partial J}{\partial \mathbf{z}_f^{(t)}} \mathbf{W}_{fx}^\top
\end{aligned}$$

These four components are then added to compute the true derivative, i.e.

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{a}^{(t-1)}} &= \left. \frac{\partial J}{\partial \mathbf{a}^{(t-1)}} \right|_{\text{Comp-1}} + \left. \frac{\partial J}{\partial \mathbf{a}^{(t-1)}} \right|_{\text{Comp-2}} + \left. \frac{\partial J}{\partial \mathbf{a}^{(t-1)}} \right|_{\text{Comp-3}} + \left. \frac{\partial J}{\partial \mathbf{a}^{(t-1)}} \right|_{\text{Comp-4}} \\ &\quad (\text{see } \oplus \text{ labeled as } 3, 4, \text{ and } 5 \text{ in figure-2}) \\ \frac{\partial J}{\partial \mathbf{x}^{(t)}} &= \left. \frac{\partial J}{\partial \mathbf{x}^{(t)}} \right|_{\text{Comp-1}} + \left. \frac{\partial J}{\partial \mathbf{x}^{(t)}} \right|_{\text{Comp-2}} + \left. \frac{\partial J}{\partial \mathbf{x}^{(t)}} \right|_{\text{Comp-3}} + \left. \frac{\partial J}{\partial \mathbf{x}^{(t)}} \right|_{\text{Comp-4}} \\ &\quad (\text{see } \oplus \text{ labeled as } 6, 7, \text{ and } 8 \text{ in figure-2})\end{aligned}$$

### 3 Gradient or Jacobian?

In the above derivations, we have used the numerator layout while performing matrix-derivatives. One of the consequences of this decision is that the derivatives that we have computed are in-fact jacobians and not gradients. Fortunately, gradients are just transpose of jacobians.

### 4 Appendix-A:

In this section we will derive an expression for the derivative of a Hadamard product of two vectors. Let  $\mathbf{x}$  and  $\mathbf{y}$  be two vectors, and let  $\mathbf{w} = \mathbf{x} \circ \mathbf{y}$  such that

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}; \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad \mathbf{w} = \mathbf{x} \circ \mathbf{y} = \begin{bmatrix} x_1 y_1 \\ x_2 y_2 \\ \vdots \\ x_n y_n \end{bmatrix}$$

then, we have

$$\begin{aligned}\frac{\partial \mathbf{w}}{\partial \mathbf{x}} &= \begin{bmatrix} \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_2} & \cdots & \frac{\partial}{\partial x_n} \end{bmatrix} \otimes \begin{bmatrix} x_1 y_1 \\ x_2 y_2 \\ \vdots \\ x_n y_n \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial(x_1 y_1)}{\partial x_1} & \frac{\partial(x_1 y_1)}{\partial x_2} & \cdots & \frac{\partial(x_1 y_1)}{\partial x_n} \\ \frac{\partial(x_2 y_2)}{\partial x_1} & \frac{\partial(x_2 y_2)}{\partial x_2} & \cdots & \frac{\partial(x_2 y_2)}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial(x_n y_n)}{\partial x_1} & \frac{\partial(x_n y_n)}{\partial x_2} & \cdots & \frac{\partial(x_n y_n)}{\partial x_n} \end{bmatrix} \\ &= \begin{bmatrix} y_1 & 0 & \cdots & 0 \\ 0 & y_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y_n \end{bmatrix} \\ &= \text{diag}(\mathbf{y})\end{aligned} \tag{1}$$