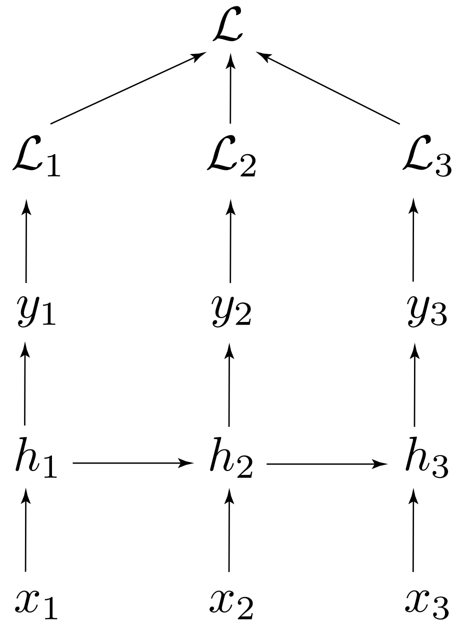
Exercises for the course
Deep Learning 1
Winter Semester 2022/23

Machine Learning Group
Faculty IV – Electrical Engineering and Computer Science
Technische Universität Berlin

Exercise Sheet 7

Exercise 1: Computing Gradients in RNN's (50 P)



$$\begin{align} h_t &= \sigma(\tilde{h}_t) = \sigma(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{1} \\ y_t &= \sigma(\tilde{y}_t) = \sigma(W_{hy}h_t + b_y) \tag{2} \\ \mathcal{L}_t(t_t, y_t) &= (t_t - y_t)^2 \tag{3} \end{align}$$

$$\begin{align} \mathcal{L} &= \sum_{t=1}^{T} \mathcal{L}_t \tag{4} \\ &= (\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3) \tag{5} \\ &= (t_1 - y_1)^2 + (t_2 - y_2)^2 + (t_3 - y_3)^2 \tag{6} \end{align}$$

You can abbreviate the derivative of the activation function $\sigma(x)$ as $\sigma'(x)$.

(a) Compute the following derivatives in the recurrent neural network.

1. $\dfrac{\partial \mathcal{L}}{\partial y_3}$
2. $\dfrac{\partial \mathcal{L}}{\partial y_2}$
3. $\dfrac{\partial \mathcal{L}_3}{\partial h_3}$
4. $\dfrac{\partial h_t}{\partial h_{t-1}}$
5. $\dfrac{\partial \mathcal{L}}{\partial h_2}$

6. $\dfrac{\partial \mathcal{L}}{\partial h_1}$

There are 10 subtasks 1-6 and 1-4, each with 5 points

1.

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial y_3} &= \frac{\partial \mathcal{L}}{\partial \mathcal{L}_3} \cdot \frac{\partial \mathcal{L}_3}{\partial y_3} && (7) \\
&= 1 \cdot 2(t_3 - y_3)(-1) && (8) \\
& && (9)
\end{aligned}
$$

2.

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial y_2} &= \frac{\partial \mathcal{L}}{\partial \mathcal{L}_2} \cdot \frac{\partial \mathcal{L}_2}{\partial y_2} && (10) \\
&= 1 \cdot 2(t_2 - y_2)(-1) && (11) \\
& && (12)
\end{aligned}
$$

3.

$$
\begin{aligned}
\frac{\partial \mathcal{L}_3}{\partial h_3} &= \frac{\partial \mathcal{L}_3}{\partial y_3} \cdot \frac{\partial y_3}{\partial h_3} && (13) \\
&= \frac{\partial \mathcal{L}_3}{\partial y_3} \cdot \frac{\partial \sigma(W_{hy} h_3 + b_y)}{\partial h_3} && (14) \\
&= \frac{\partial \mathcal{L}_3}{\partial y_3} \cdot \sigma'(W_{hy} h_3 + b_y) W_{hy} && (15)
\end{aligned}
$$

4.

$$
\begin{aligned}
\frac{\partial h_t}{\partial h_{t-1}} &= \frac{\partial \sigma(W_{xh} x_t + W_{hh} h_{t-1} + b_h)}{\partial h_{t-1}} && (16) \\
&= \sigma'(W_{xh} x_t + W_{hh} h_{t-1} + b_h) W_{hh} && (17)
\end{aligned}
$$

5.

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial h_2} &= \frac{\partial (\mathcal{L}_2 + \mathcal{L}_3)}{\partial h_2} && (18) \\
&= \frac{\partial \mathcal{L}_2}{\partial y_2} \cdot \frac{\partial y_2}{\partial h_2} + \frac{\partial \mathcal{L}_3}{\partial y_3} \cdot \frac{\partial y_3}{\partial h_2} && (19) \\
&= \frac{\partial \mathcal{L}_2}{\partial y_2} \cdot \frac{\partial y_2}{\partial h_2} + \frac{\partial \mathcal{L}_3}{\partial y_3} \cdot \frac{\partial y_3}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_2} && (20)
\end{aligned}
$$

6.

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial h_1} &= \frac{\partial (\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3)}{\partial h_1} && (21) \\
&= \frac{\partial \mathcal{L}_1}{\partial y_1} \cdot \frac{\partial y_1}{\partial h_1} && (22) \\
&\quad + \frac{\partial \mathcal{L}_2}{\partial y_2} \cdot \frac{\partial y_2}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} && (23) \\
&\quad\quad + \frac{\partial \mathcal{L}_3}{\partial y_3} \cdot \frac{\partial y_3}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} && (24)
\end{aligned}
$$

(b) Building upon the intermediate gradients computed in the exercise above, compute the following gradients

1. $\dfrac{\partial h_t}{\partial W_{xh}}$

2. $\dfrac{\partial h_t}{\partial W_{hh}}$

3. $\frac{\partial \mathcal{L}}{\partial W_{xh}}$

4. $\frac{\partial \mathcal{L}}{\partial W_{hh}}$

1.

$$\begin{aligned}\frac{\partial h_t}{\partial W_{xh}} &= \frac{\partial \sigma(W_{xh}x_t + W_{hh}h_{t-1} + b_h)}{\partial W_{xh}} && (25)\\ &= \sigma'(W_{xh}x_t + W_{hh}h_{t-1} + b_h)x_t && (26)\end{aligned}$$

2.

$$\begin{aligned}\frac{h_t}{W_{hh}} &= \frac{\partial \sigma(W_{xh}x_t + W_{hh}h_{t-1} + b_h)}{\partial W_{hh}} && (27)\\ &= \sigma'(W_{xh}x_t + W_{hh}h_{t-1} + b_h)h_{t-1} && (28)\end{aligned}$$

3.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial W_{xh}} &= \frac{\partial(\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3)}{\partial W_{xh}} && (29)\\ &= \frac{\partial \mathcal{L}_1}{\partial y_1} \cdot \frac{\partial y_1}{\partial h_1} \cdot \frac{\partial h_1}{\partial W_{xh}} && (30)\\ &\quad + \frac{\partial \mathcal{L}_2}{\partial y_2} \cdot \frac{\partial y_2}{\partial h_2} \cdot (\frac{\partial h_2}{\partial W_{xh}} + \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial W_{xh}}) && (31)\\ &\quad + \frac{\partial \mathcal{L}_3}{\partial y_3} \cdot \frac{\partial y_3}{\partial h_3} \cdot (\frac{\partial h_3}{\partial W_{xh}} + \frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial W_{xh}} + \frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial W_{xh}}) && (32)\end{aligned}$$

4.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial W_{hh}} =& \frac{\partial(\mathcal{L}_2 + \mathcal{L}_3)}{\partial W_{hh}} && (33)\\ =& \frac{\partial \mathcal{L}_2}{\partial y_2} \cdot \frac{\partial y_2}{\partial h_2} \cdot \frac{\partial h_2}{\partial W_{hh}} && (34)\\ &+ \frac{\partial \mathcal{L}_3}{\partial y_3} \cdot \frac{\partial y_3}{\partial h_3} \cdot (\frac{\partial h_3}{\partial W_{hh}} + \frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial W_{hh}}) && (35)\end{aligned}$$

Exercise 2: Vanishing and Exploding Gradient Problem (20 P)

Extrapolating the gradient computation from 4. from just three steps to very long sequences, what can you deduce about the behaviour of the chained derivatives $\frac{\partial h_t}{\partial h_{t-1}}$ with respect to very small and very large values in W_{hh}? (Hint: repeated multiplications)
Answer: The derivative $\frac{\partial h_t}{\partial h_{t-1}} = \sigma'(W_{xh}x_t + W_{hh}h_{t-1} + b_h)W_{hh}$ is repeatedly multiplied with W_{hh} such that the repeated multiplication with W_{hh} results in either the original gradient $\frac{\partial \mathcal{L}_T}{\partial h_T}$ to be greatly depressed for small values of W_{hh} or increased/exploding for large values in W_{hh}.

Exercise 3: Programming (30 P)

Download the programming files on ISIS and follow the instructions.