# Advanced Human Language Technologies
# Exercises on Recurrent Neural Networks

## Recurrent Neural Networks

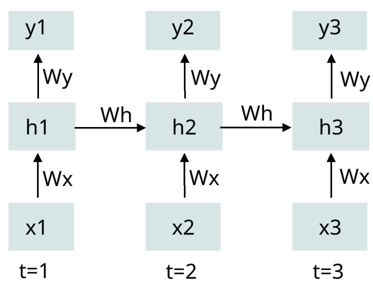
### Exercise 1.

Your friend's mood depends on the weather of the last few days. You've collected data about the weather for the past 365 days which you represent as a sequence as $x_1, \ldots, x_{365}$. You've also collected data on your friends's mood, which you represent as $y_1, \ldots, y_{365}$. You'd like to build a model to map from $x \to y$.

1. Should you use a Unidirectional RNN or Bidirectional RNN for this problem?
2. Explain why vanishing gradients affect more severerly RNNs than deep FNNs.

### Exercise 2.

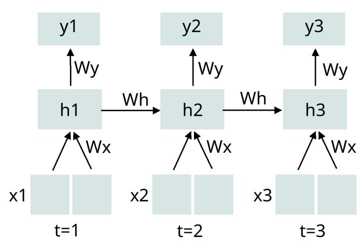
Given the unfolding of the first time steps of a RNN, where both hidden and output layers are linear:




1. Write the RNN equations.
2. Compute the values for hidden units $h_1, h_2, h_3$ and output units $y_1, y_2, y_3$, for input values $x_1 = 2, x_2 = -0.5, x_3 = 1$, assuming that the weight matrices are $W_x = W_h = W_y = 1$.
3. Explain what is the network computing.

### Exercise 3.

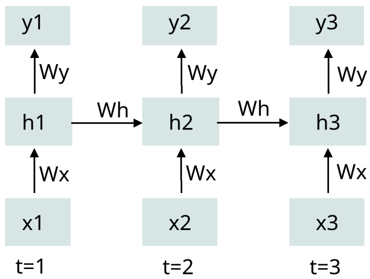
Given the unfolding of the first time steps of a RNN, where hidden layer is linear and output layer has a sigmoid activation function $\sigma(z) = \frac{1}{1+e^{-z}}$:



1. Write the RNN equations.
2. Compute the values for hidden units $h_1, h_2, h_3$ and output units $y_1, y_2, y_3$, for input values $x_1 = (2, -2), x_2 = (0, 3.5), x_3 = (1, 2.2)$, assuming that the weight matrices are $W_x = (1, -1), W_h = W_y = 1$.
3. Explain what is the network computing.

## Exercise 4.

Given the unfolding of the first time steps of a RNN, where hidden layer has a sigmoid activation function $\sigma(z) = \frac{1}{1+e^{-z}}$, and output layer is linear.




Assuming weight matrices are and we input

1. Write the RNN equations.
2. Compute the values for hidden units $h_1, h_2, h_3$ and output units $y_1, y_2, y_3$, for input values $x_1 = 18, x_2 = 9, x_3 = -8$, assuming that the weight matrices are $W_x = -0.1, W_h = 0.5, W_y = 0.25$ and biases are $b_h = 0.4, b_y = 0$.

## Exercise 5.

Consider a simple RNN with the following equations:

$$h_t = \tanh(W_x x_t + W_h h_{t-1} + b_h)$$

$$y_t = W_y h_t + b_y$$

where:

- $x_t$ is the input at time $t$,
- $h_t$ is the hidden state at time $t$, set to zero for $t = 0$
- $W_h, W_x, W_y$ are weight matrices, and
- $b_h, b_y$ are biases.

1. Given the following values, compute the hidden states $h_1$ and $h_2$ for the input sequence $x_1 = 1$, $x_2 = 2$

$$W_x = \begin{bmatrix} 0.6 \\ 0.1 \end{bmatrix} \qquad W_h = \begin{bmatrix} 0.5 & 0.2 \\ 0.3 & 0.7 \end{bmatrix} \qquad b_h = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad b_y = 0.1$$

2. Given the values above, which is the dimensionality of $h_t$, $y_t$, and $W_y$ in this RNN? Justify your answer.

## Exercise 6.

Given an LSTM cell with the following equations:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$
$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$
$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$
$$h_t = o_t \odot \tanh(c_t)$$

where:
$\sigma(x) = \frac{1}{1+e^{-x}}$ (sigmoid activation)
$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ (tanh activation)
$\odot$ represents element-wise multiplication.

1. Manually compute $f_t, i_t, o_t, c_t, h_t$, assuming the following values

   - $x_t = 1 \qquad h_{t-1} = \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix} \qquad c_{t-1} = \begin{bmatrix} 0.05 \\ 0.1 \end{bmatrix}$
   - Forget gate: $W_f = \begin{bmatrix} 0.4 \\ -0.2 \end{bmatrix} \qquad U_f = \begin{bmatrix} 0.1 & 0.3 \\ 0.2 & 0.5 \end{bmatrix} \qquad b_f = \begin{bmatrix} 0.1 \\ -0.1 \end{bmatrix}$
   - Input gate: $W_i = \begin{bmatrix} 0.2 \\ -0.3 \end{bmatrix} \qquad U_i = \begin{bmatrix} 0.4 & 0.2 \\ -0.1 & 0.3 \end{bmatrix} \qquad b_i = \begin{bmatrix} 0.05 \\ -0.05 \end{bmatrix}$
   - Output gate: $W_o = \begin{bmatrix} -0.3 \\ 0.5 \end{bmatrix} \qquad U_o = \begin{bmatrix} 0.2 & -0.2 \\ 0.1 & 0.4 \end{bmatrix} \qquad b_o = \begin{bmatrix} 0.02 \\ 0.01 \end{bmatrix}$
   - Candidate cell state: $W_c = \begin{bmatrix} 0.3 \\ 0.1 \end{bmatrix} \qquad U_c = \begin{bmatrix} -0.2 & 0.3 \\ 0.4 & -0.1 \end{bmatrix} \qquad b_c = \begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix}$

## Exercise 7.

We have an LSTM trained to perform sentiment analysis. At each step, the LSTM cell has different values for the **forget gate** ($f_t$) which decide what fraction of the previous cell state should be kept, the **input gate** ($i_t$), that controls how much new information is added, and the **output gate** ($o_t$), which controls what is sent to the hidden state.

Given the input sentence *I love this movie*, gate values at each time step are:

| $t$ | Word | $f_t$ | $i_t$ | $o_t$ |
|---|---|---|---|---|
| 1 | I | 0.9 | 0.1 | 0.4 |
| 2 | love | 0.1 | 0.9 | 0.8 |
| 3 | this | 0.7 | 0.3 | 0.5 |
| 4 | movie | 0.8 | 0.5 | 0.6 |

1. If the initial cell state is $c_0 = 1.0$, compute the final cell state $c_t$ after processing *I love this movie* using only the forget gate values and ignoring the input gate values. Interpret the impact of these values on memory retention.
2. Now take into account the input gate information at step $t = 2$, assuming that he candidate cell state computed from the input is $0.8$, and compute the new $c_t$. What does this tell about how LSTMs handle important words?
3. Assume the final $c_t$ is $2.0$ before the output gate is applied. Compute the final hidden state $h_t$. How does the output gate affect the final representation?
4. Discussion Questions

   (a) **Forget vs. Input Gates**: How do forget and input gates work together to control memory in an LSTM?

(b) **Effect of Sentiment Words**: How does the LSTM handle sentiment words like *love*?

(c) **Limitations**: What are some weaknesses of LSTMs compared to Transformers in sequence processing?