IFT 6135 - Homework 2

Adel Nabli

15/03/2019

1 Question 1

We have:

$$h^{(t)} = g(a^{(t)}) \quad ; \quad a^{(t)} = W^{(t)}h^{(t-1)} + b^{(t)} \; with \; W^{(t)} \in \mathbb{R}^{d^{(t)} \times d^{(t-1)}}, \; b^{(t)} = \underbrace{(c, ..., c)^T}_{d^{(t)} \; times}, \; c \in \mathbb{R}$$

We want $\forall i \in [\![1, d^{(t)}]\!]$, $\mathbb{E}\big[a_i^{(t)}\big] = 0$ and $\mathbb{V}\big[a_i^{(t)}\big] = 1$. But, we can write:

$$\forall i \in [\![1, d^{(t)}]\!], \; \mathbb{E}\big[a_i^{(t)}\big] = 0 \Leftrightarrow \mathbb{E}\Big[c + \sum_{j=1}^{d^{(t-1)}} w_{ij}^{(t)} h_j^{(t-1)}\Big] = 0 \tag{1}$$

$$by \; linearity \; \rightarrow \Leftrightarrow c = -\sum_{j=1}^{d^{(t-1)}} \mathbb{E}\big[w_{ij}^{(t)} h_j^{(t-1)}\big] \tag{2}$$

$$W^{(t)} \perp\!\!\!\perp h^{(t-1)} \; \rightarrow \Leftrightarrow c = -\sum_{j=1}^{d^{(t-1)}} \mathbb{E}\big[w_{ij}^{(t)}\big]\mathbb{E}\big[h_j^{(t-1)}\big] \tag{3}$$

$$w_{ij}^{(t)} \; are \; i.i.d \rightarrow \Leftrightarrow c = -\mathbb{E}\big[w_{i1}^{(t)}\big] \sum_{j=1}^{d^{(t-1)}} \mathbb{E}\big[h_j^{(t-1)}\big] \tag{4}$$

And for the variance, we have:

$$\forall i \in [\![1, d^{(t)}]\!], \; \mathbb{V}\big[a_i^{(t)}\big] = 1 \Leftrightarrow \mathbb{V}\Big[c + \sum_{j=1}^{d^{(t-1)}} w_{ij}^{(t)} h_j^{(t-1)}\Big] = 1 \tag{5}$$

$$\Leftrightarrow 1 = 0 + \mathbb{V}\Big[\sum_{j=1}^{d^{(t-1)}} w_{ij}^{(t)} h_j^{(t-1)}\Big] \tag{6}$$

$$\Leftrightarrow 1 = \sum_{j=1}^{d^{(t-1)}} \mathbb{V}\big[w_{ij}^{(t)} h_j^{(t-1)}\big] + 2 \sum_{1 \leq j_1 < j_2 \leq d^{(t-1)}} Cov\big[w_{ij_1}^{(t)} h_{j_1}^{(t-1)}, w_{ij_2}^{(t)} h_{j_2}^{(t-1)}\big] \tag{7}$$

1. First, we assume that $\forall j \in [\![1, d^{(t-1)}]\!]$, $\mathbb{E}\big[h_j^{(t-1)}\big] = 0$ and $\mathbb{V}\big[h_j^{(t-1)}\big] = 1$. Thus, taking (4), we have:

$$c = -\mathbb{E}\big[w_{i1}^{(t)}\big] \sum_{j=1}^{d^{(t-1)}} \underbrace{\mathbb{E}\big[h_j^{(t-1)}\big]}_{=0} = 0$$

Moreover, we assume that the entries of $h^{(t-1)}$ are uncorrelated and we know that the w_{ij} are i.i.d, which means that we can rewrite (7) as follows:

$$1 = \sum_{j=1}^{d^{(t-1)}} \mathbb{V}\big[w_{ij}^{(t)} h_j^{(t-1)}\big] + 2 \sum_{1 \le j_1 < j_2 \le d^{(t-1)}} \underbrace{Cov\big[w_{ij_1}^{(t)} h_{j_1}^{(t-1)}, w_{ij_2}^{(t)} h_{j_2}^{(t-1)}\big]}_{=0} \tag{8}$$

$$1 = \sum_{j=1}^{d^{(t-1)}} \mathbb{V}[w_{ij}^{(t)}]\Big(\underbrace{\mathbb{V}[h_j^{(t-1)}]}_{=1} + \underbrace{\mathbb{E}[h_j^{(t-1)}]^2}_{=0}\Big) + \underbrace{\mathbb{V}[h_j^{(t-1)}]}_{=1}\mathbb{E}[w_{ij}^{(t)}]^2 \tag{9}$$

$$1 = \sum_{j=1}^{d^{(t-1)}} \mathbb{V}[w_{ij}^{(t)}] + \mathbb{E}[w_{ij}^{(t)}]^2 = d^{(t-1)}\Big(\mathbb{V}[w_{i1}^{(t)}] + \mathbb{E}[w_{i1}^{(t)}]^2\Big) \tag{10}$$

(a) If we assume that the $w_{ij} \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$, then we can rewrite (10) and have:

$$1 = d^{(t-1)}\Big(\mathbb{V}[w_{i1}^{(t)}] + \mathbb{E}[w_{i1}^{(t)}]^2\Big) = d^{(t-1)}\big[\mu^2 + \sigma^2\big]$$

Thus, here, taking $\mu = 0$ and $\sigma = \frac{1}{\sqrt{d^{(t-1)}}}$ works *(recall that we already showed that $c = 0$)*.

(b) If we assume that the $w_{ij} \overset{i.i.d}{\sim} \mathcal{U}(\alpha, \beta)$, then we can rewrite (10) and have:

$$1 = d^{(t-1)}\Big(\mathbb{V}[w_{i1}^{(t)}] + \mathbb{E}[w_{i1}^{(t)}]^2\Big) = d^{(t-1)}\big[\frac{(\beta-\alpha)^2}{12} + \frac{(\alpha+\beta)^2}{4}\big] = d^{(t-1)}\frac{\alpha^2+\beta^2+\alpha\beta}{3}$$

Thus, if we decide to take $\alpha = -\beta$, taking $\beta = \sqrt{\frac{3}{d^{(t-1)}}}$ works *(we still have $c = 0$)*.

2. Now, we assume that $\forall j \in [\![1, d^{(t-1)}]\!]$, $\mathbb{V}\big[a_j^{(t-1)}\big] = 1$ and $\mathbb{E}\big[a_j^{(t-1)}\big] = 0$, that $a_j^{(t-1)}$ has a symmetric distribution, and that $g(x) = RELU(x) = x\mathbb{1}_{x>0}(x)$. Then, we have:

$$\begin{aligned}
\forall j \in [\![1, d^{(t-1)}]\!],\ 1 = \mathbb{V}\big[a_j^{(t-1)}\big] &\Leftrightarrow 1 = \mathbb{E}\big[(a_j^{(t-1)})^2\big] - \underbrace{\mathbb{E}\big[a_j^{(t-1)}\big]^2}_{=0} \\
&\Leftrightarrow 1 = \int_{-\infty}^{\infty} x^2 \underbrace{p_{a_j^{(t-1)}}(x)}_{symmetric}\, dx \\
&\Leftrightarrow 1 = 2\int_{0}^{\infty} x^2 p_{a_j^{(t-1)}}(x) dx \\
&\Leftrightarrow 1 = 2\int_{-\infty}^{\infty} \Big(x\mathbb{1}_{x>0}(x)\Big)^2 p_{a_j^{(t-1)}}(x) dx \\
&\Leftrightarrow \frac{1}{2} = \mathbb{E}\big[g(a_j^{(t-1)})^2\big]
\end{aligned}$$

We know that the w_{ij} are i.i.d. As the $a^{(t-1)}$ are uncorrelated and $h_j^{(t-1)}$ is either equal to 0 or $a_j^{(t-1)}$, we still have that the covariance term is equal to 0 in (7). Thus, we can write:

$$1 = \sum_{j=1}^{d^{(t-1)}} \mathbb{V}[w_{ij}^{(t)}] \underbrace{\Big(\mathbb{V}[h_j^{(t-1)}] + \mathbb{E}[h_j^{(t-1)}]^2\Big)}_{=\mathbb{E}\left[g(a_j^{(t-1)})^2\right]=1/2} + \mathbb{V}[h_j^{(t-1)}]\mathbb{E}[w_{ij}^{(t)}]^2 \tag{11}$$

$$1 = \frac{d^{(t-1)}\mathbb{V}[w_{i1}^{(t)}]}{2} + \mathbb{E}[w_{i1}^{(t)}]^2 \sum_{j=1}^{d^{(t-1)}} \mathbb{V}[h_j^{(t-1)}] \tag{12}$$

Now, as $a_j^{(t-1)}$ has a symmetric distribution (which is not a dirac on 0 as $\mathbb{V}\big[a_j^{(t-1)}\big] = 1$), we have that $\forall j \in [\![1, d^{(t-1)}]\!]$, $\mathbb{E}\big[h_j^{(t-1)}\big] = \mathbb{E}\big[g(a_j^{(t-1)})\big] > 0$. Then, we deduce from (4) that $\mathbb{E}[w_{i1}^{(t)}] = 0 \Leftrightarrow c = 0$.

(a) In the case of the gaussian, if we decide that $c = 0$ or equivalently $\mu = 0$, we set to 0 the second term of the right-hand side of (12) and deduce:

$$c = 0 \quad ; \quad \mu = 0 \quad ; \quad \sigma = \sqrt{\frac{2}{d^{(t-1)}}}$$

(b) In the case of the uniform distribution, if we decide that $c = 0$ or equivalently $\alpha = -\beta$, we set to 0 the second term of the right-hand side of (12) and deduce:

$$c = 0 \quad ; \quad \alpha = -\beta \quad ; \quad \beta = \sqrt{\frac{6}{d^{(t-1)}}}$$

(c) This kind of scheme reminds us of the Glorot initialization we already encountered in the previous homework.

(d) Making sure that the distribution at each layer is normalized and standardized helps propagate the information necessary for the model to converge and learn, especially in the case of "deep" models where we have to take particular care at the initialization if we want to avoid problems of vanishing/exploding gradients.

2 Question 2

We have a linear regression problem with a design matrix $X \in \mathbb{R}^{n \times d}$ and a vector $y \in \mathbb{R}^n$. We apply a dropout mask $R \in \{0, 1\}^{n \times d}$ on the input and want to learn the set of weights $w \in \mathbb{R}^d$.

1. If we note $\odot$ the Hadamard product of matrices, the squared error loss is expressed as:

$$L(w) = \|(R \odot X)w - y\|_2^2$$

2. Let's write the expectation of the loss L with respect to R:

$$\begin{aligned}
\mathbb{E}_R\big[L(w)\big] &= \mathbb{E}_R\Big[\|(R \odot X)w - y\|_2^2\Big] && (13)\\
&= \mathbb{E}_R\Big[\big((R \odot X)w - y\big)^T\big((R \odot X)w - y\big)\Big] && (14)\\
&= \mathbb{E}_R\Big[w^T(R \odot X)^T(R \odot X)w - 2y^T(R \odot X)w + y^Ty\Big] && (15)\\
&= w^T\mathbb{E}_R\Big[(R \odot X)^T(R \odot X)\Big]w - 2y^T\mathbb{E}_R\Big[R \odot X\Big]w + y^Ty && (16)
\end{aligned}$$

But we know that $\forall i, j \in [\![1, n]\!] \times [\![1, d]\!]$, $R_{ij} \overset{i.i.d}{\sim} \mathcal{B}(p)$. Then, we can write:

$$\mathbb{E}_R\big[R \odot X\big]_{ij} = \mathbb{E}_R\big[R_{ij}x_{ij}\big] = px_{ij}$$

And then, $\mathbb{E}_R\big[R \odot X\big] = pX$. We also have:

$$\begin{aligned}
\mathbb{E}_R\big[(R \odot X)^T(R \odot X)\big]_{ij} &= \mathbb{E}_R\Big[\sum_{k=1}^{n} R_{ki}x_{ki}R_{kj}x_{kj}\Big]\\
&= \sum_{k=1}^{n} x_{ki}x_{kj}\mathbb{E}_R\big[R_{ki}R_{kj}\big]
\end{aligned}$$

But we have:

$$\begin{cases} \forall i \neq j,\ \mathbb{E}_R\big[R_{ki}R_{kj}\big] = \mathbb{E}_R\big[R_{ki}\big]\mathbb{E}_R\big[R_{kj}\big] = p^2 \\ if\ i = j,\ \mathbb{E}_R\big[R_{ki}R_{kj}\big] = \mathbb{E}_R\big[R_{ki}^2\big] = p = p^2 + p(1-p) \end{cases}$$

And then $\mathbb{E}_R\Big[(R \odot X)^T(R \odot X)\Big] = p^2X^TX + p(1-p)\underbrace{I_d \odot (X^TX)}_{=\Gamma^2}$.

Putting all that together in (16), we obtain:

$$\begin{aligned} \mathbb{E}_R\big[L(w)\big] &= w^T\Big[p^2X^TX + p(1-p)\Gamma^2\Big]w - 2py^TXw + y^Ty \\ &= \underbrace{p^2w^TX^TXw - 2py^TXw + y^Ty}_{=\|y-pXw\|_2^2} + \underbrace{p(1-p)w^T\Gamma^2w}_{=p(1-p)\|\Gamma w\|_2^2} \end{aligned}$$

3. As we have X^TX and Γ^2 are semi-definite positive, $\mathbb{E}_R\big[L(w)\big]$ is convex in w. Then, to minimize it, we can find an w^* that sets its gradient to 0:

$$\begin{aligned} \frac{\partial \mathbb{E}_R\big[L(w)\big]}{\partial w} = 0 &\Leftrightarrow 0 = 2p^2X^TXw - 2pX^Ty + 2p(1-p)\Gamma^2w \\ &\Leftrightarrow X^Ty = \big(pX^TX + (1-p)\Gamma^2\big)w \\ &\Leftrightarrow X^Ty = \big(X^TX + \frac{(1-p)}{p}\Gamma^2\big)pw \\ &\Leftrightarrow \big(X^TX + \underbrace{\frac{(1-p)}{p}}_{=\lambda^{dropout}}\Gamma^2\big)^{-1}X^Ty = pw^{dropout} \end{aligned}$$

Thus, we observe that $\lim_{p\to 0}\lambda^{dropout} = \infty$ and $\lim_{p\to 1}\lambda^{dropout} = 0$.

4. With L_2 regularization, we have:

$$L_{L_2}(w) = \|y - Xw\|_2^2 + \lambda^{L_2}\|w\|_2^2$$

And setting the gradient to 0 leads to:

$$\begin{aligned} \frac{\partial L_{L_2}(w)}{\partial w} = 0 &\Leftrightarrow 0 = -2X^Ty + 2X^TXw + 2\lambda^{L_2}w \\ &\Leftrightarrow \big(X^TX + \lambda^{L_2}I_d\big)^{-1}X^Ty = w^{L_2} \end{aligned}$$

5. One major difference between the two results is that the one obtained on dropout is an expectation whereas the one with L_2 regularization is deterministic. We could say that λ^{L_2} and $\lambda^{dropout}$ play a similar role, but the fact is $\lambda^{dropout}$ is parametrized by p which also appears elsewhere in the equation, making it harder to interpret. An other difference between the two equations is that in dropout, due to Γ^2, we penalize more the rows of X which has a greater mass/norm, whereas in the L_2 regularization scheme the penalization is uniform on all the rows.

3 Question 3

1. For $t \geq 1$, we have:

   - **SGD momentum:**

   $$v_t = \alpha v_{t-1} + \epsilon g_t \quad ; \quad \Delta\theta_t = -v_t \; with \; \epsilon > 0, \; \alpha \in [0, 1]$$

   Which leads to:

   $$\Delta\theta_t = -\alpha v_{t-1} - \epsilon g_t = \alpha \Delta\theta_{t-1} - \epsilon g_t \tag{17}$$

   - **SGD running average:**

   $$v_t = \beta v_{t-1} + (1-\beta) g_t \quad ; \quad \Delta\theta_t = -\delta v_t \; with \; \delta > 0, \; \beta \in [0, 1]$$

   Which leads to:

   $$\Delta\theta_t = -\delta\beta v_{t-1} - \delta(1-\beta) g_t = \beta \Delta\theta_{t-1} - \delta(1-\beta) g_t \tag{18}$$

   Identifying (17) with (18) leads us to write that if $\alpha = \beta$ and $\epsilon = \delta(1-\beta)$, then the two schemes are equivalent.

2. Using the definition, we have:

   $$v_1 = \beta \underbrace{v_0}_{=0} + (1-\beta) g_1 = (1-\beta) g_1$$
   $$v_2 = \beta v_1 + (1-\beta) g_2 = \beta(1-\beta) g_1 + (1-\beta) g_2$$

   From that, we suppose that the formula is, $\forall n \geq 1, \; v_n = \sum_{t=1}^{n} \beta^{n-t}(1-\beta) g_t$. Let's prove it by recurrence:

   - **For** $n = 1$**,** $v_1 = \sum_{t=1}^{1} \beta^{1-1}(1-\beta) g_1 = (1-\beta) g_1$ which is good.
   - Let's suppose the formula true at rank n and let's show it for $n+1$:

   $$v_{n+1} = \beta v_n + (1-\beta) g_{n+1}$$
   $$v_{n+1} = \beta \sum_{t=1}^{n} \beta^{n-t}(1-\beta) g_t + (1-\beta) g_{n+1}$$
   $$v_{n+1} = \sum_{t=1}^{n+1} \beta^{n+1-t}(1-\beta) g_t$$

   And the formula is also true at rank $n+1$.

   Then, we can write: $\forall n \in \mathbb{N}^*, \; v_n = \sum_{t=1}^{n} \beta^{n-t}(1-\beta) g_t$.

3. From the formula derived in the previous question, and using the fact that g_t has a stationary distribution independent of t, we have $\forall n \geq 1$:

   $$\mathbb{E}[v_n] = \mathbb{E}[g_n](1-\beta) \sum_{t=1}^{n} \beta^{n-t}$$
   $$\mathbb{E}[v_n] = \mathbb{E}[g_n](1-\beta)\beta^n \sum_{t=1}^{n} \left(\frac{1}{\beta}\right)^t$$
   $$\mathbb{E}[v_n] = \mathbb{E}[g_n](1-\beta)\beta^n \frac{1}{\beta} \frac{1 - (1/\beta)^n}{1 - (1/\beta)}$$
   $$\mathbb{E}[v_n] = \mathbb{E}[g_n](1-\beta^n) \neq \mathbb{E}[g_n]$$

   Then, if we rescale v_n by $1-\beta^n$ we'll have $v'_n = \dfrac{v_n}{1-\beta^n}$ and $\forall n \geq 1, \; \mathbb{E}[v'_n] = \mathbb{E}[g_n]$.

4 Question 4

1. We have $w = \gamma \frac{u}{\|u\|}$ and $y = u^T x$. Then, we can write:

$$w^T x + \beta = \gamma \frac{u^T}{\|u\|} x + \beta = \gamma \frac{y}{\|u\|} + \beta \tag{19}$$

But we know that $\mathbb{E}[x] = 0$ then we have:

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

Let's call n the dimension of the random vector x. Then, we have $\forall i \in [\![1, n]\!]$, $\mathbb{V}[x_i] = 1$. Moreover, if we **suppose that the entries of x are uncorrelated**, we can write:

$$\sigma_y^2 = \mathbb{V}[y] = \mathbb{V}[u^T x] = \mathbb{V}\big[\sum_{i=1}^{n} u_i x_i\big] = \sum_{i=1}^{n} u_i^2 \underbrace{\mathbb{V}[x_i]}_{i=1} + 2 \sum_{1 \le i < j \le n} \underbrace{Cov(u_i x_i, u_j x_j)}_{=0} = \|u\|^2$$

Thus, taking back (19), we finally can identify $\hat{y}$:

$$w^T x + \beta = \gamma \frac{y}{\|u\|} + \beta = \gamma \frac{y - \mu_y}{\sigma_y} + \beta = \hat{y}$$

2. Using the chain-rule, and the fact that $w(u) = \gamma \frac{u}{\|u\|}$, we derive:

$$\nabla_u L = \frac{\partial L}{\partial u} = \frac{\partial w(u)}{\partial u} \frac{\partial L}{\partial w} = \Big(\frac{\gamma}{\|u\|} I - \frac{1}{2} \frac{\gamma u}{\|u\|^3} 2u^T \Big) \nabla_w L = \underbrace{\frac{\gamma}{\|u\|}}_{=s} \underbrace{\Big(I - \frac{uu^T}{\|u\|^2} \Big)}_{=W^\perp} \nabla_w L$$

3. First, let's write the gradient descent update *(we define α as the stepsize)*:

$$u_{t+1} = u_t - \alpha \nabla_{u_t} L$$
$$u_{t+1} = u_t - \alpha s \underbrace{W^\perp \nabla_{w_t} L}_{\in Vect(u_t)^\perp}$$

To justify what we wrote under the brace, let's note that $E = Vect(u_t) \oplus Vect(u_t)^\perp$ with E the vector space in which we are working *(as we are working in vector spaces of finite dimension, this is always true)*. Then, every vector a can be decomposed as $a = a_{Vect(u_t)} + a_{Vect(u_t)^\perp}$. As $W^\perp$ is a projection on $Vect(w_t)^\perp = Vect(u_t)^\perp$ *(as $w_t \propto u_t$)*, we have that for every vector a, $W^\perp a = W^\perp (a_{Vect(u_t)} + a_{Vect(u_t)^\perp}) = W^\perp a_{Vect(u_t)^\perp} + 0 \in Vect(u_t)^\perp$. In particular, this is true for $a = \nabla_{w_t} L$.
Then, as $u_t \perp (\alpha s W^\perp \nabla_{w_t} L)$, by the pythagorean theorem we can write:

$$\|u_{t+1}\|^2 = \|u_t\|^2 + \|\alpha s W^\perp \nabla_{w_t} L\|^2 \ge \|u_t\|^2$$

Then, we showed that the norm of u increases at each update. From that last equation, it is also obvious that larger values of α implies larger values of $\|u_{t+1}\|$ and leads to faster growth.

5 Question 5

1. Let's re-write g_t to see if we can gain some intuition:

$$g_t = \sigma\Big(Wg_{t-1} + Ux_t + b\Big) \tag{20}$$
$$\sigma^{-1}(g_t) = Wg_{t-1} + Ux_t + b \tag{21}$$

But we know that $h_t = W\sigma(h_{t-1}) + Ux_t + b$, thus, if we make an identification with (21), we are tempted to write $\forall t,\ h_t = \sigma^{-1}(g_t)$ and $g_{t-1} = \sigma(h_{t-1})$. Those two conditions being equivalent, setting $\forall t,\ g_t = \sigma(h_t)$ suffices to make the recurrence for h_t and g_t equivalent.

2. We have that $\forall t,\ h_t = W\sigma(h_{t-1}) + Ux_t + b$. Thus, we have:

$$\frac{\partial h_t}{\partial h_{t-1}} = \frac{\partial \sigma(h_{t-1})}{\partial h_{t-1}} \frac{\partial h_t}{\partial \sigma(h_{t-1})} = diag\big(\sigma'(h_{t-1})\big)W^T$$

Thus, using the chain-rule, we can write:

$$\forall T,\ \|\frac{\partial h_T}{\partial h_0}\| = \|\frac{\partial h_1}{\partial h_0}\frac{\partial h_2}{\partial h_1}...\frac{\partial h_T}{\partial h_{T-1}}\| \tag{22}$$
$$= \|diag\big(\sigma'(h_0)\big)W^T...diag\big(\sigma'(h_{T-1})\big)W^T\| \tag{23}$$
$$\|AB\| \leq \|A\|\|B\| \rightarrow \leq \|diag\big(\sigma'(h_0)\big)\|\|W^T\|...\|diag\big(\sigma'(h_{T-1})\big)\|\|W^T\| \tag{24}$$

But, we have:

$$\forall t,\ \|diag\big(\sigma'(h_t)\big)\| = \sqrt{\lambda_1\Big(diag\big(\sigma'(h_t)\big)^2\Big)}$$
$$\forall x\ |\sigma'(x)| \leq \gamma \rightarrow \leq \sqrt{\lambda_1\Big(diag\big(\gamma\big)^2\Big)}$$
$$\gamma > 0 \rightarrow \leq \sqrt{\gamma^2} = \gamma$$

Lemma 5.1. *For every matrix W, $Sp(WW^T)$ the set of eigenvalues of WW^T is equal to $Sp(W^TW)$ the set of eigenvalues of W^TW.*

Proof.
- $Sp(WW^T) \subset Sp(W^TW)$:

$$\lambda \in Sp(WW^T) \Rightarrow \exists u\ s.t\ WW^Tu = \lambda u$$
$$\Rightarrow W^TW\underbrace{W^Tu}_{=v} = \lambda\underbrace{W^Tu}_{=v}$$
$$\Rightarrow \lambda \in Sp(W^TW)$$

- $Sp(W^TW) \subset Sp(WW^T)$: This is shown the same way.

□

Then, we can show:

$$\|W^T\| = \sqrt{\lambda_1\big(WW^T\big)}$$
$$Sp(W^TW) = Sp(WW^T) \rightarrow \|W^T\| = \sqrt{\lambda_1\big(W^TW\big)}$$
$$\lambda_1\big(W^TW\big) \leq \frac{\delta^2}{\gamma^2} \rightarrow \|W^T\| \leq \frac{\delta}{\gamma}$$

Then, we can continue the derivations of (24):

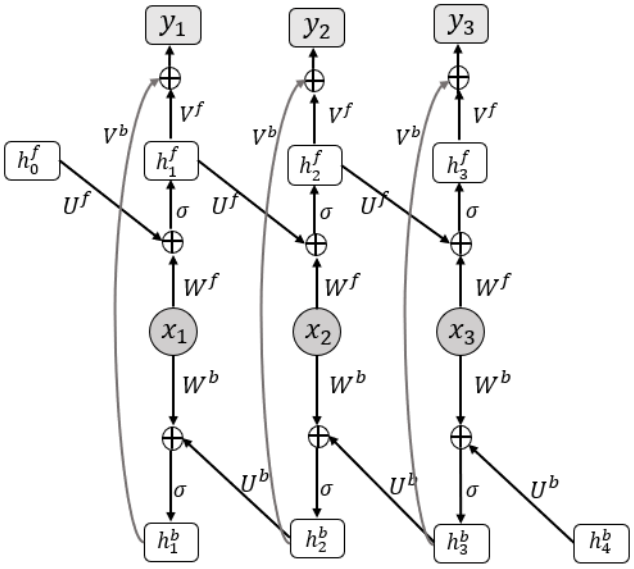
$$\forall T,\ \|\frac{\partial h_T}{\partial h_0}\| \leq \gamma^T \|W^T\|^T \leq \gamma^T \Big(\frac{\delta}{\gamma}\Big)^T = \delta^T \tag{25}$$

And as $0 \leq \delta < 1$, we have $\lim_{T\to\infty} \delta^T = 0$ and then $\lim_{T\to\infty} \|\frac{\partial h_T}{\partial h_0}\| = 0$.

3. If the largest eigenvalue of the weights is greater than $\frac{\delta^2}{\gamma^2}$, then we can suppose that instead of vanishing, the gradients will explode. This condition will be **necessary** but **not sufficient**: as (25) tells us, in that case we can only say that the gradients will be smaller than something that is greater than 0, but this doesn't imply that the gradients will tend to infinity. On the other hand, we showed in (25) that if this condition is not met (i.e the largest eigenvalue of the weights is smaller than $\frac{\delta^2}{\gamma^2}$), we are sure that the gradients won't explode.

6 Question 6

1. The computational graph we derived from the equations is the following:



Computational graph for the RNN

2. Let's name $f^{(f)}$ and $f^{(b)}$ the two functions $f^{(f)}(x_t, h_{t-1}^{(f)}) = W^{(f)}x_t + U^{(f)}h_{t-1}^{(f)}$ and $f^{(b)}(x_t, h_{t+1}^{(b)}) = W^{(b)}x_t + U^{(b)}h_{t+1}^{(b)}$. Then, we have that $h_t^{(f)} = \sigma\big(f^{(f)}(x_t, h_{t-1}^{(f)})\big)$ and $h_t(b) = \sigma\big(f^{(b)}(x_t, h_{t+1}^{(b)})\big)$. Thus, we can write:

$$\nabla_{h_t^{(f)}} L = \frac{\partial L}{\partial h_t^{(f)}} = \frac{\partial h_{t+1}^{(f)}}{\partial h_t^{(f)}} \frac{\partial L}{\partial h_{t+1}^{(f)}} = \frac{\partial f^{(f)}}{\partial h_t^{(f)}} \frac{\partial h_{t+1}^{(f)}}{\partial f^{(f)}} \frac{\partial L}{\partial h_{t+1}^{(f)}} = \frac{\partial f^{(f)}}{\partial h_t^{(f)}} \frac{\partial h_{t+1}^{(f)}}{\partial f^{(f)}} \nabla_{h_{t+1}^{(f)}} L \tag{26}$$

$$\nabla_{h_t^{(b)}} L = \frac{\partial L}{\partial h_t^{(b)}} = \frac{\partial h_{t-1}^{(b)}}{\partial h_t^{(b)}} \frac{\partial L}{\partial h_{t-1}^{(b)}} = \frac{\partial f^{(b)}}{\partial h_t^{(b)}} \frac{\partial h_{t-1}^{(b)}}{\partial f^{(b)}} \frac{\partial L}{\partial h_{t-1}^{(b)}} = \frac{\partial f^{(b)}}{\partial h_t^{(b)}} \frac{\partial h_{t-1}^{(b)}}{\partial f^{(b)}} \nabla_{h_{t-1}^{(b)}} L \tag{27}$$

As the derivations to make explicit the expression (26) are similar to the one for (27), let's concentrate on (26). We have, by definition of $f^{(f)}$:

$$\frac{\partial f^{(f)}}{\partial h_t^{(f)}} = U^{(f)T} \tag{28}$$

and using the derivative of the sigmoid function, we find:

$$\frac{\partial h_{t+1}^{(f)}}{\partial f^{(f)}(x_t, h_t^{(f)})} = \frac{\partial \sigma\big(f^{(f)}(x_t, h_t^{(f)})\big)}{\partial f^{(f)}(x_t, h_t^{(f)})} = diag\Big(\sigma\big(f^{(f)}(x_t, h_t^{(f)})\big) \odot \Big(I - \sigma\big(f^{(f)}(x_t, h_t^{(f)})\big)\Big)\Big) \tag{29}$$

$$= diag\Big(h_{t+1}^{(f)} \odot (I - h_{t+1}^{(f)})\Big) \tag{30}$$

Thus, using (28) and (30), we can rewrite (26) as:

$$\nabla_{h_t^{(f)}} L = U^{(f)T} diag\Big(h_{t+1}^{(f)} \odot (I - h_{t+1}^{(f)})\Big) \nabla_{h_{t+1}^{(f)}} L \tag{31}$$

Similarly, we can rewrite (27) as:

$$\nabla_{h_t^{(b)}} L = U^{(b)T} diag\Big(h_{t-1}^{(b)} \odot (I - h_{t-1}^{(b)})\Big) \nabla_{h_{t-1}^{(b)}} L \tag{32}$$