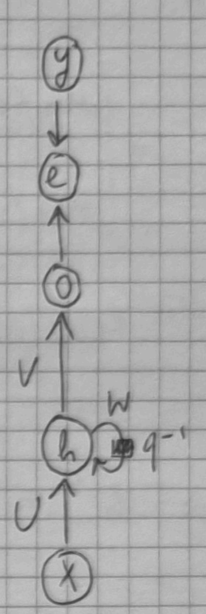
TUTOR LESSON 2 - DEEP LEARNING (RTRL + BPTT)

<u>Exercise 1</u>: Given the recurrent network with equations and loss function L:



$$h^{(0)} = 0$$
$$a_h^{(t)} = Ux^{(t)} + Wh^{(t-1)} + b$$
$$h^{(t)} = ReLU(a_h^{(t)})$$
$$a_o^{(t)} = Vh^{(t)} + c$$
$$o^{(t)} = ReLU(a_o^{(t)})$$
$$L = \sum_t e^{(t)} = \sum_t (y^{(t)} - o^{(t)})^2$$

Consider the following learning task involving a single sequence:

INPUT SEQUENCE

$$\underset{t=1}{\begin{bmatrix} 1 \\ -1 \end{bmatrix}} \longleftarrow \underset{t=2}{\begin{bmatrix} -1 \\ 1 \end{bmatrix}}$$

TARGET SEQUENCE

$$\underset{t=1}{[1]} \longleftarrow \underset{t=2}{[-1]}$$

Assume the following values for the parameters:

$$U = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad W = \begin{bmatrix} -1 & 1 \\ 2 & 1 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & -1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad c = 3.$$

Compute the gradient $\dfrac{\partial o^{(t=1)}}{\partial U}$ using RTRL, reporting all the derivation steps.

RTLR is an ONLINE learning algorithm for RNNs, meaning that the gradient of the error wrt to the weights is computed as the data arrives, one timestep at a time.

This is different to BPTT, which unfolds the recurrent network in time and backpropagates the error across many time steps after seeing the whole sequence.

Remember that RTLR has a very high computational complexity wrt BPTT, which makes it unfeasible for larger networks.

SPACE COMPLEXITY: $O(N^3)$ and TIME COMPLEXITY: $O(N^4)$

where N is the # of UNITS.

FORWARD PASS ($t=1$):

$$a_h^{(1)} = UX^{(1)} + \underbrace{Wh^{(0)}}_{=0} + b = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix} \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

$$= \begin{bmatrix} u_{11} x_1^{(1)} + u_{12} x_2^{(1)} + b_1 \\ u_{21} x_1^{(1)} + u_{22} x_2^{(1)} + b_2 \end{bmatrix} = \begin{bmatrix} 1 \cdot 1 + 1(-1) + 1 \\ 1 \cdot 1 + (-1) \cdot (-1) + 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} a_{h,1}^{(1)} \\ a_{h,2}^{(1)} \end{bmatrix}$$

$$h^{(1)} = \mathrm{ReLU}(a_h^{(1)}) = \begin{bmatrix} \mathrm{ReLU}(1) \\ \mathrm{ReLU}(3) \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} h_1^{(1)} \\ h_2^{(1)} \end{bmatrix}$$

$$\boxed{\mathrm{ReLU}(x) = \max(0, x)}$$

$$\rightarrow \quad h^{(1)} \equiv a_h^{(1)} \quad \Rightarrow \quad \frac{\partial h^{(1)}}{\partial a_h^{(1)}} = 1$$

We can start computing the first derivative: $\frac{\partial h^{(1)}}{\partial U} \in \mathbb{R}^{2 \times (2 \times 2)}$

which we can decompose in $\frac{\partial h_1^{(1)}}{\partial U}$ and $\frac{\partial h_2^{(1)}}{\partial U}$, both 2×2 matrices.

$$(1) \quad \frac{\partial h_1^{(1)}}{\partial U} = \begin{bmatrix} \frac{\partial h_1^{(1)}}{\partial u_{11}} & \frac{\partial h_1^{(1)}}{\partial u_{12}} \\ \frac{\partial h_1^{(1)}}{\partial u_{21}} & \frac{\partial h_1^{(1)}}{\partial u_{22}} \end{bmatrix} \overset{(*)}{=} \begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}$$

(*) Remember that $\frac{\partial h^{(1)}}{\partial a_h^{(1)}} = 1$, so $\frac{\partial h_1^{(1)}}{\partial U} = \frac{\partial a_{h,1}^{(1)}}{\partial U}$

$$(2) \quad \frac{\partial h_2^{(1)}}{\partial U} = \begin{bmatrix} \frac{\partial h_2^{(1)}}{\partial u_{11}} & \frac{\partial h_2^{(1)}}{\partial u_{12}} \\ \frac{\partial h_2^{(1)}}{\partial u_{21}} & \frac{\partial h_2^{(1)}}{\partial u_{22}} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ x_1^{(1)} & x_2^{(1)} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix}$$

Continue with the FORWARD PASS:

$$a_o^{(1)} = Vh^{(1)} + c = \begin{bmatrix} v_1 & v_2 \end{bmatrix} \begin{bmatrix} h_1^{(1)} \\ h_2^{(1)} \end{bmatrix} + c = v_1 h_1^{(1)} + v_2 h_2^{(1)} + c$$

$$= 1 \cdot 1 + (-1) \cdot 3 + 3 = 1$$

$$o^{(1)} = \mathrm{ReLU}(a_o^{(1)}) = \mathrm{ReLU}(1) = 1 \quad \Rightarrow \quad o^{(1)} \equiv a_o^{(1)}$$

Now we can compute the derivative $\frac{\partial o^{(1)}}{\partial U}$

$$\frac{\partial o^{(1)}}{\partial U} = \frac{\partial o^{(1)}}{\partial h^{(1)}} \cdot \underbrace{\frac{\partial h^{(1)}}{\partial U}}_{\text{already computed!}}$$

$$\underbrace{\frac{\partial o^{(1)}}{\partial h^{(1)}}}_{\in \mathbb{R}^2} = V = \begin{bmatrix} 1 & -1 \end{bmatrix}$$

In the end:

$$\underbrace{\frac{\partial o^{(1)}}{\partial U}}_{\in \mathbb{R}^{2\times 2}} = \underbrace{\underbrace{V}_{\in \mathbb{R}^{1\times 2}} \cdot \underbrace{\frac{\partial h^{(1)}}{\partial U}}_{\in \mathbb{R}^{2\times(2\times 2)}}}_{\mathbb{R}^{1\times(2\times 2)}} = \begin{bmatrix} 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

Exploiting the sparsity of the tensor, we can equivalently write:

$$\frac{\partial o^{(1)}}{\partial U} = v_1 \odot \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} + v_2 \odot \begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix}.$$

Exercise 2: Let's consider a variation of the previous exercise. The equations are all the same except for

$$o^{(t)} = \tanh\left(a_o^{(t)}\right)$$

and compute $\frac{\partial L}{\partial U}$ using BPTT.

SOLUTION

Let's start with the forward pass.

$\boxed{t=1}$

$$a_h^{(1)} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \quad ; \quad h^{(1)} = \mathrm{ReLU}\left(a_h^{(1)}\right) = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

$$a_o^{(1)} = 1 \quad ; \quad o^{(1)} = \tanh\left(a_o^{(1)}\right) = \tanh(1) \simeq 0.76$$

$\boxed{t=2}$

$$a_h^{(2)} = Ux^{(2)} + Wh^{(1)} + b = \begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\begin{bmatrix}-1\\ 1\end{bmatrix} + \begin{bmatrix}-1 & 1\\ 2 & 1\end{bmatrix}\begin{bmatrix}1\\ 3\end{bmatrix} + \begin{bmatrix}1\\ 1\end{bmatrix}$$

$$= \begin{bmatrix}\cancel{-1}+\cancel{1}-\cancel{1}+3+\cancel{1}\\ \cancel{-1}-\cancel{1}+\cancel{2}+3+1\end{bmatrix} = \begin{bmatrix}3\\ 4\end{bmatrix}$$

$$h^{(2)} = ReLU\left(a_h^{(2)}\right) = \begin{bmatrix}3\\ 4\end{bmatrix} \implies h^{(2)} \equiv a_h^{(2)}$$

$$a_o^{(2)} = V \cdot h^{(2)} + c = \begin{bmatrix}1 & -1\end{bmatrix} \cdot \begin{bmatrix}3\\ 4\end{bmatrix} + 3 = 3 - 4 + 3 = 2.$$

$$o^{(2)} = \tanh\left(a_o^{(2)}\right) = \tanh(2) \approx 0.96$$

Now we compute the loss:

$$L = \sum_t e^{(t)} = \sum_t \left(y^{(t)} - o^{(t)}\right)^2 = \underbrace{(1-0.76)^2}_{t=1} + \underbrace{(-1-0.96)^2}_{t=2} = 0.06 + 3.84 = 3.90.$$

Now we can compute the derivative.

We know that:

$$\frac{\partial L}{\partial U} = \frac{\partial \sum_t e^{(t)}}{\partial U} = \sum_t \frac{\partial \left(y^{(t)} - o^{(t)}\right)^2}{\partial U} = \sum_t -2\left(y^{(t)} - o^{(t)}\right) \cdot \frac{\partial o^{(t)}}{\partial U}$$

In our case (2 timesteps):

$$\frac{\partial L}{\partial U} = -2\left(1 - o^{(1)}\right) \cdot \frac{\partial o^{(1)}}{\partial U} - 2\left(1 - o^{(2)}\right) \cdot \frac{\partial o^{(2)}}{\partial U}$$

$$= -2 \cdot \left[\left(1 - o^{(1)}\right) \cdot \frac{\partial o^{(1)}}{\partial U} + \left(1 - o^{(2)}\right) \cdot \frac{\partial o^{(2)}}{\partial U}\right]$$

Now: $\dfrac{\partial o^{(t)}}{\partial U} = \dfrac{\partial o^{(t)}}{\partial h^{(t)}} \cdot \dfrac{\partial h^{(t)}}{\partial U}$, where $\dfrac{\partial o^{(t)}}{\partial h^{(t)}} = \dfrac{\partial o^{(t)}}{\partial a_o^{(t)}} \cdot \dfrac{\partial a_o^{(t)}}{\partial h^{(t)}}$

Recalling that $o^{(t)} = \tanh\left(a_o^{(t)}\right) \Rightarrow \dfrac{\partial o^{(t)}}{\partial a_o^{(t)}} = 1 - \left(o^{(t)}\right)^2$

(because $\tanh'(x) = 1 - \tanh^2(x)$)

So:

$$\frac{\partial o^{(t)}}{\partial U} = \underbrace{\frac{\partial o^{(t)}}{\partial a_o^{(t)}}}_{= 1-(o^{(t)})^2} \cdot \underbrace{\frac{\partial a_o^{(t)}}{\partial h^{(t)}}}_{\text{① } = V} \cdot \underbrace{\frac{\partial h^{(t)}}{\partial a_h^{(t)}}}_{\text{② } = I_{2\times 2}} \cdot \underbrace{\frac{\partial a_h^{(t)}}{\partial U}}_{\text{③}}$$

For ① remember that $a_o^{(t)} = V \cdot h^{(t)} + c \Rightarrow \frac{\partial a_o^{(t)}}{\partial h^{(t)}} = V$

For ② remember that $h^{(t)} \equiv a_h^{(t)}$.

③ We compute this partial derivative looking at each time step separately.

$\frac{\partial a_h^{(1)}}{\partial U} \in \mathbb{R}^{2\times(2\times 2)}$, we decompose the calculation for simplicity.

$$\frac{\partial a_{h,1}^{(1)}}{\partial U} = \begin{bmatrix} \frac{\partial a_{h,1}^{(1)}}{\partial u_{11}} & \frac{\partial a_{h,1}^{(1)}}{\partial u_{12}} \\ \frac{\partial a_{h,1}^{(1)}}{\partial u_{21}} & \frac{\partial a_{h,1}^{(1)}}{\partial u_{22}} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}$$

Remember that $a_{h,1}^{(1)} = u_{11} x_1^{(1)} + u_{12} x_2^{(1)} + b$

$$\frac{\partial a_{h,2}^{(1)}}{\partial U} = \begin{bmatrix} \frac{\partial a_{h,2}^{(1)}}{\partial u_{11}} & \frac{\partial a_{h,2}^{(1)}}{\partial u_{12}} \\ \frac{\partial a_{h,2}^{(1)}}{\partial u_{21}} & \frac{\partial a_{h,2}^{(1)}}{\partial u_{22}} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ x_1^{(1)} & x_2^{(1)} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix}$$

Now we compute $\frac{\partial a_h^{(2)}}{\partial U} \in \mathbb{R}^{2\times(2\times 2)}$

$$a_h^{(2)} = Ux^{(2)} + Wh^{(1)} + b = Ux^{(2)} + W(Ux^{(1)} + b) + b$$

with $h^{(1)} = a_h^{(1)} = Ux^{(1)} + b$

$$= \begin{bmatrix} u_{11}x_1^{(2)} + u_{12}x_2^{(2)} \\ u_{21}x_1^{(2)} + u_{22}x_2^{(2)} \end{bmatrix} + \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \cdot \begin{bmatrix} u_{11}x_1^{(1)} + u_{12}x_2^{(1)} + b_1 \\ u_{21}x_1^{(1)} + u_{22}x_2^{(1)} + b_2 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

$$a_{h,1}^{(2)} = u_{11}x_1^{(2)} + u_{12}x_2^{(2)} + w_{11}(u_{11}x_1^{(1)} + u_{12}x_2^{(1)} + b_1) +$$
$$+ w_{12}(u_{21}x_1^{(1)} + u_{22}x_2^{(1)} + b_2) + b_1$$

$$a_{h,2}^{(2)} = u_{21} x_1^{(2)} + u_{22} x_2^{(2)} + b_2 + w_{21} \cdot \left(u_{11} x_1^{(1)} + u_{12} x_2^{(1)} + b_1\right) +$$

$$+ w_{22} \cdot \left(u_{21} x_1^{(1)} + u_{22} x_2^{(1)} + b_2\right)$$

$$\frac{\partial a_{h,1}^{(2)}}{\partial U} = \begin{bmatrix} \frac{\partial a_{h,1}^{(2)}}{\partial u_{11}} & \frac{\partial a_{h,1}^{(2)}}{\partial u_{12}} \\ \frac{\partial a_{h,1}^{(2)}}{\partial u_{21}} & \frac{\partial a_{h,1}^{(2)}}{\partial u_{22}} \end{bmatrix} = \begin{bmatrix} x_1^{(2)} + w_{11} x_1^{(1)} & x_2^{(2)} + w_{11} x_2^{(1)} \\ w_{12} x_1^{(1)} & w_{12} x_2^{(1)} \end{bmatrix}$$

$$= \begin{bmatrix} -1-1 & 1+1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} -2 & 2 \\ 1 & -1 \end{bmatrix}.$$

$$\frac{\partial a_{h,2}^{(2)}}{\partial U} = \begin{bmatrix} \frac{\partial a_{h,2}^{(2)}}{\partial u_{11}} & \frac{\partial a_{h,2}^{(2)}}{\partial u_{12}} \\ \frac{\partial a_{h,2}^{(2)}}{\partial u_{21}} & \frac{\partial a_{h,2}^{(2)}}{\partial u_{22}} \end{bmatrix} = \begin{bmatrix} w_{21} x_1^{(1)} & w_{21} x_2^{(1)} \\ x_1^{(2)} + w_{22} x_1^{(1)} & x_2^{(2)} + w_{22} x_2^{(1)} \end{bmatrix}$$

$$= \begin{bmatrix} 2 & -2 \\ -1+1 & 1-1 \end{bmatrix} = \begin{bmatrix} 2 & -2 \\ 0 & 0 \end{bmatrix}.$$

Notice that this is the same quantity computed with RTRL

Putting all together:

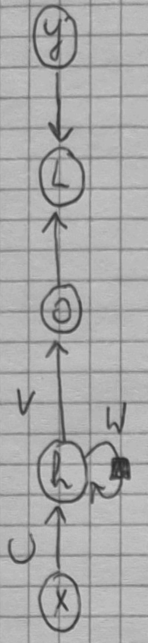
$$\frac{\partial L}{\partial U} = \overbrace{-2\left(1-o^{(1)}\right) \cdot \left(1-\left(o^{(1)}\right)^2\right)}^{= c_1} \cdot \overbrace{\left[v_1 \odot \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} + v_2 \odot \begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix} \right]} +$$

$$+ \underbrace{(-2)\left(1-o^{(2)}\right) \cdot \left(1-\left(o^{(2)}\right)^2\right)}_{= c_2} \cdot \left[v_1 \odot \begin{bmatrix} -2 & 2 \\ 1 & -1 \end{bmatrix} + v_2 \odot \begin{bmatrix} 2 & -2 \\ 0 & 0 \end{bmatrix} \right]$$

$$= c_1 \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} + c_1 \begin{bmatrix} 0 & 0 \\ -1 & 1 \end{bmatrix} + c_2 \begin{bmatrix} -2 & 2 \\ 1 & -1 \end{bmatrix} + c_2 \begin{bmatrix} -2 & 2 \\ 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} c_1 - 2c_2 - 2c_2 & -c_1 + 2c_2 + 2c_2 \\ -c_1 + c_2 & c_1 - c_2 \end{bmatrix} = \begin{bmatrix} c_1 - 4c_2 & -c_1 + 4c_2 \\ -c_1 + c_2 & c_1 - c_2 \end{bmatrix}.$$

Exercise 3:



y is the TARGET, L the LOSS, o is the OUTPUT,
h is the HIDDEN STATE, x is the INPUT at time t
and $\blacksquare$ is the time-shift operator q^{-1}.

U, W, V weights matrices.

Suppose to use BPTT with MINIBATCH = 1.
Given an input sequence of 3 elements, how many terms should be summed up to compute $\frac{\partial L}{\partial U}$?

RISPOSTA: $6 = 3 + 2 + 1$.

$$L = \sum_{t=1}^{3} e^{(t)} \implies \frac{\partial L}{\partial U} = \sum_{t=1}^{3} \frac{\partial e^{(t)}}{\partial U}$$

Each $\frac{\partial e^{(t)}}{\partial U}$ involves unrolling the RNN backward from time t up to time 1:

$$\frac{\partial L}{\partial U} = \underbrace{\frac{\partial e^{(3)}}{\partial U}}_{\text{3 TERMS}} + \underbrace{\frac{\partial e^{(2)}}{\partial U}}_{\text{2 TERMS}} + \underbrace{\frac{\partial e^{(1)}}{\partial U}}_{\text{1 TERM}} = 3 + 2 + 1 = 6.$$

- What if the sequence is long 4? $4 + 3 + 2 + 1 = 10$.
- What if the MINIBATCH = 2 and the sequence is long 4?

$$2 \cdot (4 + 3 + 2 + 1) = 2 \cdot 10 = 20.$$

For each sequence in the minibatch, you unroll the RNN through time and compute the gradients. So the total gradient wrt U is the sum of the gradients over:
- all time steps
- all sequences in the batch.