RNN Training with Back Propagation Through Time [BBTT]

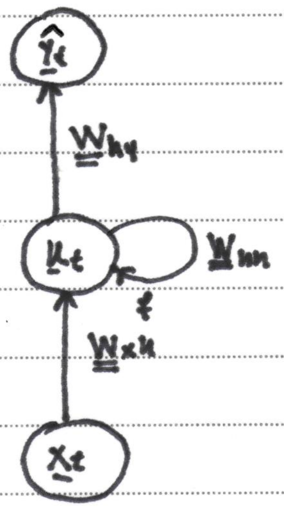
Gradient Descend: (α): Take the derivative of the loss (gradient) with respect to each parameter

(β): Shift parameters to the opposite direction in order to minimize loss.

▸ let $\mathcal{X} = \{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_\tau\}$ a sequence of ℓ-dimensional input vectors $\underline{x}_t \in \mathbb{R}^\ell$ for $1 \le t \le \tau$.

▸ let $\mathcal{Y} = \{\underline{y}_1, y_2, \dots, \underline{y}_\tau\}$ a sequence of m-dimensional output vectors $\underline{y}_t \in \mathbb{R}^m$ for $1 \le t \le \tau$.

▸ The estimated output of the RNN cell at each time step may be computed as:



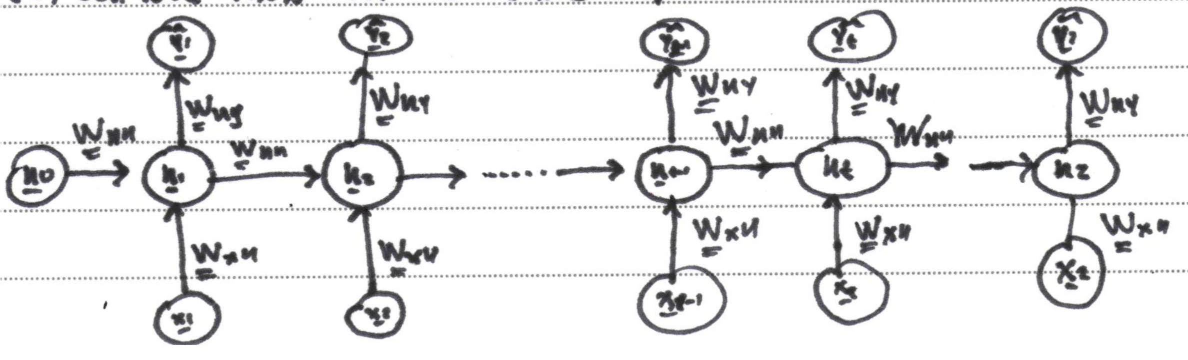
(A): Update Hidden State:

$$\boxed{\underline{h}_t = f(\underline{W}_{xh} \cdot \underline{x}_t + \underline{W}_{hh} \cdot \underline{h}_{t-1}) \quad 1 \le t \le \tau \quad (1)}$$

(B): Update Output Vector:

$$\boxed{\hat{\underline{y}}_t = \underline{W}_{hy} \cdot \underline{h}_t \quad 1 \le t \le \tau \quad (2)}$$

(*) Unfold RNN for τ time steps:

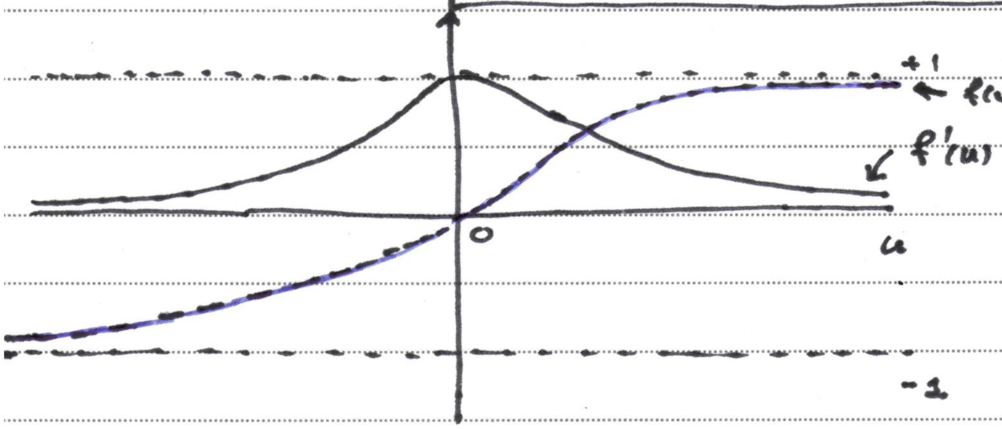


$\underline{h}_0$: Some Random Initialization

▸ The activation function utilized to update the hidden state is given by:

$$\boxed{f(u) = \tanh(u) = \frac{e^{u} - e^{-u}}{e^{u} + e^{-u}} \quad (3)}$$

SIGMOID FORM



(α): $f(u) \in [-1, +1]$

(β): $f'(u) \in [0, +1]$

▸ A very helpfull technique involves expressing the derivative of a given activation function as a function of itself.

▸ Try expressing $f'(x)$ as a function of $f(x)$:

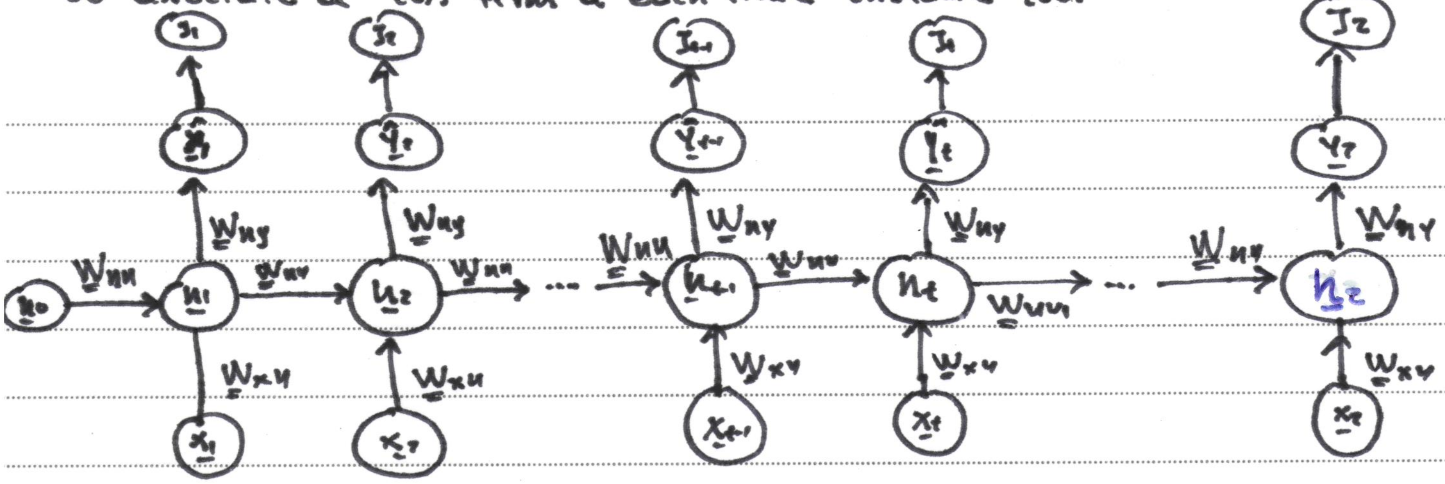
(i): Let $p(x) = e^{x} - e^{-x}$ and $q(x) = e^{x} + e^{-x}$ so that (iii): $f(x) = \frac{p(x)}{q(x)}$

▸ We may write that: (iv): $\begin{cases} p'(x) = e^{x} + e^{-x} \\ p'(x) = q(x) \end{cases}$ and $\begin{cases} q'(x) = e^{x} - e^{-x} \\ q'(x) = p(x) \end{cases}$

▸ Thus, we may write that: $f'(x) = \frac{p'(x)q(x) - p(x)q'(x)}{q^2(x)} = \frac{q^2(x) - p^2(x)}{q^2(x)} \Rightarrow$

$$f'(x) = 1 - \frac{p^2(x)}{q^2(x)} = 1 - \left(\frac{p(x)}{q(x)}\right)^2 = 1 - f^2(x). \quad (4)$$

(*) Since we are making a prediction at each time step we have to associate a loss term a each time instance too:



(2) Updating Equations: $(1 \le t \le T)$:

$$\begin{cases} \underline{h}_t = f(\underline{\underline{W}}_{xh} \cdot \underline{x}_t + \underline{\underline{W}}_{hh} \cdot \underline{h}_{t-1}) & (5) \\ \hat{\underline{y}}_t = \underline{\underline{W}}_{hy} \cdot \underline{h}_t & (6) \end{cases}$$

(3) Form the expression of the total loss for the RNN throughout the complete sequence $\underline{x}$:

$$\boxed{J(\theta) = \sum_t J_t(\theta) \quad (7)}$$

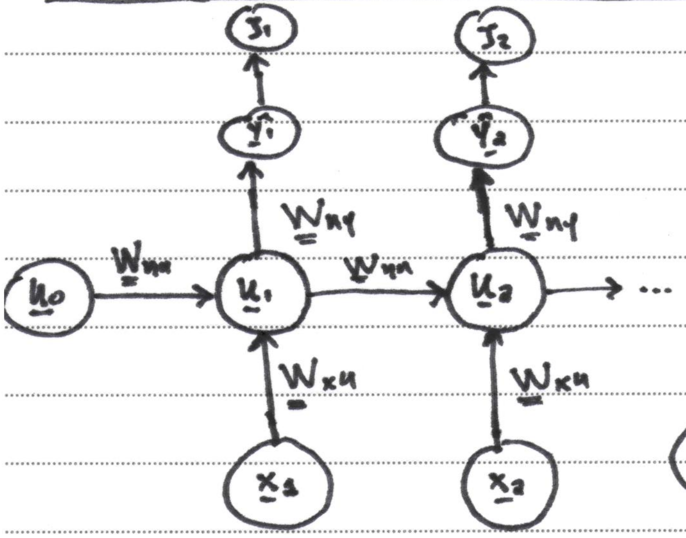
(4) Compute the total gradient by summing the partial derivative with respect to each parameter θ of the RNN:

$$\boxed{\frac{\partial J}{\partial \theta} = \sum_t \frac{\partial J_t(\theta)}{\partial \theta} \quad (8)}$$

- Let's initially focus on the weight-matrix $\underline{\underline{W}}_{xh}$ which is used to update the internal state of the cell according to the current input signal at each time-step.

$$\boxed{\frac{\partial J}{\partial \underline{\underline{W}}_{xh}} = \sum_t \frac{\partial J_t}{\partial \underline{\underline{W}}_{xh}} \quad (9)}$$

- <u>We may focus on time-step $t=2$:</u>



(*) CHAIN RULE FOR DERIVATIVES:

$$\boxed{\frac{\partial J_2}{\partial \underline{\underline{W}}_{xh}} = \frac{\partial J_2}{\partial \hat{\underline{y}}_2} \cdot \frac{\partial \hat{\underline{y}}_2}{\partial \underline{h}_2} \cdot \frac{\partial \underline{h}_2}{\partial \underline{\underline{W}}_{xh}}} \quad (10)$$

(**)

(**): But, $\underline{h}_2$ may be written as:

$$\underline{h}_2 = f(\underline{\underline{W}}_{xh} \cdot \underline{x}_2 + \underline{\underline{W}}_{hh} \cdot \underline{h}_1) \quad (11)$$

where $\underline{h}_1$ also depends on $\underline{\underline{W}}_{xh}$. Therefore, the term $\partial \underline{h}_2 / \partial \underline{\underline{W}}_{xh}$ cannot be treated as a constant.

- We need to know how exactly the cell-state at $t=2$ depends on the weight-matrix $\underline{\underline{W}}_{xh}$.

- We need to expand the term $\frac{\partial \underline{h}_2}{\partial \underline{\underline{W}}_{xh}}$ by considering/substituting the corresponding expression for the total derivative w.r.t. $\underline{\underline{W}}_{xh}$, $\frac{d \underline{h}_2}{d \underline{\underline{W}}_{xh}}$.

⊙ Equation (5) suggests that: $\boxed{\underline{h}_2 = \underline{h}_2(\underline{h}_1(\underline{\underline{W}}_{xh}), \underline{h}_0(\underline{\underline{W}}_{xh}); \underline{\underline{W}}_{xh}) \quad (12)}$

Thus, the total derivative of $\underline{h}_2$ w.r.t $\underline{\underline{W}}_{xh}$ may be written according to the chain rule as:

$$\boxed{\frac{d\underline{h}_2}{d\underline{\underline{W}}_{xh}} = \frac{\partial \underline{h}_2}{\partial \underline{\underline{W}}_{xh}} + \frac{\partial \underline{h}_2}{\partial \underline{h}_1} \cdot \frac{\partial \underline{h}_1}{\partial \underline{\underline{W}}_{xh}} + \frac{\partial \underline{h}_2}{\partial \underline{h}_0} \cdot \frac{\partial \underline{h}_0}{\partial \underline{\underline{W}}_{xh}} \quad (13)}$$

⊙ Eq.(13) suggest the following compact form:

$$\boxed{\frac{d\underline{h}_2}{d\underline{\underline{W}}_{xh}} = \sum_{k=0}^{2} \frac{\partial \underline{h}_2}{\partial \underline{h}_k} \cdot \frac{\partial \underline{h}_k}{\partial \underline{\underline{W}}_{xh}} \quad (14)}$$

(*) Apparently, the summation term for $k=2$ yields:

$$\boxed{\frac{\partial \underline{h}_2}{\partial \underline{h}_2} \cdot \frac{\partial \underline{h}_2}{\partial \underline{\underline{W}}_{xh}} = \frac{\partial \underline{h}_2}{\partial \underline{\underline{W}}_{xh}} \quad (a)}$$

⊙ Finally, Eq.(10) may be written as:

$$\boxed{\frac{\partial J_2}{\partial \underline{\underline{W}}_{xh}} = \sum_{k=0}^{2} \frac{\partial J_2}{\partial \hat{\underline{Y}}_2} \cdot \frac{\partial \hat{\underline{Y}}_2}{\partial \underline{h}_2} \cdot \underbrace{\frac{\partial \underline{h}_2}{\partial \underline{h}_k} \cdot \frac{\partial \underline{h}_k}{\partial \underline{\underline{W}}_{xh}}} \quad (15)}$$

↓ Contributions of $\underline{\underline{W}}_{xh}$ in previous time-steps to the error at time-step $t=2$.

▶ In this setting, the general form of Eq.(15) may be written as:

$$\boxed{\frac{\partial J_t}{\partial \underline{\underline{W}}_{xh}} = \sum_{k=0}^{t} \frac{\partial J_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial \underline{h}_t} \cdot \underbrace{\frac{\partial \underline{h}_t}{\partial \underline{h}_k} \cdot \frac{\partial \underline{h}_k}{\partial \underline{\underline{W}}_{xh}}} \quad (16)}$$

Contributions of $\underline{\underline{W}}_{xh}$ in previous timesteps to the error at timestamp t.

⊛ What about $\frac{\partial J}{\partial \underline{\underline{W}}_{hh}}$ and $\frac{\partial J}{\partial \underline{\underline{W}}_{hy}}$?

(*) Keep in mind that:

(a): $\underline{h}_t$ does depend on $\underline{\underline{W}}_{hh}$

(b): $\underline{h}_t$ does not depend on $\underline{\underline{W}}_{hy}$.

(#7)

⊛ Gradient of the loss function at time t:

$$\boxed{\frac{\partial J_t}{\partial \underline{\underline{W}}_{xh}} = \sum_{k=0}^{t} \frac{\partial J_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial \underline{h}_t} \cdot \boxed{\frac{\partial \underline{h}_t}{\partial \underline{h}_k}} \cdot \frac{\partial \underline{h}_k}{\partial \underline{\underline{W}}_{xh}} \quad [16]}$$

requires the computation of the term:

$$\boxed{\frac{\partial \underline{h}_t}{\partial \underline{h}_k} \quad [17].}$$

⊛ For example at $t=2$ and for $k=0$, we would have to compute the contribution of the term $\frac{\partial \underline{h}_2}{\partial \underline{h}_0}$ which according to the interdependence of the previous hidden states of the RNN can be written as:

$$\boxed{\frac{\partial \underline{h}_2}{\partial \underline{h}_0} = \frac{\partial \underline{h}_2}{\partial \underline{h}_1} \cdot \frac{\partial \underline{h}_1}{\partial \underline{h}_0} \quad [18]}$$

⊛ But, if we are focusing on a timestep very far away in the future?

$$\boxed{\frac{\partial \underline{h}_t}{\partial \underline{h}_{\underset{k}{0}}} = \overset{m=t}{\frac{\partial \underline{h}_t}{\partial \underline{h}_{\underset{t-1}{t-1}}}} \cdot \frac{\partial \underline{h}_{t-1}}{\partial \underline{h}_{t-2}} \cdot \ldots \cdot \frac{\partial \underline{h}_2}{\partial \underline{h}_1} \cdot \overset{m=k+1}{\frac{\partial \underline{h}_1}{\partial \underline{h}_{\cancel{0}\,k}}} \quad [19]}$$

⊛ The general form of Eq. (19) can be written as:

$$\boxed{\frac{\partial \underline{h}_t}{\partial \underline{h}_k} = \prod_{m=k+1}^{m=t} \frac{\partial \underline{h}_m}{\partial \underline{h}_{m-1}} \quad [20]}$$

which reveals the necessity to compute the term:

$$\boxed{\partial \underline{h}_t / \partial \underline{h}_{t-1} \quad [21]}$$

* For the case of a trivial (vanilla) RNN with a single neuron, we may assume that:

(1): $\underline{x}_t \in \mathbb{R}^{\ell \times 1}$, $1 \leq t \leq \tau$

(2): $y_t \in \mathbb{R}$, $1 \leq t \leq \tau$

(3): $\underline{h}_t \in \mathbb{R}$, $1 \leq t \leq \tau$ $(\underline{h}_t \equiv h_t)$

(4): $\underline{\underline{W}}_{xh} \in \mathbb{R}^{\ell \times 1}$, $1 \leq t \leq \tau$ $(\underline{\underline{W}}_{xh} \equiv \underline{W}_{xh})$

(5): $\underline{\underline{W}}_{hh} \in \mathbb{R}$, $1 \leq t \leq \tau$ $(\underline{\underline{W}}_{hh} \equiv W_{hh})$

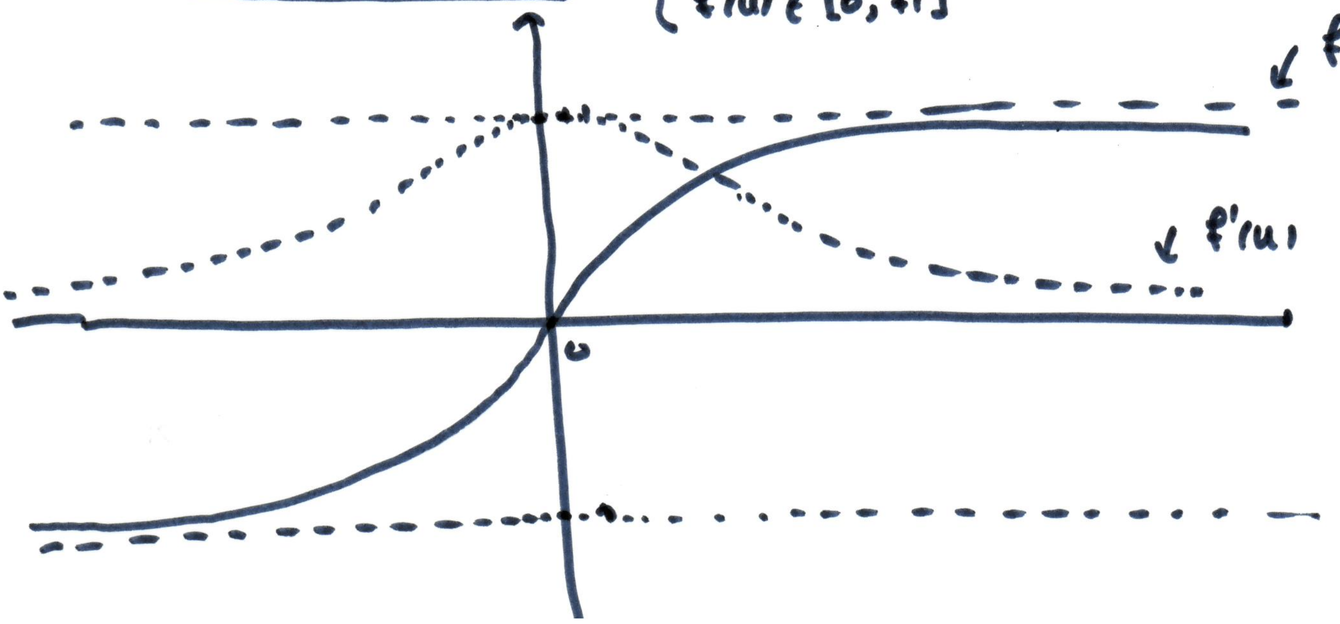
(6): $\underline{\underline{W}}_{hy} \in \mathbb{R}$, $1 \leq t \leq \tau$ $(\underline{\underline{W}}_{hy} \equiv W_{hy})$

* According to the previous definitions, we may write for the forward pass of the information within the network that:

$$\boxed{\begin{cases} h_t = f(\underline{W}_{xh}^T \cdot \underline{x}_t + W_{hh} \cdot h_{t-1}) & [27] \\ y_t = W_{hy} \cdot h_t & [28] \end{cases}}$$

* At this point mention (Page #2) of previous notes in order to prove that for: $\boxed{f(u) = \tanh(u) = \dfrac{e^u - e^{-u}}{e^u + e^{-u}}}$ we have

that $\boxed{f'(u) = 1 - f^2(u)}$. $\begin{cases} f(u) \in [-1, +1] \\ f'(u) \in [0, +1] \end{cases}$



(*) In light of the previous declarations we may write that:

[i]: Let $u \in \mathbb{R}$ be the expression that forms the input argument for the transfer function $f(\cdot)$ as:

$$\boxed{u = \underbrace{\underset{[1\times p]}{\underline{W}^T_{xh}} \cdot \underset{[p\times 1]}{\underline{x}_t}} + \underset{[1\times 1]}{W_{hh}} \cdot \underset{[1\times 1]}{h_{t-1}} \qquad [1\times 1] + [1\times 1] = [1]}$$

[ii]: $$\boxed{\frac{\partial h_t}{\partial h_{t-1}} = \frac{\partial f}{\partial u} \cdot \frac{\partial u}{\partial h_{t-1}} \quad [24]}$$

[iii]: $$\boxed{\frac{\partial u}{\partial h_{t-1}} = W_{hh} \quad [25]}$$

(*) Combining Eqs. (24) and (25) yields:

$$\boxed{\frac{\partial h_t}{\partial h_{t-1}} = f'(u) \cdot W_{hh} \quad [26] \text{ which gives:}}$$

$$\frac{\partial h_t}{\partial h_{t-1}} = \left[1 - f^2(\underline{W}^T_{xh} \cdot \underline{x}_t + W_{hh} \cdot h_{t-1})\right] \cdot W_{hh} \quad [27]$$

(*) Eq. (26) suggest that:

(i): Since, $f(u) = \tanh(u) \Rightarrow f'(u) \in [0,1]$

(ii): W_{hh} are sampled from the standardized normal distribution (for the 1-d case, $\mu = 0$, $\sigma^2 = 1$) so that $W_{hh} < 1$.

(*) We are multiplying a lot small numbers together.

(I) Errors due to further back timesteps have increasingly smaller gradients.

(II) Weight-parameters become biased towards capturing shorter-term dependencies.