Recurrent Neural Network

mursalimov.emil

August 27, 2019

1 Recurrent Neural Network

Elman RNN model with tanh:

$$h^{[t]} = \tanh(W_x x^{[t]} + b_x + W_h h^{[t-1]} + b_h)$$
$$y^{[t]} = W_y h^{[t]} + b_y$$
$$p^{[t]} = softmax(y^{[t]})$$
$$E^{[t]} = -\log p_\lambda^{[t]}$$
$$E = \sum_{t=1}^{T} E^{[t]}$$

$$x^{[t]} \in \mathbb{R}^{N\times 1},\ h^{[t]} \in \mathbb{R}^{H\times 1},\ y^{[t]} \in \mathbb{R}^{K\times 1},\ p^{[t]} \in \mathbb{R}^{K\times 1},\ E^{[t]} \in \mathbb{R},\ E \in \mathbb{R}$$
$$W_x \in \mathbb{R}^{H\times N},\ b_x \in \mathbb{R}^{H\times 1},\ W_h \in \mathbb{R}^{H\times H},\ b_h \in \mathbb{R}^{H\times 1},\ W_y \in \mathbb{R}^{K\times H},\ b_y \in \mathbb{R}^{K\times 1}$$

Notes:

1. $\lambda \colon \mathbb{R}^{N\times 1} \to \mathbb{R}$ - feature vector to class label: $\lambda = \lambda(x^{[t]})$
2. $l^{[t]} = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{K\times 1}$ - one-hot target vector, where $l_i^{[t]} = \begin{cases} 1, & \text{if } i = \lambda \\ 0, & \text{otherwise} \end{cases}$
3. σ_{ij} - the Kronecker delta, where $\sigma_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$
4. $\text{vec} \colon \mathbb{R}^{m\times n} \to \mathbb{R}^{mn\times 1}$, if $A \in \mathbb{R}^{m\times n}$ then $\text{vec}\ A = \begin{bmatrix} A^{(1)} \\ A^{(2)} \\ \vdots \\ A^{(n)} \end{bmatrix} \in \mathbb{R}^{mn\times 1}$, where $A^{(j)}$ is *j-th* column
5. $\text{mat}_{m\times n} \colon \mathbb{R}^{1\times mn} \to \mathbb{R}^{m\times n}$, if $a \in \mathbb{R}^{1\times mn}$ then $\text{mat}\quad a = \begin{bmatrix} a_1 & a_{m+1} & \dots & a_{(n-1)m+1} \\ a_2 & a_{m+2} & \dots & a_{(n-1)m+2} \\ \vdots & \vdots & \ddots & \vdots \\ a_m & a_{2m} & \dots & a_{nm} \end{bmatrix} \in \mathbb{R}^{m\times n}$
6. $I \in \mathbb{R}^{H\times H}$ - identity matrix

Update weights equation:

$$\Theta = \{W_x, b_x, W_h, b_h, W_y, b_y\}$$

$\theta_j = \theta_j - \gamma DE(\theta_j)$, where $\theta_j \in \Theta$ is a matrix, $DE(\theta_j) = \frac{\partial E}{\partial \theta_j}$ - Jacobian matrix

$\theta_j = \theta_j - \gamma \nabla E(\theta_j)$, where $\theta_j \in \Theta$ is a vector, $\nabla E(\theta_j) = (\frac{\partial E}{\partial \theta_j})^T$ - Gradient vector

2 Differentials and Jacobian matrices

2.1 $\boldsymbol{\theta_j = W_x}$

$$\frac{\partial E}{\partial W_x} = \sum_{t=1}^{T} \frac{\partial E^{[t]}}{\partial W_x} \in \mathbb{R}^{H \times N}$$

$$\frac{\partial E^{[t]}}{\partial W_x} = \mathrm{mat}_{H \times N}(\frac{\partial E^{[t]}}{\partial(\mathrm{vec}\ W_x)}) \in \mathbb{R}^{H \times N}$$

$$\frac{\partial E^{[t]}}{\partial(\mathrm{vec}\ W_x)} = \frac{\partial E^{[t]}}{\partial p^{[t]}} \frac{\partial p^{[t]}}{\partial y^{[t]}} \frac{\partial y^{[t]}}{\partial h^{[t]}} \frac{\partial h^{[t]}}{\partial(\mathrm{vec}\ W_x)} \in \mathbb{R}^{1 \times HN}$$

Recursive ratio:

$$\begin{aligned} z^{[t]} &= z_x^{[t]} + z_h^{[t-1]} \\ z_x^{[t]} &= W_x x^{[t]} + b_x \\ z_h^{[t-1]} &= W_h h^{[t-1]} + b_h \end{aligned}$$

$$\begin{aligned} \frac{\partial h^{[t]}}{\partial(\mathrm{vec}\ W_x)} &= \frac{\partial h^{[t]}}{\partial z^{[t]}} \frac{\partial z^{[t]}}{\partial(\mathrm{vec}\ W_x)} = \frac{\partial h^{[t]}}{\partial z^{[t]}} \frac{\partial(z_x^{[t]} + z_h^{[t-1]})}{\partial(\mathrm{vec}\ W_x)} \\ &= \frac{\partial h^{[t]}}{\partial z^{[t]}} (\frac{\partial z_x^{[t]}}{\partial(\mathrm{vec}\ W_x)} + \frac{\partial z_h^{[t-1]}}{\partial(\mathrm{vec}\ W_x)}) \\ &= \frac{\partial h^{[t]}}{\partial z^{[t]}} (\frac{\partial z_x^{[t]}}{\partial(\mathrm{vec}\ W_x)} + \frac{\partial z_h^{[t-1]}}{\partial h^{[t-1]}} \frac{\partial h^{[t-1]}}{\partial(\mathrm{vec}\ W_x)}) \end{aligned}$$

$$\frac{\partial h^{[t]}}{\partial(\mathrm{vec}\ W_x)} = \frac{\partial h^{[t]}}{\partial z^{[t]}} (\frac{\partial z_x^{[t]}}{\partial(\mathrm{vec}\ W_x)} + \frac{\partial z_h^{[t-1]}}{\partial h^{[t-1]}} \frac{\partial h^{[t-1]}}{\partial(\mathrm{vec}\ W_x)})$$

Recursive formula:

$$\frac{\partial E^{[t]}}{\partial(\mathrm{vec}\ W_x)} = \frac{\partial E^{[t]}}{\partial p^{[t]}} \frac{\partial p^{[t]}}{\partial y^{[t]}} \frac{\partial y^{[t]}}{\partial h^{[t]}} \frac{\partial h^{[t]}}{\partial z^{[t]}} (\frac{\partial z_x^{[t]}}{\partial(\mathrm{vec}\ W_x)} + \frac{\partial z_h^{[t-1]}}{\partial h^{[t-1]}} \frac{\partial h^{[t-1]}}{\partial(\mathrm{vec}\ W_x)}) \in \mathbb{R}^{1 \times HN}$$

Jacobians:

$$\frac{\partial E^{[t]}}{\partial p^{[t]}} = \frac{\partial(-\log p_\lambda^{[t]})}{\partial p^{[t]}} = \begin{bmatrix} 0 \dots (-\frac{1}{p_\lambda^{[t]}}) \dots 0 \end{bmatrix} \in \mathbb{R}^{1 \times K}$$

$$\frac{\partial p^{[t]}}{\partial y^{[t]}} = \frac{\partial(softmax(y^{[t]}))}{\partial y^{[t]}} = \begin{bmatrix} p_1^{[t]}(\sigma_{11} - p_1^{[t]}) & p_1^{[t]}(\sigma_{12} - p_2^{[t]}) & \dots & p_1^{[t]}(\sigma_{1K} - p_K^{[t]}) \\ p_2^{[t]}(\sigma_{21} - p_1^{[t]}) & p_2^{[t]}(\sigma_{22} - p_2^{[t]}) & \dots & p_2^{[t]}(\sigma_{2K} - p_K^{[t]}) \\ \vdots & \vdots & \ddots & \vdots \\ p_K^{[t]}(\sigma_{K1} - p_1^{[t]}) & p_K^{[t]}(\sigma_{K2} - p_2^{[t]}) & \dots & p_K^{[t]}(\sigma_{KK} - p_K^{[t]}) \end{bmatrix} \in \mathbb{R}^{K \times K}$$

$$\frac{\partial y^{[t]}}{\partial h^{[t]}} = \frac{\partial(W_y h^{[t]} + b_y)}{\partial h^{[t]}} = \begin{bmatrix} w_{y_{11}} & w_{y_{12}} & \dots & w_{y_{1H}} \\ w_{y_{21}} & w_{y_{22}} & \dots & w_{y_{2H}} \\ \vdots & \vdots & \ddots & \vdots \\ w_{y_{K1}} & w_{y_{K2}} & \dots & w_{y_{KH}} \end{bmatrix} = W_y \in \mathbb{R}^{K \times H}$$

$$\frac{\partial h^{[t]}}{\partial z^{[t]}} = \frac{\partial \tanh(z^{[t]})}{\partial z^{[t]}} = \begin{bmatrix} 1 - (h_1^{[t]})^2 & 0 & \dots & 0 \\ 0 & 1 - (h_2^{[t]})^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 - (h_H^{[t]})^2 \end{bmatrix} = \mathrm{diag}(\mathbf{1} - (h^{[t]})^2) = \Lambda^{[t]} \in \mathbb{R}^{H \times H}$$

$$\frac{\partial z_x^{[t]}}{\partial (\text{vec } W_x)} = \frac{\partial (W_x x^{[t]} + b_x)}{\partial (\text{vec } W_x)} = \begin{bmatrix} x_1^{[t]} & 0 & \dots & 0 & \dots & x_N^{[t]} & 0 & \dots & 0 \\ 0 & x_1^{[t]} & \dots & 0 & \dots & 0 & x_N^{[t]} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x_1^{[t]} & \dots & 0 & 0 & \dots & x_N^{[t]} \end{bmatrix} = (x^{[t]})^T \otimes I \in \mathbb{R}^{H \times HN}$$

$$\frac{\partial z_h^{[t]}}{\partial h^{[t]}} = \frac{\partial (W_h h^{[t]} + b_h)}{\partial h^{[t]}} = \begin{bmatrix} w_{h_{11}} & w_{h_{12}} & \dots & w_{h_{1H}} \\ w_{h_{21}} & w_{h_{22}} & \dots & w_{h_{2H}} \\ \vdots & \vdots & \ddots & \vdots \\ w_{h_{H1}} & w_{h_{H2}} & \dots & w_{h_{HH}} \end{bmatrix} = W_h \in \mathbb{R}^{H \times H}$$

Deployment:

$$\begin{aligned}
\frac{\partial h^{[1]}}{\partial (\text{vec } W_x)} &= \frac{\partial h^{[1]}}{\partial z^{[1]}} \Big(\frac{\partial z_x^{[1]}}{\partial (\text{vec } W_x)} + \overbrace{\frac{\partial z_h^{[0]}}{\partial h^{[0]}} \frac{\partial h^{[0]}}{\partial (\text{vec } W_x)}}^{\mathbf{0} \in \mathbb{R}^{H \times HN}} \Big) \\
&= \frac{\partial h^{[1]}}{\partial z^{[1]}} \frac{\partial z_x^{[1]}}{\partial (\text{vec } W_x)} \\
\frac{\partial h^{[2]}}{\partial (\text{vec } W_x)} &= \frac{\partial h^{[2]}}{\partial z^{[2]}} \frac{\partial z_x^{[2]}}{\partial (\text{vec } W_x)} + \frac{\partial h^{[2]}}{\partial z^{[2]}} \frac{\partial z_h^{[1]}}{\partial h^{[1]}} \frac{\partial h^{[1]}}{\partial z^{[1]}} \frac{\partial z_x^{[1]}}{\partial (\text{vec } W_x)} \\
\frac{\partial h^{[3]}}{\partial (\text{vec } W_x)} &= \frac{\partial h^{[3]}}{\partial z^{[3]}} \frac{\partial z_x^{[3]}}{\partial (\text{vec } W_x)} + \frac{\partial h^{[3]}}{\partial z^{[3]}} \frac{\partial z_h^{[2]}}{\partial h^{[2]}} \frac{\partial h^{[2]}}{\partial z^{[2]}} \frac{\partial z_x^{[2]}}{\partial (\text{vec } W_x)} + \frac{\partial h^{[3]}}{\partial z^{[3]}} \frac{\partial z_h^{[2]}}{\partial h^{[2]}} \frac{\partial h^{[2]}}{\partial z^{[2]}} \frac{\partial z_h^{[1]}}{\partial h^{[1]}} \frac{\partial h^{[1]}}{\partial z^{[1]}} \frac{\partial z_x^{[1]}}{\partial (\text{vec } W_x)} \\
&\vdots \\
\frac{\partial h^{[t]}}{\partial (\text{vec } W_x)} &= \frac{\partial h^{[t]}}{\partial z^{[t]}} \frac{\partial z_x^{[t]}}{\partial (\text{vec } W_x)} + \frac{\partial h^{[t]}}{\partial z^{[t]}} \frac{\partial z_h^{[t-1]}}{\partial h^{[t-1]}} \frac{\partial h^{[t-1]}}{\partial z^{[t-1]}} \frac{\partial z_x^{[t-1]}}{\partial (\text{vec } W_x)} + \dots + \frac{\partial h^{[t]}}{\partial z^{[t]}} \frac{\partial z_h^{[t-1]}}{\partial h^{[t-1]}} \frac{\partial h^{[t-1]}}{\partial z^{[t-1]}} \frac{\partial z_h^{[t-2]}}{\partial h^{[t-2]}} \times \dots \times \frac{\partial h^{[2]}}{\partial z^{[2]}} \frac{\partial z_h^{[1]}}{\partial h^{[1]}} \frac{\partial h^{[1]}}{\partial z^{[1]}} \frac{\partial z_x^{[1]}}{\partial (\text{vec } W_x)}
\end{aligned}$$

Intermediate formula:

$$\frac{\partial E^{[t]}}{\partial (\text{vec } W_x)} = \frac{\partial E^{[t]}}{\partial p^{[t]}} \frac{\partial p^{[t]}}{\partial y^{[t]}} \frac{\partial y^{[t]}}{\partial h^{[t]}} \sum_{k=0}^{t-1} \Big(\prod_{s=0}^{k-1} \frac{\partial h^{[t-s]}}{\partial z^{[t-s]}} \frac{\partial z_h^{[t-s-1]}}{\partial h^{[t-s-1]}} \Big) \frac{\partial h^{[t-k]}}{\partial z^{[t-k]}} \frac{\partial z_x^{[t-k]}}{\partial (\text{vec } W_x)}$$

Matrix formula:

$$\frac{\partial E^{[t]}}{\partial (\text{vec } W_x)} = (p^{[t]} - l^{[t]})^T W_y \sum_{k=0}^{t-1} \Big(\prod_{s=0}^{k-1} \Lambda^{[t-s]} W_h \Big) \, \Lambda^{[t-k]} ((x^{[t-k]})^T \otimes I) \tag{1}$$

Examples:

$$\begin{aligned}
\frac{\partial E^{[1]}}{\partial (\text{vec } W_x)} &= (p^{[1]} - l^{[1]})^T W_y \big(\Lambda^{[1]} ((x^{[1]})^T \otimes I) \big) \\
\frac{\partial E^{[2]}}{\partial (\text{vec } W_x)} &= (p^{[2]} - l^{[2]})^T W_y \big(\Lambda^{[2]} ((x^{[2]})^T \otimes I) + \Lambda^{[2]} W_h \Lambda^{[1]} ((x^{[1]})^T \otimes I) \big) \\
\frac{\partial E^{[3]}}{\partial (\text{vec } W_x)} &= (p^{[3]} - l^{[3]})^T W_y \big(\Lambda^{[3]} ((x^{[3]})^T \otimes I) + \Lambda^{[3]} W_h \Lambda^{[2]} ((x^{[2]})^T \otimes I) + \Lambda^{[3]} W_h \Lambda^{[2]} W_h \Lambda^{[1]} ((x^{[1]})^T \otimes I) \big)
\end{aligned}$$

Matrix from Vec:

Let $x \in \mathbb{R}^{n \times 1}$, $y \in \mathbb{R}^{m \times 1}$, $I \in \mathbb{R}^{m \times m}$

$$\begin{aligned}
y^T (x^T \otimes I) &= \begin{bmatrix} y_1 & y_2 & \dots & y_m \end{bmatrix} \begin{bmatrix} x_1 & 0 & \dots & 0 & \dots & x_N & 0 & \dots & 0 \\ 0 & x_1 & \dots & 0 & \dots & 0 & x_N & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x_1 & \dots & 0 & 0 & \dots & x_N \end{bmatrix} \\
&= \begin{bmatrix} x_1 y_1 & x_1 y_2 & \dots & x_1 y_m & x_2 y_1 & x_2 y_2 & \dots & x_2 y_m & \dots & x_n y_1 & x_n y_2 & \dots & x_n y_m \end{bmatrix} \in \mathbb{R}^{1 \times mn}
\end{aligned}$$

$$\text{mat}_{m \times n} (y^T (x^T \otimes I)) = \begin{bmatrix} x_1 y_1 & x_2 y_1 & \dots & x_n y_1 \\ x_1 y_2 & x_2 y_2 & \dots & x_n y_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1 y_m & x_2 y_m & \dots & x_n y_m \end{bmatrix} = y x^T \in \mathbb{R}^{m \times n}$$

Simplified formula:

$$\begin{aligned}
\frac{\partial E^{[t]}}{W_x} &= \textstyle\sum_{k=0}^{t-1}((p^{[t]}-l^{[t]})^T W_y \ (\prod_{s=0}^{k-1}\Lambda^{[t-s]}W_h)\ \Lambda^{[t-k]})^T (x^{[t-k]})^T \\
&= \textstyle\sum_{k=0}^{t-1}\Lambda^{[t-k]}\ (\prod_{s=0}^{k-1}\Lambda^{[t-s]}W_h)^T\ W_y^T(p^{[t]}-l^{[t]})(x^{[t-k]})^T \\
&= \textstyle\sum_{k=1}^{t}\Lambda^{[k]}\ (\prod_{s=k+1}^{t}W_h^T\Lambda^{[s]})\ W_y^T(p^{[t]}-l^{[t]})(x^{[k]})^T
\end{aligned}$$

$$\frac{\partial E^{[t]}}{\partial W_x} = \sum_{k=1}^{t}\Lambda^{[k]}\ (\prod_{s=k+1}^{t}W_h^T\Lambda^{[s]}\ )\ W_y^T(p^{[t]}-l^{[t]})(x^{[k]})^T$$

Examples:

$$\begin{aligned}
\tfrac{\partial E^{[1]}}{\partial W_x} &= \Lambda^{[1]}W_y^T(p^{[1]}-l^{[1]})(x^{[1]})^T \\
\tfrac{\partial E^{[2]}}{\partial W_x} &= \Lambda^{[1]}W_h^T\Lambda^{[2]}W_y^T(p^{[2]}-l^{[2]})(x^{[1]})^T + \Lambda^{[2]}W_y^T(p^{[2]}-l^{[2]})(x^{[2]})^T \\
\tfrac{\partial E^{[3]}}{\partial W_x} &= \Lambda^{[1]}W_h^T\Lambda^{[2]}W_h^T\Lambda^{[3]}W_y^T(p^{[3]}-l^{[3]})(x^{[1]})^T + \Lambda^{[2]}W_h^T\Lambda^{[3]}W_y^T(p^{[3]}-l^{[3]})(x^{[2]})^T + \Lambda^{[3]}W_y^T(p^{[3]}-l^{[3]})(x^{[3]})^T
\end{aligned}$$

Final formula:

$$DE(W_x) = \frac{\partial E}{\partial W_x} = \sum_{t=1}^{T}\sum_{k=1}^{t}\Lambda^{[k]}\ (\prod_{s=k+1}^{t}W_h^T\Lambda^{[s]}\ )\ W_y^T(p^{[t]}-l^{[t]})(x^{[k]})^T$$

Calculation algorithm:

1: **For t from T downto 1 do**
2: $\quad e = \Lambda^{[t]}(W_y^T(p^{[t]}-l^{[t]}) + W_h^T e)$
3: $\quad dW_x = dW_x + e(x^{[t]})^T$

4: $W_x = W_x - \gamma dW_x$

2.2 $\theta_j = b_x$

$$\frac{\partial E}{\partial b_x} = \sum_{t=1}^{T}\frac{\partial E^{[t]}}{\partial b_x} \in \mathbb{R}^{1\times H}$$

$$\frac{\partial E^{[t]}}{\partial b_x} = \frac{\partial E^{[t]}}{\partial p^{[t]}}\frac{\partial p^{[t]}}{\partial y^{[t]}}\frac{\partial y^{[t]}}{\partial h^{[t]}}\frac{\partial h^{[t]}}{\partial b_x} \in \mathbb{R}^{1\times H}$$

Recursive ratio:

$$\begin{aligned}
\tfrac{\partial h^{[t]}}{\partial b_x} &= \tfrac{\partial h^{[t]}}{\partial z^{[t]}}\tfrac{\partial z^{[t]}}{\partial b_x} = \tfrac{\partial h^{[t]}}{\partial z^{[t]}}\tfrac{\partial(z_x^{[t]}+z_h^{[t-1]})}{\partial b_x} \\
&= \tfrac{\partial h^{[t]}}{\partial z^{[t]}}(\tfrac{\partial z_x^{[t]}}{\partial b_x} + \tfrac{\partial z_h^{[t-1]}}{\partial b_x}) \\
&= \tfrac{\partial h^{[t]}}{\partial z^{[t]}}(\tfrac{\partial z_x^{[t]}}{\partial b_x} + \tfrac{\partial z_h^{[t-1]}}{\partial h^{[t-1]}}\tfrac{\partial h^{[t-1]}}{\partial b_x})
\end{aligned}$$

$$\frac{\partial h^{[t]}}{\partial b_x} = \frac{\partial h^{[t]}}{\partial z^{[t]}}(\frac{\partial z_x^{[t]}}{\partial b_x} + \frac{\partial z_h^{[t-1]}}{\partial h^{[t-1]}}\frac{\partial h^{[t-1]}}{\partial b_x})$$

Recursive formula:

$$\frac{\partial E^{[t]}}{\partial b_x} = \frac{\partial E^{[t]}}{\partial p^{[t]}}\frac{\partial p^{[t]}}{\partial y^{[t]}}\frac{\partial y^{[t]}}{\partial h^{[t]}}\frac{\partial h^{[t]}}{\partial z^{[t]}}(\frac{\partial z_x^{[t]}}{\partial b_x} + \frac{\partial z_h^{[t-1]}}{\partial h^{[t-1]}}\frac{\partial h^{[t-1]}}{\partial b_x}) \in \mathbb{R}^{1\times H}$$

Jacobians:

$$\frac{\partial z_x^{[t]}}{\partial b_x} = \frac{\partial(W_x x^{[t]} + b_x)}{\partial b_x} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = I \in \mathbb{R}^{H\times H}$$

Deployment:

$$
\begin{aligned}
\frac{\partial h^{[1]}}{\partial b_x} &= \frac{\partial h^{[1]}}{\partial z^{[1]}}\left(\frac{\partial z_x^{[1]}}{\partial b_x} + \overbrace{\frac{\partial z_h^{[0]}}{\partial h^{[0]}}\frac{\partial h^{[0]}}{\partial b_x}}^{\mathbf{0}\in\mathbb{R}^{H\times H}}\right) \\
&= \frac{\partial h^{[1]}}{\partial z^{[1]}}\frac{\partial z_x^{[1]}}{\partial b_x} \\
\frac{\partial h^{[2]}}{\partial b_x} &= \frac{\partial h^{[2]}}{\partial z^{[2]}}\frac{\partial z_x^{[2]}}{\partial b_x} + \frac{\partial h^{[2]}}{\partial z^{[2]}}\frac{\partial z_h^{[1]}}{\partial h^{[1]}}\frac{\partial h^{[1]}}{\partial z^{[1]}}\frac{\partial z_x^{[1]}}{\partial b_x} \\
\frac{\partial h^{[3]}}{\partial b_x} &= \frac{\partial h^{[3]}}{\partial z^{[3]}}\frac{\partial z_x^{[3]}}{\partial b_x} + \frac{\partial h^{[3]}}{\partial z^{[3]}}\frac{\partial z_h^{[2]}}{\partial h^{[2]}}\frac{\partial h^{[2]}}{\partial z^{[2]}}\frac{\partial z_x^{[2]}}{\partial b_x} + \frac{\partial h^{[3]}}{\partial z^{[3]}}\frac{\partial z_h^{[2]}}{\partial h^{[2]}}\frac{\partial h^{[2]}}{\partial z^{[2]}}\frac{\partial z_h^{[1]}}{\partial h^{[1]}}\frac{\partial h^{[1]}}{\partial z^{[1]}}\frac{\partial z_x^{[1]}}{\partial b_x} \\
&\vdots \\
\frac{\partial h^{[t]}}{\partial b_x} &= \frac{\partial h^{[t]}}{\partial z^{[t]}}\frac{\partial z_x^{[t]}}{\partial b_x} + \frac{\partial h^{[t]}}{\partial z^{[t]}}\frac{\partial z_h^{[t-1]}}{\partial h^{[t-1]}}\frac{\partial h^{[t-1]}}{\partial z^{[t-1]}}\frac{\partial z_x^{[t-1]}}{\partial b_x} + \cdots + \frac{\partial h^{[t]}}{\partial z^{[t]}}\frac{\partial z_h^{[t-1]}}{\partial h^{[t-1]}}\frac{\partial h^{[t-1]}}{\partial z^{[t-1]}}\frac{\partial z_h^{[t-2]}}{\partial h^{[t-2]}}\times\cdots\times\frac{\partial h^{[2]}}{\partial z^{[2]}}\frac{\partial z_h^{[1]}}{\partial h^{[1]}}\frac{\partial h^{[1]}}{\partial z^{[1]}}\frac{\partial z_x^{[1]}}{\partial b_x}
\end{aligned}
$$

Intermediate formula:

$$
\frac{\partial E^{[t]}}{\partial b_x} = \frac{\partial E^{[t]}}{\partial p^{[t]}}\frac{\partial p^{[t]}}{\partial y^{[t]}}\frac{\partial y^{[t]}}{\partial h^{[t]}}\sum_{k=0}^{t-1}\left(\prod_{s=0}^{k-1}\frac{\partial h^{[t-s]}}{\partial z^{[t-s]}}\frac{\partial z_h^{[t-s-1]}}{\partial h^{[t-s-1]}}\right)\frac{\partial h^{[t-k]}}{\partial z^{[t-k]}}\frac{\partial z_x^{[t-k]}}{\partial b_x}
$$

Matrix formula:

$$
\frac{\partial E^{[t]}}{\partial b_x} = (p^{[t]} - l^{[t]})^T W_y \sum_{k=0}^{t-1}\left(\prod_{s=0}^{k-1}\Lambda^{[t-s]}W_h\right)\Lambda^{[t-k]} \tag{2}
$$

Examples:

$$
\begin{aligned}
\frac{\partial E^{[1]}}{\partial b_x} &= (p^{[1]} - l^{[1]})^T W_y(\Lambda^{[1]}) \\
\frac{\partial E^{[2]}}{\partial b_x} &= (p^{[2]} - l^{[2]})^T W_y(\Lambda^{[2]} + \Lambda^{[2]}W_h\Lambda^{[1]}) \\
\frac{\partial E^{[3]}}{\partial b_x} &= (p^{[3]} - l^{[3]})^T W_y(\Lambda^{[3]} + \Lambda^{[3]}W_h\Lambda^{[2]} + \Lambda^{[3]}W_h\Lambda^{[2]}W_h\Lambda^{[1]})
\end{aligned}
$$

Simplified formula:

$$
\begin{aligned}
\left(\frac{\partial E^{[t]}}{\partial b_x}\right)^T &= \textstyle\sum_{k=0}^{t-1}((p^{[t]} - l^{[t]})^T W_y\ (\prod_{s=0}^{k-1}\Lambda^{[t-s]}W_h)\ \Lambda^{[t-k]})^T \\
&= \textstyle\sum_{k=0}^{t-1}\Lambda^{[t-k]}\ (\prod_{s=0}^{k-1}\Lambda^{[t-s]}W_h)^T\ W_y^T(p^{[t]} - l^{[t]}) \\
&= \textstyle\sum_{k=1}^{t}\Lambda^{[k]}\ (\prod_{s=k+1}^{t}W_h^T\Lambda^{[s]})\ W_y^T(p^{[t]} - l^{[t]})
\end{aligned}
$$

$$
\left(\frac{\partial E^{[t]}}{\partial b_x}\right)^T = \sum_{k=1}^{t}\Lambda^{[k]}\left(\prod_{s=k+1}^{t}W_h^T\Lambda^{[s]}\right)W_y^T(p^{[t]} - l^{[t]})
$$

Examples:

$$
\begin{aligned}
\left(\frac{\partial E^{[1]}}{\partial b_x}\right)^T &= \Lambda^{[1]}W_y^T(p^{[1]} - l^{[1]}) \\
\left(\frac{\partial E^{[2]}}{\partial b_x}\right)^T &= \Lambda^{[1]}W_h^T\Lambda^{[2]}W_y^T(p^{[2]} - l^{[2]}) + \Lambda^{[2]}W_y^T(p^{[2]} - l^{[2]}) \\
\left(\frac{\partial E^{[3]}}{\partial b_x}\right)^T &= \Lambda^{[1]}W_h^T\Lambda^{[2]}W_h^T\Lambda^{[3]}W_y^T(p^{[3]} - l^{[3]}) + \Lambda^{[2]}W_h^T\Lambda^{[3]}W_y^T(p^{[3]} - l^{[3]}) + \Lambda^{[3]}W_y^T(p^{[3]} - l^{[3]})
\end{aligned}
$$

Final formula:

$$
\nabla E(b_x) = \left(\frac{\partial E}{\partial b_x}\right)^T = \sum_{t=1}^{T}\sum_{k=1}^{t}\Lambda^{[k]}\left(\prod_{s=k+1}^{t}W_h^T\Lambda^{[s]}\right)W_y^T(p^{[t]} - l^{[t]})
$$

Calculation algorithm:

1: **For t from T downto 1 do**
2: $\quad e = \Lambda^{[t]}(W_y^T(p^{[t]} - l^{[t]}) + W_h^T e)$
3: $\quad db_x = db_x + e$

4: $b_x = b_x - \gamma db_x$

2.3 $\boldsymbol{\theta_j = W_h}$

$$\frac{\partial E}{\partial W_h} = \sum_{t=1}^{T} \frac{\partial E^{[t]}}{\partial W_h} \in \mathbb{R}^{H\times H}$$

$$\frac{\partial E^{[t]}}{\partial W_h} = \text{mat}_{H\times H}\left(\frac{\partial E^{[t]}}{\partial(\text{vec } W_h)}\right) \in \mathbb{R}^{H\times H}$$

$$\frac{\partial E^{[t]}}{\partial(\text{vec } W_h)} = \frac{\partial E^{[t]}}{\partial p^{[t]}} \frac{\partial p^{[t]}}{\partial y^{[t]}} \frac{\partial y^{[t]}}{\partial h^{[t]}} \frac{\partial h^{[t]}}{\partial(\text{vec } W_h)} \in \mathbb{R}^{1\times HH}$$

Recursive ratio:

$$\begin{aligned} \frac{\partial h^{[t]}}{\partial(\text{vec } W_h)} &= \frac{\partial h^{[t]}}{\partial z^{[t]}} \frac{\partial z^{[t]}}{\partial(\text{vec } W_h)} = \frac{\partial h^{[t]}}{\partial z^{[t]}} \frac{\partial(z_x^{[t]} + z_h^{[t-1]})}{\partial(\text{vec } W_h)} \\ &= \frac{\partial h^{[t]}}{\partial z^{[t]}} \left(\frac{\partial z_x^{[t]}}{\partial(\text{vec } W_h)} + {\color{red}\frac{\partial z_h^{[t-1]}}{\partial(\text{vec } W_h)}}\right) \end{aligned}$$

2.4 $\boldsymbol{\theta_j = b_h}$

2.5 $\boldsymbol{\theta_j = W_y}$

2.6 $\boldsymbol{\theta_j = b_y}$