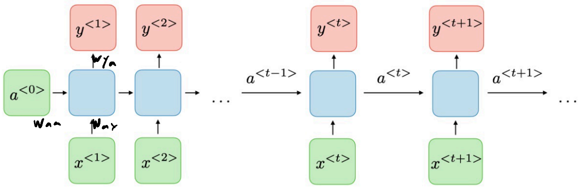
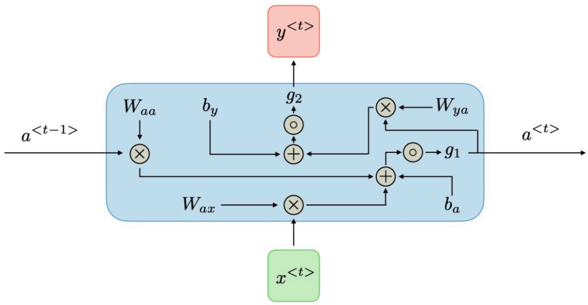
□ **Architecture of a traditional RNN** — Recurrent neural networks, also known as RNNs, are a class of neural networks that allow previous outputs to be used as inputs while having hidden states. They are typically as follows:



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:

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

□ **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:

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

$o \rightarrow L$
$\uparrow$
a_T

$a_0 \quad a_1 \quad a_2 \; . \; . \; .$

$L \rightarrow o \rightarrow a_T \rightarrow W_{aa}$
$\hookrightarrow a_{t-1} \rightarrow W_{aa}$
$\hookrightarrow a_{t-2} \rightarrow W_{aa}$
$\vdots$
$a_1 \rightarrow W_{aa}$

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

CE loss, (K class) , label y = $\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{pmatrix}$ (one-hot)

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

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

logit $o \in R^K$

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

$$S = \sum_k \exp(o_k)$$

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

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

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

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

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

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

Kronecker delta

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

$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$ $\therefore$ $$\boxed{\frac{\partial \hat{y}_i}{\partial o_j} = \hat{y}_i(\delta_{ij} - \hat{y}_j)}$$

$$f(x+\Delta x) - f(x) = f'(x)\,\Delta x = \Delta f$$

$$\Delta \hat{y}_i = \frac{\partial \hat{y}_i}{\partial o_j}\,\Delta o_j$$

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

$$\Delta L \approx \frac{\partial L}{\partial \hat{y}_n}\,\Delta \hat{y}_n = -\frac{1}{\hat{y}_n}\,\Delta \hat{y}_n$$

$$\hat{y}_n = \frac{e^{o_n}}{S}$$

$$\Delta \hat{y}_n \approx \frac{\partial \hat{y}_n}{\partial o_j}\,\Delta o_j$$

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

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

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

$$\Delta L \simeq \hat{y}_j \cdot \Delta o_j$$

$$L(\hat{y}) = L(\hat{y}_1, \hat{y}_2, \cdots \hat{y}_K)$$

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

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

$$\boxed{\Delta L \simeq -\frac{1}{\hat{y}_m}\,\Delta \hat{y}_n}$$

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

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

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

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

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

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

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

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

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

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

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

$$\frac{\partial L_t}{\partial (W_{yh})_{ij}} = \sum \frac{\partial L_t}{\partial (o_t)_k} \frac{\partial (o_t)_k}{\partial (W_{yh})_{ij}}$$

$$\frac{\partial (o_t)_k}{\partial (W_{yh})_{ij}} = \delta_{ki}(h_t)_j \quad , \quad \frac{\partial L_t}{\partial (o_t)_k} = \left(\delta_t^y\right)_k$$

$$\frac{\partial L_t}{\partial (W_{ya})_{ij}} = \sum_k \frac{\partial L_t}{\partial (o_t)_k} \delta_{ki} (h_t)_j$$

$$= (h_t)_j \sum_k \frac{\partial L_t}{\partial (o_t)_k} \delta_{ki}$$

$$= (h_t)_j \frac{\partial L_t}{\partial (o_t)_i} = (h_t)_j \left(\delta_t^2\right)_i$$

$$\frac{\partial L_t}{\partial W_{ya}} = \delta_t^2 \, h_t^T$$

$$\boxed{\frac{\partial L}{\partial W_{ya}} = \sum_t \delta_t^2 \, h_t^T}$$

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

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

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

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

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

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

$(A, B$ is matrix$)$

$$\|AB\| \leq \|A\|\|B\|$$

$T = 3$

$$\delta_1^h = \cdots + \cdots + (D_1 W_{hh}^T D_2 W_{hh}^T D_3)(W_{yh}^T \delta_3^y)$$

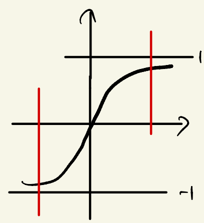
$J_j = W_{hh}^T D_j$ (Jacobian)

$$D_1 W_{hh}^T D_2 W_{hh}^T D_3 = D_1 J_2 J_3$$

$$\|J_2 J_3\| \leq \|J_2\|\|J_3\|$$

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

$0 \leq 1 - h_{j,i}^2 \leq 1$, $h = \tanh(u)$,



$(h \simeq \pm 1)$

$$(1 - h_{j,i}^2) \simeq 0$$

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

$\boxed{\text{exploding}}$

- Gradient clipping

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