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

Exercise Sheet 9

Exercise 1: Computing Gradients in RNNs (5×10 + 5×10 = 100 P)

We consider the task of binary classifying univariate time series (only two time steps for the purpose of the exercise) using a recurrent neural network. Let (x_1, x_2) be the time series given as input. The recurrent neural network is given by the equations:

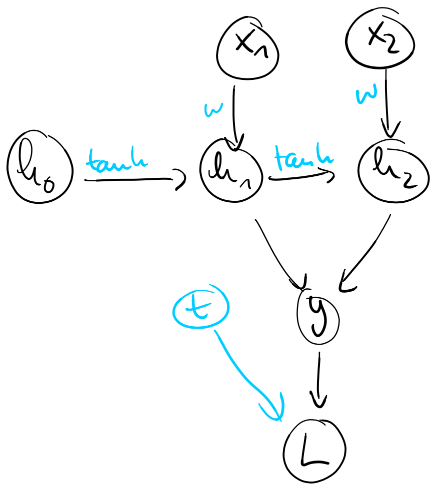
$$
\begin{aligned}
h_1 &= w \cdot x_1 + \tanh(h_0) \\
h_2 &= w \cdot x_2 + \tanh(h_1) \\
y &= h_1 + h_2,
\end{aligned}
$$

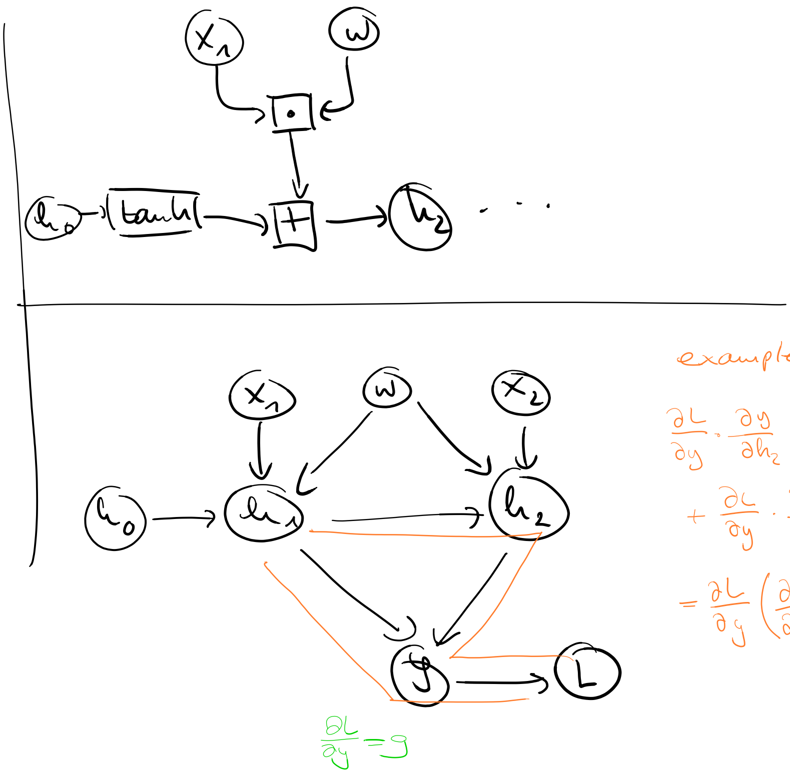
and we assume that the neural network has initial state $h_0 = 0$. The variable y is the neural network output and w is the model parameter. We further assume that the univariate time series (x_1, x_2) comes with a binary target label $t \in \{-1, 1\}$ and the prediction error for this data point is modeled via the log-loss function

$$
\mathcal{L}(y, t) = \log(1 + \exp(-yt)).
$$

We would like to extract the gradient of the objective w.r.t. the parameter w.

(a) Draw the neural network graph, and annotate it with relevant variables (inputs, activations, and parameters).





example $\frac{\partial L}{\partial h_1}$

$$
\frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} + \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial h_1} = \frac{\partial L}{\partial y}\left(\frac{\partial y}{\partial h_1} + \frac{\partial y}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1}\right)
$$

(b) *Compute $\partial \mathcal{L}/\partial y$.*

$$
\frac{\partial L}{\partial y} = \frac{\partial \log(1+\exp(-yt))}{\partial y} = \frac{\partial \log(1+\exp(-yt))}{\partial(1+\exp(-yt))} \cdot \frac{\partial(1+\exp(-yt))}{\partial(-yt)} \cdot \frac{\partial(-yt)}{\partial y}
$$

$$
= \frac{1}{1+\exp(-yt)} \cdot \exp(-yt) \cdot (-t) = -\frac{t \cdot \exp(-yt)}{1+\exp(-yt)}
$$

$\overset{\text{opt.}}{=} -t \cdot \mathrm{sigm}(-yt)$

(c) Assuming the last computation was stored in g, *compute* $\partial\mathcal{L}/\partial h_2$ as a function of g.

$$\frac{\partial L}{\partial h_2} = \frac{\partial L}{\partial y}\cdot\frac{\partial y}{\partial h_2} = g\cdot\frac{\partial(h_1+h_2)}{\partial h_2} = g\left(\underbrace{\frac{\partial h_1}{\partial h_2}}_{=0} + \underbrace{\frac{\partial h_2}{\partial h_2}}_{=1}\right)$$

$$= g = \delta_2$$

(d) Assuming the last computation was stored in δ_2, *compute* $\partial\mathcal{L}/\partial h_1$ as a function of g and δ_2.

$$\frac{\partial L}{\partial h_1} = \frac{\partial L}{\partial y}\cdot\frac{\partial y}{\partial h_1} = \frac{\partial L}{\partial y}\cdot\frac{\partial(h_1+h_2)}{\partial h_1} = \underbrace{\frac{\partial L}{\partial y}}_{=g}\cdot\left(\underbrace{\frac{\partial h_1}{\partial h_1}}_{=1} + \frac{\partial h_2}{\partial h_1}\right)$$

$$= g + g\cdot\frac{\partial(x_2 w + \tanh(h_1))}{\partial h_1} = g + g\cdot\left(\underbrace{\frac{\partial x_2 w}{\partial h_1}}_{=0} + \underbrace{\frac{\partial \tanh(h_1)}{\partial h_1}}_{=\tanh'(h_1)}\right)$$

$$= g + g\cdot\tanh'(h_1) = g\left(1+\tanh'(h_1)\right)$$

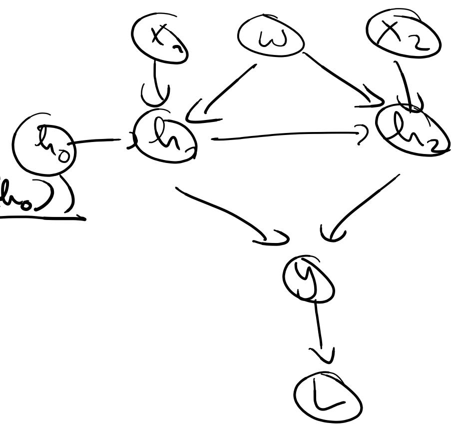
$$= \delta_2\left(1+\tanh'(h_1)\right)$$

(e) Assuming the last computation was stored in δ_1, *compute* $\partial\mathcal{L}/\partial w$ as a function of g, δ_2 and δ_1.

$$\frac{\partial L}{\partial w} = \delta_2\cdot\frac{\partial^+ h_2}{\partial w} + \delta_1\cdot\frac{\partial^+ h_1}{\partial w}$$

$$= \delta_2\cdot\frac{\partial^+(x_2 w + \tanh(h_1))}{\partial w} + \delta_1\cdot\frac{\partial^+(x_1 w + \tanh(h_0))}{\partial w}$$

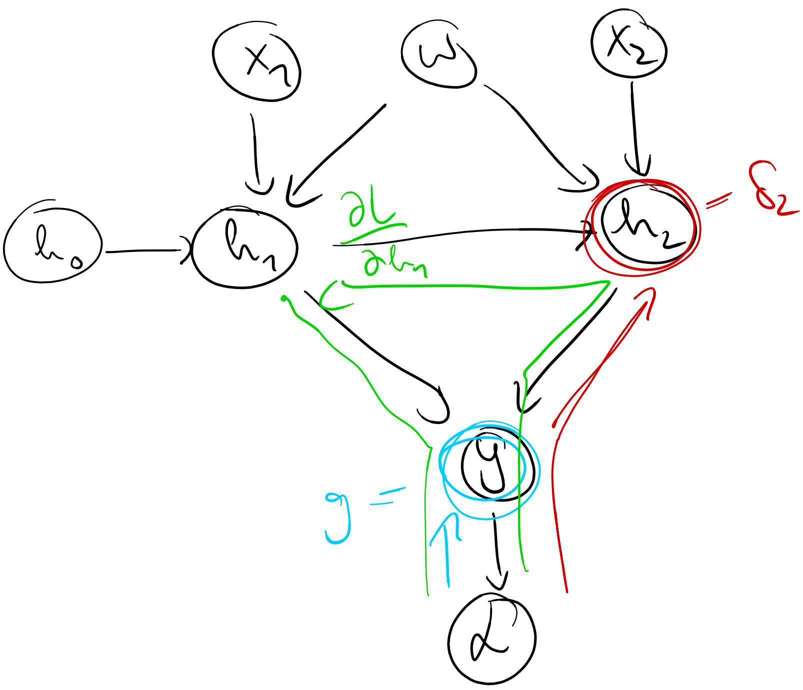
$$= \delta_2\cdot x_2 + \delta_1\cdot x_1$$



d) with direct derivative ∂^+

$$\frac{\partial y}{\partial h_1} = \frac{\partial h_1 + h_2}{\partial h_1} = \underbrace{1}_{y \to h_1} + \underbrace{\tanh'(h_1)}_{y \to h_2 \to h_1}$$

$$\frac{\partial^+ y}{\partial h_1} = \frac{\partial^+ (h_1 + h_2)}{\partial h_1} = \frac{\partial h_1}{\partial h_1}$$



$$\frac{\partial L}{\partial y} = \delta_2 \cdot \frac{\partial^+ h_2}{\partial h_1} + g \cdot \frac{\partial^+ y}{\partial h_1}$$

$$= \delta_2 \cdot \tanh'(h_1)$$

$$+ g \cdot 1$$

$$= g \cdot \tanh'(h_1) + g$$

$$= g(1 + \tanh'(h_1))$$

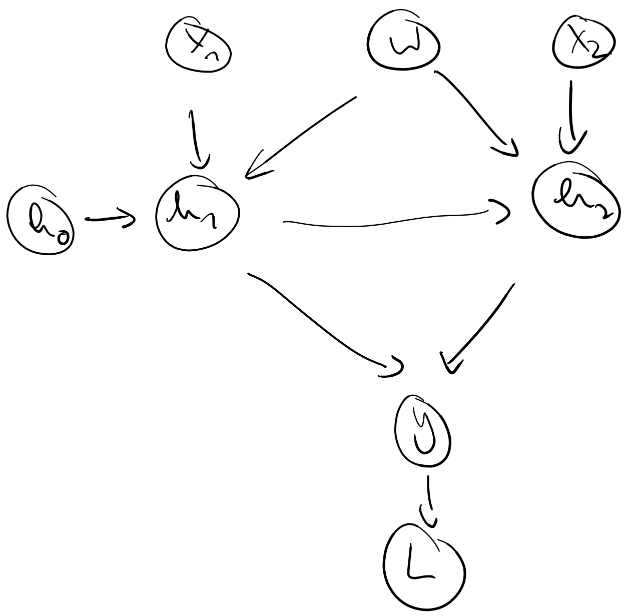
(f) Repeat the steps above (a–e) for the case where the recurrent neural network is given by the equations:

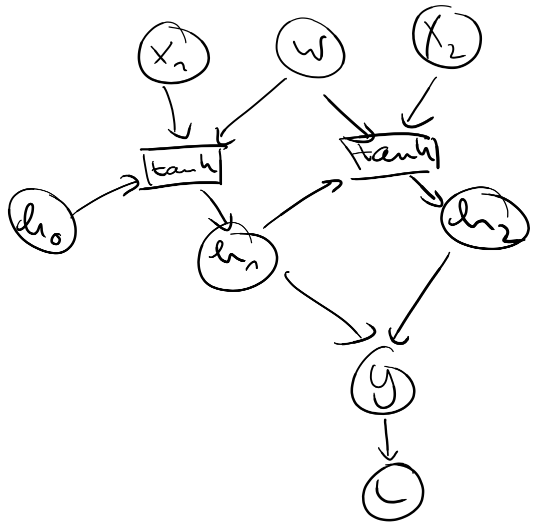
$$\begin{aligned} h_1 &= \tanh(x_1 + w + h_0) \\ h_2 &= \tanh(x_2 + w + h_1) \\ y &= h_1 + h_2, \end{aligned}$$

where the initial state is set to $h_0 = 0$, the target is real-valued ($t \in \mathbb{R}$), and the error function is given by

$$\mathcal{L}(y,t) = \log\cosh(y - t).$$

f_a)





f_b)

$$\frac{\partial L}{\partial y} = \underbrace{\frac{\partial \log(\cosh(y-t))}{\partial \cosh(y-t)}}_{= \frac{1}{\cosh\ldots}} \cdot \underbrace{\frac{\partial \cosh(y-t)}{\partial (y-t)}}_{=\sinh(\ldots)} \cdot \underbrace{\frac{\partial (y-t)}{\partial y}}_{=1}$$

$$= \frac{1}{\cosh(y-t)} \cdot \sinh(y-t) \cdot 1 = \frac{\sinh(y-t)}{\cosh(y-t)}$$

opt: $\tanh(y-t) = g$

f_c)

$$\frac{\partial L}{\partial h_2} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial h_2} = g \cdot \frac{\partial h_1 + h_2}{\partial h_2} = g\left(\underbrace{\frac{\partial h_1}{\partial h_2}}_{=0} + \underbrace{\frac{\partial h_2}{\partial h_2}}_{=1}\right)$$

$$= g =: \delta_2$$

f_d)

$$\frac{\partial L}{\partial h_1} = g \cdot \underbrace{\frac{\partial^+ y}{\partial h_1}}_{=1} + \delta_2 \cdot \frac{\partial^+ h_2}{\partial h_1} = g \cdot 1 + \delta_2 \cdot \frac{\partial^+ \tanh(x_2 + w + h_1)}{\partial h_1}$$

$$= g + \delta_2 \cdot \frac{\partial \tanh(x_2 + w + h_1)}{\partial(x_2 + w + h_1)} \cdot \underbrace{\frac{\partial^+ (x_2 + w + h_1)}{\partial h_1}}_{=1}$$

$$= g + \underbrace{\delta_2}_{=g} \cdot \tanh'(x_2 + w + h_1)$$

$$= g(1 + \tanh'(x_2 + w + h_1)) = \delta_1$$

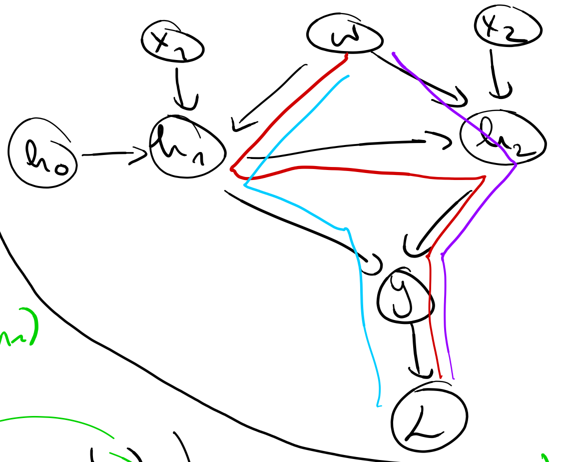
f_e)

$$\frac{\partial L}{\partial w} = \delta_1 \cdot \frac{\partial^+ h_1}{\partial w} + \delta_2 \cdot \frac{\partial^+ h_2}{\partial w}$$

$$= \delta_1 \cdot \frac{\partial^+ \tanh(x_1 + w + h_0)}{\partial w} + \delta_2 \cdot \frac{\partial^+ \tanh(x_2 + w + h_1)}{\partial w}$$

$$= \delta_1 \cdot \tanh'(x_1 + w + h_0) + \delta_2 \cdot \tanh'(x_2 + w + h_1)$$

f.e) with normal derivative



$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial y}\cdot\frac{\partial y}{\partial w} = \frac{\partial L}{\partial y}\cdot\frac{\partial y}{\partial(h_1+h_2)}\cdot\frac{\partial h_1+h_2}{\partial w}$$

$$= \frac{\partial L}{\partial y}\cdot\frac{\partial y}{\partial(h_1+h_2)}\left(\frac{\partial h_1}{\partial w}+\frac{\partial h_2}{\partial w}\right)$$

$$= \frac{\partial L}{\partial y}\cdot\frac{\partial y}{\partial(h_1+h_2)}\left(\frac{\partial h_1}{\partial w}+\overset{= \tanh(x_2+w+h_1)}{\frac{\partial h_2}{\partial(x_2+w+h_1)}}\cdot\frac{\partial(x_2+w+h_1)}{\partial w}\right) \Rightarrow = \underbrace{\frac{\partial x_2}{\partial w}}_{=0}+\frac{\partial w}{\partial w}+\frac{\partial h_1}{\partial w}$$

$$= \frac{\partial L}{\partial y}\cdot\frac{\partial y}{\partial h_1+h_2}\left(\frac{\partial h_1}{\partial w}+\frac{\partial h_2}{\partial(x_2+w+h_1)}\left(\frac{\partial w}{\partial w}+\frac{\partial h_1}{\partial w}\right)\right)$$

$$= \underbrace{\frac{\partial L}{\partial y}\cdot\frac{\partial y}{\partial h_1+h_2}\cdot\frac{\partial h_1}{\partial w}}_{L \to y \to h_1 \to w} + \underbrace{\frac{\partial L}{\partial y}\cdot\frac{\partial y}{\partial h_1+h_2}\cdot\frac{\partial h_2}{\partial(x_2+w+h_1)}\cdot\frac{\partial w}{\partial w}}_{L \to y \to h_2 \to w}$$

$$+ \underbrace{\frac{\partial L}{\partial y}\cdot\frac{\partial y}{\partial h_1+h_2}\cdot\frac{\partial h_2}{\partial(x_2+w+h_1)}\cdot\frac{\partial h_1}{\partial w}}_{L \to y \to h_2 \to h_1 \to w}$$

$$\frac{\partial L}{\partial h_1}\cdot\frac{\partial h_1}{\partial w}+\frac{\partial L}{\partial h_2}\cdot\frac{\partial h_2}{\partial w}$$

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial y}\cdot\frac{\partial y}{\partial w} = g\cdot\frac{\partial(h_1+h_2)}{\partial w} = g\left(\frac{\partial h_1}{\partial w}+\frac{\partial h_2}{\partial w}\right)$$

$$= g\left(\underbrace{\frac{\partial \tanh(x_1+w+h_0)}{\partial w}}_{\tanh'(x_1+w+h_0)} + \underbrace{\frac{\partial \tanh(x_2+w+h_1)}{\partial w}}_{=\tanh'(x_2+w+h_1)\cdot\left(\frac{\partial h_1}{\partial w}+\frac{\partial w}{\partial w}\right)}\right)$$

$$= g\left(\tanh'(x_1+w+h_0) + \tanh'(x_2+w+h_1)\left(\frac{\partial h_1}{\partial w}+\underbrace{\frac{\partial w}{\partial w}}_{=1}\right)\right)$$

$$= g\left(\tanh'(x_1+w+h_0)+\tanh'(x_2+w+h_1)\left(\tanh'(x_1+w+h_0)+1\right)\right)$$

$$= \overset{=\delta_2}{\widetilde{g}} \cdot \tanh'(x_2 + w + h_1)$$
$$+ g \cdot \tanh'(x_1 + w + h_0)$$
$$+ g \cdot \tanh'(x_2 + w + h_1)\tanh'(x_1 + w + h_0)$$

$$= \delta_2 \cdot \tanh'(x_2 + w + h_1) + \tanh'(x_1 + w + h_0)\big(g + g \tanh'(x_2 + w + h_1)\big)$$

$$= \delta_2 \cdot \tanh'(x_2 + w + h_1) + \tanh'(x_1 + w + h_0)\Big(\underbrace{g\big(1 + \tanh'(x_2 + w + h_1)\big)}_{=\delta_1}\Big)$$

$$= \delta_2 \cdot \tanh'(x_2 + w + h_1) + \delta_1 \cdot \tanh'(x_1 + w + h_0)$$

$$= \delta_2 \cdot \frac{\partial^+ h_2}{\partial w} + \delta_1 \cdot \frac{\partial^+ h_1}{\partial w}$$

// compare with solution using direct derivative