

Derivation of (Deep) RNN

Lifu Huang

Mar. 2016

1 Loss Function

$$\begin{aligned} J &= \frac{1}{T} \sum_{t=1}^T J^{(t)} \\ &= \frac{1}{T} \sum_{t=1}^T \sum_{j=1}^{|V|} -y_j^{(t)} \log \hat{y}_j^{(t)} \end{aligned} \tag{1}$$

$$\begin{aligned} &= \frac{1}{T} \sum_{t=1}^T -\log \hat{y}_k^{(t)} \\ J^{(t)} &= -\log \hat{y}_k^{(t)} \end{aligned} \tag{2}$$

2 Forward Propagation

Assume that we have a $n + 1$ layer Recurrent Neural Network,

$$\hat{y}^{(t)} = g(z^{(n)(t)}) \tag{3}$$

$$h^{(l)(t)} = \begin{cases} f(z^{(l-1)(t)}) & 0 < l < n \\ x^{(t)} & l = 0 \end{cases} \tag{4}$$

$$z^{(l)(t)} = \begin{cases} W^{(l)} h^{(l-1)(t)} + b^{(l)} & l = n \\ H^{(l)} h^{(l)(t-1)} + W^{(l)} h^{(l-1)(t)} + b^{(l)} & 0 < l < n \end{cases} \tag{5}$$

3 Backward Propagation

For the purpose of simplicity, all derivation in this chapter will be made with respect to the loss function of a single time step. Moreover, we will presume that $g(x)$ here is the softmax function, which can be easily substituted with any other functions as needed.

3.1 Derivation of δ s

3.1.1 The Output Layer

$$\begin{aligned}\frac{\partial J^{(t)}}{\partial z^{(n)}(t)} &= \frac{\partial}{\partial z^{(n)}(t)} (-\log \hat{y}_k^{(t)}) \\ &= \frac{\partial}{\partial z^{(n)}(t)} (-z_k^{(n)}(t) + \log \sum_{j=1}^{|V|} z_j^{(n)}(t)) \\ &= (\hat{y}^{(t)} - y^{(t)})^T\end{aligned}\tag{6}$$

$$\begin{aligned}\delta^{(n)}(t) &= \nabla_{z^{(n)}(t)} J^{(t)} \\ &= \hat{y}^{(t)} - y^{(t)}\end{aligned}\tag{7}$$

3.1.2 Top Hidden Layer

$$\begin{aligned}\frac{\partial J^{(t)}}{\partial z^{(n-1)}(t)} &= \frac{\partial J^{(t)}}{\partial z^{(n)}(t)} \frac{\partial z^{(n)}(t)}{\partial z^{(n-1)}(t)} \\ &= (\delta^{(n)}(t))^T \frac{\partial}{\partial z^{(n-1)}(t)} (W^{(n)} f(z^{(n-1)}(t)) + b^{(n)}) \\ &= (\delta^{(n)}(t))^T W^{(n)} \text{diag}[f'(z^{(n-1)}(t))]\end{aligned}\tag{8}$$

$$\begin{aligned}\delta^{(n-1)}(t) &= \nabla_{z^{(n-1)}(t)} J^{(t)} \\ &= \text{diag}[f'(z^{(n-1)}(t))](W^{(n)})^T \delta^{(n)}(t)\end{aligned}\tag{9}$$

$$\begin{aligned}\frac{\partial J^{(t)}}{\partial z^{(n-1)}(t-c)} &= \frac{\partial J^{(t)}}{\partial z^{(n-1)}(t-c+1)} \frac{\partial z^{(n-1)}(t-c+1)}{\partial z^{(n-1)}(t-c)} \\ &= (\delta^{(n-1)}(t-c+1))^T \frac{\partial}{\partial z^{(n-1)}(t-c)} (H^{(n-1)} h^{(n-1)}(t-c) + W^{(n-1)} h^{(n-2)}(t-c+1) + b^{(n-1)}) \\ &= (\delta^{(n-1)}(t-c+1))^T \frac{\partial}{\partial z^{(n-1)}(t-c)} (H^{(n-1)} f(z^{(n-1)}(t-c))) \\ &= (\delta^{(n-1)}(t-c+1))^T H^{(n-1)} \text{diag}[f'(z^{(n-1)}(t-c))]\end{aligned}\tag{10}$$

$$\begin{aligned}\delta^{(n-1)}(t-c) &= \nabla_{z^{(n-1)}(t-c)} J^{(t)} \\ &= \text{diag}[f'(z^{(n-1)}(t-c))](H^{(n-1)})^T \delta^{(n-1)}(t-c+1)\end{aligned}\tag{11}$$

3.1.3 Other Hidden Layers

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial z^{(l)}(t)} &= \frac{\partial J^{(t)}}{\partial z^{(l+1)}(t)} \frac{\partial z^{(l+1)}(t)}{\partial z^{(l)}(t)} \\
&= (\delta^{(l+1)}(t))^T \frac{\partial}{\partial z^{(l)}(t)} (H^{(l+1)} h^{(l+1)}(t-1) + W^{(l+1)} h^{(l)}(t) + b^{(l+1)}) \\
&= (\delta^{(l+1)}(t))^T \frac{\partial}{\partial z^{(l)}(t)} (W^{(l+1)} f(z^{(l)}(t))) \\
&= (\delta^{(l+1)}(t))^T W^{(l+1)} \text{diag}[f'(z^{(l)}(t))]
\end{aligned} \tag{12}$$

$$\begin{aligned}
\delta^{(l)}(t) &= \nabla_{z^{(l)}(t)} J^{(t)} \\
&= \text{diag}[f'(z^{(l)}(t))](W^{(l+1)})^T \delta^{(l+1)}(t)
\end{aligned} \tag{13}$$

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial z^{(l)}(t-c)} &= \left(\frac{\partial J^{(t)}}{\partial z^{(l)}(t-c+1)} \frac{\partial z^{(l)}(t-c+1)}{\partial z^{(l)}(t-c)} \right) + \left(\frac{\partial J^{(t)}}{\partial z^{(l+1)}(t-c)} \frac{\partial z^{(l+1)}(t-c)}{\partial z^{(l)}(t-c)} \right) \\
&= \left((\delta^{(l)}(t-c+1))^T \frac{\partial}{\partial z^{(l)}(t-c)} (H^{(l)} h^{(l)}(t-c) + W^{(l)} h^{(l-1)}(t-c+1) + b^{(l)}) \right) + \\
&\quad \left((\delta^{(l+1)}(t-c))^T \frac{\partial}{\partial z^{(l)}(t-c)} (H^{(l+1)} h^{(l+1)}(t-c-1) + W^{(l+1)} h^{(l)}(t-c) + b^{(l+1)}) \right) \\
&= \left((\delta^{(l)}(t-c+1))^T \frac{\partial}{\partial z^{(l)}(t-c)} (H^{(l)} f(z^{(l)}(t-c))) \right) + \\
&\quad \left((\delta^{(l+1)}(t-c))^T \frac{\partial}{\partial z^{(l)}(t-c)} (W^{(l+1)} f(z^{(l)}(t-c))) \right) \\
&= (\delta^{(l)}(t-c+1))^T H^{(l)} \text{diag}[f'(z^{(l)}(t-c))] + (\delta^{(l+1)}(t-c))^T W^{(l+1)} \text{diag}[f'(z^{(l)}(t-c))] \\
&= ((\delta^{(l)}(t-c+1))^T H^{(l)} + (\delta^{(l+1)}(t-c))^T W^{(l+1)}) \text{diag}[f'(z^{(l)}(t-c))]
\end{aligned} \tag{14}$$

$$\begin{aligned}
\delta^{(l)}(t-c) &= \nabla_{z^{(l)}(t-c)} J^{(t)} \\
&= \text{diag}[f'(z^{(l)}(t-c))](H^{(l)})^T \delta^{(l)}(t-c+1) + W^{(l+1)} (\delta^{(l+1)}(t-c))^T
\end{aligned} \tag{15}$$

3.2 Gradients of Parameters

H – horizontally propagation matrix

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial H^{(l)}} &= \sum_{k=1}^t \frac{\partial J^{(t)}}{\partial H^{(l)}} \Big|_{(k)} \\
&= \sum_{k=1}^t \frac{\partial J^{(t)}}{\partial z^{(l)(k)}} \frac{\partial z^{(l)(k)}}{\partial H^{(l)}} \Big|_{(k)} \\
&= \sum_{k=1}^t (\delta^{(l)(k)})^T \left[\frac{\partial}{\partial H^{(l)}} (\epsilon_i)^T (H^{(l)} h^{(l)(k-1)} + W^{(l)} h^{(l-1)(k)} + b^{(l)}) \right] \quad (16) \\
&= \sum_{k=1}^t (\delta^{(l)(k)})^T \left[h^{(l)(k-1)} (\epsilon_i)^T \right] \\
&= \sum_{k=1}^t h^{(l)(k-1)} (\delta^{(l)(k)})^T
\end{aligned}$$

$$\nabla_{H^{(l)}} J^{(t)} = \sum_{k=1}^t \delta^{(l)(k)} (h^{(l)(k-1)})^T \quad (17)$$

W – vertically propagation matrix

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial W^{(n)}} &= \frac{\partial J^{(t)}}{\partial z^{(n)(t)}} \frac{\partial z^{(n)(t)}}{\partial W^{(n)}} \\
&= (\delta^{(n)(t)})^T \left[\frac{\partial z_i^{(n)(t)}}{\partial W^{(n)}} \right] \\
&= (\delta^{(n)(t)})^T \left[\frac{\partial}{\partial W^{(n)}} ((\epsilon_i)^T W^{(n)} h^{(n-1)(t)} + b_i^{(n)}) \right] \quad (18) \\
&= (\delta^{(n)(t)})^T \left[h^{(n-1)(t)} (\epsilon_i)^T \right] \\
&= h^{(n-1)(t)} (\delta^{(n)(t)})^T
\end{aligned}$$

$$\nabla_{W^{(n)}} J^{(t)} = \delta^{(n)(t)} (h^{(n-1)(t)})^T \quad (19)$$

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial W^{(l)}} &= \sum_{k=1}^t \frac{\partial J^{(t)}}{\partial W^{(l)}} \Big|_{(k)} \\
&= \sum_{k=1}^t \frac{\partial J^{(t)}}{\partial z^{(l)(k)}} \frac{\partial z^{(l)(k)}}{\partial W^{(l)}} \Big|_{(k)} \\
&= \sum_{k=1}^t (\delta^{(l)(k)})^T \left[\frac{\partial}{\partial W^{(l)}} (\epsilon_i)^T (H^{(l)} h^{(l)(k-1)} + W^{(l)} h^{(l-1)(k)} + b^{(l)}) \right] \quad (20) \\
&= \sum_{k=1}^t (\delta^{(l)(k)})^T \left[h^{(l-1)(k)} (\epsilon_i)^T \right] \\
&= \sum_{k=1}^t h^{(l-1)(k)} (\delta^{(l)(k)})^T
\end{aligned}$$

$$\nabla_{W^{(l)}} J^{(t)} = \sum_{k=1}^t \delta^{(l)(k)} (h^{(l-1)(k)})^T \quad (21)$$

b – bias term

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial b^{(n)}} &= \frac{\partial J^{(t)}}{\partial z^{(n)(t)}} \frac{\partial z^{(n)(t)}}{\partial b^{(n)}} \\
&= (\delta^{(n)(t)})^T I \\
&= (\delta^{(n)(t)})^T \quad (22)
\end{aligned}$$

$$\nabla_{b^{(n)}} J^{(t)} = \delta^{(n)(t)} \quad (23)$$

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial b^{(l)}} &= \sum_{k=1}^t \frac{\partial J^{(t)}}{\partial b^{(l)}} \Big|_{(k)} \\
&= \sum_{k=1}^t \frac{\partial J^{(t)}}{\partial z^{(l)(k)}} \frac{\partial z^{(l)(k)}}{\partial b^{(l)}} \Big|_{(k)} \\
&= \sum_{k=1}^t (\delta^{(l)(k)})^T I \\
&= \sum_{k=1}^t (\delta^{(l)(k)})^T \quad (24)
\end{aligned}$$

$$\nabla_{b^{(l)}} J^{(t)} = \sum_{k=1}^t \delta^{(l)(k)} \quad (25)$$

3.3 Conclusion

$$\delta^{(n)(t)} = \hat{y}^{(t)} - y^{(t)} \quad (26)$$

$$\delta^{(n-1)(t)} = \text{diag}[f'(z^{(n-1)(t)})](W^{(n)})^T \delta^{(n)(t)} \quad (27)$$

$$\delta^{(n-1)(t-c)} = \text{diag}[f'(z^{(n-1)(t-c)})](H^{(n-1)})^T \delta^{(n-1)(t-c+1)} \quad (28)$$

$$\delta^{(l)(t)} = \text{diag}[f'(z^{(l)(t)})](W^{(l+1)})^T \delta^{(l+1)(t)} \quad (29)$$

$$\delta^{(l)(t-c)} = \text{diag}[f'(z^{(l)(t-c)})]((H^{(l)})^T \delta^{(l)(t-c+1)} + (W^{(l+1)})^T \delta^{(l+1)(t-c)}) \quad (30)$$

$$\nabla_{H^{(l)}} J^{(t)} = \sum_{k=1}^t \delta^{(l)(t)} (h^{(l)(k-1)})^T \quad (31)$$

$$\nabla_{W^{(n)}} J^{(t)} = \delta^{(n)(t)} (h^{(n-1)(t)})^T \quad (32)$$

$$\nabla_{W^{(l)}} J^{(t)} = \sum_{k=1}^t \delta^{(l)(k)} (h^{(l-1)(k)})^T \quad (33)$$

$$\nabla_{b^{(n)}} J^{(t)} = \delta^{(n)(t)} \quad (34)$$

$$\nabla_{b^{(l)}} J^{(t)} = \sum_{k=1}^t \delta^{(l)(k)} \quad (35)$$

4 Trick for Implementation

Derivation above is made with respect to the loss function of a single time step, (i.e. $\nabla J^{(t)}$). When taking derivative of the final loss function ∇J with respect to $H^{(l)}$, $W^{(l)}$, and $b^{(l)}$ for $l < n$, instead of calculate $\nabla J^{(t)}$ for every time step and sum them up at last, which is correct but inefficient, we backprop only once by keeping record of $\gamma^{(l)(t)}$, which is the accumulating sum of $\delta^{(l)(t)(i)}$ over range t to T .¹

It is easy to find that, when $t = T - 1$:

$$\gamma^{(l)(t)} = \delta^{(l)(t)(t)} \quad (36)$$

when $1 \leq t \leq T - 1$:

$$\begin{aligned} \gamma^{(n-1)(t)} &= \sum_{i=t}^T \delta^{(n-1)(t)(i)} \\ &= \delta^{(n-1)(t)(t)} + \sum_{i=t+1}^T \delta^{(n-1)(t)(i)} \\ &= \delta^{(n-1)(t)(t)} + \sum_{i=t+1}^T \text{diag}[f'(z^{(n-1)(t)})](H^{(n-1)})^T \delta^{(n-1)(t+1)(i)} \quad (37) \\ &= \delta^{(n-1)(t)(t)} + \text{diag}[f'(z^{(n-1)(t)})](H^{(n-1)})^T \sum_{i=t+1}^T \delta^{(n-1)(t+1)(i)} \\ &= \delta^{(n-1)(t)(t)} + \text{diag}[f'(z^{(n-1)(t)})](H^{(n-1)})^T \gamma^{(n-1)(t+1)} \\ &= \delta^{(n-1)(t)(t)} + f'(z^{(n-1)(t)}) \circ ((H^{(n-1)})^T \gamma^{(n-1)(t+1)}) \end{aligned}$$

¹We will slightly change our notation here by adding one more superscript to δ so as to differentiate δ from different $J^{(t)}$, now $\delta^{(l)(t)(i)} = \nabla_{z^{(l)(t)}} J^{(i)}$.

when $1 \leq t \leq T-1$ and $1 \leq l \leq n-2$:

$$\begin{aligned}
\gamma^{(l)(t)} &= \sum_{i=t}^T \delta^{(l)(t)(i)} \\
&= \sum_{i=t}^T \text{diag}[f'(z^{(l)(t)})]((H^{(l)})^T \delta^{(l)(t+1)(i)} + (W^{(l+1)})^T \delta^{(l+1)(t)(i)}) \\
&= \delta^{(l)(t)(t)} + \sum_{i=t+1}^T \delta^{(l)(t)(i)} \\
&= \delta^{(l)(t)(t)} + \sum_{i=t+1}^T \text{diag}[f'(z^{(l)(t)})]((H^{(l)})^T \delta^{(l)(t+1)(i)} + (W^{(l+1)})^T \delta^{(l+1)(t)(i)}) \\
&= \delta^{(l)(t)(t)} + \text{diag}[f'(z^{(l)(t)})] \\
&\quad \left((H^{(l)})^T \left(\sum_{i=t+1}^T \delta^{(l)(t+1)(i)} \right) + (W^{(l+1)})^T \left(\sum_{i=t+1}^T \delta^{(l+1)(t)(i)} \right) \right) \\
&= \delta^{(l)(t)(t)} + f'(z^{(l)(t)}) \circ ((H^{(l)})^T \gamma^{(l)(t+1)} + (W^{(l+1)})^T (\gamma^{(l+1)(t)} - \delta^{(l+1)(t)(t)})) \\
&= f'(z^{(l)(t)}) \circ ((H^{(l)})^T \gamma^{(l)(t+1)} + (W^{(l+1)})^T \gamma^{(l+1)(t)})
\end{aligned} \tag{38}$$

We can now simplify our gradients by reorganizing terms and using γ s instead of δ s,

$$\begin{aligned}
\nabla_{H^{(l)}} J &= \frac{1}{T} \sum_{i=1}^T \sum_{j=1}^i \delta^{(l)(j)(i)} (h^{(j-1)})^T \\
&= \frac{1}{T} \sum_{i=1}^T \sum_{j=i}^T \delta^{(l)(i)(j)} (h^{(l)(i-1)})^T \\
&= \frac{1}{T} \sum_{i=1}^T \gamma^{(l)(i)} (h^{(l)(i-1)})^T
\end{aligned} \tag{39}$$

$$\begin{aligned}
\nabla_{W^{(l)}} J &= \frac{1}{T} \sum_{i=1}^T \sum_{j=1}^i \delta^{(l)(j)(i)} (h^{(l-1)(j)})^T \\
&= \frac{1}{T} \sum_{i=1}^T \sum_{j=i}^T \delta^{(l)(i)(j)} (h^{(l-1)(i)})^T \\
&= \frac{1}{T} \sum_{i=1}^T \gamma^{(l)(i)} (h^{(l-1)(i)})^T
\end{aligned} \tag{40}$$

$$\begin{aligned}
\nabla_{b^{(l)}} J &= \frac{1}{T} \sum_{i=1}^T \sum_{j=1}^i \delta^{(l)(j)(i)} \\
&= \frac{1}{T} \sum_{i=1}^T \sum_{j=i}^T \delta^{(l)(i)(j)} \\
&= \frac{1}{T} \sum_{i=1}^T \gamma^{(l)(i)}
\end{aligned} \tag{41}$$

Hmm, now we can put every thing together,

$$\delta^{(l)(t)(t)} = \begin{cases} \hat{y}^{(t)} - y^{(t)} & l = n \\ f'(z^{(l)(t)}) \circ ((W^{(l+1)})^T \delta^{(l+1)(t)(t)}) & l < n \end{cases} \tag{42}$$

$$\gamma^{(l)(t)} = \begin{cases} \delta^{(l)(t)(t)} & t = T \\ \delta^{(l)(t)(t)} + f'(z^{(l)(t)}) \circ ((H^{(l)})^T \gamma^{(l)(t+1)}) & l = n - 1 \\ f'(z^{(l)(t)}) \circ ((H^{(l)})^T \gamma^{(l)(t+1)} + (W^{(l+1)})^T \gamma^{(l+1)(t)}) & l < n - 1 \end{cases} \tag{43}$$

$$\nabla_{H^{(l)}} J = \frac{1}{T} \sum_{t=1}^T \gamma^{(l)(t)} (h^{(l)(t-1)})^T \tag{44}$$

$$\nabla_{W^{(l)}} J = \begin{cases} \frac{1}{T} \sum_{t=1}^T \delta^{(n)(t)(t)} (h^{(n-1)(t)})^T & l = n \\ \frac{1}{T} \sum_{t=1}^T \gamma^{(l)(t)} (h^{(l-1)(t)})^T & l < n \end{cases} \tag{45}$$

$$\nabla_{b^{(l)}} J = \begin{cases} \frac{1}{T} \sum_{t=1}^T \delta^{(n)(t)(t)} & l = n \\ \frac{1}{T} \sum_{t=1}^T \gamma^{(l)(t)} & l < n \end{cases} \tag{46}$$