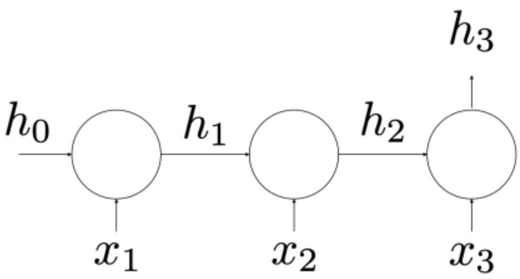
RNN Questions

Problem 6 Recurrent Neural Networks and Backpropagation (9 credits)

Consider a vanilla RNN cell of the form $h_t = \tanh(V \cdot h_{t-1} + W \cdot x_t + b)$. The figure below shows the input sequence x_1, x_2, and x_3.



a) Given the dimensions $x_t \in \mathbb{R}^3$ and $h_t \in \mathbb{R}^5$, what is the number of parameters in the RNN cell? (Calculate final number)

$3 \times 5 + 5 \times 5 + 5(bias) = 15 + 25 + 5 = 45$ (1p for 45, else 0)

b) If x_t and b are the 0 vector, then $h_t = h_{t-1}$ for any value of h_t. Discuss whether this statement is correct.

False: (0.5p)
After transformation with V and non-linearity $x_t = 0$ does not lead to $h_t = h_{t-1}$ (0.5p) , i.e. h_t can be changed.
Note: simply repeating the formula $h_t = \tanh(V \cdot h_{t-1}))$ does not give any points. If you only mention V or *tanh* then this is also correct, though giving an incorrect formula invalidates that half point.

Now consider the following **one-dimensional** ReLU-RNN cell without bias b.

$$h_t = \text{ReLU}(V \cdot h_{t-1} + W \cdot x_t)$$

(Hidden state, input, and weights are scalars)

c) Calculate h_2 and h_3 where

$$V = -3, \quad W = 3, \quad h_0 = 0, \quad x_1 = 2, \quad x_2 = 3 \quad \text{and } x_3 = 1.$$

$$\begin{aligned} h_0 &= 0 \\ h_1 &= \text{relu}\,(-3 \cdot 0 + 3 \cdot 2) = 6 \\ h_2 &= \text{relu}\,(-3 \cdot 6 + 3 \cdot 3) = 0 && \text{(1 p)} \\ h_3 &= \text{relu}\,(-3 \cdot 0 + 3 \cdot 1) = 3 && \text{(1 p)} \end{aligned}$$

Note: Only points for correct solutions, no points for intermediate steps (even if you have an incorrect h_1)

d) Calculate the derivatives $\frac{\partial h_3}{\partial V}$, $\frac{\partial h_3}{\partial W}$, and $\frac{\partial h_3}{\partial x_1}$ for the forward pass of the ReLU-RNN where

$$V = -2, \quad W = 1, \quad h_0 = 2, \quad x_1 = 2, \quad x_2 = \frac{3}{2} \quad \text{and } x_3 = 4.$$

0
1
2
3

for the forward outputs

$$h_1 = 0, \quad h_2 = \frac{2}{3}, \quad h_3 = 1.$$

Use that $\frac{\partial}{\partial x}\mathrm{ReLU}(x)\Big|_{x=0} = 0$.

Generally:

$$\frac{\partial h_t}{\partial V} = h_{t-1} + V \cdot \frac{\partial h_{t-1}}{\partial V}$$
$$\frac{\partial h_t}{\partial W} = \frac{\partial \mathrm{ReLU}(z_t)}{\partial z_t} \cdot \left(V \cdot \frac{\partial h_{t-1}}{\partial W} + x_t\right)$$
$$\frac{\partial h_t}{\partial x_\tau} = \frac{\partial \mathrm{ReLU}(z_t)}{\partial z_t} \cdot \left(V \cdot \frac{\partial h_t}{\partial x_\tau} + W \cdot \delta_{t\tau}\right)$$

$$\frac{\partial h_3}{\partial V} = h_2 + V \cdot h_1 = \frac{2}{3} + 0 = \frac{2}{3} \qquad \text{(1p)}$$
$$\frac{\partial h_3}{\partial W} = V \cdot x_2 + x_3 = -2 \cdot \frac{3}{2} + 4 = 1 \qquad \text{(1p)}$$
$$\frac{\partial h_3}{\partial x_1} = 0 \text{ (dead ReLU)} \qquad \text{(1p)}$$

Note: alternatively $\frac{\partial h_3}{\partial V} = \frac{3}{2}$ if student correctly identified that h_2 should have been flipped to be a correct forward pass.
For $\frac{\partial h_3}{\partial x_1}$, it's okay even if no formula, but some explanation is given (dead relu after first layer)

0 1 2

e) A Long-Short Term Memory (LSTM) unit is defined as

$$\begin{aligned}
g_1 &= \sigma\left(W_1 \cdot x_t + U_1 \cdot h_{t-1}\right), \\
g_2 &= \sigma\left(W_2 \cdot x_t + U_2 \cdot h_{t-1}\right), \\
g_3 &= \sigma\left(W_3 \cdot x_t + U_3 \cdot h_{t-1}\right), \\
\tilde{c}_t &= \tanh\left(W_c \cdot x_t + u_c \cdot h_{t-1}\right), \\
c_t &= g_2 \circ c_{t-1} + g_3 \circ \tilde{c}_t, \\
h_t &= g_1 \circ c_t,
\end{aligned}$$

where g_1, g_2, and g_3 are the gates of the LSTM cell.
1) Assign these gates correctly to the **forget** f, **update** u, and **output** o gates. (1p)
2) What does the value c_t represent in a LSTM? (1p)

g_1 = output gate
g_2 = forget gate
g_3 = update gate/input gate
(1p for all three, zero otherwise)
c_t: cell state/memory (1p)
Note: if students interpreted c_t as "what does it do?" half a point was awarded. Possible half point: “Intermediate value, check what to forget and what to add from input”

Problem 10 Recurrent Neural Networks and Backpropagation (8 credits)

Recurrent neural networks, also known as RNNs, are a class of neural networks that allow an arbitrary number of inputs and, thus, are often used for sequences of data, e.g., in the fields of natural language processing and speech recognition.

a) Mathematically explain the reason for exploding and vanishing gradients when using a classic RNN, i.e., $A_t = \theta_c A_{t-1} + \theta_x x_t$, where both θ_c and θ_x are orthogonal. (2 points)

0 1 2

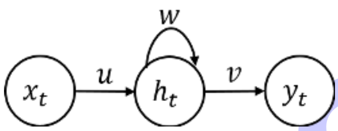
1. Show backpropagation explicitly to calculate gradients (1 point):

$$\frac{\delta h^{(t)}}{\delta h^{(1)}} = \frac{\delta h^{(t)}}{\delta h^{(t-1)}} \cdots \frac{\delta h^{(2)}}{\delta h^{(1)}}$$

2. After eigen-decomposition of θ_c, the **largest** eigenvalue of $\theta_c > 1$ means explosion(0.5 point) and < 1 means vanishing (0.5 point).
   P.s. if the answer discussed eigenvalues of A_t instead of θ_c: 0 point
   if the answer discussed two cases of eigenvalues but didn't show explicitly which matrix should be decomposed: 0.5 point

b) Now consider the following RNN

0 1 2 3



which uses the one-dimensional ReLU-RNN cell

$$h_t = ReLU(u * h_{t-1} + w * x_t).$$

Compute the forward propagation y_2, h_2 and the gradient $\mathbf{dy}_2 := \frac{\delta y_2}{\delta u}$ where

$$h_0 = 3,\ w = 2,\ v = -1,\ u = 3,\ x_1 = 1,\ x_2 = 2.$$

Since there is a mismatch between the graph and the given equation, answers based either on the graph or on the equation are accepted. Solution based on the graph:

$$h_1 = ReLU(u * x_1 + w * h_0) = 9$$
$$h_2 = ReLU(u * x_2 + w * h_1) = 24(1p)$$
$$y_2 = v * h_2 = -24(1p)$$
$$\frac{\delta y_2}{\delta u} = v * w * x_1 + v * x_2 = -4(1p)$$

Solution based on the equation:

$$h_1 = ReLU(u * h_0 + w * x_1) = 11$$
$$h_2 = ReLU(u * h_1 + w * x_2) = 37(1p)$$
$$y_2 = v * h_2 = -37(1p)$$
$$\frac{\delta y_2}{\delta u} = v * (h_1 + u * h_0) = -20(1p)$$

Each equation and final result counts 0.5 points.

0
1
2

c) To circumvent the vanishing gradient problem, the Long-Short Term Memory (LSTM) unit was proposed. It is defined as

$$
\begin{aligned}
g_1 &= \sigma\left(W_1 \cdot x_t + U_1 \cdot h_{t-1}\right), \\
g_2 &= \sigma\left(W_2 \cdot x_t + U_2 \cdot h_{t-1}\right), \\
g_3 &= \sigma\left(W_3 \cdot x_t + U_3 \cdot h_{t-1}\right), \\
\tilde{c}_t &= \tanh\left(W_c \cdot x_t + u_c \cdot h_{t-1}\right), \\
c_t &= g_2 \circ c_{t-1} + g_3 \circ \tilde{c}_t, \\
h_t &= g_1 \circ c_t,
\end{aligned}
$$

where g_1, g_2, and g_3 are the gates of the LSTM cell.
1) Assign these gates correctly to the **forget** f, **update** u, and **output** o gates. (1p)
2) What does the value c_t represent in a LSTM? (1p)

g_1 = output gate
g_2 = forget gate
g_3 = update gate/input gate
(1 pt)
c_t: cell state
(1 pt)

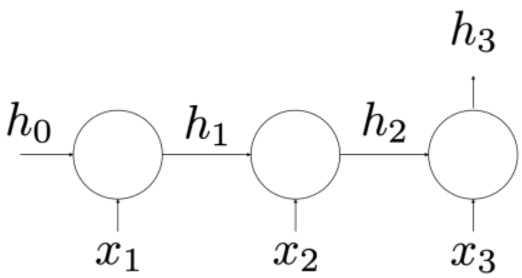
0
1

d) Why does the LSTM unit solve the vanishing gradient problem that is present in the default definition of an RNN cell?

gradient highway through the cell state (0.5pt) and gate system to remove or add information to the cell state (0.5pt)
(another alternative: from the perspectives of weights and activation functions as in the slides)

Problem 6 Recurrent Neural Networks and LSTMs (12 credits)

a) Consider a vanilla RNN cell of the form $h_t = \tanh(V \cdot h_{t-1} + W \cdot x_t)$. The figure below shows the input sequence x_1, x_2, and x_3.



Given the dimensions $x_t \in \mathbb{R}^4$ and $h_t \in \mathbb{R}^{12}$, what is the number of parameters in the RNN cell? Neglect the bias parameter.

$4 \times 12 + 12 \times 12$ (1 pt) $= 48 + 144 = 192$ (1 pt)

b) ~~If x_t is the 0 vector, then $h_t = h_{t-1}$.~~ Discuss whether this statement is correct.

False: (1 pt)
After transformation with V and non-linearity $x_t = 0$ does not lead to $h_t = h_{t-1}$ (1 pt) . Full points require explanation, solely equation not sufficient.

0 ☐ 1 ☐ 2 ☐ 3 ☐

~~c) Now consider the following **one-dimensional** ReLU-RNN cell.~~

$$\cancel{h_t = \mathrm{ReLU}(V \cdot h_{t-1} + W \cdot x_t)}$$

~~(Hidden state, input, and weights are scalars)~~
~~Calculate h_1, h_2 and h_3 where~~ $V = 1$, $W = 2$, $h_0 = -3$, $x_1 = 1$, $x_2 = 2$ and $x_3 = 0$.

$h_0 = -3$
$h_1 = \mathrm{relu}(1 \cdot (-3) + 2 \cdot 1) = 0$ (1 pt)
$h_2 = \mathrm{relu}(1 \cdot 0 + 2 \cdot 2) = 4$ (1 pt)
$h_3 = \mathrm{relu}(1 \cdot 4 + 2 \cdot 0) = 4$ (1 pt)

d) Calculate the derivatives $\frac{\partial h_3}{\partial V}$, $\frac{\partial h_3}{\partial W}$, and $\frac{\partial h_3}{\partial x_1}$ for the forward pass of the ReLU-RNN Cell of (c). Use that $\frac{\partial}{\partial x}\mathrm{ReLU}(x)\Big|_{x=0} = 1$.

0 1 2 3

$$h_t = \mathrm{ReLU}(V \cdot h_{t-1} + W \cdot x_t) = \mathrm{ReLU}(z_t)$$

$$\frac{\partial h_3}{\partial V} = \frac{\partial}{\partial x}\mathrm{ReLU}(x)\Big|_{x=z_3} \cdot h_2 + \frac{\partial}{\partial x}\mathrm{ReLU}(x)\Big|_{x=z_2} \cdot V \cdot h_1 + \frac{\partial}{\partial x}\mathrm{ReLU}(x)\Big|_{x=z_1} \cdot V^2 \cdot h_0 =$$

$$= 1 \cdot 4 + 1 \cdot 1 \cdot 0 + 0 \cdot 1 \cdot (-3) = 4$$ ~~(1 pt)~~

~~$\frac{\partial h_3}{\partial W} = \frac{\partial}{\partial x}\mathrm{ReLU}(x)\Big|_{x=z_3} \cdot x_3 + \frac{\partial}{\partial x}\mathrm{ReLU}(x)\Big|_{x=z_2} \cdot V \cdot x_2 + \frac{\partial}{\partial x}\mathrm{ReLU}(x)\Big|_{x=z_1} \cdot V^2 \cdot x_1 =$~~

~~$1 \cdot 0 + 1 \cdot 2 + 0 \cdot 0 = 2$~~ ~~(1 pt)~~

~~$\frac{\partial h_3}{\partial x_1} = \mathrm{ReLU}(x)\Big|_{x=z_3} \cdot V \cdot \mathrm{ReLU}(x)\Big|_{x=z_2} \cdot V \cdot \mathrm{ReLU}(x)\Big|_{x=z_1} \cdot W = 1 \cdot 1 \cdot 1 \cdot 1 \cdot 0 \cdot 2 = 0$~~ ~~(1 pt)~~

~~Only correct and calculated result gives point.~~

0
1
2

e) A Long-Short Term Memory (LSTM) unit is defined as

$$
\begin{aligned}
g_1 &= \sigma\left(W_1 \cdot x_t + U_1 \cdot h_{t-1}\right),\\
g_2 &= \sigma\left(W_2 \cdot x_t + U_2 \cdot h_{t-1}\right),\\
g_3 &= \sigma\left(W_3 \cdot x_t + U_3 \cdot h_{t-1}\right),\\
\tilde{c}_t &= \tanh\left(W_c \cdot x_t + u_c \cdot h_{t-1}\right),\\
c_t &= g_2 \circ c_{t-1} + g_3 \circ \tilde{c}_t,\\
h_t &= g_1 \circ c_t,
\end{aligned}
$$

where g_1, g_2, and g_3 are the gates of the LSTM cell.
1) Assign these gates correctly to the **forget** f, **update** u, and **output** o gates. (1p)
~~2) What does the value c_t represent in a LSTM? (1p)~~

g_1 = output gate
g_2 = forget gate
g_3 = update gate
(1 pt)
c_t: cell state
(1 pt)