IFT6135-H2019 - Representation learning

Assignment 2, Theoretical Part

Sanae Lotfi

Num de matricule (Poly) : 1968682
Num de matricule (Udem) : 20147309

25th March, 2019

Due Date: March 22nd 23:59, 2019

Instructions

- *For all questions, show your work!*
- *Starred questions are* ***hard*** *questions, not* ***bonus*** *questions.*
- *Please use a document preparation system such as LaTeX, unless noted otherwise.*
- *Unless noted that questions are related, assume that notation and defintions for each question are self-contained and independent*
- *Submit your answers electronically via Gradescope.*
- ***TAs for this assignment are David Krueger, Tegan Maharaj, and Chin-Wei Huang.***

Question 1 (6-10)**.** The goal of this question is for you to understand the reasoning behind different parameter initializations for deep networks, particularly to think about the ways that the initialization affects the activations (and therefore the gradients) of the network. Consider the following equation for the t-th layer of a deep network:

$$\boldsymbol{h}^{(t)} = g(\boldsymbol{a}^{(t)}) \qquad\qquad \boldsymbol{a}^{(t)} = \boldsymbol{W}^{(t)}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}^{(t)}$$

where $\boldsymbol{a}^{(t)}$ are the preactivations and $\boldsymbol{h}^{(t)}$ are the activations for layer t, g is an activation function, $\boldsymbol{W}^{(t)}$ is a $d^{(t)} \times d^{(t-1)}$ matrix, and $\boldsymbol{b}^{(t)}$ is a $d^{(t)} \times 1$ bias vector. The bias is initialized as a constant vector $\boldsymbol{b}^{(t)} = [c, .., c]^\top$ for some $c \in \mathbb{R}$, and the entries of the weight matrix are initialized by sampling i.i.d. from either (a) a Gaussian distribution $\boldsymbol{W}^{(t)}_{ij} \sim \mathcal{N}(\mu, \sigma^2)$, or (b) a Uniform distribution $\boldsymbol{W}^{(t)}_{ij} \sim U(\alpha, \beta)$.

For both of the assumptions (1 and 2) about the distribution of the inputs to layer t listed below, and for both (a) Gaussian, and (b) Uniform sampling, design an initialization scheme that would achieve preactivations with zero-mean and unit variance at layer t, i.e.: $\mathbb{E}[\boldsymbol{a}^{(t)}_i] = 0$ and $\mathrm{Var}(\boldsymbol{a}^{(t)}_i) = 1$, for $1 \leq i \leq d^{(t)}$.

(Hint: if $X \perp Y$, $\mathrm{Var}(XY) = \mathrm{Var}(X)\mathrm{Var}(Y) + \mathrm{Var}(X)\mathbb{E}[Y]^2 + \mathrm{Var}(Y)\mathbb{E}[X]^2$)

1. Assume $\mathbb{E}[\boldsymbol{h}^{(t-1)}_i] = 0$ and $\mathrm{Var}(\boldsymbol{h}^{(t-1)}_i) = 1$ for $1 \leq i \leq d^{(t-1)}$. Assume entries of $\boldsymbol{h}^{(t-1)}$ are uncorrelated (the answer should not depend on g).
   (a) Gaussian: give values for c, μ, and σ^2 as a function of $d^{(t-1)}$.
   (b) Uniform: give values for c, α, and β as a function of $d^{(t-1)}$.
2. Assume that the preactivations of the previous layer satisfy $\mathbb{E}[\boldsymbol{a}^{(t-1)}_i] = 0$, $\mathrm{Var}(\boldsymbol{a}^{(t-1)}_i) = 1$ and $\boldsymbol{a}^{(t-1)}_i$ has a symmetric distribution for $1 \leq i \leq d^{(t-1)}$. Assume entries of $\boldsymbol{a}^{(t-1)}$ are uncorrelated. Consider the case of ReLU activation: $g(x) = \max\{0, x\}$.
   (a) Gaussian: give values for c, μ, and σ^2 as a function of $d^{(t-1)}$.
   (b) Uniform: give values for c, α, and β as a function of $d^{(t-1)}$.
   (c) What popular initialization scheme has this form?
   (d) Why do you think this initialization would work well in practice? Answer in 1-2 sentences.

Answer 1.

1. In this question, we assume $\mathbb{E}[\boldsymbol{h}_i^{(t-1)}] = 0$ and $\mathrm{Var}(\boldsymbol{h}_i^{(t-1)}) = 1$ for $1 \leq i \leq d^{(t-1)}$. We also assume entries of $\boldsymbol{h}^{(t-1)}$ are uncorrelated.

   We will first derive general expressions then answer the particular cases of questions (1.a) and (1.b)

   For $1 \leq i \leq d^{(t-1)}$ and fixed t-th layer, we have that: $\boldsymbol{a}_i^{(t)} = \sum_{j=1}^{d^{(t-1)}} \boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)} + \boldsymbol{b}_i^{(t)}$. Thus, we have:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{a}_i^{(t)}] &= \mathbb{E}[\sum_{j=1}^{d^{(t-1)}} \boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)} + \boldsymbol{b}_i^{(t)}] \\
&= \sum_{j=1}^{d^{(t-1)}} \mathbb{E}[\boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)}] + \mathbb{E}[vb_i^{(t)}] \quad \text{(linearity of the expectation)} \\
&= \sum_{j=1}^{d^{(t-1)}} \mathbb{E}[\boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)}] + c \quad \text{(by definition of the bias } \mathbb{E}[\boldsymbol{b}_i^{(t)}] = \mathbb{E}[c] = c)
\end{aligned}$$

   Since the entries of the weight matrix are sampled independently for the t-th layer, we have that : $\boldsymbol{W}_{ij}^{(t)} \perp \boldsymbol{h}_j^{(t-1)}$, thus, using the property of the expectation of the product of two independent variables, we have: $\mathbb{E}[\boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)}] = \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]\mathbb{E}[\boldsymbol{h}_j^{(t-1)}]$. Then:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{a}_i^{(t)}] &= \sum_{j=1}^{d^{(t-1)}} \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]\mathbb{E}[\boldsymbol{h}_j^{(t-1)}] + c \\
&= c \quad \text{(because } \mathbb{E}[\boldsymbol{h}_i^{(t-1)}] = 0)
\end{aligned}$$

   Thus:

$$\mathbb{E}[\boldsymbol{a}_i^{(t)}] = c \tag{1}$$

   For the variance, we have:

$$\begin{aligned}
\mathrm{Var}[\boldsymbol{a}_i^{(t)}] &= \mathrm{Var}[\sum_{j=1}^{d^{(t-1)}} \boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)} + \boldsymbol{b}_i^{(t)}] \\
&= \mathrm{Var}[\sum_{j=1}^{d^{(t-1)}} \boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)} + c] \\
&= \mathrm{Var}[\sum_{j=1}^{d^{(t-1)}} \boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)}] \quad \text{(Variance of a constant is equal to zero)}
\end{aligned}$$

   Since the entries of the weight matrix are sampled independently for the t-th layer, we have that they are independent of each other and of the $\boldsymbol{h}_j^{(t-1)}$: $\mathbb{E}[(\boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)})(\boldsymbol{W}_{kl}^{(t)} \boldsymbol{h}_l^{(t-1)})] = \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]\mathbb{E}[\boldsymbol{W}_{kl}^{(t)}]\mathbb{E}[\boldsymbol{h}_j^{(t-1)} \boldsymbol{h}_l^{(t-1)}]$. Since we assume that entries of $\boldsymbol{h}^{(t-1)}$ are uncorrelated, we have $\boldsymbol{h}_j^{(t-1)} \perp \boldsymbol{h}_l^{(t-1)}$, thus:

$$\begin{aligned}\mathbb{E}[(\boldsymbol{W}_{ij}^{(t)}\boldsymbol{h}_j^{(t-1)})(\boldsymbol{W}_{kl}^{(t)}\boldsymbol{h}_l^{(t-1)})] &= \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]\mathbb{E}[\boldsymbol{W}_{kl}^{(t)}]\mathbb{E}[\boldsymbol{h}_j^{(t-1)}]\mathbb{E}[\boldsymbol{h}_l^{(t-1)}] \\ &= \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}\boldsymbol{h}_j^{(t-1)}]\mathbb{E}[\boldsymbol{W}_{kl}^{(t)}\boldsymbol{h}_l^{(t-1)}]\end{aligned}$$

We proved by this that $\boldsymbol{W}_{ij}^{(t)}\boldsymbol{h}_j^{(t-1)}$ and $\boldsymbol{W}_{kl}^{(t)}\boldsymbol{h}_l^{(t-1)}$ are uncorrelated, thus we can write:

$$\mathrm{Var}[\boldsymbol{a}_i^{(t)}] = \sum_{j=1}^{d^{(t-1)}} \mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}\boldsymbol{h}_j^{(t-1)}]$$

Using the hint and $\boldsymbol{W}_{ij}^{(t)} \perp \boldsymbol{h}_j^{(t-1)}$, we obtain

$$\mathrm{Var}[\boldsymbol{a}_i^{(t)}] = \sum_{j=1}^{d^{(t-1)}} \left[\mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}]\mathrm{Var}[\boldsymbol{h}_j^{(t-1)}] + \mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}]\mathbb{E}[\boldsymbol{h}_j^{(t-1)}]^2 + \mathrm{Var}[\boldsymbol{h}_j^{(t-1)}]\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]^2\right] \tag{2}$$

(a) Gaussian:

Using the two equations (1) and (2), having $\mathbb{E}[\boldsymbol{a}_i^{(t)}] = 0$ and $\mathrm{Var}(\boldsymbol{a}_i^{(t)}) = 1$ is equivalent to having:

$$\begin{cases}\mathbb{E}[\boldsymbol{a}_i^{(t)}] = c = 0 \\ \mathrm{Var}[\boldsymbol{a}_i^{(t)}] = \sum_{j=1}^{d^{(t-1)}} \left[\mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}]\mathrm{Var}[\boldsymbol{h}_j^{(t-1)}] + \mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}]\mathbb{E}[\boldsymbol{h}_j^{(t-1)}]^2 + \mathrm{Var}[\boldsymbol{h}_j^{(t-1)}]\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]^2\right] = 1\end{cases}$$

with: $\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] = \mu$, $\mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}] = \sigma^2$, $\mathbb{E}[\boldsymbol{h}_i^{(t-1)}] = 0$ and $\mathrm{Var}(\boldsymbol{h}_i^{(t-1)}) = 1$, which gives:

$$\begin{cases}c = 0 \\ \sum_{j=1}^{d^{(t-1)}}[\sigma^2 + 0 + \mu^2] = 1\end{cases} \implies \begin{cases}c = 0 \\ [\sigma^2 + \mu^2]d^{(t-1)} = 1\end{cases}$$

We can choose: $\boxed{c = 0,\ \mu = 0,\ \sigma^2 = \frac{1}{d^{(t-1)}}}$.

(b) Uniform:

Using the two equations (1) and (2), having $\mathbb{E}[\boldsymbol{a}_i^{(t)}] = 0$ and $\mathrm{Var}(\boldsymbol{a}_i^{(t)}) = 1$ is equivalent to having:

$$\begin{cases}\mathbb{E}[\boldsymbol{a}_i^{(t)}] = c = 0 \\ \mathrm{Var}[\boldsymbol{a}_i^{(t)}] = \sum_{j=1}^{d^{(t-1)}} \left[\mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}]\mathrm{Var}[\boldsymbol{h}_j^{(t-1)}] + \mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}]\mathbb{E}[\boldsymbol{h}_j^{(t-1)}]^2 + \mathrm{Var}[\boldsymbol{h}_j^{(t-1)}]\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]^2\right] = 1\end{cases}$$

with: $\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] = \frac{\alpha+\beta}{2}$, $\mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}] = \frac{1}{12}(\beta-\alpha)^2$, $\mathbb{E}[\boldsymbol{h}_i^{(t-1)}] = 0$ and $\mathrm{Var}(\boldsymbol{h}_i^{(t-1)}) = 1$, which gives:

$$\begin{cases}c = 0 \\ \sum_{j=1}^{d^{(t-1)}}[\frac{1}{12}(\beta-\alpha)^2 + 0 + \frac{(\alpha+\beta)^2}{4}] = 1\end{cases} \implies \begin{cases}c = 0 \\ [\alpha^2 + \beta^2 + \alpha\beta]d^{(t-1)} = 3\end{cases}$$

We can choose: $\boxed{c = 0,\ \alpha = -\sqrt{\frac{3}{d^{(t-1)}}},\ \beta = \sqrt{\frac{3}{d^{(t-1)}}}}$.

2. In this question, we assume that the preactivations of the previous layer satisfy $\mathbb{E}[\boldsymbol{a}_i^{(t-1)}] = 0$, $\text{Var}(\boldsymbol{a}_i^{(t-1)}) = 1$ and $\boldsymbol{a}_i^{(t-1)}$ has a symmetric distribution for $1 \leq i \leq d^{(t-1)}$. We also assume entries of $\boldsymbol{a}^{(t-1)}$ are uncorrelated and consider the case of ReLU activation: $g(x) = \max\{0, x\}$.

   We will first derive general expressions then answer the particular cases of questions (2.a) and (2.b)

   For $1 \leq i \leq d^{(t-1)}$ and fixed t-th layer, we have that: $\boldsymbol{a}_i^{(t)} = \sum_{j=1}^{d^{(t-1)}} \boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)} + \boldsymbol{b}_i^{(t)}$. Thus, we have:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{a}_i^{(t)}] &= \mathbb{E}[\sum_{j=1}^{d^{(t-1)}} \boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)} + \boldsymbol{b}_i^{(t)}] \\
&= \sum_{j=1}^{d^{(t-1)}} \mathbb{E}[\boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)}] + \mathbb{E}[vb_i^{(t)}] \quad \text{(linearity of the expectation)} \\
&= \sum_{j=1}^{d^{(t-1)}} \mathbb{E}[\boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)}] + c \quad \text{(by definition of the bias } \mathbb{E}[\boldsymbol{b}_i^{(t)}] = \mathbb{E}[c] = c\text{)}
\end{aligned}$$

   Since the entries of the weight matrix are sampled independently for the t-th layer, we have that : $\boldsymbol{W}_{ij}^{(t)} \perp \boldsymbol{h}_j^{(t-1)}$, thus, using the property of the expectation of the product of two independent variables, we have: $\mathbb{E}[\boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)}] = \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]\mathbb{E}[\boldsymbol{h}_j^{(t-1)}]$. Then:

$$\mathbb{E}[\boldsymbol{a}_i^{(t)}] = \sum_{j=1}^{d^{(t-1)}} \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]\mathbb{E}[\boldsymbol{h}_j^{(t-1)}] + c \tag{3}$$

   For the variance, we have:

$$\begin{aligned}
\text{Var}[\boldsymbol{a}_i^{(t)}] &= \text{Var}[\sum_{j=1}^{d^{(t-1)}} \boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)} + \boldsymbol{b}_i^{(t)}] \\
&= \text{Var}[\sum_{j=1}^{d^{(t-1)}} \boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)} + c] \\
&= \text{Var}[\sum_{j=1}^{d^{(t-1)}} \boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)}] \quad \text{(Variance of a constant is equal to zero)}
\end{aligned}$$

   We want to show that the entries of $h^{(t-1)}$ are uncorrelated.
   Using the law of total expectation we have for all i:

$$\mathbb{E}\big[h_i^{(t-1)}\big] = \mathbb{P}(a_i^{(t-1)} \geq 0)\mathbb{E}\big[h_i^{(t-1)}|a_i^{(t-1)} \geq 0\big] + \mathbb{P}(a_i^{(t-1)} \leq 0)\mathbb{E}\big[h_i^{(t-1)}|a_i^{(t-1)} \leq 0\big]$$

   And using the fact that:

$$h_i^{(t-1)} = \begin{cases} a_i^{(t-1)} & \text{if } a_i^{(t-1)} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

We obtain:

$$\mathbb{E}\big[h_i^{(t-1)}\big] = \mathbb{P}(a_i^{(t-1)} \geq 0)\mathbb{E}\big[h_i^{(t-1)}|a_i^{(t-1)} \geq 0\big]$$

However, the distribution of a_i is symmetric therefore $\mathbb{P}(a_i^{(t-1)} \geq 0) = \frac{1}{2}$, thus:

$$\mathbb{E}\big[h_i^{(t-1)}\big] = \frac{1}{2}\mathbb{E}\big[a_i^{(t-1)}|a_i^{(t-1)} \geq 0\big]$$

We apply the same approach to $\mathbb{E}\big[h_i^{(t-1)}h_j^{(t-1)}\big]$ for all $i \neq j$:

$$\begin{aligned}
\mathbb{E}\big[h_i^{(t-1)}h_j^{(t-1)}\big] &= \mathbb{P}(a_i^{(t-1)} \geq 0)\mathbb{E}\big[h_i^{(t-1)}h_j^{(t-1)}|a_i^{(t-1)} \geq 0\big] + \mathbb{P}(a_i^{(t-1)} \leq 0)\mathbb{E}\big[h_i^{(t-1)}h_j^{(t-1)}|a_i^{(t-1)} \leq 0\big] \\
&= \frac{1}{2}\mathbb{E}\big[a_i^{(t-1)}h_j^{(t-1)}|a_i^{(t-1)} \geq 0\big] \\
&= \frac{1}{2}\Big(\mathbb{P}(a_j^{(t-1)} \geq 0)\mathbb{E}\big[a_i^{(t-1)}h_j^{(t-1)}|a_i^{(t-1)} \geq 0, a_j^{(t-1)} \geq 0\big] \\
&+ \mathbb{P}(a_j^{(t-1)} \leq 0)\mathbb{E}\big[a_i^{(t-1)}h_j^{(t-1)}|a_i^{(t-1)} \leq 0, a_j^{(t-1)} \leq 0\big]\Big) \\
&= \frac{1}{4}\mathbb{E}\big[a_i^{(t-1)}a_j^{(t-1)}|a_i^{(t-1)} \geq 0, a_j^{(t-1)} \geq 0\big]
\end{aligned}$$

Now we can use the assumption that the entries of $a_i^{(t-1)}$ are uncorrelated $i \neq j$, therefore $(a_i^{(t-1)}|a_i^{(t-1)})$ and $(a_j^{(t-1)}|a_j^{(t-1)})$ are uncorrelated for all $i \neq j$ ie.

$$\mathbb{E}\big[a_i^{(t-1)}a_j^{(t-1)}|a_i^{(t-1)} \geq 0, a_j^{(t-1)} \geq 0\big] = \mathbb{E}\big[a_i^{(t-1)}|a_i^{(t-1)} \geq 0\big]\mathbb{E}\big[a_j^{(t-1)}|a_j^{(t-1)} \geq 0\big]$$

This allows to conclude that:

$$\mathbb{E}\big[h_i^{(t-1)}h_j^{(t-1)}\big] = \mathbb{E}\big[h_i^{(t-1)}\big]\mathbb{E}\big[h_j^{(t-1)}\big]$$

So the entries of $h^{(t-1)}$ are indeed uncorrelated. We also have that $\boldsymbol{W}_{ij}^{(t)}$ are iid. Combining the two information, we can prove in the same fashion as before that $\boldsymbol{W}_{ij}^{(t)}\boldsymbol{h}_j^{(t-1)}$ and $\boldsymbol{W}_{kl}^{(t)}\boldsymbol{h}_l^{(t-1)}$ are uncorrelated, thus we can write:

$$\mathrm{Var}[\boldsymbol{a}_i^{(t)}] = \sum_{j=1}^{d^{(t-1)}} \mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}\boldsymbol{h}_j^{(t-1)}]$$

Using the hint and $\boldsymbol{W}_{ij}^{(t)} \perp \boldsymbol{h}_j^{(t-1)}$, we obtain

$$\begin{aligned}
\mathrm{Var}[\boldsymbol{a}_i^{(t)}] &= \sum_{j=1}^{d^{(t-1)}} \Big[\mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}]\mathrm{Var}[\boldsymbol{h}_j^{(t-1)}] + \mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}]\mathbb{E}[\boldsymbol{h}_j^{(t-1)}]^2 + \mathrm{Var}[\boldsymbol{h}_j^{(t-1)}]\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]^2\Big] \\
&= \sum_{j=1}^{d^{(t-1)}} \Big[\mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}]\mathbb{E}[(\boldsymbol{h}_j^{(t-1)})^2] + \mathrm{Var}[\boldsymbol{h}_j^{(t-1)}]\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]^2\Big]
\end{aligned} \tag{4}$$

(using that: $\text{Var}[\boldsymbol{h}_j^{(t-1)}] = \mathbb{E}[(\boldsymbol{h}_j^{(t-1)})^2] - \mathbb{E}[\boldsymbol{h}_j^{(t-1)}]^2$)
Now we want to prove that $\mathbb{E}\big[h_i^{(t-1)2}\big] = \frac{1}{2}$ for all i.
Let $1 \leq i \leq d^{(t-1)}$ and f_i the density of $a_i^{(t-1)}$.
We have :

$$\mathbb{E}\big[(h_i^{(t-1)})^2\big] = \int_{\mathbb{R}} \max(0,x)^2 f_i(x)dx = \int_0^{+\infty} x^2 f_i(x)dx$$

On the other hand, we also have:

$$\begin{aligned}
\mathbb{E}\big[(a_i^{(t-1)})^2\big] &= \int_{\mathbb{R}} x^2 f_i(x)dx \\
&= \int_0^{+\infty} x^2 f_i(x)dx + \int_{-\infty}^0 x^2 f_i(x)dx \\
&= \int_0^{+\infty} x^2 f_i(x)dx + \int_0^{+\infty} x^2 f_i(-x)dx \quad \text{(change of variable t=-x)} \\
&= 2\int_0^{+\infty} x^2 f_i(x)dx \quad \text{(symmetry of the distribution ie. } f_i(x) = f_i(-x)\text{)} \\
&= 2\mathbb{E}\big[(h_i^{(t-1)})^2\big]
\end{aligned}$$

Since $\text{Var}\big[a_i^{(t-1)}\big] = 1$, we conclude that $\mathbb{E}\big[h_i^{(t-1)2}\big] = \frac{1}{2}$.
Thus:

$$\text{Var}[\boldsymbol{a}_i^{(t)}] = \sum_{j=1}^{d^{(t-1)}} \left[\frac{1}{2}\text{Var}[\boldsymbol{W}_{ij}^{(t)}] + \text{Var}[\boldsymbol{h}_j^{(t-1)}]\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]^2\right] \tag{5}$$

(a) Gaussian:

Using the two equations (3) and (5), having $\mathbb{E}[\boldsymbol{a}_i^{(t)}] = 0$ and $\text{Var}(\boldsymbol{a}_i^{(t)}) = 1$ is equivalent to having:

$$\begin{cases} \sum_{j=1}^{d^{(t-1)}} \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]\mathbb{E}[\boldsymbol{h}_j^{(t-1)}] + c = 0 \\ \sum_{j=1}^{d^{(t-1)}} \left[\frac{1}{2}\text{Var}[\boldsymbol{W}_{ij}^{(t)}] + \text{Var}[\boldsymbol{h}_j^{(t-1)}]\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]^2\right] = 1 \end{cases}$$

We know $\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] = \mu$, $\text{Var}[\boldsymbol{W}_{ij}^{(t)}] = \sigma^2$. We take $\boxed{\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] = \mu = 0}$, which gives:

$$\begin{cases} c = 0 \\ \sum_{j=1}^{d^{(t-1)}} \left[\frac{1}{2}\sigma^2 + 0\right] = 1 \end{cases}$$

We can choose: $\boxed{c = 0,\ \mu = 0,\ \sigma^2 = \dfrac{2}{d^{(t-1)}}}$.

(b) Uniform:

Using the two equations (3) and (5), having $\mathbb{E}[\boldsymbol{a}_i^{(t)}] = 0$ and $\text{Var}(\boldsymbol{a}_i^{(t)}) = 1$ is equivalent to having:

$$\begin{cases} \sum_{j=1}^{d^{(t-1)}} \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]\mathbb{E}[\boldsymbol{h}_j^{(t-1)}] + c = 0 \\ \sum_{j=1}^{d^{(t-1)}} \left[\frac{1}{2}\text{Var}[\boldsymbol{W}_{ij}^{(t)}] + \text{Var}[\boldsymbol{h}_j^{(t-1)}]\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]^2\right] = 1 \end{cases}$$

With $\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] = \frac{\alpha+\beta}{2}$ and $\mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}] = \frac{1}{12}(\beta-\alpha)^2$. We take $\boxed{\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] = \dfrac{\alpha+\beta}{2} = 0}$, thus $\alpha = -\beta$ which gives:

$$\begin{cases} c = 0 \\ \sum_{j=1}^{d^{(t-1)}} [\frac{1}{6}\alpha^2] = 1 \end{cases}$$

We choose: $\boxed{c = 0,\ \alpha = -\sqrt{\dfrac{6}{d^{(t-1)}}},\ \beta = \sqrt{\dfrac{6}{d^{(t-1)}}}}$.

(c) This initialization corresponds to the He initialization that is implemented in keras. The initialization we found in question (2.2.(a)) corresponds exactly to what keras names *he_normal* and the initialization we found in question (2.2.(b)) corresponds exactly to what keras names *he_uniform*.

(d) This initialization would work well in practice because it ensures that the pre-activations have zero-mean and unit variance at layer t, i.e: $\mathbb{E}[\boldsymbol{a}_i^{(t)}] = 0$ and $\mathrm{Var}(\boldsymbol{a}_i^{(t)}) = 1$, for $1 \le i \le d^{(t)}$, which might stabilise the network. In fact, in question (1), our answer didn't depend on the non-linearity g and we found the Xavier-Glorot initialization, while here it does use information on g being the ReLU function to get the initialization. This initialization is thus able to make the extremely deep model converge in practice when using Relu because it takes into account its expression (and reduces the variance accordingly).

Question 2 (4-6-4-4-3)**.** The point of this question is to understand and compare the effects of different regularizers (specifically dropout and weight decay) on the weights of a network. Consider a linear regression problem with input data $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, weights $\boldsymbol{w} \in \mathbb{R}^{d\times 1}$ and targets $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$. Suppose that dropout is applied to the input (with probability $1-p$ of dropping the unit i.e. setting it to 0). Let $\boldsymbol{R} \in \mathbb{R}^{n\times d}$ be the dropout mask such that $\boldsymbol{R}_{ij} \sim \mathrm{Bern}(p)$ is sampled i.i.d. from the Bernoulli distribution.

1. For squared error loss, express the loss function $L(\boldsymbol{w})$ in matrix form (in terms of $\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w}$, and $\boldsymbol{R}$).
2. Let Γ be a diagonal matrix with $\Gamma_{ii} = (\boldsymbol{X}^\top \boldsymbol{X})_{ii}^{1/2}$. Show that the *expectation (over $\boldsymbol{R}$)* of the loss function can be rewritten as $\mathbb{E}[L(\boldsymbol{w})] = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma\boldsymbol{w}||^2$.
3. Show that the solution $\boldsymbol{w}^{\text{dropout}}$ that minimizes the expected loss from question 2.2 satisfies

$$p\boldsymbol{w}^{\text{dropout}} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{\text{dropout}}\Gamma^2)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

   where λ^{dropout} is a regularization coefficient depending on p. How does the value of p affect the regularization coefficient, λ^{dropout} ?
4. Express the solution $\boldsymbol{w}^{L_2}$ for a linear regression problem without dropout and with L^2 regularization, with regularization coefficient λ^{L_2} in closed form.
5. Compare the results of 2.3 and 2.4: identify specific differences in the equations you arrive at, and discuss qualitatively what the equations tell you about the similarities and differences in the effects of weight decay and dropout (1-3 sentences).

Answer 2. 1. For squared error loss, the loss function $L(\boldsymbol{w})$ can be expressed in matrix form as follows:

$$L(\boldsymbol{w}) = ((\boldsymbol{X}\odot\boldsymbol{R})\boldsymbol{w} - \boldsymbol{y})^\top((\boldsymbol{X}\odot\boldsymbol{R})\boldsymbol{w} - \boldsymbol{y}) \tag{6}$$

where $(\boldsymbol{X} \odot \boldsymbol{R})$ means that we apply the dropout mask to $\boldsymbol{X}$ (an elements-wise operation) before using it to predict the output.

2. Let Γ be a diagonal matrix with $\Gamma_{ii} = (\boldsymbol{X}^\top \boldsymbol{X})_{ii}^{1/2}$. Let's show that the *expectation (over* $\boldsymbol{R}$*)* of the loss function can be rewritten as $\mathbb{E}[L(\boldsymbol{w})] = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma\boldsymbol{w}||^2$.

In fact, we have that:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{R}}[L(\boldsymbol{w})] &= E_{\boldsymbol{R}}[\boldsymbol{w}^\top (\boldsymbol{X} \odot \boldsymbol{R})^\top (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w} + \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w} - \boldsymbol{w}^\top (\boldsymbol{X} \odot \boldsymbol{R})^\top \boldsymbol{y}] \\ &= \boldsymbol{w}^\top E_{\boldsymbol{R}}[(\boldsymbol{X} \odot \boldsymbol{R})^\top (\boldsymbol{X} \odot \boldsymbol{R})]\,\boldsymbol{w} + \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{y}^\top E_{\boldsymbol{R}}[\boldsymbol{X} \odot \boldsymbol{R}]\boldsymbol{w}\end{aligned} \tag{7}$$

Let's calculate $E_{\boldsymbol{R}}[(\boldsymbol{X} \odot \boldsymbol{R})^\top (\boldsymbol{X} \odot \boldsymbol{R})]$ and $E_{\boldsymbol{R}}[\boldsymbol{X} \odot \boldsymbol{R}]$. For this latter, we have by definition, for all $(i,j) \in \{1,\dots,n\} \times \{1,\dots,d\}$, that:

$$\begin{aligned}(E_{\boldsymbol{R}}[\boldsymbol{X} \odot \boldsymbol{R}])_{ij} &= E_{\boldsymbol{R}}[(\boldsymbol{X} \odot \boldsymbol{R})_{ij}] \\ &= E_{\boldsymbol{R}}[(\boldsymbol{X})_{ij}(\boldsymbol{R})_{ij}] \\ &= (\boldsymbol{X})_{ij} E_{\boldsymbol{R}}[(\boldsymbol{R})_{ij}] \\ &= p(\boldsymbol{X})_{ij}\end{aligned} \tag{8}$$

Thus, we conclude that: $E_{\boldsymbol{R}}[\boldsymbol{X} \odot \boldsymbol{R}] = p\boldsymbol{X}$.

Now we need to calculate $E_{\boldsymbol{R}}[(\boldsymbol{X} \odot \boldsymbol{R})^\top (\boldsymbol{X} \odot \boldsymbol{R})]$. We have, for all $(i,j) \in \{1,\dots,d\} \times \{1,\dots,d\}$, that:

$$\begin{aligned}(E_{\boldsymbol{R}}[(\boldsymbol{X} \odot \boldsymbol{R})^\top (\boldsymbol{X} \odot \boldsymbol{R})])_{ij} &= E_{\boldsymbol{R}}[((\boldsymbol{X} \odot \boldsymbol{R})^\top (\boldsymbol{X} \odot \boldsymbol{R}))_{ij}] \\ &= E_{\boldsymbol{R}}[\sum_{k=1}^{n} ((\boldsymbol{X} \odot \boldsymbol{R})^\top)_{ik} (\boldsymbol{X} \odot \boldsymbol{R})_{kj}] \\ &= E_{\boldsymbol{R}}[\sum_{k=1}^{n} (\boldsymbol{X})_{ki}(\boldsymbol{R})_{ki} (\boldsymbol{X})_{kj}(\boldsymbol{R})_{kj}] \\ &= \sum_{k=1}^{n} (\boldsymbol{X})_{ki}(\boldsymbol{X})_{kj} E_{\boldsymbol{R}}[(\boldsymbol{R})_{ki} (\boldsymbol{R})_{kj}]\end{aligned} \tag{9}$$

To calculate $E_{\boldsymbol{R}}[(\boldsymbol{R})_{ki} (\boldsymbol{R})_{kj}]$, we need to examine two cases that will determine whether the variables $(\boldsymbol{R})_{ki}$ and $(\boldsymbol{R})_{kj}$ are independent or not:

- if $i \neq j$, then $(\boldsymbol{R})_{ki}$ and $(\boldsymbol{R})_{kj}$ are independent (because they are different random variables) and we have that:

$$E_{\boldsymbol{R}}[(\boldsymbol{R})_{ki} (\boldsymbol{R})_{kj}] = E_{\boldsymbol{R}}[(\boldsymbol{R})_{ki}] E_{\boldsymbol{R}}(\boldsymbol{R})_{kj}] = p^2 \tag{10}$$

  Thus, we have:

$$(E_{\boldsymbol{R}}[(\boldsymbol{X} \odot \boldsymbol{R})^\top (\boldsymbol{X} \odot \boldsymbol{R})])_{ij} = p^2 (\boldsymbol{X}^\top \boldsymbol{X})_{ij} \tag{11}$$

- if $i = j$, then:

$$E_{\boldsymbol{R}}[((\boldsymbol{R})_{ki})^2] = \mathcal{P}((\boldsymbol{R})_{ki} = 0) \times 0^2 + \mathcal{P}((\boldsymbol{R})_{ki} = 1) \times 1^2 = p \tag{12}$$

  Thus, we have:

$$(E_{\boldsymbol{R}}[(\boldsymbol{X} \odot \boldsymbol{R})^\top (\boldsymbol{X} \odot \boldsymbol{R})])_{ij} = p(\boldsymbol{X}^\top \boldsymbol{X})_{ij} \tag{13}$$

Combining both cases, we obtain, for all $(i, j) \in \{1, \ldots, d\} \times \{1, \ldots, d\}$, that:

$$\begin{aligned}(E_{\boldsymbol{R}}[(\boldsymbol{X} \odot \boldsymbol{R})^\top (\boldsymbol{X} \odot \boldsymbol{R})])_{ij} &= \delta_{ij} p(\boldsymbol{X}^\top \boldsymbol{X})_{ij} + (1 - \delta_{ij}) p^2 (\boldsymbol{X}^\top \boldsymbol{X})_{ij} \\ &= p^2 (\boldsymbol{X}^\top \boldsymbol{X})_{ij} + \delta_{ij} p(1-p)(\boldsymbol{X}^\top \boldsymbol{X})_{ij} \\ &= p^2 (\boldsymbol{X}^\top \boldsymbol{X})_{ij} + p(1-p)\Gamma^2_{ij}\end{aligned} \tag{14}$$

Conclusion of this part is that:

$$E_{\boldsymbol{R}}[(\boldsymbol{X} \odot \boldsymbol{R})^\top (\boldsymbol{X} \odot \boldsymbol{R})] = p^2(\boldsymbol{X}^\top \boldsymbol{X}) + p(1-p)\Gamma^2 \tag{15}$$

Back to the calculation of $\mathbb{E}_{\boldsymbol{R}}[L(\boldsymbol{w})]$:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{R}}[L(\boldsymbol{w})] &= \boldsymbol{w}^\top \left(p^2(\boldsymbol{X}^\top \boldsymbol{X}) + p(1-p)\Gamma^2\right) \boldsymbol{w} + \boldsymbol{y}^\top \boldsymbol{y} - 2p\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w} \\ &= \boldsymbol{y}^\top \boldsymbol{y} - 2p\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w} + p^2 \boldsymbol{w}^\top (\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{w} + p(1-p)\boldsymbol{w}^\top \Gamma^2 \boldsymbol{w} \\ &= \boldsymbol{y}^\top \boldsymbol{y} - 2p\boldsymbol{y}^\top (\boldsymbol{X}\boldsymbol{w}) + p^2 (\boldsymbol{X}\boldsymbol{w})^\top \boldsymbol{X}\boldsymbol{w} + p(1-p)(\Gamma\boldsymbol{w})^\top \Gamma\boldsymbol{w} \\ &= ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma\boldsymbol{w}||^2\end{aligned} \tag{16}$$

which represents the expression we are looking for.

3. Let's show that the solution $\boldsymbol{w}^{\text{dropout}}$ that minimizes the expected loss from question 2.2 satisfies

$$p\boldsymbol{w}^{\text{dropout}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda^{\text{dropout}}\Gamma^2)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

where λ^{dropout} is a regularization coefficient depending on p that we should specify.
$\boldsymbol{w}^{\text{dropout}}$ satisfies that the gradient of $\mathbb{E}_{\boldsymbol{R}}[L(\boldsymbol{w})]$ with respect $\boldsymbol{w}$ is null in $\boldsymbol{w}^{\text{dropout}}$. It is a necessary first order condition. Let's write the gradient first:

$$\begin{aligned}\nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{R}}[L(\boldsymbol{w})] &= -2p\boldsymbol{X}^\top (\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}) + 2p(1-p)\Gamma^2 \boldsymbol{w} \\ &= -2p\boldsymbol{X}^\top \boldsymbol{y} + 2p^2 \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} + 2p(1-p)\Gamma^2 \boldsymbol{w}\end{aligned} \tag{17}$$

The case where $p = 0$ is not interesting because we set all the units to 0 with probability 1. So we assume for the following that $p \neq 0$.

$$\begin{aligned}\nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{R}}[L(\boldsymbol{w}^{\text{dropout}})] = 0 &\implies p^2 \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w}^{\text{dropout}} + p(1-p)\Gamma^2 \boldsymbol{w}^{\text{dropout}} = p\boldsymbol{X}^\top \boldsymbol{y} \\ &\implies p^2(\boldsymbol{X}^\top \boldsymbol{X} + \frac{1-p}{p}\Gamma^2)\boldsymbol{w}^{\text{dropout}} = p\boldsymbol{X}^\top \boldsymbol{y} \\ &\implies (\boldsymbol{X}^\top \boldsymbol{X} + \frac{1-p}{p}\Gamma^2)(p\,\boldsymbol{w}^{\text{dropout}}) = \boldsymbol{X}^\top \boldsymbol{y}\end{aligned} \tag{18}$$

We put $\lambda^{\text{dropout}} = \frac{1-p}{p}$. The matrix $\boldsymbol{X}^\top \boldsymbol{X} + \lambda^{\text{dropout}}\Gamma^2$ can be invertible with an adjustment of λ^{dropout}. We conclude that:

$$p\boldsymbol{w}^{\text{dropout}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda^{\text{dropout}}\Gamma^2)^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{19}$$

We are sure this $\boldsymbol{w}$ gives the minimum of the expectation of the loss since its Hessian is given by:

$$H = 2p^2\boldsymbol{X}^\top\boldsymbol{X} + 2p(1-p)\Gamma^2 \tag{20}$$

For a vector $\mathbf{u} \in \mathbb{R}^d$, we have that: $\mathbf{u}^\top H\mathbf{u} = 2p^2||\boldsymbol{X}\mathbf{u}||^2 + 2p(1-p)||\mathbf{u}\Gamma||^2 \geq 0$. Thus the hessian is positive semi-definite and $\boldsymbol{w}^{\text{dropout}}$ gives indeed the minimum.
The regularization coefficient $\lambda^{\text{dropout}} = \frac{1-p}{p} = \frac{1}{p} - 1$ increases when p decreases, thus when $1-p$ increases. In other words, the more dropout we apply by increasing the dropout probability, the more regularization we introduce, by increasing the regularization coefficient. In particuler, when $p = 1$ we don't apply the dropout, thus $\lambda^{\text{dropout}} = 0$, i.e there is no regularization for the weights.

4. For a linear regression problem without dropout and with L^2 regularization, with regularization coefficient λ^{L_2}, we can write the loss function as follows:

$$L(\boldsymbol{w}) = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + \frac{1}{2}\lambda^{L_2}||\boldsymbol{w}||_2^2 \tag{21}$$

The gradient of the loss w.r.t $\boldsymbol{w}$ is given by:

$$\begin{aligned}
\nabla_{\boldsymbol{w}} L(\boldsymbol{w}) &= -\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \lambda^{L_2}\boldsymbol{w} \\
&= -\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + \lambda^{L_2}\boldsymbol{w} \\
&= -\boldsymbol{X}^\top\boldsymbol{y} + (\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{L_2}\mathbb{I}_d)\boldsymbol{w}
\end{aligned} \tag{22}$$

Thus:

$$\begin{aligned}
\nabla_{\boldsymbol{w}} L(\boldsymbol{w}^{L_2}) = 0 &\implies (\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{L_2}\mathbb{I}_d)\boldsymbol{w}^{L_2} = \boldsymbol{X}^\top\boldsymbol{y} \\
&\implies \boldsymbol{w}^{L_2} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{L_2}\mathbb{I}_d)^{-1}\,\boldsymbol{X}^\top\boldsymbol{y}
\end{aligned} \tag{23}$$

where $\mathbb{I}_d$ is the identity matrix with d lines and d columns. We consider that we can choose λ^{L_2} such as $\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{L_2}\mathbb{I}_d$ is an invertible matrix.
Conclusion:

$$\boldsymbol{w}^{L_2} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{L_2}\mathbb{I}_d)^{-1}\,\boldsymbol{X}^\top\boldsymbol{y} \tag{24}$$

We are sure this $\boldsymbol{w}$ gives the minimum of the loss since its Hessian is given by:

$$H = \boldsymbol{X}^\top\boldsymbol{X} + \lambda^{L_2}\mathbb{I}_d \tag{25}$$

For a vector $\mathbf{u} \in \mathbb{R}^d$, we have that: $\mathbf{u}^\top H\mathbf{u} = ||\boldsymbol{X}\mathbf{u}||^2 + ||\mathbf{u}||^2 \geq 0$. Thus the hessian is positive semi-definite and $\boldsymbol{w}^{L_2}$ gives indeed the minimum.

5. The expression we get for the dropout is:

$$p\boldsymbol{w}^{\text{dropout}} = (\boldsymbol{X}\boldsymbol{X}^\top + \lambda^{\text{dropout}}\Gamma^2)^{-1}\boldsymbol{X}^\top\boldsymbol{y} \tag{26}$$

and the expression we get for weight decay is:

$$\boldsymbol{w}^{L_2} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{L_2}\mathbb{I}_d)^{-1}\,\boldsymbol{X}^\top\boldsymbol{y} \tag{27}$$

We notice that the regularization with weight decay is uniform and affects all elements of the weight the same way since the regularization coefficient is multiplied by the identity matrix. From the other side, the regularization using dropout is not uniform and we do penalize some weights more than others. The weights are also scaled by a factor p, that is inferior to one.

Question 3 (5-5-5)**.** In this question you will demonstrate that an estimate of the first moment of the gradient using an (exponential) running average is equivalent to using momentum, and is biased by a scaling factor. The goal of this question is for you to consider the relationship between different optimization schemes, and to practice noting and quantifying the effect (particularly in terms of bias/variance) of *estimating* a quantity.

Let $\boldsymbol{g}_t$ be an unbiased sample of gradient at time step t and $\Delta\boldsymbol{\theta}_t$ be the update to be made. Initialize $\boldsymbol{v}_0$ to be a vector of zeros.

1. For $t \geq 1$, consider the following update rules:
   - SGD with momentum:
   $$\boldsymbol{v}_t = \alpha \boldsymbol{v}_{t-1} + \epsilon \boldsymbol{g}_t \qquad \Delta\boldsymbol{\theta}_t = -\boldsymbol{v}_t$$
   where $\epsilon > 0$ and $\alpha \in (0, 1)$.
   - SGD with running average of $\boldsymbol{g}_t$:
   $$\boldsymbol{v}_t = \beta \boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t \qquad \Delta\boldsymbol{\theta}_t = -\delta \boldsymbol{v}_t$$
   where $\beta \in (0, 1)$ and $\delta > 0$.

   Express the two update rules recursively ($\Delta\boldsymbol{\theta}_t$ as a function of $\Delta\boldsymbol{\theta}_{t-1}$). Show that these two update rules are equivalent ; i.e. express (α, ϵ) as a function of (β, δ).
2. Unroll the running average update rule, i.e. express $\boldsymbol{v}_t$ as a linear combination of $\boldsymbol{g}_i$'s ($1 \leq i \leq t$).
3. Assume $\boldsymbol{g}_t$ has a stationary distribution independent of t. Show that the running average is biased, i.e. $\mathbb{E}[\boldsymbol{v}_t] \neq \mathbb{E}[\boldsymbol{g}_t]$. Propose a way to eliminate such a bias by rescaling $\boldsymbol{v}_t$.

Answer 3. 1. For $t \geq 1$, we have the following update rules recursively:

- For SGD with momentum:
$$\begin{aligned}\Delta\boldsymbol{\theta}_t &= -\boldsymbol{v}_t \\ &= -\alpha \boldsymbol{v}_{t-1} - \epsilon \boldsymbol{g}_t \\ &= \alpha \Delta\boldsymbol{\theta}_{t-1} - \epsilon \boldsymbol{g}_t\end{aligned} \tag{28}$$
where $\epsilon > 0$ and $\alpha \in (0, 1)$.
- For SGD with running average of $\boldsymbol{g}_t$:
$$\begin{aligned}\Delta\boldsymbol{\theta}_t &= -\delta \boldsymbol{v}_t \\ &= -\delta\beta \boldsymbol{v}_{t-1} - \delta(1-\beta)\boldsymbol{g}_t \\ &= \beta \Delta\boldsymbol{\theta}_{t-1} - \delta(1-\beta)\boldsymbol{g}_t\end{aligned} \tag{29}$$
where $\beta \in (0, 1)$ and $\delta > 0$.

We notice that both update rules are equivalent if we take: $\alpha = \beta$ and $\epsilon = \delta(1-\beta)$.

2. Let $t \geq 1$ and let's unroll the running average update rule:

$$\begin{aligned}
\boldsymbol{v}_t &= \beta \boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t \\
&= \beta(\beta \boldsymbol{v}_{t-2} + (1-\beta)\boldsymbol{g}_{t-1}) + (1-\beta)\boldsymbol{g}_t \\
&= \beta^2 \boldsymbol{v}_{t-2} + (1-\beta)(\beta \boldsymbol{g}_{t-2} + \boldsymbol{g}_t) \\
&= \beta^2(\beta \boldsymbol{v}_{t-3} + (1-\beta)\boldsymbol{g}_{t-2}) + (1-\beta)(\beta \boldsymbol{g}_{t-2} + \boldsymbol{g}_t) \\
&= \beta^3 \boldsymbol{v}_{t-3} + (1-\beta)(\beta^2 \boldsymbol{g}_{t-2} + \beta \boldsymbol{g}_{t-2} + \boldsymbol{g}_t) \\
&= \beta^{t-(t-3)} \boldsymbol{v}_{t-3} + (1-\beta) \sum_{k=t-2}^{t} \beta^{t-k} \boldsymbol{g}_k \\
&= \beta^t \boldsymbol{v}_0 + (1-\beta) \sum_{k=1}^{t} \beta^{t-k} \boldsymbol{g}_k
\end{aligned} \tag{30}$$

The proof can be done more properly using recurrence to prove that:

$$\boldsymbol{v}_t = \beta^t \boldsymbol{v}_0 + (1-\beta) \sum_{k=1}^{t} \beta^{t-k} \boldsymbol{g}_k, \quad \forall t \geq \tag{31}$$

Let's do the recurrence:

- For t=1, using the update rule, we have:

$$\boldsymbol{v}_1 = \beta \boldsymbol{v}_0 + (1 - \beta \boldsymbol{g}_1)$$

. Thus property (31) is true.

- Let's fix $t \geq 1$ and consider that (31) is true for t. We need to prove that (31) is true for $t+1$. We have, using the update rule and our recurrence hypothesis, that:

$$\begin{aligned}
\boldsymbol{v}_{t+1} &= \beta \boldsymbol{v}_t + (1-\beta)\boldsymbol{g}_{t+1} \\
&= \beta(\beta^t \boldsymbol{v}_0 + (1-\beta) \sum_{k=1}^{t} \beta^{t-k} \boldsymbol{g}_k) + (1-\beta)\boldsymbol{g}_{t+1} \\
&= \beta^{t+1} \boldsymbol{v}_0 + (1-\beta) \sum_{k=1}^{t} \beta \times \beta^{t-k} \boldsymbol{g}_k + (1-\beta)\boldsymbol{g}_{t+1} \\
&= \beta^{t+1} \boldsymbol{v}_0 + (1-\beta) \sum_{k=1}^{t} \beta^{t+1-k} \boldsymbol{g}_k + (1-\beta)\beta^{t+1-(t+1)}\boldsymbol{g}_{t+1} \\
&= \beta^{t+1} \boldsymbol{v}_0 + (1-\beta) \sum_{k=1}^{t+1} \beta^{t+1-k} \boldsymbol{g}_k
\end{aligned} \tag{32}$$

Thus (31) is true for $t+1$.
Conclusion (since $\boldsymbol{v}_0$ is a vector of zeros):

$$\boldsymbol{v}_t = (1-\beta) \sum_{k=1}^{t} \beta^{t-k} \boldsymbol{g}_k, \quad \forall t \geq \tag{33}$$

3. We assume $\boldsymbol{g}_t$ has a stationary distribution independent of t and let $t \geq 1$, we have:

$$\begin{aligned}\mathbb{E}[\boldsymbol{v}_t] &= \mathbb{E}[(1-\beta)\sum_{k=1}^{t}\beta^{t-k}\boldsymbol{g}_k] \\ &= (1-\beta)\sum_{k=1}^{t}\beta^{t-k}\,\mathbb{E}[\boldsymbol{g}_k] \\ &= (1-\beta^t)\,\mathbb{E}[\boldsymbol{g}_k]\end{aligned} \tag{34}$$

We assume that $\beta \neq 0$ and $\beta \neq 1$, because both represent special cases that are not the target of this exercise. Since $\beta \neq 0$, we have that $\mathbb{E}[\boldsymbol{v}_t] \neq \mathbb{E}[\boldsymbol{g}_t]$. Thus the running average is biased. We can rescale $\boldsymbol{v}_t$ and consider a new update rule :

$$\boldsymbol{v}_t^{\text{new}} = \frac{1}{(1-\beta^t)}(\beta\boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t)$$

The expectation term becomes:

$$\mathbb{E}[\boldsymbol{v}_t^{\text{new}}] = \frac{1}{(1-\beta^t)}\,\mathbb{E}[\boldsymbol{v}_t] = \frac{1}{(1-\beta^t)} \times (1-\beta^t)\,\mathbb{E}[\boldsymbol{g}_k] = \mathbb{E}[\boldsymbol{g}_k]$$

The rescaled $\boldsymbol{v}_t$ makes the running average unbiased.

Both objectives of the question are achieved. We showed that an estimate of the first moment of the gradient using an running average is equivalent to using momentum and that is biased by a scaling factor.

Question 4 (5-5-5)**.** This question is about weight normalization.[1] We consider the following parameterization of a weight vector $\boldsymbol{w}$:

$$\boldsymbol{w} := \gamma\frac{\boldsymbol{u}}{||\boldsymbol{u}||}$$

where γ is scalar parameter controlling the magnitude and $\boldsymbol{u}$ is a vector controlling the direction of $\boldsymbol{w}$.

1. Consider one layer of a neural network, and omit the bias parameter. To carry out batch normalization, one normally standardizes the preactivation and performs elementwise scale and shift $\hat{y} = \gamma \cdot \frac{y-\mu_y}{\sigma_y} + \beta$ where $y = \boldsymbol{u}^\top\boldsymbol{x}$. Assume the data $\boldsymbol{x}$ (a random vector) is whitened ($\text{Var}(\boldsymbol{x}) = \boldsymbol{I}$) and centered at 0 ($\mathbb{E}[\boldsymbol{x}] = \boldsymbol{0}$). Show that $\hat{y} = \boldsymbol{w}^\top\boldsymbol{x} + \beta$.
2. Show that the gradient of a loss function $L(\boldsymbol{u},\gamma,\beta)$ with respect to $\boldsymbol{u}$ can be written in the form $\nabla_{\boldsymbol{u}}L = s\boldsymbol{W}^\perp\nabla_{\boldsymbol{w}}L$ for some s, where $\boldsymbol{W}^\perp = \left(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^\top}{||\boldsymbol{u}||^2}\right)$. Note that[2] $\boldsymbol{W}^\perp\boldsymbol{u} = \boldsymbol{0}$.
3. Figure 1 shows the norm of $\boldsymbol{u}$ as a function of number of updates made to a two-layer MLP using gradient descent. Different curves correspond to models trained with different log-learning rate. Explain why (1) the norm is increasing, and (2) why larger learning rate corresponds to faster growth. (Hint: Use the Pythagorean theorem and the fact that $\boldsymbol{W}^\perp\boldsymbol{u} = 0$ from question 4.2).

[1] See https://arxiv.org/abs/1602.07868

[2] As a side note: $\boldsymbol{W}^\perp$ is an orthogonal complement that projects the gradient away from the direction of $\boldsymbol{w}$, which is usually (empirically) close to a dominant eigenvector of the covariance of the gradient. This helps to condition the landscape of the objective that we want to optimize.

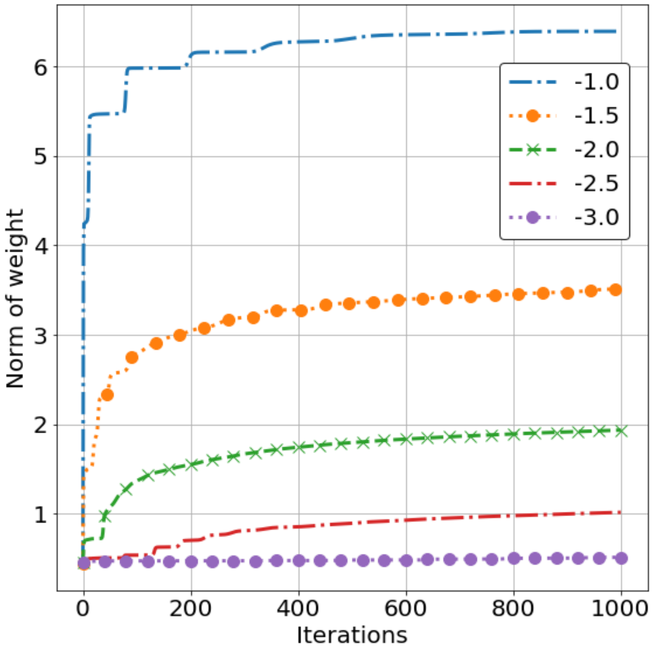

FIGURE 1 – Norm of parameters with different learning rate.

Answer 4. 1. We standardize the preactivation and perform elementwise scale and shift $\hat{y} = \gamma \cdot \frac{y-\mu_y}{\sigma_y} + \beta$ where $y = \boldsymbol{u}^\top \boldsymbol{x}$.

We have

$$\mu_y = \mathbb{E}[y] = \mathbb{E}[\boldsymbol{u}^\top \boldsymbol{x}] = \boldsymbol{u}^\top \mathbb{E}[\boldsymbol{x}] = 0$$

because the data $\boldsymbol{x}$ is centered, and

$$\sigma_y^2 = \text{Var}(y) = \text{Var}(\boldsymbol{u}^\top \boldsymbol{x}) = \boldsymbol{u}^\top \text{Var}(\boldsymbol{x}) \boldsymbol{u} = \boldsymbol{u}^\top \boldsymbol{I} \boldsymbol{u} = ||u||^2$$

because $\boldsymbol{x}$ is whitened. Thus:

$$\hat{y} = \gamma \cdot \frac{y-0}{||\boldsymbol{u}||} + \beta = (\gamma \cdot \frac{\boldsymbol{u}^\top}{||\boldsymbol{u}||})\boldsymbol{x} + \beta = \boldsymbol{w}^\top \boldsymbol{x} + \beta$$

2. We want to calculae the gradient of a loss function $L(\boldsymbol{u}, \gamma, \beta)$ with respect to $\boldsymbol{u}$. Let's suppose that $\boldsymbol{u}$ has a length of n. For all $i\{1, \dots, n\}$, we have:

$$\frac{\partial L}{\partial \boldsymbol{u}_i} = \sum_{j=1}^{n} \frac{\partial L}{\partial \boldsymbol{w}_j} \cdot \frac{\partial \boldsymbol{w}_j}{\partial \boldsymbol{u}_i}$$

and, we have that:

$$\begin{aligned}\frac{\partial \boldsymbol{w}_j}{\partial \boldsymbol{u}_i} &= \gamma \cdot \frac{\partial}{\partial \boldsymbol{u}_i}\frac{\boldsymbol{u}_j}{\sqrt{\sum_{k=1}^n \boldsymbol{u}_k^2}} \\ &= \gamma \cdot \Big(\frac{\partial \boldsymbol{u}_j}{\partial \boldsymbol{u}_i} \cdot ||\boldsymbol{u}|| - \boldsymbol{u}_j \frac{\partial}{\partial \boldsymbol{u}_i}\sqrt{\sum_{k=1}^n \boldsymbol{u}_k^2}\Big)/||\boldsymbol{u}||^2 \\ &= \gamma \cdot (\delta i, j||\boldsymbol{u}|| - \boldsymbol{u}_j \frac{\boldsymbol{u}_i}{||\boldsymbol{u}||})/||\boldsymbol{u}||^2 \\ &= \gamma \cdot (\delta i, j\frac{1}{||\boldsymbol{u}||} - \frac{\boldsymbol{u}_j \boldsymbol{u}_i}{||\boldsymbol{u}||^3})\end{aligned} \tag{35}$$

Then we have that:

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{u}_i} &= \gamma \cdot \sum_{j=1}^n \Big(\delta i, j\frac{1}{||\boldsymbol{u}||} - \frac{\boldsymbol{u}_j \boldsymbol{u}_i}{||\boldsymbol{u}||^3}\Big)\frac{\partial L}{\partial \boldsymbol{w}_j} \\ &= \gamma \cdot \Big(\frac{1}{||\boldsymbol{u}||}\frac{\partial L}{\partial \boldsymbol{w}_i} - \frac{\boldsymbol{u}_i}{||\boldsymbol{u}||^3}\sum_{j=1}^n \boldsymbol{u}_j \frac{\partial L}{\partial \boldsymbol{w}_j}\Big)\end{aligned} \tag{36}$$

We obtain that the gradient is given by:

$$\begin{aligned}\nabla_{\boldsymbol{u}} L &= \gamma \cdot \Big[\frac{1}{||\boldsymbol{u}||}\nabla_{\boldsymbol{w}} L - \frac{1}{||\boldsymbol{u}||^3}\boldsymbol{u}(\boldsymbol{u}^\top \nabla_{\boldsymbol{w}} L)\Big] \\ &= \frac{\gamma}{||\boldsymbol{u}||} \cdot \Big[\boldsymbol{I} - \frac{1}{||\boldsymbol{u}||^2}\boldsymbol{u}\boldsymbol{u}^\top\Big]\nabla_{\boldsymbol{w}} L \\ &= s\boldsymbol{W}^\perp \nabla_{\boldsymbol{w}} L\end{aligned} \tag{37}$$

where $s = \frac{\gamma}{||\boldsymbol{u}||}$ and $\boldsymbol{W}^\perp = \Big(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^\top}{||\boldsymbol{u}||^2}\Big)$.

3. We have that $y = \boldsymbol{u}^\top \boldsymbol{x}$, thus the update of $\boldsymbol{u}^{(t+1)}$ at iteration $t+1$ using gradient descent and a learning rate α is given by $\boldsymbol{u}^{(t+1)} = \boldsymbol{u}^{(t)} - \alpha \nabla_{\boldsymbol{u}} L$. Thus, we have:

$$\begin{aligned}||\boldsymbol{u}^{(t+1)}||^2 &= ||\boldsymbol{u}^{(t)} - \alpha \nabla_{\boldsymbol{u}^{(t)}} L||^2 \\ &= ||\boldsymbol{u}^{(t)}||^2 + \alpha^2 ||\nabla_{\boldsymbol{u}^{(t)}} L||^2 - 2\alpha(\boldsymbol{u}^{(t)})^\top \nabla_{\boldsymbol{u}} L \\ &= ||\boldsymbol{u}^{(t)}||^2 + \alpha^2 ||\nabla_{\boldsymbol{u}^{(t)}} L||^2 - 2\alpha(\boldsymbol{u}^{(t)})^\top s(\boldsymbol{W}^\perp)^{(t)} \nabla_{\boldsymbol{w}} L \\ &= ||\boldsymbol{u}^{(t)}||^2 + \alpha^2 ||\nabla_{\boldsymbol{u}^{(t)}} L||^2 - 2\alpha s((\boldsymbol{u}^{(t)})^\top (\boldsymbol{W}^\perp)^{(t)}) \nabla_{\boldsymbol{w}^{(t)}} L, \quad \text{and} \quad ((\boldsymbol{u}^{(t)})^\top (\boldsymbol{W}^\perp)^{(t)}) = \boldsymbol{0} \\ &= ||\boldsymbol{u}^{(t)}||^2 + \alpha^2 ||\nabla_{\boldsymbol{u}^{(t)}} L||^2\end{aligned} \tag{38}$$

Thus:

$$||\boldsymbol{u}^{(t+1)}||^2 = ||\boldsymbol{u}^{(t)}||^2 + \alpha^2 ||\nabla_{\boldsymbol{u}^{(t)}} L||^2, \quad \forall t \geq 1 \tag{39}$$

This last equation explains the behaviour we see in Figure 1:

- (1) the norm is increasing because according to (39), it is written as the same of its previous value and another positive value that is non null in general,
- (2) the contribution of $||\nabla_{\boldsymbol{u}^{(t)}} L||^2$ in (39) is multiplied by the square of the learning rate, which means that the larger learning rate corresponds to a bigger contribution of $||\nabla_{\boldsymbol{u}^{(t)}} L||^2$, thus to a faster growth.

Question 5 (5-5-5)**.** This question is about activation functions and vanishing/exploding gradients in recurrent neural networks (RNNs). Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function. When the argument is a vector, we apply σ element-wise. Consider the following recurrent unit:

$$\boldsymbol{h}_t = \boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}$$

1. Show that applying the activation function in this way is equivalent to the conventional way of applying the activation function: $\boldsymbol{g}_t = \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$ (i.e. express $\boldsymbol{g}_t$ in terms of $\boldsymbol{h}_t$). More formally, you need to prove it using mathematical induction. You only need to prove the induction step in this question, assuming your expression holds for time step $t-1$.

*2. Let $||\boldsymbol{A}||$ denote the L_2 operator norm [3] of matrix $\boldsymbol{A}$ ($||\boldsymbol{A}|| := \max_{\boldsymbol{x}:||\boldsymbol{x}||=1} ||\boldsymbol{A}\boldsymbol{x}||$). Assume $\sigma(x)$ has bounded derivative, i.e. $|\sigma'(x)| \leq \gamma$ for some $\gamma > 0$ and for all x. We denote as $\lambda_1(\cdot)$ the largest eigenvalue of a symmetric matrix. Show that if the largest eigenvalue of the weights is upper-bounded by $\frac{\delta^2}{\gamma^2}$ for some $0 \leq \delta < 1$, gradients of the hidden state will vanish over time, i.e.

$$\lambda_1(\boldsymbol{W}^\top \boldsymbol{W}) \leq \frac{\delta^2}{\gamma^2} \quad \Longrightarrow \quad \left|\left|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0}\right|\right| \to 0 \text{ as } T \to \infty$$

Use the following properties of the L_2 operator norm

$$||\boldsymbol{A}\boldsymbol{B}|| \leq ||\boldsymbol{A}||\,||\boldsymbol{B}|| \qquad \text{and} \qquad ||\boldsymbol{A}|| = \sqrt{\lambda_1(\boldsymbol{A}^\top \boldsymbol{A})}$$

3. What do you think will happen to the gradients of the hidden state if the condition in the previous question is reversed, i.e. if the largest eigenvalue of the weights is larger than $\frac{\delta^2}{\gamma^2}$? Is this condition *necessary* or *sufficient* for the gradient to explode? (Answer in 1-2 sentences).

Answer 5.

1. Let's show that applying the activation function in this way: $\boldsymbol{h}_t = \boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}$ is equivalent to the conventional way of applying the activation function: $\boldsymbol{g}_t = \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$. First, let's try to get an idea about the relation between the two:

$$\boldsymbol{h}_t = \boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b} \implies \sigma(\boldsymbol{h}_t) = \sigma(\boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$$

So we can think that $g_t = \sigma(h_t)$, for each $t \geq 0$. However, if we want to prove this supposition by induction, we will need it to hold for $t = 0$, which is not particularly specified in this exercise. This is why we only need to prove the induction step in this question, assuming your expression holds for time step $t-1$.

Let's fix $t \geq 1$ and suppose that $g_{t-1} = \sigma(h_{t-1})$. We need to prove that : $g_t = \sigma(h_t)$.

[3] The L_2 operator norm of a matrix $\boldsymbol{A}$ is is an *induced norm* corresponding to the L_2 norm of vectors. You can try to prove the given properties as an exercise.

We have:

$$\begin{aligned}\boldsymbol{g}_t &= \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}) \qquad \text{(by definition)}\\ &= \sigma(\boldsymbol{W}\sigma(h_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}) \qquad \text{(recurrence assumption)}\\ &= \sigma(h_t) \qquad \text{(by definition)}\end{aligned} \tag{40}$$

We proved the induction step. This, if $g_0 = \sigma(h_0)$, we have that the two ways of applying the activation function are equivalent (with a transformation between the two).

2. First, let's calculate $\left|\left|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}}\right|\right|$ for $t \geq 0$. Let $(i,j) \in \{1,\dots,d\} \times \{1,\dots,d\}$, if we consider that d is the length of both vectors $\boldsymbol{h}_t$ and $\boldsymbol{h}_{t-1}$, we have:

$$\begin{aligned}\left(\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}}\right)_{ij} &= \left(\frac{\partial}{\partial \boldsymbol{h}_{t-1}}\Big(\boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}\Big)\right)_{ij}\\ &= \left(\frac{\partial}{\partial \boldsymbol{h}_{t-1}}\Big(\boldsymbol{W}\sigma(\boldsymbol{h}_{t-1})\Big)\right)_{ij} \qquad (\boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b} \text{ is not a function of } \boldsymbol{h}_{t-1})\\ &= \frac{\partial}{\partial(\boldsymbol{h}_{t-1})_j}\left(\sum_{k=1}^{d} \boldsymbol{W}_{ik}\sigma((\boldsymbol{h}_{t-1})_k)\right) \qquad \text{(by definition of the jacobian)}\\ &= \boldsymbol{W}_{ik}\,\frac{\partial}{\partial(\boldsymbol{h}_{t-1})_j}\left(\sum_{k=1}^{d}\sigma((\boldsymbol{h}_{t-1})_k)\right)\\ &= \boldsymbol{W}_{ik}\,\frac{\partial}{\partial(\boldsymbol{h}_{t-1})_j}\Big(\sigma((\boldsymbol{h}_{t-1})_j)\Big)\\ &= \boldsymbol{W}_{ij}\,\sigma'((\boldsymbol{h}_{t-1})_j)\end{aligned} \tag{41}$$

Thus, we have that:

$$\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} = \boldsymbol{W}\,diag(\sigma'(\boldsymbol{h}_{t-1})) \tag{42}$$

Using this property $||\boldsymbol{AB}|| \leq ||\boldsymbol{A}||\,||\boldsymbol{B}||$, we have that:

$$\left|\left|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}}\right|\right| \leq ||\boldsymbol{W}||\,||diag(\sigma'(\boldsymbol{h}_{t-1}))|| \tag{43}$$

Using $||\boldsymbol{A}|| = \sqrt{\lambda_1(\boldsymbol{A}^\top\boldsymbol{A})}$, we have that :

- $$||\boldsymbol{W}|| = \sqrt{\lambda_1(\boldsymbol{W}^\top\boldsymbol{W})} \leq \frac{\delta}{\gamma} \tag{44}$$
- and $||diag(\sigma'(\boldsymbol{h}_{t-1}))|| = \sqrt{\lambda_1(diag(\sigma'(\boldsymbol{h}_{t-1}))^\top diag(\sigma'(\boldsymbol{h}_{t-1})))}$.

In addition, we know that $diag(\sigma'(\boldsymbol{h}_{t-1}))^\top diag(\sigma'(\boldsymbol{h}_{t-1}))$ is a diagonal matrix and:

$$(diag(\sigma'(\boldsymbol{h}_{t-1}))^\top diag(\sigma'(\boldsymbol{h}_{t-1})))_{ii} = (\sigma'((\boldsymbol{h}_{t-1})_i))^2 \quad \forall i \in \{1,\dots,d\}$$

Thus, we know that the spectrum of this matrix is exactly the set: $\{(\sigma'((\boldsymbol{h}_{t-1})_i))^2, i \in \{1,\dots,d\}\}$ and since σ has a bounded derivative, we have that:

$$\underset{i\in\{1,\dots,d\}}{max}((\sigma'((\boldsymbol{h}_{t-1})_i))^2) \leq \gamma^2$$

Since $\lambda_1(diag(\sigma'(\boldsymbol{h}_{t-1}))^\top diag(\sigma'(\boldsymbol{h}_{t-1}))) = \underset{i\in\{1,\dots,d\}}{max}((\sigma'((\boldsymbol{h}_{t-1})_i))^2)$, we conclude that:

$$||diag(\sigma'(\boldsymbol{h}_{t-1}))|| = \sqrt{\lambda_1(diag(\sigma'(\boldsymbol{h}_{t-1}))^\top diag(\sigma'(\boldsymbol{h}_{t-1})))} \leq \gamma \tag{45}$$

From equations (43), (44) and (45), we conclude that:

$$\left|\left|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}}\right|\right| \leq \delta \tag{46}$$

Let $T \in \mathbb{N}^*$, using the chain rule, we have that:

$$\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0} = \frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_{T-1}} \frac{\partial \boldsymbol{h}_{T-1}}{\partial \boldsymbol{h}_{T-2}} \dots \frac{\partial \boldsymbol{h}_1}{\partial \boldsymbol{h}_0}$$

Using this equation and equation (46), we get that:

$$\left|\left|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0}\right|\right| \leq \delta^T$$

Since $0 \leq \delta < 1$, then $\delta^T \to 0$ as $T \to \infty$, thus: $\left|\left|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0}\right|\right| \to 0$ as $T \to \infty$.

3. For the gradient to explode, we need to necessarily have that the largest eigenvalue of the weights is larger than $\frac{\delta^2}{\gamma^2}$. Otherwise, the gradient will vanish as we saw in the previous question (its norm will tend to zero when T tends to infinity). Thus, it is a necessary condition. Objectively, we can't say whether this condition is sufficient without further calculations, because the concerned term only does provide an upper bound and not a lower bound on the norm.

Question 6 (6-12)**.** Denote by σ the logistic sigmoid function. Consider the following Bidirectional RNN:

$$\begin{aligned}\boldsymbol{h}_t^{(f)} &= \sigma(\boldsymbol{W}^{(f)}\boldsymbol{x}_t + \boldsymbol{U}^{(f)}\boldsymbol{h}_{t-1}^{(f)})\\ \boldsymbol{h}_t^{(b)} &= \sigma(\boldsymbol{W}^{(b)}\boldsymbol{x}_t + \boldsymbol{U}^{(b)}\boldsymbol{h}_{t+1}^{(b)})\\ \boldsymbol{y}_t &= \boldsymbol{V}^{(f)}\boldsymbol{h}_t^{(f)} + \boldsymbol{V}^{(b)}\boldsymbol{h}_t^{(b)}\end{aligned}$$

where the superscripts f and b correspond to the forward and backward RNNs respectively.

1. Draw the computational graph for this RNN, unrolled for 3 time steps (from $t = 1$ to $t = 3$) Include and label the initial hidden states for both the forward and backward RNNs, $h_0^{(f)}$ and $h_4^{(b)}$ respectively. You may draw this by hand ; you may also use a computer rendering package such as TikZ, but you are not required to do so. Label each node and edge with the corresponding hidden unit or weight.

*2. Let $\boldsymbol{z}_t$ be the true target of the prediction $\boldsymbol{y}_t$ and consider the sum of squared loss $L = \sum_t L_t$ where $L_t = ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2$. Express the gradients $\nabla_{\boldsymbol{h}_t^{(f)}} L$ and $\nabla_{\boldsymbol{h}_t^{(b)}} L$ recursively (in terms of $\nabla_{\boldsymbol{h}_{t+1}^{(f)}} L$ and $\nabla_{\boldsymbol{h}_{t-1}^{(b)}} L$ respectively). Then derive $\nabla_{\boldsymbol{W}^{(f)}} L$ and $\nabla_{\boldsymbol{U}^{(b)}} L$.

Answer 6.

1. Figure (2) gives the computational graph for this bidirectional RNN, unrolled for 3 time steps (from $t = 1$ to $t = 3$). We include the initial hidden states $h_0^{(f)}$ and $h_4^{(b)}$, the true target $\boldsymbol{z}_t$ of the

prediction $\boldsymbol{y}_t$ and the loss at time t: L_t because we will refer to the graph in the next question to derive the right dependencies between the variables and the right chain rules.

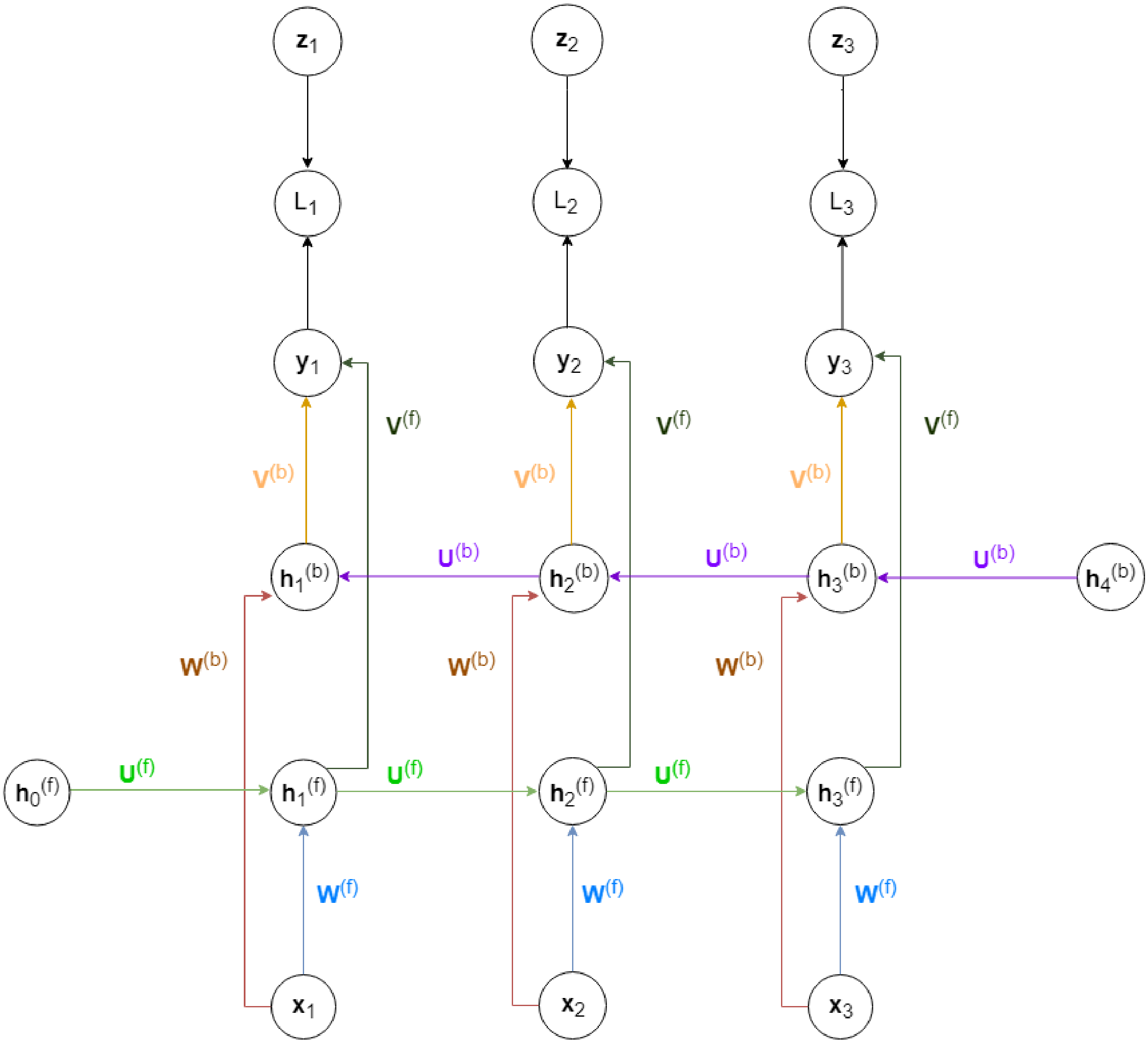

FIGURE 2 – Computational graph for the bidirectional RNN, unrolled for 3 time steps (from $t = 1$ to $t = 3$)

2. We divide the answer of this question to 4 sub-sections.

Expression of $\nabla_{\boldsymbol{h}_t^{(f)}} L$ using $\nabla_{\boldsymbol{h}_{t+1}^{(f)}} L$:

We have that:

$$\nabla_{\boldsymbol{h}_t^{(f)}} L = \frac{\partial L}{\partial \boldsymbol{h}_t^{(f)}} = \frac{\partial}{\partial \boldsymbol{h}_t^{(f)}} \sum_k L_k$$

For a given t, only the loss terms L_k with $k \geq t$ depend on $\boldsymbol{h}_t^{(f)}$ (as shown in the drawing of question 1), thus we can simplify the previous expression using this remark this we use the chain rule to introduce $\boldsymbol{h}_{t+1}^{(f)}$

$$\begin{aligned}
\nabla_{\boldsymbol{h}_t^{(f)}} L &= \frac{\partial}{\partial \boldsymbol{h}_t^{(f)}} \sum_{k \geq t} L_k \\
&= \frac{\partial L_t}{\partial \boldsymbol{h}_t^{(f)}} + \frac{\partial}{\partial \boldsymbol{h}_t^{(f)}} \sum_{k \geq t+1} L_k \\
&= \nabla_{\boldsymbol{h}_t^{(f)}} L_t + \left(\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}} \right)^\top \frac{\partial}{\partial \boldsymbol{h}_{t+1}^{(f)}} \sum_{k \geq t+1} L_k \\
&= \nabla_{\boldsymbol{h}_t^{(f)}} L_t + \left(\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}} \right)^\top \frac{\partial L}{\partial \boldsymbol{h}_{t+1}^{(f)}}, \quad \text{(using same argument as above)} \\
&= \nabla_{\boldsymbol{h}_t^{(f)}} L_t + \left(\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}} \right)^\top \nabla_{\boldsymbol{h}_{t+1}^{(f)}} L
\end{aligned} \tag{47}$$

Now, let's find the expressions of both $\nabla_{\boldsymbol{h}_t^{(f)}} L_t$ and $\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}}$.
We have that:

$$\begin{aligned}
\nabla_{\boldsymbol{h}_t^{(f)}} L_t &= \frac{\partial L_t}{\partial \boldsymbol{h}_t^{(f)}} \\
&= \frac{\partial}{\partial \boldsymbol{h}_t^{(f)}} ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2 \\
&= \left(\frac{\partial \boldsymbol{y}_t}{\partial \boldsymbol{h}_t^{(f)}} \right)^\top \frac{\partial}{\partial \boldsymbol{y}_t} ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2 \\
&= \left(\frac{\partial}{\partial \boldsymbol{h}_t^{(f)}} (\boldsymbol{V}^{(f)} \boldsymbol{h}_t^{(f)} + \boldsymbol{V}^{(b)} \boldsymbol{h}_t^{(b)}) \right)^\top \frac{\partial}{\partial \boldsymbol{y}_t} (\boldsymbol{y}_t^\top \boldsymbol{y}_t - 2\boldsymbol{y}_t^\top \boldsymbol{z}_t + \boldsymbol{z}_t^\top \boldsymbol{z}_t) \\
&= (\boldsymbol{V}^{(f)})^\top (2\boldsymbol{y}_t - 2\boldsymbol{z}_t) \\
&= 2(\boldsymbol{V}^{(f)})^\top (\boldsymbol{y}_t - \boldsymbol{z}_t)
\end{aligned} \tag{48}$$

And we have that:

$$\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}} = \frac{\partial}{\partial \boldsymbol{h}_t^{(f)}} \sigma(\boldsymbol{W}^{(f)} \boldsymbol{x}_t + \boldsymbol{U}^{(f)} \boldsymbol{h}_t^{(f)}) \tag{49}$$

For given i and j, we have:

$$\begin{aligned}
\left(\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}}\right)_{ij} &= \frac{\partial}{\partial (\boldsymbol{h}_t^{(f)})_j}\sigma((\boldsymbol{W}^{(f)}\boldsymbol{x}_t + \boldsymbol{U}^{(f)}\boldsymbol{h}_t^{(f)})_i) \\
&= \frac{\partial}{\partial (\boldsymbol{h}_t^{(f)})_j}\sigma((\boldsymbol{W}^{(f)}\boldsymbol{x}_t)_i + \sum_k (\boldsymbol{U}^{(f)})_{ik}(\boldsymbol{h}_t^{(f)})_k) \\
&= (\boldsymbol{U}^{(f)})_{ij}\sigma'((\boldsymbol{W}^{(f)}\boldsymbol{x}_t + \boldsymbol{U}^{(f)}\boldsymbol{h}_t^{(f)})_i) \\
&= (diag(\sigma'(\boldsymbol{W}^{(f)}\boldsymbol{x}_t + \boldsymbol{U}^{(f)}\boldsymbol{h}_t^{(f)})\boldsymbol{U}^{(f)})_{ij} \\
&= (diag(\sigma'(\boldsymbol{W}^{(f)}\boldsymbol{x}_t + \boldsymbol{U}^{(f)}\boldsymbol{h}_t^{(f)})\boldsymbol{U}^{(f)})_{ij}
\end{aligned} \tag{50}$$

Thus we conclude that:

$$\begin{aligned}
\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}} &= diag(\sigma'(\boldsymbol{W}^{(f)}\boldsymbol{x}_t + \boldsymbol{U}^{(f)}\boldsymbol{h}_t^{(f)})\boldsymbol{U}^{(f)} \\
&= diag(\sigma(\boldsymbol{W}^{(f)}\boldsymbol{x}_t + \boldsymbol{U}^{(f)}\boldsymbol{h}_t^{(f)}))diag(1 - \sigma(\boldsymbol{W}^{(f)}\boldsymbol{x}_t + \boldsymbol{U}^{(f)}\boldsymbol{h}_t^{(f)}))\boldsymbol{U}^{(f)} \\
&= diag(\boldsymbol{h}_{t+1}^{(f)})diag(\mathbf{1} - \boldsymbol{h}_{t+1}^{(f)})\,\boldsymbol{U}^{(f)}
\end{aligned} \tag{51}$$

where $\mathbf{1}$ is a vector with the same length as $\boldsymbol{h}_{t+1}^{(f)}$.
In conclusion, we express $\nabla_{\boldsymbol{h}_t^{(f)}} L$ using $\nabla_{\boldsymbol{h}_{t+1}^{(f)}} L$ as follows:

$$\boxed{\nabla_{\boldsymbol{h}_t^{(f)}} L = 2(\boldsymbol{V}^{(f)})^\top(\boldsymbol{y}_t - \boldsymbol{z}_t) + (\boldsymbol{U}^{(f)})^\top diag(\boldsymbol{h}_{t+1}^{(f)})diag(\mathbf{1} - \boldsymbol{h}_{t+1}^{(f)})\nabla_{\boldsymbol{h}_{t+1}^{(f)}} L} \tag{52}$$

Expression of $\nabla_{\boldsymbol{h}_t^{(b)}} L$ using $\nabla_{\boldsymbol{h}_{t-1}^{(b)}} L$:

We have that:

$$\begin{aligned}
\nabla_{\boldsymbol{h}_t^{(b)}} L &= \frac{\partial L}{\partial \boldsymbol{h}_t^{(b)}} \\
&= \frac{\partial}{\partial \boldsymbol{h}_t^{(b)}} \sum_k L_k
\end{aligned} \tag{53}$$

For a given t, only the loss terms L_k with $k \leq t$ dependq on $\boldsymbol{h}_t^{(b)}$ (as shown in the drawing of question 1, for the backward direction), thus we can simplify the previous expression using this remark this we use the chain rule to introduce $\boldsymbol{h}_{t-1}^{(b)}$

$$\begin{aligned}
\nabla_{\boldsymbol{h}_t^{(b)}} L &= \frac{\partial}{\partial \boldsymbol{h}_t^{(b)}} \sum_{k \leq t} L_k \\
&= \frac{\partial L_t}{\partial \boldsymbol{h}_t^{(b)}} + \frac{\partial}{\partial \boldsymbol{h}_t^{(b)}} \sum_{k \leq t-1} L_k \\
&= \nabla_{\boldsymbol{h}_t^{(b)}} L_t + \left(\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}} \right)^\top \frac{\partial}{\partial \boldsymbol{h}_{t-1}^{(b)}} \sum_{k \leq t-1} L_k \\
&= \nabla_{\boldsymbol{h}_t^{(b)}} L_t + \left(\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}} \right)^\top \frac{\partial L}{\partial \boldsymbol{h}_{t-1}^{(b)}}, \quad \text{(using same argument as above)} \\
&= \nabla_{\boldsymbol{h}_t^{(b)}} L_t + \left(\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}} \right)^\top \nabla_{\boldsymbol{h}_{t-1}^{(b)}} L
\end{aligned} \tag{54}$$

Now, let's find the expressions of both $\nabla_{\boldsymbol{h}_t^{(b)}} L_t$ and $\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}}$.
We have that:

$$\begin{aligned}
\nabla_{\boldsymbol{h}_t^{(b)}} L_t &= \frac{\partial L_t}{\partial \boldsymbol{h}_t^{(b)}} \\
&= \frac{\partial}{\partial \boldsymbol{h}_t^{(b)}} ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2 \\
&= \left(\frac{\partial \boldsymbol{y}_t}{\partial \boldsymbol{h}_t^{(b)}} \right)^\top \frac{\partial}{\partial \boldsymbol{y}_t} ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2 \\
&= \left(\frac{\partial}{\partial \boldsymbol{h}_t^{(b)}} (\boldsymbol{V}^{(f)} \boldsymbol{h}_t^{(f)} + \boldsymbol{V}^{(b)} \boldsymbol{h}_t^{(b)}) \right)^\top \frac{\partial}{\partial \boldsymbol{y}_t} (\boldsymbol{y}_t^\top \boldsymbol{y}_t - 2\boldsymbol{y}_t^\top \boldsymbol{z}_t + \boldsymbol{z}_t^\top \boldsymbol{z}_t) \\
&= (\boldsymbol{V}^{(b)})^\top (2\boldsymbol{y}_t - 2\boldsymbol{z}_t) \\
&= 2(\boldsymbol{V}^{(b)})^\top (\boldsymbol{y}_t - \boldsymbol{z}_t)
\end{aligned} \tag{55}$$

And we have that:

$$\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}} = \frac{\partial}{\partial \boldsymbol{h}_t^{(b)}} \sigma(\boldsymbol{W}^{(b)} \boldsymbol{x}_t + \boldsymbol{U}^{(b)} \boldsymbol{h}_t^{(b)}) \tag{56}$$

For given i and j, we have:

$$\begin{aligned}\left(\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}}\right)_{ij} &= \frac{\partial}{\partial(\boldsymbol{h}_t^{(b)})_j}\sigma((\boldsymbol{W}^{(b)}\boldsymbol{x}_t + \boldsymbol{U}^{(b)}\boldsymbol{h}_t^{(b)})_i) \\ &= \frac{\partial}{\partial(\boldsymbol{h}_t^{(b)})_j}\sigma((\boldsymbol{W}^{(b)}\boldsymbol{x}_t)_i + \sum_k (\boldsymbol{U}^{(b)})_{ik}(\boldsymbol{h}_t^{(b)})_k) \\ &= (\boldsymbol{U}^{(b)})_{ij}\sigma'((\boldsymbol{W}^{(b)}\boldsymbol{x}_t + \boldsymbol{U}^{(b)}\boldsymbol{h}_t^{(b)})_i) \\ &= (diag(\sigma'(\boldsymbol{W}^{(b)}\boldsymbol{x}_t + \boldsymbol{U}^{(b)}\boldsymbol{h}_t^{(b)})\boldsymbol{U}^{(b)})_{ij} \\ &= (diag(\sigma'(\boldsymbol{W}^{(b)}\boldsymbol{x}_t + \boldsymbol{U}^{(f)}\boldsymbol{h}_t^{(b)})\boldsymbol{U}^{(b)})_{ij}\end{aligned} \tag{57}$$

Thus we conclude that:

$$\begin{aligned}\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}} &= diag(\sigma'(\boldsymbol{W}^{(b)}\boldsymbol{x}_t + \boldsymbol{U}^{(b)}\boldsymbol{h}_t^{(b)})\boldsymbol{U}^{(b)} \\ &= diag(\sigma(\boldsymbol{W}^{(b)}\boldsymbol{x}_t + \boldsymbol{U}^{(b)}\boldsymbol{h}_t^{(b)}))diag(1 - \sigma(\boldsymbol{W}^{(b)}\boldsymbol{x}_t + \boldsymbol{U}^{(b)}\boldsymbol{h}_t^{(b)}))\boldsymbol{U}^{(b)} \\ &= diag(\boldsymbol{h}_{t-1}^{(b)})diag(\boldsymbol{1} - \boldsymbol{h}_{t-1}^{(b)})\,\boldsymbol{U}^{(b)}\end{aligned} \tag{58}$$

where $\boldsymbol{1}$ is a vector with the same length as $\boldsymbol{h}_{t-1}^{(b)}$.
In conclusion, we express $\nabla_{\boldsymbol{h}_t^{(b)}} L$ using $\nabla_{\boldsymbol{h}_{t-1}^{(b)}} L$ as follows:

$$\boxed{\nabla_{\boldsymbol{h}_t^{(b)}} L = 2(\boldsymbol{V}^{(b)})^\top(\boldsymbol{y}_t - \boldsymbol{z}_t) + (\boldsymbol{U}^{(b)})^\top diag(\boldsymbol{h}_{t-1}^{(b)})diag(\boldsymbol{1} - \boldsymbol{h}_{t-1}^{(b)})\nabla_{\boldsymbol{h}_{t-1}^{(b)}} L} \tag{59}$$

Equations (52) and (59) allow us to answer the first part of the question. The second part consists of deriving $\nabla_{\boldsymbol{W}^{(f)}} L$ and $\nabla_{\boldsymbol{U}^{(b)}} L$.

Deriving $\nabla_{\boldsymbol{W}^{(f)}} L$:

The derivative of a scalar with respect to a matrix is simply a matrix where each element (i, j) is the partial derivative of the scalar with respect to the element (i, j) of the matrix. Formally, we can write this the following way: $(\nabla_{\boldsymbol{W}^{(f)}} L)_{ij} = \frac{\partial L}{\partial \boldsymbol{W}_{ij}^{(f)}}$. Thus, if we consider $\boldsymbol{W}_j^{(f)}$ to be the j-th column of $\boldsymbol{W}^{(f)}$, we can write the $\nabla_{\boldsymbol{W}^{(f)}} L$ as follows: $\nabla_{\boldsymbol{W}^{(f)}} L = [\nabla_{\boldsymbol{W}_1^{(f)}} L, \ldots, \nabla_{\boldsymbol{W}_d^{(f)}} L]$ (if we suppose that $\boldsymbol{W}^{(f)}$ has d columns). In other words, $\nabla_{\boldsymbol{W}^{(f)}} L$ is the matrix with columns $\nabla_{\boldsymbol{W}_j^{(f)}} L$. From the other side, we know that:

$$\nabla_{\boldsymbol{W}_j^{(f)}} L = \sum_t \left(\frac{\partial \boldsymbol{h}_t^{(f)}}{\partial \boldsymbol{W}_j^{(f)}}\right)^\top \frac{\partial L}{\partial \boldsymbol{h}_t^{(f)}} \tag{60}$$

It is important to note that instead of using $\boldsymbol{W}^{(f)}$, we can use $(\boldsymbol{W}_j^{(f)})^{(t)}$ to specify that the concerned vector is the one at time t, thus the expression becomes:

$$\nabla_{\boldsymbol{W}_j^{(f)}} L = \sum_t \left(\frac{\partial \boldsymbol{h}_t^{(f)}}{\partial (\boldsymbol{W}_j^{(f)})^{(t)}}\right)^\top \frac{\partial L}{\partial \boldsymbol{h}_t^{(f)}} \tag{61}$$

We don't do that because the notation becomes heavy. However, we keep this in mind during the calculations steps, since $\boldsymbol{h}_{t-1}^{(f)}$ for example is not a function of $(\boldsymbol{W}_j^{(f)})^{(t)}$ (see the computational graph) and it might be confusing not to note $(\boldsymbol{W}_j^{(f)})^{(t)}$ instead of $\boldsymbol{W}^{(f)}$.
The i-th element is given by:

$$\begin{aligned}(\nabla_{\boldsymbol{W}_j^{(f)}} L)_i &= \sum_t \left(\left(\frac{\partial \boldsymbol{h}_t^{(f)}}{\partial \boldsymbol{W}_j^{(f)}} \right)^\top \frac{\partial L}{\partial \boldsymbol{h}_t^{(f)}} \right)_i \\ &= \sum_t \sum_k \frac{\partial (\boldsymbol{h}_t^{(f)})_k}{\partial \boldsymbol{W}_{ij}^{(f)}} \frac{\partial L}{\partial (\boldsymbol{h}_t^{(f)})_k}\end{aligned} \tag{62}$$

We have the expression of $\frac{\partial L}{\partial (\boldsymbol{h}_t^{(f)})_k}$ thanks to the previous question. Let's calculate $\frac{\partial (\boldsymbol{h}_t^{(f)})_k}{\partial \boldsymbol{W}_{ij}^{(f)}}$:

$$\begin{aligned}\frac{\partial (\boldsymbol{h}_t^{(f)})_k}{\partial \boldsymbol{W}_{ij}^{(f)}} &= \frac{\partial}{\partial \boldsymbol{W}_{ij}^{(f)}} \sigma(\boldsymbol{W}^{(f)} \boldsymbol{x}_t + \boldsymbol{U}^{(f)} \boldsymbol{h}_{t-1}^{(f)})_k \\ &= \frac{\partial}{\partial \boldsymbol{W}_{ij}^{(f)}} \sigma \left(\sum_p \boldsymbol{W}_{kp}^{(f)} (\boldsymbol{x}_t)_p + (\boldsymbol{U}^{(f)} \boldsymbol{h}_{t-1}^{(f)})_k \right) \\ &= \delta_{i,k}\, \sigma'(\boldsymbol{W}^{(f)} \boldsymbol{x}_t + \boldsymbol{U}^{(f)} \boldsymbol{h}_{t-1}^{(f)})_k (\boldsymbol{x}_t)_j \\ &= \delta_{i,k}\, \sigma(\boldsymbol{W}^{(f)} \boldsymbol{x}_t + \boldsymbol{U}^{(f)} \boldsymbol{h}_{t-1}^{(f)})_k (1 - \sigma(\boldsymbol{W}^{(f)} \boldsymbol{x}_t + \boldsymbol{U}^{(f)} \boldsymbol{h}_{t-1}^{(f)})_k) (\boldsymbol{x}_t)_j \\ &= \delta_{i,k}\, (\boldsymbol{h}_t^{(f)})_k (1 - (\boldsymbol{h}_t^{(f)})_k) (\boldsymbol{x}_t)_j\end{aligned} \tag{63}$$

We obtain:

$$\begin{aligned}(\nabla_{\boldsymbol{W}_j^{(f)}} L)_i &= \sum_t \sum_k \delta_{i,k}\, (\boldsymbol{h}_t^{(f)})_k (1 - (\boldsymbol{h}_t^{(f)})_k) \frac{\partial L}{\partial (\boldsymbol{h}_t^{(f)})_k} (\boldsymbol{x}_t)_j \\ &= \sum_t (\boldsymbol{h}_t^{(f)})_i (1 - (\boldsymbol{h}_t^{(f)})_i) \frac{\partial L}{\partial (\boldsymbol{h}_t^{(f)})_i} (\boldsymbol{x}_t)_j \\ &= \sum_t (\boldsymbol{h}_t^{(f)})_i (1 - (\boldsymbol{h}_t^{(f)})_i) \left(\nabla_{\boldsymbol{h}_t^{(f)}} L\ \boldsymbol{x}_t^\top \right)_{ij} \\ &= \sum_t \left(diag(\boldsymbol{h}_t^{(f)})\, diag(\mathbf{1} - \boldsymbol{h}_t^{(f)})\, \nabla_{\boldsymbol{h}_t^{(f)}} L\ \boldsymbol{x}_t^\top \right)_{ij}\end{aligned} \tag{64}$$

where $\mathbf{1}$ is a vector of the same length as $\boldsymbol{h}_t^{(f)}$.
We conclude that:

$$\boxed{\nabla_{\boldsymbol{W}^{(f)}} L = \sum_t diag(\boldsymbol{h}_t^{(f)})\, diag(\mathbf{1} - \boldsymbol{h}_t^{(f)})\, \nabla_{\boldsymbol{h}_t^{(f)}} L\ \boldsymbol{x}_t^\top} \tag{65}$$

Taking into account the expression we got for $\nabla_{\boldsymbol{h}_t^{(f)}} L$ in equation (52), we have:

$$\nabla_{\boldsymbol{W}^{(f)}} L = \sum_t diag(\boldsymbol{h}_t^{(f)})\, diag(\mathbf{1} - \boldsymbol{h}_t^{(f)}) \left(2(\boldsymbol{V}^{(f)})^\top (\boldsymbol{y}_t - \boldsymbol{z}_t) + (\boldsymbol{U}^{(f)})^\top diag(\boldsymbol{h}_{t+1}^{(f)}) diag(\mathbf{1} - \boldsymbol{h}_{t+1}^{(f)}) \nabla_{\boldsymbol{h}_{t+1}^{(f)}} L \right) \boldsymbol{x}_t^\top$$

Deriving $\nabla_{\boldsymbol{U}^{(b)}} L$:

For the same reasons we explained above, can write this the following way: $(\nabla_{\boldsymbol{U}^{(b)}} L)_{ij} = \frac{\partial L}{\partial \boldsymbol{U}^{(b)}_{ij}}$. Thus, if we consider $\boldsymbol{U}^{(b)}_j$ to be the j-th column of $\boldsymbol{U}^{(b)}$, we can write the $\nabla_{\boldsymbol{U}^{(b)}} L$ as follows: $\nabla_{\boldsymbol{U}^{(b)}} L = [\nabla_{\boldsymbol{U}^{(b)}_1} L, \ldots, \nabla_{\boldsymbol{U}^{(b)}_m} L]$ (if we suppose that $\boldsymbol{U}^{(b)}$ has m columns). In other words, $\nabla_{\boldsymbol{U}^{(b)}} L$ is the matrix with columns $\nabla_{\boldsymbol{U}^{(b)}_j} L$.
From the other side, we know that:

$$\nabla_{\boldsymbol{U}^{(b)}_j} L = \sum_t \left(\frac{\partial \boldsymbol{h}^{(b)}_t}{\partial \boldsymbol{U}^{(b)}_j} \right)^\top \frac{\partial L}{\partial \boldsymbol{h}^{(b)}_t} \tag{66}$$

In the same fashion, it is important to note that instead of using $\boldsymbol{U}^{(b)}$, we can use $(\boldsymbol{U}^{(b)}_j)^{(t)}$ to specify that the concerned vector is the one at time t, thus the expression becomes:

$$\nabla_{\boldsymbol{U}^{(b)}_j} L = \sum_t \left(\frac{\partial \boldsymbol{h}^{(b)}_t}{\partial (\boldsymbol{U}^{(b)}_j)^t} \right)^\top \frac{\partial L}{\partial \boldsymbol{h}^{(b)}_t} \tag{67}$$

We don't do that because the notation becomes heavy. However, we keep this in mind during the calculations steps as well.
The i-th element is given by:

$$\begin{aligned} (\nabla_{\boldsymbol{U}^{(b)}_j} L)_i &= \sum_t \left(\left(\frac{\partial \boldsymbol{h}^{(b)}_t}{\partial \boldsymbol{U}^{(b)}_j} \right)^\top \frac{\partial L}{\partial \boldsymbol{h}^{(b)}_t} \right)_i \\ &= \sum_t \sum_k \frac{\partial (\boldsymbol{h}^{(b)}_t)_k}{\partial \boldsymbol{U}^{(b)}_{ij}} \frac{\partial L}{\partial (\boldsymbol{h}^{(b)}_t)_k} \end{aligned} \tag{68}$$

We have the expression of $\frac{\partial L}{\partial (\boldsymbol{h}^{(b)}_t)_k}$ thanks to the previous question. Let's calculate $\frac{\partial (\boldsymbol{h}^{(b)}_t)_k}{\partial \boldsymbol{U}^{(b)}_{ij}}$:

$$\begin{aligned} \frac{\partial (\boldsymbol{h}^{(b)}_t)_k}{\partial \boldsymbol{U}^{(b)}_{ij}} &= \frac{\partial}{\partial \boldsymbol{U}^{(b)}_{ij}} \sigma(\boldsymbol{W}^{(b)} \boldsymbol{x}_t + \boldsymbol{U}^{(b)} \boldsymbol{h}^{(b)}_{t+1})_k \\ &= \frac{\partial}{\partial \boldsymbol{U}^{(b)}_{ij}} \sigma \left(\sum_p \boldsymbol{U}^{(b)}_{kp} (\boldsymbol{h}^{(b)}_{t+1})_p + (\boldsymbol{W}^{(b)} \boldsymbol{x}_t)_k \right) \\ &= \delta_{i,k}\, \sigma'(\boldsymbol{W}^{(b)} \boldsymbol{x}_t + \boldsymbol{U}^{(b)} \boldsymbol{h}^{(b)}_{t+1})_k (\boldsymbol{h}^{(b)}_{t+1})_j \\ &= \delta_{i,k}\, \sigma(\boldsymbol{W}^{(b)} \boldsymbol{x}_t + \boldsymbol{U}^{(b)} \boldsymbol{h}^{(b)}_{t+1})_k (1 - \sigma(\boldsymbol{W}^{(b)} \boldsymbol{x}_t + \boldsymbol{U}^{(b)} \boldsymbol{h}^{(b)}_{t+1}) (\boldsymbol{h}^{(b)}_{t+1})_j \\ &= \delta_{i,k}\, (\boldsymbol{h}^{(b)}_t)_k (1 - (\boldsymbol{h}^{(b)}_t)_k) (\boldsymbol{h}^{(b)}_{t+1})_j \end{aligned} \tag{69}$$

We obtain:

$$\begin{aligned}
(\nabla_{\boldsymbol{U}_j^{(b)}} L)_i &= \sum_t \sum_k \delta_{i,k} \, (\boldsymbol{h}_t^{(b)})_k (1 - (\boldsymbol{h}_t^{(b)})_k) \frac{\partial L}{\partial (\boldsymbol{h}_t^{(b)})_k} (\boldsymbol{h}_{t+1}^{(b)})_j \\
&= \sum_t (\boldsymbol{h}_t^{(b)})_i (1 - (\boldsymbol{h}_t^{(b)})_i) \frac{\partial L}{\partial (\boldsymbol{h}_t^{(b)})_i} (\boldsymbol{h}_{t+1}^{(b)})_j \\
&= \sum_t (\boldsymbol{h}_t^{(b)})_i (1 - (\boldsymbol{h}_t^{(b)})_i) \left(\nabla_{\boldsymbol{h}_t^{(b)}} L \; (\boldsymbol{h}_{t+1}^{(b)})^\top \right)_{ij} \\
&= \sum_t \left(diag(\boldsymbol{h}_t^{(b)}) \, diag(\mathbf{1} - \boldsymbol{h}_t^{(b)}) \, \nabla_{\boldsymbol{h}_t^{(b)}} L \; (\boldsymbol{h}_{t+1}^{(b)})^\top \right)_{ij}
\end{aligned} \tag{70}$$

where $\mathbf{1}$ is a vector of the same length as $\boldsymbol{h}_t^{(b)}$.
We conclude that:

$$\boxed{\nabla_{\boldsymbol{U}^{(b)}} L = \sum_t diag(\boldsymbol{h}_t^{(b)}) \, diag(\mathbf{1} - \boldsymbol{h}_t^{(b)}) \, \nabla_{\boldsymbol{h}_t^{(b)}} L \; (\boldsymbol{h}_{t+1}^{(b)})^\top} \tag{71}$$

Taking into account the expression of $\nabla_{\boldsymbol{h}_t^{(b)}} L$ that we found in (59), we can write:

$$\nabla_{\boldsymbol{U}^{(b)}} L = \sum_t diag(\boldsymbol{h}_t^{(b)}) \, diag(\mathbf{1} - \boldsymbol{h}_t^{(b)}) \left(2(\boldsymbol{V}^{(b)})^\top (\boldsymbol{y}_t - \boldsymbol{z}_t) + (\boldsymbol{U}^{(b)})^\top diag(\boldsymbol{h}_{t-1}^{(b)}) diag(\mathbf{1} - \boldsymbol{h}_{t-1}^{(b)}) \nabla_{\boldsymbol{h}_{t-1}^{(b)}} L \right) (\boldsymbol{h}_{t+1}^{(b)})^\top$$