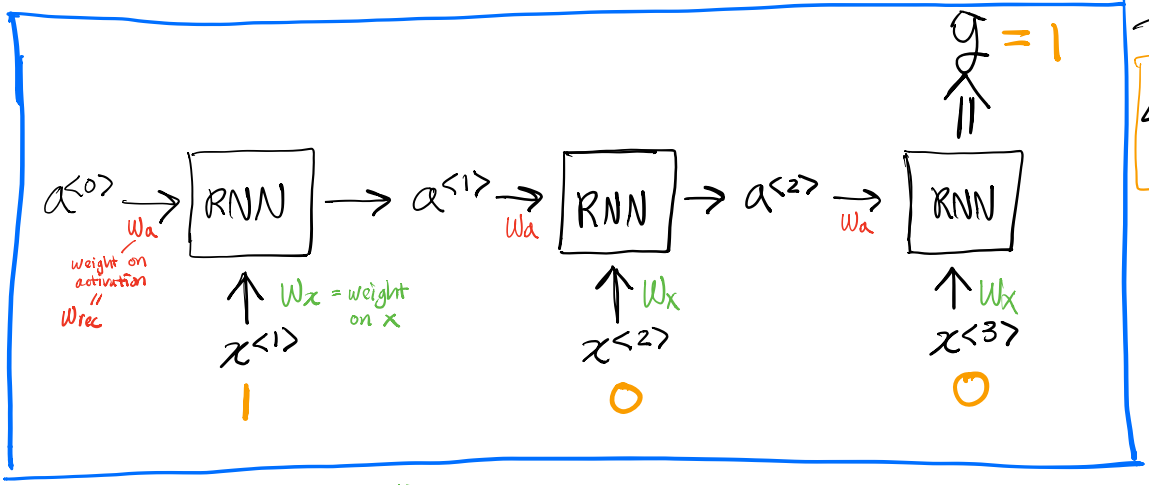
# < RNN by Hand >

Data = ex1 [1 0 0] → [1] (y), ex2 [0 1 1] → [2], ex3 [1 1 1] → [3]

3 examples, length of 3

target (sum of input)

**Goal:** have RNN sum the input sequence as the output.




→ cost (loss) function

$$L = \frac{\sum^{n} (y^{[i]} - \hat{y}^{[i]})^2}{n}$$

↳ Mean Squared Error.

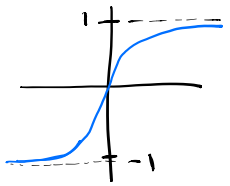
[i] denotes ith example.

n = # of examples

others refer it as $S^{<t>}$ for state

$$a^{<t>} = \tanh(W_a a^{<t-1>} + W_x x^{<t>} + b_a)$$

※ $\tanh(x) = \frac{\sinh(x)}{\cosh(x)}$




we will ignore bias for

**Step 1**: initialize variables. + first forward prop.

$W_a = 1$, $W_x = 1$, $a^{<0>} = 0$

ex1 →
- $a^{<1>} = \tanh(1 \cdot 0 + 1 \cdot 1) = 0.76$
- $a^{<2>} = \tanh(1 \cdot 0.76 + 1 \cdot 0) = 0.64$
- $a^{<3>} = \tanh(1 \cdot 0.64 + 1 \cdot 0) = 0.57 = \hat{y}^{[1]} \rightarrow y^{[1]} = 1$

ex2 →
- $a^{<1>} = \tanh(1 \cdot 0 + 1 \cdot 0) = 0$
- $a^{<2>} = \tanh(1 \cdot 0 + 1 \cdot 1) = 0.76$
- $a^{<3>} = \tanh(1 \cdot 0.76 + 1 \cdot 1) = 0.94 = \hat{y}^{[2]} \rightarrow y^{[2]} = 2$

ex3 →
- $a^{<1>} = \tanh(1 \cdot 0 + 1 \cdot 1) = 0.76$
- $a^{<2>} = \tanh(1 \cdot 0.76 + 1 \cdot 1) = 0.94$
- $a^{<3>} = \tanh(1 \cdot 0.94 + 1 \cdot 1) = 0.96 = \hat{y}^{[3]} \rightarrow y^{[3]} = 3$

1st Epoch

$$\text{Loss} = L = \frac{(1-0.57)^2 + (2-0.94)^2 + (3-0.96)^2}{3} = 1.82$$

## Step 2: Backpropagation

MSE loss function for some parameters $\theta$ and function $h_\theta$

$$L(\theta) = \frac{1}{n}\sum_i^n \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 = \frac{1}{n}\sum_i^n \left(a^{<3>(i)} - y^{(i)}\right)^2$$

To keep things simple, we simplify the activation function to be:

$$a^{<t>} = W_a\, a^{<t-1>} + W_x\, x^{<t>}$$

($a^{<t>}$ → $h_\theta$, $W_a$ → $\theta_0$, $W_x$ → $\theta_1$)

$$\frac{\partial}{\partial \theta_0} L(\theta_0, \theta_1) = \frac{\partial}{\partial \theta_0}\left(\frac{1}{n}\sum_i^n \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2\right)$$

$$= \frac{1}{n}\sum_i^n \frac{\partial}{\partial \theta_0}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$= \frac{1}{n}\sum_i^n 2\left(h_\theta(x^{(i)}) - y^{(i)}\right)\cdot\frac{\partial}{\partial \theta_0}\left(h_\theta(x^{(i)}) - y^{(i)}\right)$$

$$= \frac{a^{<t-1>}}{n}\sum_i^n \left(h_\theta(x^{(i)}) - y^{(i)}\right)$$

### State 3 (Last State)

$$\frac{\partial L}{\partial W_x} = \frac{\partial L}{\partial a^{<3>}}\cdot\frac{\partial a^{<3>}}{\partial W_x} = \frac{2}{n}\sum_i^n \left(a^{<3>(i)} - y^{(i)}\right)\cdot x^{<3>(i)} \quad -(a)$$

$$\frac{\partial L}{\partial W_a} = \frac{\partial L}{\partial a^{<3>}}\cdot\frac{\partial a^{<3>}}{\partial W_a} = \frac{2}{n}\sum_i^n \left(a^{<3>(i)} - y^{(i)}\right)\cdot a^{<2>(i)} \quad -(b)$$

helper:

$$\frac{\partial a^{<2>}}{\partial W_x} = x^{<2>}, \quad \frac{\partial a^{<2>}}{\partial W_a} = a^{<1>}$$

$$\frac{\partial a^{<3>}}{\partial a^{<2>}} = \frac{\partial}{\partial a^{<2>}}\left(W_a\, a^{<2>} + W_x\, x^{<3>}\right) = W_a$$

$$\frac{\partial L}{\partial a^{<3>}} = \frac{2}{n}\sum^n \left(a^{<3>(i)} - y^{(i)}\right)$$

### State 2 ($a^{<2>}$)

$$\frac{\partial L}{\partial W_x} = \frac{\partial L}{\partial a^{<3>}}\cdot\frac{\partial a^{<3>}}{\partial a^{<2>}}\cdot\frac{\partial a^{<2>}}{\partial W_x} = \frac{2}{n}\sum_i^n \left(a^{<3>(i)} - y^{(i)}\right)\left(W_a\right)\left(x^{<2>(i)}\right) \quad -(c)$$

$$\frac{\partial L}{\partial W_a} = \frac{\partial L}{\partial a^{<3>}}\cdot\frac{\partial a^{<3>}}{\partial a^{<2>}}\cdot\frac{\partial a^{<2>}}{\partial W_a} = \frac{2}{n}\sum^n \left(a^{<3>(i)} - y^{(i)}\right)\left(W_a\right)\left(a^{<1>(i)}\right) \quad -(d)$$

helper:

$$\frac{\partial a^{<2>}}{\partial a^{<1>}} = W_a$$

$$\frac{\partial a^{<1>}}{\partial W_x} = x^{<1>}, \quad \frac{\partial a^{<1>}}{\partial W_a} = a^{<0>}$$

### State 1 ($a^{<1>}$)

$$\frac{\partial L}{\partial W_x} = \frac{\partial L}{\partial a^{<3>}}\cdot\frac{\partial a^{<3>}}{\partial a^{<2>}}\cdot\frac{\partial a^{<2>}}{\partial a^{<1>}}\cdot\frac{\partial a^{<1>}}{\partial W_x} = \frac{2}{n}\sum^n \left(a^{<3>(i)} - y^{(i)}\right)\cdot\left(W_a\right)^2\cdot\left(x^{<1>(i)}\right) \quad -(e)$$

$$\frac{\partial L}{\partial W_a} = \text{"} \cdot\frac{\partial a^{<1>}}{\partial W_a} = \frac{2}{n}\sum^n \left(a^{<3>(i)} - y^{(i)}\right)\left(W_a\right)^2\cdot\left(a^{<0>(i)}\right) \quad -(f)$$

Remember that I did not use consistent activation function in forward prop & backward prop. But the flow of calculations should be same

(a) $\frac{2}{3}\left[(0.57-1)(0) + (0.94-2)(1) + (0.96-3)(1)\right] = -1.36$ ← probably need to multiply by learning rate since $-1.36$ is too big

(b) $\frac{2}{3}\left[(0.57-1)(0.64) + (0.94-2)(0.76) + (0.96-3)(0.94)\right] = \cdots$

(c) (d) (e) (f) → do the same for these equations

## Step 3: Update Weights

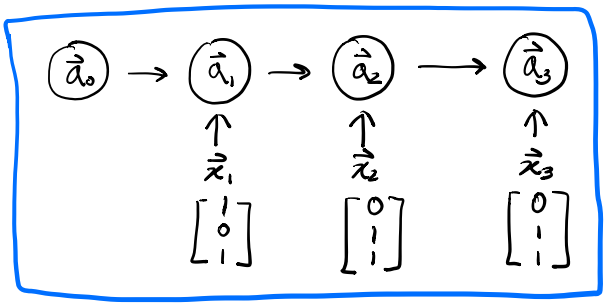
$$W_x = W_x - ((a)+(c)+(e))$$
$$W_a = W_a - ((b)+(d)+(f))$$

## Step 4: Repeat!

Repeat Step 1 until the model converges to minimum loss.

# ⟨ Vectorization ⟩ vector implementation of above

$$x = \begin{bmatrix} [1,0,0] \\ [0,1,1] \\ [1,1,1] \end{bmatrix} \quad \vec{y} = [1,2,3]$$




where $\vec{a}_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

$$\vec{a}_1 = \vec{a}_0 \cdot w_a + \vec{x}_1 \cdot w_x = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} w_a + \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} w_x = \begin{bmatrix} w_x \\ 0 \\ w_x \end{bmatrix}$$

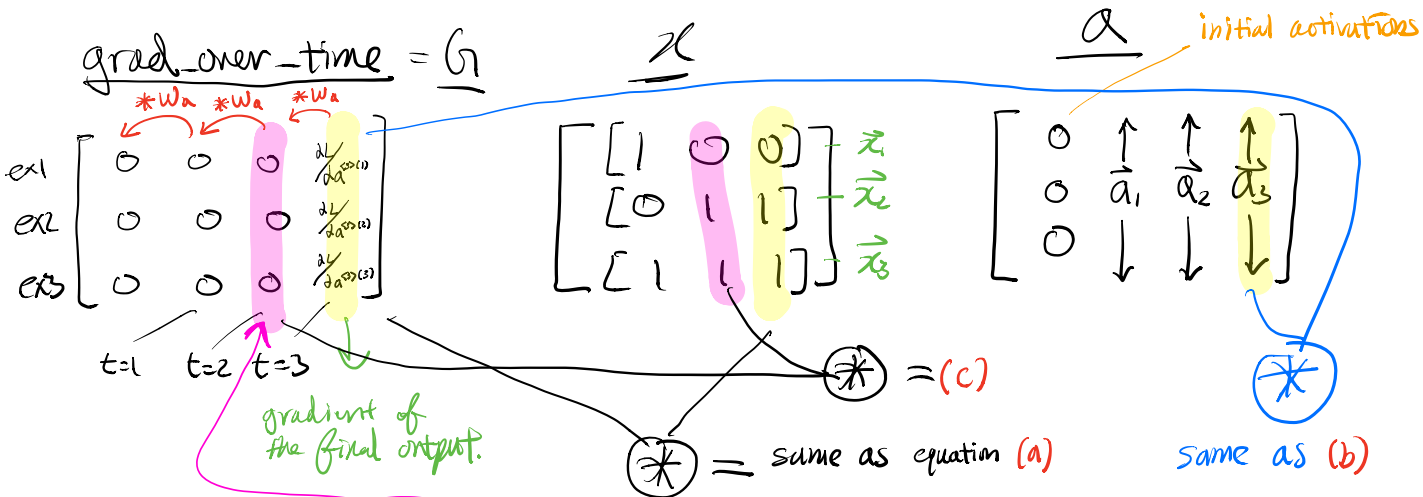
$$\vec{a}_2 = \vec{a}_1 \cdot w_a + \vec{x}_2 \cdot w_x = \begin{bmatrix} w_x \\ 0 \\ w_x \end{bmatrix} w_a + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} w_x = \begin{bmatrix} w_x w_a + 0 \\ 0 + w_x \\ w_x w_a + w_x \end{bmatrix} = \begin{bmatrix} w_x w_a \\ w_x \\ w_x w_a + w_x \end{bmatrix}$$

$$\vec{a}_3 = \vec{a}_2 w_a + \vec{x}_3 \cdot w_x = \cdots$$

$$L = \frac{1}{n}(\vec{a}_3 - \vec{y})^2 = \frac{1}{n}\left( \begin{bmatrix} a_{31} \\ a_{32} \\ a_{33} \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \right)^2 = \square$$

---

Below is in reference to the 'RNN from Scratch' Jupyter Notebook.




$$\frac{\partial L}{\partial a^{<2>(i)}} = \frac{2}{n}\sum^{n}\left(a^{<2>(i)} - y^{(i)}\right)$$

update value

$$\frac{\partial L}{\partial W} = \left(\frac{\partial L}{\partial a^{<2>}} \cdot \frac{\partial a^{<2>}}{\partial a^{<1>}}\right) \cdot \frac{\partial a^{<1>}}{\partial W}$$

$$= \frac{2}{n}\sum^{n}\left(a^{<2>(i)} - y^{(i)}\right) \cdot W_a$$