Section 1. Demonstration 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 is to consider the relationship between different optimization schemes, and to note and quantify 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 $\delta > 0$.

   By expressing the two update rules recursively ($\Delta\boldsymbol{\theta}_t$ as a function of $\Delta\boldsymbol{\theta}_{t-1}$), I show that these two update rules are equivalent ; i.e. express (α, ϵ) as a function of (β, δ).
2. Prepare for the next step by unrolling 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. Assuming $\boldsymbol{g}_t$ has a stationary distribution independent of t. I show that the running average is biased, i.e. $\mathbb{E}[\boldsymbol{v}_t] \neq \mathbb{E}[\boldsymbol{g}_t]$ and I propose a way to eliminate such a bias by rescaling $\boldsymbol{v}_t$.

Steps 1.

1. SGD with momentum
   $$\Delta\boldsymbol{\theta}_t = -\alpha\boldsymbol{v}_{t-1} - \epsilon\boldsymbol{g}_t = \alpha\Delta\boldsymbol{\theta}_{t-1} - \epsilon\boldsymbol{g}_t \tag{1}$$
   SGD with running average
   $$\boldsymbol{v}_t = -\frac{\Delta\boldsymbol{\theta}_t}{\delta} = \beta\boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t = -\beta\frac{\Delta\boldsymbol{\theta}_{t-1}}{\delta} + (1-\beta)\boldsymbol{g}_t$$
   $$\Longrightarrow$$
   $$\Delta\boldsymbol{\theta}_t = \beta\Delta\boldsymbol{\theta}_{t-1} - \delta(1-\beta)\boldsymbol{g}_t \tag{2}$$
   (1) and (2) are equivalent $\iff \alpha\Delta\boldsymbol{\theta}_{t-1} - \epsilon\boldsymbol{g}_t = \beta\Delta\boldsymbol{\theta}_{t-1} - \delta(1-\beta)\boldsymbol{g}_t$
   and it's true with $\alpha = \beta$ and $\epsilon = \delta(1-\beta)$
2.
   $$\begin{aligned}\boldsymbol{v}_t &= \beta\boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t = (1-\beta)\boldsymbol{g}_t + \beta(\beta\boldsymbol{v}_{t-2} + (1-\beta)\boldsymbol{g}_{t-1})\\ &= (1-\beta)\boldsymbol{g}_t + \beta(1-\beta)\boldsymbol{g}_{t-1} + \beta^2(\beta\boldsymbol{v}_{t-3} + (1-\beta)\boldsymbol{g}_{t-2})\\ &= (1-\beta)\left(\boldsymbol{g}_t + \beta\boldsymbol{g}_{t-1} + \beta^2\boldsymbol{g}_{t-2} + \cdots + \beta^{t-2}\boldsymbol{g}_{t-(t-2)}\right) + \beta^{t-1}(\beta\boldsymbol{v}_0 + (1-\beta)\boldsymbol{g}_1)\\ &= (1-\beta)\left(\boldsymbol{g}_t + \beta\boldsymbol{g}_{t-1} + \beta^2\boldsymbol{g}_{t-2} + \cdots + \beta^{t-2}\boldsymbol{g}_2 + \beta^{t-1}\boldsymbol{g}_1\right)\end{aligned}$$

3. Because $\boldsymbol{g}_t$ is defined as unbiased and as having a stationary distribution independent of t, let's say $\mathbb{E}[\boldsymbol{g}_t] = h$.

$$\begin{aligned}
\mathbb{E}[\boldsymbol{v}_t] &= (1-\beta)\left(\mathbb{E}[\boldsymbol{g}_t] + \beta\mathbb{E}[\boldsymbol{g}_{t-1}] + \beta^2\mathbb{E}[\boldsymbol{g}_{t-2}] + \cdots + \beta^{t-2}\mathbb{E}[\boldsymbol{g}_2] + \beta^{t-1}\mathbb{E}[\boldsymbol{g}_1]\right) \\
&= (1-\beta)\left(h + \beta h + \beta^2 h + \cdots + \beta^{t-2}h + \beta^{t-1}h\right) \\
&= (1-\beta)h\left(\sum_{i=0}^{t-1}\beta^i\right) \\
&= h(1-\beta)\frac{1-\beta^t}{1-\beta} \\
&= h(1-\beta^t) \\
&\neq h = \mathbb{E}[\boldsymbol{g}_t]
\end{aligned}$$

To eliminate the bias, it's possible to rescale $\boldsymbol{v}_t$ with $\frac{1}{(1-\beta^t)}$

Section 2. The goal is to understand and compare the effects of different regularizers (specifically dropout and weight decay) on the weights of a network. Let's 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 $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$ and 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 $\boldsymbol{R}_{ij} \sim \text{Bern}(p)$ is sampled i.i.d. from the Bernoulli distribution.

For a squared error loss function with dropout, we then have :

$$L(\boldsymbol{w}) = ||\boldsymbol{y} - (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w}||^2$$

1. Let Γ be a diagonal matrix with $\Gamma_{ii} = (\boldsymbol{X}^\top\boldsymbol{X})_{ii}^{1/2}$. I show that the *expectation (over $\boldsymbol{R}$)* of the loss function can be rewritten as $\mathbb{E}[L(\boldsymbol{w})] = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma\boldsymbol{w}||^2$. *Note : we are trying to find the expectation over a squared term and use* $\text{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2$.
2. I 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. Also, I briefly discuss how the value of p affect the regularization coefficient, λ^{dropout} ?
3. I express the loss function for a linear regression problem without dropout and with L^2 regularization, with regularization coefficient λ^{L_2} and derive its closed form solution $\boldsymbol{w}^{L_2}$.
4. I compare the results of 2.3 and 2.4 (identify specific differences in the equations, and discuss qualitatively what the equations tell you about the similarities and differences in the effects of weight decay and dropout).

Steps 2.

1.

$$\boldsymbol{R}_{ij}\text{i.i.d} \implies Cov(\boldsymbol{R}_{ij}, \boldsymbol{F}_{ik}) = 0 \tag{3}$$

$$\Gamma^2 = diag(\boldsymbol{X}^\top \boldsymbol{X})^{2/2} = \begin{bmatrix} \sum_{i=0}^n \boldsymbol{X}_{i1}^2 & 0 & \dots & 0 \\ 0 & \sum_{i=0}^n \boldsymbol{X}_{i2}^2 & \dots & 0 \\ \vdots & \dots & \ddots & \vdots \\ 0 & 0 & \dots & \sum_{i=0}^n \boldsymbol{X}_{id}^2 \end{bmatrix}$$

$$\implies \Gamma_{jj}^2 = \sum_{i=1}^n \boldsymbol{X}_{ij}^2$$

$$||\Gamma \boldsymbol{w}||^2 = \sum_{j=1}^d \Gamma_{jj}^2 \boldsymbol{w}_j^2 = \sum_{j=1}^d \sum_{i=1}^n \boldsymbol{X}_{ij}^2 \boldsymbol{w}_j^2 \tag{4}$$

With (3),

$$Var(\boldsymbol{y}_i - \sum_{j=1}^d \boldsymbol{w}_j \boldsymbol{X}_{ij} \boldsymbol{R}_{ij}) = \sum_{j=1}^d \boldsymbol{w}_j^2 \boldsymbol{X}_{ij}^2 Var(\boldsymbol{R}_{ij}) + 0 = \sum_{j=1}^d \boldsymbol{w}_j^2 \boldsymbol{X}_{ij}^2 p(1-p) \tag{5}$$

$$\mathbb{E}[\boldsymbol{y}_i - \sum_{j=1}^d \boldsymbol{w}_j \boldsymbol{X}_{ij} \boldsymbol{R}_{ij}] = \boldsymbol{y}_i - \sum_{j=1}^d \boldsymbol{w}_j \boldsymbol{X}_{ij} \mathbb{E}[\boldsymbol{R}_{ij}] = \boldsymbol{y}_i - \sum_{j=1}^d \boldsymbol{w}_j \boldsymbol{X}_{ij} p \tag{6}$$

$$L(\boldsymbol{w}) = ||\boldsymbol{y} - (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w}||^2 = \sum_{i=1}^n (\boldsymbol{y}_i - \sum_{j=1}^d \boldsymbol{w}_j \boldsymbol{X}_{ij} \boldsymbol{R}_{ij})^2 \tag{7}$$

With (7), (6), (5), and (4),

$$\begin{aligned}
\mathbb{E}[L(\boldsymbol{w})] &= \sum_{i=1}^n \mathbb{E}[(\boldsymbol{y}_i - \sum_{j=1}^d \boldsymbol{w}_j \boldsymbol{X}_{ij} \boldsymbol{R}_{ij})^2] = \sum_{i=1}^n \left(Var(\boldsymbol{y}_i - \sum_{j=1}^d \boldsymbol{w}_j \boldsymbol{X}_{ij} \boldsymbol{R}_{ij}) + \mathbb{E}[\boldsymbol{y}_i - \sum_{j=1}^d \boldsymbol{w}_j \boldsymbol{X}_{ij} \boldsymbol{R}_{ij}]^2 \right) \\
&= \sum_{i=1}^n \left(p(1-p) \sum_{j=1}^d \boldsymbol{w}_j^2 \boldsymbol{X}_{ij}^2 + (\boldsymbol{y}_i - p \sum_{j=1}^d \boldsymbol{X}_{ij} \boldsymbol{w}_j)^2 \right) \\
&= \sum_{i=1}^n (\boldsymbol{y}_i - p \sum_{j=1}^d \boldsymbol{X}_{ij} \boldsymbol{w}_j)^2 + p(1-p) \sum_{j=1}^d \sum_{i=1}^n \boldsymbol{X}_{ij}^2 \boldsymbol{w}_j^2 \\
&= ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma \boldsymbol{w}||^2
\end{aligned}$$

2.

$$
\begin{aligned}
\frac{\partial \mathbb{E}[L(\boldsymbol{w})]}{\partial \boldsymbol{w}_k} &= \frac{\partial \sum_{i=1}^{n}(\boldsymbol{y}_i - p\sum_{j=1}^{d}\boldsymbol{X}_{ij}\boldsymbol{w}_j)^2 + p(1-p)\sum_{j=1}^{d}\sum_{i=1}^{n}\boldsymbol{X}_{ij}^2\boldsymbol{w}_j^2}{\partial \boldsymbol{w}_k} \\
&= 2p\sum_{i=1}^{n}\left(-\boldsymbol{X}_{ik}(\boldsymbol{y}_i - p\sum_{j=1}^{d}\boldsymbol{X}_{ij}\boldsymbol{w}_j) + (1-p)\boldsymbol{X}_{ik}^2\boldsymbol{w}_k\right) \\
&= 2p\left(-\sum_{i=1}^{n}\boldsymbol{X}_{ik}\boldsymbol{y}_i + p\sum_{i=1}^{n}\sum_{j=1}^{d}\boldsymbol{X}_{ik}\boldsymbol{X}_{ij}\boldsymbol{w}_j + (1-p)\sum_{i=1}^{n}\boldsymbol{X}_{ik}^2\boldsymbol{w}_k\right)
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\frac{\partial \mathbb{E}[L(\boldsymbol{w})]}{\partial \boldsymbol{w}} &= 2p\left(-\boldsymbol{X}^\top\boldsymbol{y} + p\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + (1-p)\Gamma^2\boldsymbol{w}\right) = 0 \\
&\implies \left(\boldsymbol{X}^\top\boldsymbol{X} + \frac{(1-p)}{p}\Gamma^2\right)\boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{y}\frac{1}{p} \\
&\implies p\boldsymbol{w}^{\text{dropout}} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{\text{dropout}}\Gamma^2)^{-1}\boldsymbol{X}^\top\boldsymbol{y} \\
&\text{where } \lambda^{\text{dropout}} = \frac{(1-p)}{p}
\end{aligned}
$$

When p increases, the regularization coefficient decreases until reaching 0 with $p = 1$ and, when p decreases, the regularization coefficient tends toward infinity.

3.

$$L(\boldsymbol{w}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2 + \lambda^{L_2}||\boldsymbol{w}||^2$$

$$\frac{\partial(\lambda^{L_2}||\boldsymbol{w}||^2)}{\partial \boldsymbol{w}} = 2\lambda^{L_2}\boldsymbol{w}$$

$$\begin{aligned}
\frac{\partial(||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2)}{\partial \boldsymbol{w}_k} &= \frac{\partial \sum_{i=1}^{n}(\boldsymbol{y}_i - \sum_{j=1}^{d}\boldsymbol{X}_{ij}\boldsymbol{w}_j)^2}{\partial \boldsymbol{w}_k} \\
&= 2\sum_{i=1}^{n}\left(-\boldsymbol{X}_{ik}(\boldsymbol{y}_i - p\sum_{j=1}^{d}\boldsymbol{X}_{ij}\boldsymbol{w}_j)\right) \\
&= 2\left(-\sum_{i=1}^{n}\boldsymbol{X}_{ik}\boldsymbol{y}_i + \sum_{i=1}^{n}\sum_{j=1}^{d}\boldsymbol{X}_{ik}\boldsymbol{X}_{ij}\boldsymbol{w}_j\right) \\
\implies \frac{\partial(||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2)}{\partial \boldsymbol{w}} &= 2\left(-\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}\right)
\end{aligned}$$

Therefore

$$\begin{aligned}
\frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}} &= 2\left(-\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + \lambda^{L_2}\boldsymbol{w}\right) = 0 \\
&\implies \left(\boldsymbol{X}^\top\boldsymbol{X} + diag(\lambda^{L_2})\right)\boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{y} \\
&\implies \boldsymbol{w}^{\mathrm{L}_2} = (\boldsymbol{X}^\top\boldsymbol{X} + diag(\lambda^{L_2}))^{-1}\boldsymbol{X}^\top\boldsymbol{y}
\end{aligned}$$

4. Both equations are similar. The main difference is that the regularization coefficient for L^2 is constant for each element of the vector $\boldsymbol{w}$ and the regularization coefficient for dropout is multiplied by different values ($\Gamma_{ii}^2 = (\boldsymbol{X}^\top\boldsymbol{X})_{ii}$) for each element of $\boldsymbol{w}$. Therefore, dropout is kind of like L^2 except that the regularization coefficient of each weight is scaled in function of the feature's values.

Section 3. The goal is 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. Let's consider the following equation for the t-th layer of a deep network :

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

where $\boldsymbol{a}^{(t)}$ are the pre-activations 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 a Gaussian distribution $W_{ij}^{(t)} \sim \mathcal{N}(\mu, \sigma^2)$.

Let's design an initialization scheme that would achieve a vector of **pre-activations** at layer t whose elements are zero-mean and unit variance (i.e. : $\mathbb{E}[a_i^{(t)}] = 0$ and $\mathrm{Var}(a_i^{(t)}) = 1$, $1 \leq i \leq d^{(t)}$)

for the assumptions about either the activations or pre-activations of layer $t-1$ listed below. Note we are not looking for a general formula, just for one setting that meets these criteria (there are many possiblities).

1. Let's assume that the activations of the previous layer satisfy $\mathbb{E}[h_i^{(t-1)}] = 0$ and $\text{Var}(h_i^{(t-1)}) = 1$ for $1 \leq i \leq d^{(t-1)}$. Also, we can assume that entries of $\boldsymbol{h}^{(t-1)}$ are uncorrelated (the result should not depend on g).
   (a) I show $\text{Var}(XY) = \text{Var}(X)\text{Var}(Y) + \text{Var}(X)\mathbb{E}[Y]^2 + \text{Var}(Y)\mathbb{E}[X]^2$ when $X \perp Y$
   (b) Then, I first write $\mathbb{E}[a_i^{(t)}]$ and $\text{Var}(a_i^{(t)})$ in terms of $c, \mu, \sigma^2, \text{Var}(h_i^{(t-1)}), \mathbb{E}[h_i^{(t-1)}]$.
   (c) To be able to then give values for c, μ, and σ^2 as a function of $d^{(t-1)}$ such that $\mathbb{E}[a_i^{(t)}] = 0$ and $\text{Var}(a_i^{(t)}) = 1$ for $1 \leq i \leq d^{(t)}$.
2. Now let's assume that the pre-activations of the previous layer satisfy $\mathbb{E}[a_i^{(t-1)}] = 0$, $\text{Var}(a_i^{(t-1)}) = 1$ and $a_i^{(t-1)}$ has a symmetric distribution for $1 \leq i \leq d^{(t-1)}$. and that the entries of $\boldsymbol{a}^{(t-1)}$ are uncorrelated. If we consider the case of ReLU activation : $g(x) = \max\{0, x\}$.
   (a) Derivation $\mathbb{E}[(h_i^{(t-1)})^2]$
   (b) Using the result from (a), we can give values for c, μ, and σ^2 as a function of $d^{(t-1)}$ such that $\mathbb{E}[a_i^{(t)}] = 0$ and $\text{Var}(a_i^{(t)}) = 1$ for $1 \leq i \leq d^{(t)}$.
3. For both assumptions (1,2), we can also give values α, β for $W_{ij}^{(t)} \sim Uniform(\alpha, \beta)$ such that $\mathbb{E}[a_i^{(t)}] = 0$ and $\text{Var}(a_i^{(t)}) = 1$.

Steps 3.

1.

(a)

$$\begin{aligned}
\text{Var}(XY) &= \mathbb{E}[(XY)^2] - \mathbb{E}[XY]^2 \\
&= \mathbb{E}[X^2]\mathbb{E}[Y^2] - \mathbb{E}[X]^2\mathbb{E}[Y]^2 \\
&= \mathbb{E}[X^2]\mathbb{E}[Y^2] - \mathbb{E}[X]^2\mathbb{E}[Y]^2 \\
&\quad + (\mathbb{E}[X]^2\mathbb{E}[Y]^2 - \mathbb{E}[X]^2\mathbb{E}[Y]^2) + (\mathbb{E}[X^2]\mathbb{E}[Y]^2 - \mathbb{E}[X^2]\mathbb{E}[Y]^2) + (\mathbb{E}[Y^2]\mathbb{E}[X]^2 - \mathbb{E}[Y^2]\mathbb{E}[X]^2) \\
&= \mathbb{E}[X^2]\mathbb{E}[Y^2] - \mathbb{E}[X^2]\mathbb{E}[Y]^2 - \mathbb{E}[Y^2]\mathbb{E}[X]^2 + \mathbb{E}[X]^2\mathbb{E}[Y]^2 \\
&\quad + \mathbb{E}[X^2]\mathbb{E}[Y]^2 - \mathbb{E}[X]^2\mathbb{E}[Y]^2 \\
&\quad + \mathbb{E}[Y^2]\mathbb{E}[X]^2 - \mathbb{E}[X]^2\mathbb{E}[Y]^2 \\
&= \text{Var}(X)\text{Var}(Y) + \text{Var}(X)\mathbb{E}[Y]^2 + \text{Var}(Y)\mathbb{E}[X]^2
\end{aligned}$$

(b)

$$\begin{aligned}
E[a_i^{(t)}] &= E[\sum_{j=1}^{d^{(t-1)}} \boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)} + \boldsymbol{b}_i^{(t)}] \\
&= c + \sum_{j=1}^{d^{(t-1)}} E[\boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)}] \\
&= c + \mu \sum_{j=1}^{d^{(t-1)}} E[\boldsymbol{h}_j^{(t-1)}]
\end{aligned}$$

$$
\begin{aligned}
Var(a_i^{(t)}) &= Var(\sum_{j=1}^{d^{(t-1)}} \boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)} + \boldsymbol{b}_i^{(t)}) \\
&= \sum_{j=1}^{d^{(t-1)}} Var(\boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)}) + 0 + 0 \\
&= \sum_{j=1}^{d^{(t-1)}} \Big[Var(\boldsymbol{W}_{ij}^{(t)}) Var(\boldsymbol{h}_j^{(t-1)}) + Var(\boldsymbol{W}_{ij}^{(t)}) E[\boldsymbol{h}_j^{(t-1)}]^2 + Var(\boldsymbol{h}_j^{(t-1)}) E[\boldsymbol{W}_{ij}^{(t)}]^2 \Big] \\
&= \sum_{j=1}^{d^{(t-1)}} \Big[\sigma^2 Var(\boldsymbol{h}_j^{(t-1)}) + \sigma^2 E[\boldsymbol{h}_j^{(t-1)}]^2 + Var(\boldsymbol{h}_j^{(t-1)}) \mu^2 \Big]
\end{aligned}
$$

(c) With $c = 0$ and $\mu = 0$

$$E[a_i^{(t)}] = 0 + 0 \sum_{j=1}^{d^{(t-1)}} E[\boldsymbol{h}_j^{(t-1)}] = 0$$

With $\sigma^2 = \frac{1}{d^{(t-1)}}$

$$
\begin{aligned}
Var(a_i^{(t)}) &= \sum_{j=1}^{d^{(t-1)}} \Big[\sigma^2 Var(\boldsymbol{h}_j^{(t-1)}) + \sigma^2 E[\boldsymbol{h}_j^{(t-1)}]^2 + Var(\boldsymbol{h}_j^{(t-1)}) \mu^2 \Big] \\
&= \sum_{j=1}^{d^{(t-1)}} \Big[\sigma^2 \times 1 + \sigma^2 \times 0 + 1 \times 0 \Big] \\
&= d^{(t-1)} \sigma^2 = 1
\end{aligned}
$$

2.

(a) $\mathbb{E}[(h_i^{(t-1)})^2] = \mathbb{E}[g(a_i^{(t-1)})^2]$

if $a_i^{(t-1)} < 0$, $\mathbb{E}[0] = 0$

if $a_i^{(t-1)} > 0$, $\mathbb{E}[(a_i^{(t-1)})^2] = Var(a_i^{(t-1)}) + \mathbb{E}[a_i^{(t-1)}]^2 = 1$

Therefore, because $a_i^{(t-1)}$ has a symmetric distribution, $\mathbb{E}[(h_i^{(t-1)})^2] = \frac{1}{2}$

(b) With $c = 0$ and $\mu = 0$

$$
\begin{aligned}
E[a_i^{(t)}] &= c + \mu \sum_{j=1}^{d^{(t-1)}} E[\boldsymbol{h}_j^{(t-1)}] \\
&= 0 + 0 \sum_{j=1}^{d^{(t-1)}} E[\boldsymbol{h}_j^{(t-1)}] = 0
\end{aligned}
$$

According to (a),

$$Var(\boldsymbol{h}_j^{(t-1)}) = E[g(\boldsymbol{a}_j^{(t-1)})^2] - E[g(\boldsymbol{a}_j^{(t-1)})]^2 = \frac{1}{2}$$

Therefore, with $\sigma^2 = \frac{2}{d^{(t-1)}}$

$$\begin{aligned} Var(a_i^{(t)}) &= \sum_{j=1}^{d^{(t-1)}} \left[\sigma^2 Var(\boldsymbol{h}_j^{(t-1)}) + \sigma^2 E[\boldsymbol{h}_j^{(t-1)}]^2 + Var(\boldsymbol{h}_j^{(t-1)})\mu^2\right] \\ &= \sum_{j=1}^{d^{(t-1)}} \left[\sigma^2 Var(\boldsymbol{h}_j^{(t-1)}) + \sigma^2 E[g(\boldsymbol{a}_j^{(t-1)})]^2 + 0\right] \\ &= \sum_{j=1}^{d^{(t-1)}} \left[\sigma^2 \times \frac{1}{2} + \sigma^2 \times 0\right] \\ &= \frac{d^{(t-1)}}{2}\sigma^2 = 1 \end{aligned}$$

He initialization is a popular initialization scheme with this form

3.
With $\alpha = -\beta$ and $c = 0$, for both assumptions

$$E[a_i^{(t)}] = c + \frac{\alpha+\beta}{2}\sum_{j=1}^{d^{(t-1)}} E[\boldsymbol{h}_j^{(t-1)}] = 0$$

For assumption #1, with $\beta = \sqrt{\frac{3}{d^{(t-1)}}}$,

$$Var(a_i^{(t)}) = \sum_{j=1}^{d^{(t-1)}} \left[\frac{(\beta-\alpha)^2}{12} Var(\boldsymbol{h}_j^{(t-1)}) + 0 + 0\right] = \sum_{j=1}^{d^{(t-1)}} \frac{\beta^2}{3} = d^{(t-1)}\frac{\beta^2}{3} = 1$$

For assumption #2, with $\beta = \sqrt{\frac{6}{d^{(t-1)}}}$,

$$Var(a_i^{(t)}) = \sum_{j=1}^{d^{(t-1)}} \left[\frac{(\beta-\alpha)^2}{12} Var(\boldsymbol{h}_j^{(t-1)}) + 0 + 0\right] = \sum_{j=1}^{d^{(t-1)}} \frac{\beta^2}{3}\frac{1}{2} = d^{(t-1)}\frac{\beta^2}{6} = 1$$

Section 4. This section is about normalization techniques.

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 $\boldsymbol{w}$.

Let's 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 $y = \boldsymbol{u}^\top \boldsymbol{x}$. Assuming the data $\boldsymbol{x}$ (a random vector) is whitened ($\text{Var}(\boldsymbol{x}) = \boldsymbol{I}$) and centered at 0 ($\mathbb{E}[\boldsymbol{x}] = \boldsymbol{0}$), we can show that $\hat{y} = \boldsymbol{w}^\top \boldsymbol{x} + \beta$.

2. Let's 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[1] $\boldsymbol{W}^\perp \boldsymbol{u} = \boldsymbol{0}$.

Steps 4.

1.

$$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{0} \implies \mathbb{E}[\mu_y] = \mathbb{E}[\boldsymbol{u}^\top \boldsymbol{x}] = \boldsymbol{u}^\top \times 0 = 0 \tag{8}$$

$$\text{Var}(\boldsymbol{x}) = \boldsymbol{I} \implies \sigma_y^2 = \text{Var}(\boldsymbol{u}^\top \boldsymbol{x}) = \boldsymbol{u}^\top \text{Var}(\boldsymbol{x}) \boldsymbol{u} = \boldsymbol{u}^\top \boldsymbol{u} = ||\boldsymbol{u}||^2 \tag{9}$$

$$\begin{aligned} \hat{y} = \gamma \cdot \frac{\boldsymbol{u}^\top \boldsymbol{x} - \mu_y}{\sigma_y} + \beta = \gamma \frac{\boldsymbol{u}^\top \boldsymbol{x}}{||\boldsymbol{u}||} + \beta = \boldsymbol{w}^\top \boldsymbol{x} + \beta & \\ \iff \frac{\boldsymbol{u}^\top \boldsymbol{x} - \mu_y}{\sigma_y} = \frac{\boldsymbol{u}^\top \boldsymbol{x}}{||\boldsymbol{u}||} & \\ \iff \frac{\boldsymbol{u}^\top \boldsymbol{x}}{\sigma_y} = \frac{\boldsymbol{u}^\top \boldsymbol{x}}{||\boldsymbol{u}||} & \qquad \text{Because of (8)} \\ \iff \frac{\boldsymbol{u}^\top \boldsymbol{x}}{||\boldsymbol{u}||} = \frac{\boldsymbol{u}^\top \boldsymbol{x}}{||\boldsymbol{u}||} & \qquad \text{Because of (9)} \end{aligned}$$

2.

$$\frac{\partial(||\boldsymbol{u}||)}{\partial \boldsymbol{u}} = \frac{\boldsymbol{u}}{||\boldsymbol{u}||}$$

$$\frac{\partial \boldsymbol{w}}{\partial \boldsymbol{u}} = \frac{\partial}{\partial \boldsymbol{u}} \left(\gamma \frac{\boldsymbol{u}}{||\boldsymbol{u}||}\right) = \gamma \left(\frac{||\boldsymbol{u}|| - \boldsymbol{u}\frac{\partial(||\boldsymbol{u}||)}{\partial \boldsymbol{u}}}{||\boldsymbol{u}||^2}\right) = \gamma \left(\frac{1}{||\boldsymbol{u}||} - \frac{\boldsymbol{u}\boldsymbol{u}^T / ||\boldsymbol{u}||}{||\boldsymbol{u}||^2}\right) = \frac{\gamma}{||\boldsymbol{u}||} \left(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^T}{||\boldsymbol{u}||^2}\right) = \frac{\gamma}{||\boldsymbol{u}||} \boldsymbol{W}^\perp$$

$$\nabla_{\boldsymbol{u}} L = \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{u}} \nabla_{\boldsymbol{w}} L = \frac{\gamma}{||\boldsymbol{u}||} \boldsymbol{W}^\perp \nabla_{\boldsymbol{w}} L = s \boldsymbol{W}^\perp \nabla_{\boldsymbol{w}} L$$

Section 5. This section 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. Let's consider the following recurrent unit :

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

[1] 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.

1. I partially 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 $\boldsymbol{h}_t$). Here, I'm assuming that it holds for time step $t-1$ and only prove the induction step.
2. Let $||\boldsymbol{A}||$ denote the L_2 operator norm of matrix $\boldsymbol{A}$ ($||\boldsymbol{A}|| := \max_{\boldsymbol{x}:||\boldsymbol{x}||=1} ||\boldsymbol{A}\boldsymbol{x}||$) and assume $\sigma(x)$ has bounded derivative, i.e. $|\sigma'| \leq \gamma$ for some $\gamma > 0$ and for all x. We denote as $\lambda_1(\cdot)$ the largest eigenvalue of a symmetric matrix. We can show that if the largest eigenvalue of the weights is bounded by $\frac{\delta^2}{\gamma^2}$ for some $0 \leq \delta < 1$, gradients of the hidden state will vanish over time, i.e.

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

The following properties of the L_2 operator norm are going to be used :

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

3. I briefly discuss about 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 $\frac{\delta^2}{\gamma^2}$

Steps 5.

1. We assume $\boldsymbol{g}_{t-1} = f(\boldsymbol{h}_{t-1})$

Assuming that σ^{-1} exists (which is often not the case with activation functions),

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

Therefore, going the other way (where we don't need the assumption that σ^{-1} exists) gives

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

2.

$$\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0} = \prod_{i=0}^{T-1} \frac{\partial \boldsymbol{h}_{i+1}}{\partial \boldsymbol{h}_i} = \prod_{i=0}^{T-1} \boldsymbol{W} \frac{\partial \sigma(\boldsymbol{h}_i)}{\partial \boldsymbol{h}_i} \tag{10}$$

$$||\frac{\partial \boldsymbol{h}_{i+1}}{\partial \boldsymbol{h}_i}|| = ||\boldsymbol{W}\frac{\partial \sigma(\boldsymbol{h}_i)}{\partial \boldsymbol{h}_i}|| \leq ||\boldsymbol{W}|| \cdot ||\frac{\partial \sigma(\boldsymbol{h}_i)}{\partial \boldsymbol{h}_i}|| \leq ||\boldsymbol{W}||\gamma \tag{11}$$

With those equations and the L_2 operator norm properties given

$$||\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0}|| = ||\prod_{i=0}^{T-1} \frac{\partial \boldsymbol{h}_{i+1}}{\partial \boldsymbol{h}_i}|| \leq \prod_{i=0}^{T-1} ||\frac{\partial \boldsymbol{h}_{i+1}}{\partial \boldsymbol{h}_i}|| \leq \prod_{i=0}^{T-1} ||\boldsymbol{W}|| \cdot \gamma = \prod_{i=0}^{T-1} \sqrt{\lambda_1(\boldsymbol{W}^\top \boldsymbol{W})} \cdot \gamma = \sqrt{\lambda_1(\boldsymbol{W}^\top \boldsymbol{W})}^T \cdot \gamma^T$$

Therefore,

$$\lambda_1(\boldsymbol{W}^\top\boldsymbol{W}) \leq \frac{\delta^2}{\gamma^2} \implies \lim_{T\to\infty} ||\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0}|| \leq \lim_{T\to\infty} \sqrt{\lambda_1(\boldsymbol{W}^\top\boldsymbol{W})}^T \cdot \gamma^T \leq \lim_{T\to\infty} \frac{\delta^T}{\gamma^T} \cdot \gamma^T = \lim_{T\to\infty} \delta^T = 0$$

Because $0 \leq \delta < 1$

3. For the gradient to explode, the largest eigenvalue of the weights would need to be larger than $\frac{\delta^2}{\gamma^2}$. The condition is necessary, but it's not sufficient for the gradient to explode.

Section 6. Let's consider the following Bidirectional RNN :

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

where the superscripts f and b correspond to the forward and backward RNNs respectively and σ denotes the logistic sigmoid function. 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 $L_t = ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2$.

The goal is to obtain an expression for the gradients $\nabla_{\boldsymbol{W}^{(f)}} L$ and $\nabla_{\boldsymbol{U}^{(b)}} L$.

1. First, let's make a computational graph for this RNN, unrolled for 3 time steps (from $t = 1$ to $t = 3$), labelling each node with the corresponding hidden unit and each edge with the corresponding weight. Note that it includes the initial hidden states for both the forward and backward RNNs.
2. Using total derivatives we can 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 $\nabla_{\boldsymbol{h}_{t-1}^{(b)}} L$ as follows :

$$\nabla_{\boldsymbol{h}_t^{(f)}} L = \nabla_{\boldsymbol{h}_t^{(f)}} L_t + \left(\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}}\right)^\top \nabla_{\boldsymbol{h}_{t+1}^{(f)}} L$$

$$\nabla_{\boldsymbol{h}_t^{(b)}} L = \nabla_{\boldsymbol{h}_t^{(b)}} L_t + \left(\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}}\right)^\top \nabla_{\boldsymbol{h}_{t-1}^{(b)}} L$$

Let's derive an expression for $\nabla_{\boldsymbol{h}_t^{(f)}} L_t$, $\nabla_{\boldsymbol{h}_t^{(b)}} L_t$, $\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}}$ and $\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}}$.

3. With those expressions, we can derive $\nabla_{\boldsymbol{W}^{(f)}} L$ and $\nabla_{\boldsymbol{U}^{(b)}} L$ as functions of $\nabla_{\boldsymbol{h}_t^{(f)}} L$ and $\nabla_{\boldsymbol{h}_t^{(b)}} L$, respectively.

Steps 6.

1.

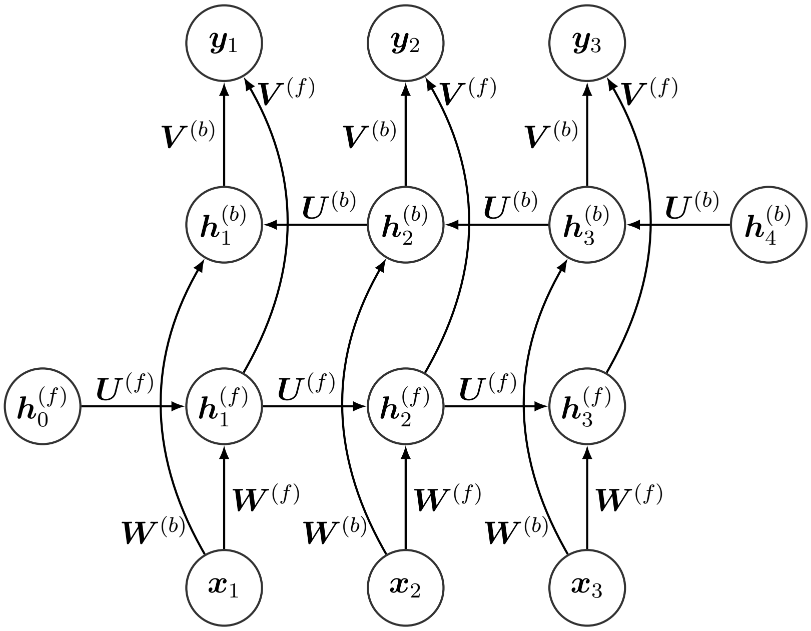

FIGURE 1 – Computational graph for the Bidirectional RNN, unrolled for 3 time steps

2.

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

Similarly,

$$\nabla_{\boldsymbol{h}_t^{(b)}} L_t = -2\boldsymbol{V}^{(b)^\top}(\boldsymbol{z}_t - \boldsymbol{y}_t)$$

$$\begin{aligned}
\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}} &= \frac{\partial(\sigma(\boldsymbol{W}^{(f)}\boldsymbol{x}_{t+1} + \boldsymbol{U}^{(f)}\boldsymbol{h}_t^{(f)}))}{\partial \boldsymbol{h}_t^{(f)}} \\
&= diag\left(\sigma(\boldsymbol{W}^{(f)}\boldsymbol{x}_{t+1} + \boldsymbol{U}^{(f)}\boldsymbol{h}_t^{(f)})(1 - \sigma(\boldsymbol{W}^{(f)}\boldsymbol{x}_{t+1} + \boldsymbol{U}^{(f)}\boldsymbol{h}_t^{(f)}))\right)\boldsymbol{U}^{(f)} \\
&= diag\left(\boldsymbol{h}_{t+1}^{(f)}(1 - \boldsymbol{h}_{t+1}^{(f)})\right)\boldsymbol{U}^{(f)}
\end{aligned}$$

Similarly,

$$\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}} = diag\left(\boldsymbol{h}_{t-1}^{(b)}(1 - \boldsymbol{h}_{t-1}^{(b)})\right)\boldsymbol{U}^{(b)}$$

3.

$$\begin{aligned}
\nabla_{\boldsymbol{W}_t^{(f)}} L &= \left(\frac{\partial \boldsymbol{h}_t^{(f)}}{\partial \boldsymbol{W}_t^{(f)}}\right)^\top \nabla_{\boldsymbol{h}_t^{(f)}} L = \Big(diag(\boldsymbol{h}_t^{(f)}(1-\boldsymbol{h}_t^{(f)}))\boldsymbol{x}_t\Big)^\top \nabla_{\boldsymbol{h}_t^{(f)}} L \\
\implies \nabla_{\boldsymbol{W}^{(f)}} L &= \sum_t diag\left(\boldsymbol{h}_t^{(f)}(1-\boldsymbol{h}_t^{(f)})\right)(\nabla_{\boldsymbol{h}_t^{(f)}} L)\boldsymbol{x}_t^\top
\end{aligned}$$

Similarly,

$$\begin{aligned}
\nabla_{\boldsymbol{U}_t^{(b)}} L &= \left(\frac{\partial \boldsymbol{h}_t^{(b)}}{\partial \boldsymbol{U}_t^{(b)}}\right)^\top \nabla_{\boldsymbol{h}_t^{(b)}} L = \Big(diag(\boldsymbol{h}_t^{(b)}(1-\boldsymbol{h}_t^{(b)}))\boldsymbol{h}_{t+1}^{(b)}\Big)^\top \nabla_{\boldsymbol{h}_t^{(b)}} L \\
\implies \nabla_{\boldsymbol{U}^{(b)}} L &= \sum_t diag\left(\boldsymbol{h}_t^{(b)}(1-\boldsymbol{h}_t^{(b)})\right)(\nabla_{\boldsymbol{h}_t^{(b)}} L)\boldsymbol{h}_{t+1}^{(b)\top}
\end{aligned}$$