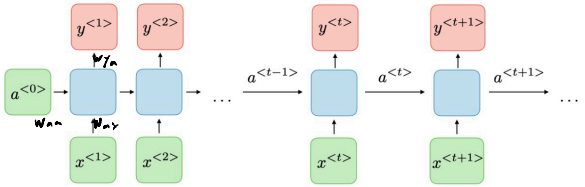
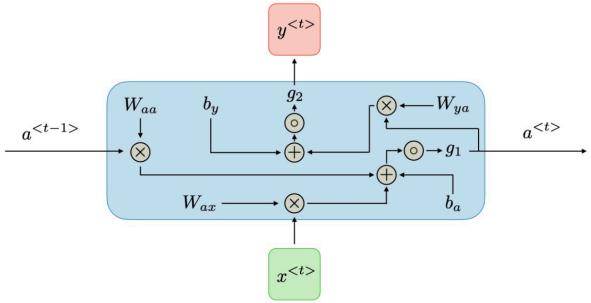
□ **Architecture of a traditional RNN** — Recurrent neural networks, also known as RNNs, are a class of neural networks that allow previous outputs to be used as inputs while having hidden states. They are typically as follows:



For each timestep t, the activation $a^{<t>}$ and the output $y^{<t>}$ are expressed as follows:

$$\boxed{a^{<t>} = g_1(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a)} \quad \text{and} \quad \boxed{y^{<t>} = g_2(W_{ya}a^{<t>} + b_y)}$$

where W_{ax}, W_{aa}, W_{ya}, b_a, b_y are coefficients that are shared temporally and g_1, g_2 activation functions.

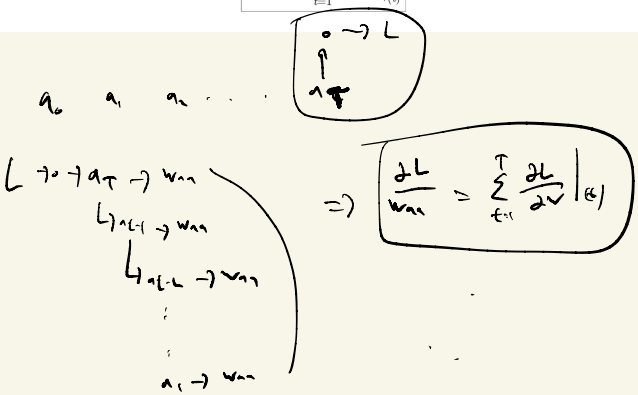


□ **Loss function** — In the case of a recurrent neural network, the loss function $\mathcal{L}$ of all time steps is defined based on the loss at every time step as follows:

$$\mathcal{L}(\hat{y}, y) = \sum_{t=1}^{T_y} \mathcal{L}(\hat{y}^{<t>}, y^{<t>})$$

□ **Backpropagation through time** — Backpropagation is done at each point in time. At timestep T, the derivative of the loss $\mathcal{L}$ with respect to weight matrix W is expressed as follows:

$$\frac{\partial \mathcal{L}^{(T)}}{\partial W} = \sum_{t=1}^{T} \left.\frac{\partial \mathcal{L}^{(T)}}{\partial W}\right|_{(t)}$$



$C \in (0,1,\ldots,K\text{-class})$, $\text{label } y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{pmatrix}$ (one-hot)

$L = -\sum_{k=1}^{K} y_i \log(\hat{y}_i) = -\log(\hat{y}_m)$, (label = m)

$$\boxed{\frac{\partial L}{\partial \hat{y}_i} = -\frac{y_i}{\hat{y}_i}}$$

logits $o \in R^K$

$\hat{y} = \text{softmax}(o)$

$S = \sum_k \exp(o_k)$

$\hat{y}_i = \frac{e^{o_i}}{S}$

$\left(\left(\frac{f}{g}\right)' = \frac{fg - fg'}{g^2}\right)$

$$\frac{\partial \hat{y}_i}{\partial o_i} = \frac{\partial}{\partial o_i}\left(\frac{e^{o_i}}{S}\right) = \frac{e^{o_i} S - e^{o_i} \cdot e^{o_i}}{S^2} = \frac{e^{o_i} S}{S^2} - \frac{e^{o_i} e^{o_i}}{S^2}$$

$$= \frac{e^{o_i}}{S} - \frac{e^{o_i} \cdot e^{o_i}}{S^2}$$

$$= \frac{e^{o_i}}{S}\left(1 - \frac{e^{o_i}}{S}\right) = \hat{y}_i(1 - \hat{y}_i)$$

$$\frac{\partial \hat{y}_i}{\partial o_j} = \frac{\partial}{\partial o_j}\left(\frac{e^{o_i}}{S}\right) = -e^{o_i}\frac{1}{S^2} \cdot \frac{\partial S}{\partial o_j} = -\frac{e^{o_i} \cdot e^{o_j}}{S^2} = -\hat{y}_i \cdot \hat{y}_j$$

Kronecker delta

$\delta_{ij} \in \{0, 1\}$

$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$ ∴

$$\boxed{\frac{\partial \hat{y}_i}{\partial o_j} = \hat{y}_i(\delta_{ij} - \hat{y}_j)}$$

$$f(x+\delta x) - f(x) = f'(x)\delta x = \delta f$$

$$\delta \hat{y}_i = \frac{\partial \hat{y}_i}{\partial o_j}\delta o_j$$

$$L = -\log(\hat{y}_m)$$

$$\delta L = \frac{\partial L}{\partial \hat{y}_m}\delta \hat{y}_m = -\frac{1}{\hat{y}_m}\delta \hat{y}_m$$

$$\hat{y}_m = \frac{e^{o_m}}{S}$$

$$\delta \hat{y}_m = \frac{\partial \hat{y}_m}{\partial o_j}\delta o_j$$

$$\frac{\partial \hat{y}_i}{\partial o_j} = \hat{y}_i(\delta_{ij} - \hat{y}_j)$$

$$\frac{\partial \hat{y}_m}{\partial o_j} = -\hat{y}_m \cdot \hat{y}_j$$

$$\delta \hat{y}_m = -\hat{y}_m \cdot \hat{y}_j \cdot \delta o_j$$

$$\delta L = \hat{y}_j \cdot \delta o_j$$

$$L(\hat{y}) = L(\hat{y}_1, \hat{y}_2, \dots \hat{y}_k)$$

$$\delta L = \sum_{i=1}^{k} \frac{\partial L}{\partial \hat{y}_i}\cdot \delta \hat{y}_i$$

$$= \sum_{i=1}^{k}\left(-\frac{y_i}{\hat{y}_i}\right)\cdot \delta \hat{y}_i$$

$$\boxed{\delta L = -\frac{1}{\hat{y}_m}\delta \hat{y}_m}$$

$$\delta L = \sum\left(\frac{\partial L}{\partial \hat{y}_i}\frac{\partial \hat{y}_i}{\partial o_j}\right)\delta o_j$$

$$\frac{\partial L}{\partial o_j} = \sum\left(\frac{\partial L}{\partial \hat{y}_i}\frac{\partial \hat{y}_i}{\partial o_j}\right)$$

$$\frac{\partial L}{\partial o_j} = \sum \frac{\partial L}{\partial \hat{y}_i}\frac{\partial \hat{y}_i}{\partial o_j}$$

$$L_t = -\sum_{i=1}^{K} y_{t,i} \log(\hat{y}_{t,i})$$

$$\frac{\partial L_t}{\partial \hat{y}_{t,i}} = -\frac{y_{t,i}}{\hat{y}_{t,i}}$$

$$\frac{\partial \hat{y}_{t,i}}{\partial o_{t,j}} = \hat{y}_{t,i}(\delta_{ij} - \hat{y}_{t,j})$$

$$\frac{\partial L_t}{\partial o_{t,j}} = \sum \frac{\partial L_t}{\partial \hat{y}_{t,i}} \frac{\partial \hat{y}_{t,i}}{\partial o_{t,j}}$$

$$= \sum_i -y_{t,i}(\delta_{ij} - \hat{y}_{t,j})$$

$$\sum_i -y_{t,i}\delta_{ij} + \sum_i y_{t,i}\hat{y}_{t,j}$$

$$= \hat{y}_{t,j} - y_{t,j}$$

$$\frac{\partial L_t}{\partial o_t} = \hat{y}_t - y_t = \delta_t^y$$

$$\frac{\partial L_t}{\partial (W_{hy})_{ij}} = \sum \frac{\partial L_t}{\partial o_{t,k}} \frac{\partial o_{t,k}}{\partial (W_{hy})_{ij}}$$

$$\frac{\partial o_{t,k}}{\partial (W_{hy})_{ij}} = \delta_{ki} (h_t)_j \quad , \quad \frac{\partial L_t}{\partial o_{t,k}} = (\delta_t^y)_k$$

$$\frac{\partial L_t}{\partial (W_{xh})_{ij}} = \sum_k \frac{\partial L_t}{\partial (s_t)_k} \delta_{ki} (h_{t-1})_j$$

$$= (h_{t-1})_j \sum_k \frac{\partial L_t}{\partial (s_t)_k} \delta_{ki}$$

$$= (h_{t-1})_j \frac{\partial L_t}{\partial (s_t)_i} = (h_{t-1})_j (\delta_t^2)_i$$

$$\frac{\partial L_t}{\partial W_{xh}} = \delta_t^2 \, h_{t-1}^T$$

$$\boxed{\frac{\partial L}{\partial W_{xh}} = \sum_t \delta_t^2 \, h_{t-1}^T}$$

tanh(x) $\frac{e^x - e^{-x}}{e^x + e^{-x}}$

$\frac{d}{dx}$ tanh(x) $1 - \tanh^2(x)$

$$u_t = W_{aa}h_{t-1} + W_{ax}x_t + b_a, h_t = \tanh(u_t), o_t = W_{ya}h_t + b_y, \hat{y}_t = \text{softmax}(o_t), L_t = -\sum_{i=1}^{K} y_{t,i}\log(\hat{y}_{t,i}), \mathcal{L} = \sum_{t=1}^{T} L_t$$

$$\delta_t^y \equiv \frac{\partial L_t}{\partial o_t} = \hat{y}_t - y_t, \delta_{T+1}^u \equiv 0, \delta_t^h \equiv \frac{\partial \mathcal{L}}{\partial h_t} = W_{ya}^\top \delta_t^y + W_{aa}^\top \delta_{t+1}^u, \delta_t^u \equiv \frac{\partial \mathcal{L}}{\partial u_t} = (1 - h_t^2) \odot \delta_t^h.$$

$$\frac{\partial \mathcal{L}}{\partial W_{ya}} = \sum_{t=1}^{T} \delta_t^y h_t^\top, \frac{\partial \mathcal{L}}{\partial b_y} = \sum_{t=1}^{T} \delta_t^y, \frac{\partial \mathcal{L}}{\partial W_{aa}} = \sum_{t=1}^{T} \delta_t^u h_{t-1}^\top, \frac{\partial \mathcal{L}}{\partial W_{ax}} = \sum_{t=1}^{T} \delta_t^u x_t^\top, \frac{\partial \mathcal{L}}{\partial b_a} = \sum_{t=1}^{T} \delta_t^u.$$

$$W_{ya} \leftarrow W_{ya} - \eta \frac{\partial \mathcal{L}}{\partial W_{ya}}, b_y \leftarrow b_y - \eta \frac{\partial \mathcal{L}}{\partial b_y}, W_{aa} \leftarrow W_{aa} - \eta \frac{\partial \mathcal{L}}{\partial W_{aa}}, W_{ax} \leftarrow W_{ax} - \eta \frac{\partial \mathcal{L}}{\partial W_{ax}}, b_a \leftarrow b_a - \eta \frac{\partial \mathcal{L}}{\partial b_a}.$$

(A, B is matrix)

$$\|AB\| \le \|A\|\|B\|$$

$T = 3$

$$\delta_1^{(n)} = \cdots + \left(\theta_1 W_{n}^{\top} D_2 W_{n}^{\top} D_3\right)\left(W_{in}^{\top} \delta_3^{(2)}\right)$$

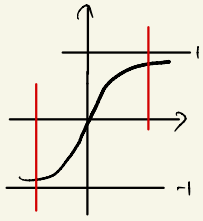
$J_j = W_{n}^{\top} D_j$ (Jacobian)

$$\theta_1 W_{n}^{\top} D_2 W_{n}^{\top} D_3 = \theta_1 J_2 J_3$$

$$\|J_2 J_3\| \le \|J_2\|\|J_3\|$$

$D_j = \mathrm{diag}(1 - h_j^2)$

$0 \le 1 - h_{ji}^2 \le 1$, $h = \tanh(u)$



$(h \simeq \pm 1)$

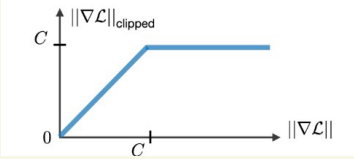
$(1 - h_{ji}^2) \simeq 0$

$\left\| \prod_{t=2}^{T} J_t \right\| \simeq 0 \Rightarrow$ gradient vanishing

explode

- Gradient clipping

$$g_{clipped} = g \cdot \min\left(1, \frac{c}{\|g\|}\right)$$



$$\frac{\partial h_{T-1}}{\partial x_1} \simeq 0$$

$$\frac{\partial L_T}{\partial x_1} = \frac{\partial L_T}{\partial h_{T-1}} \cdot \frac{\partial h_{T-1}}{\partial x_1} \simeq 0$$

Context error.

$$\boxed{\frac{\partial L}{\partial W_{aa}} \simeq \sum_{k=1}^{T} \sum_{t=k}^{T} \frac{\partial L_T}{\partial W_{aa}^{(k)}}}$$

$$\frac{\partial L_T}{\partial W_{aa}^{(k)}} = \delta_k^{u_1(T)} \cdot h_{k-1}^T$$

$$\delta_k^{u_1(T)} \simeq 0 \Rightarrow \frac{\partial L_T}{\partial W_{aa}^{(k)}} \simeq 0$$

If the gradient from L_T doesn't reach the early W_{aa}, then the early parameters can't be trained to reduce L_T. As a result, the model doesn't learn to maintain (store) information over time, and it fails to handle long-term dependencies.