Due Date: March 17th 23:00, 2020

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*
- *All norms denote Euclidean norms unless otherwise specified.*
- *Submit your answers electronically via Gradescope.*
- *TAs for this assignment are **Jessica Thompson, Jonathan Cornford and Lluis Castrejon**.*

Question 1 (4-4-4)**.** 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 1. 1. Express the two update rules recursively:

(a) $\Delta\boldsymbol{\theta}_t = -\boldsymbol{v}_t \xrightarrow{\boldsymbol{v}_t = \alpha\boldsymbol{v}_{t-1} + \epsilon\boldsymbol{g}_t} \alpha(-\boldsymbol{v}_{t-1}) + \epsilon\boldsymbol{g}_t = \alpha\Delta\boldsymbol{\theta}_{t-1} + \epsilon\boldsymbol{g}_t$

(b) $\Delta\boldsymbol{\theta}_t = -\delta\boldsymbol{v}_t \xrightarrow{\boldsymbol{v}_t = \beta\boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t} -\Delta\beta\boldsymbol{v}_{t-1} - \Delta(1-\beta)\boldsymbol{g}_t = \beta(-\Delta\boldsymbol{v}_{t-1}) - \Delta(1-\beta)\boldsymbol{g}_t$
$= \beta\Delta\boldsymbol{\theta}_t - \Delta(1-\beta)\boldsymbol{g}_t$
$\longrightarrow$ Both update rules are equivalent, with $(\alpha, \epsilon) = (\beta, \Delta(1-\beta))$ and $(\beta, \Delta) = (\alpha, \frac{\epsilon}{1-\alpha})$

2. SGD with running average:

$$\begin{aligned}
\boldsymbol{v}_t &= \beta \boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t \\
&= \beta(\beta \boldsymbol{v}_{t-2} + (1-\beta)\boldsymbol{g}_{t-1}) + (1-\beta)\boldsymbol{g}_t = \beta^2 \boldsymbol{v}_{t-2} + \beta(1-\beta)\boldsymbol{g}_{t-1} + (1-\beta)\boldsymbol{g}_t \\
&= \beta^2(\beta \boldsymbol{v}_{t-3} + (1-\beta)\boldsymbol{g}_{t-2}) + \beta(1-\beta)\boldsymbol{g}_{t-1} + (1-\beta)\boldsymbol{g}_t \\
&= \beta^3 \boldsymbol{v}_{t-3} + \beta^2(1-\beta)\boldsymbol{g}_{t-2} + \beta(1-\beta)\boldsymbol{g}_{t-1} + (1-\beta)\boldsymbol{g}_t \\
&. \\
&. \\
&. \\
&= \beta \boldsymbol{v}_0 + (1-\beta)\Sigma_{i=1}^{t} \beta^{t-i} \boldsymbol{g}_i \xrightarrow{\boldsymbol{v}_0 = 0} \\
&= (1-\beta)\Sigma_{i=1}^{t} \beta^{t-i} \boldsymbol{g}_i
\end{aligned} \tag{1}$$

3. We know that $\boldsymbol{g}_t$ has a stationary distribution independent of t and we need to show that the running average is biased. Lets start with calculating $\mathbb{E}[\boldsymbol{v}_t]$.

$$\begin{aligned}
\mathbb{E}[\boldsymbol{v}_t] &= \mathbb{E}[(1-\beta)\Sigma_{i=1}^{t} \beta^{t-i} \boldsymbol{g}_i] \\
&= (1-\beta)\Sigma_{i=1}^{t} \beta^{t-i} \mathbb{E}[\boldsymbol{g}_i] \xrightarrow{g-is-stationary} \\
&= (1-\beta)\Sigma_{i=1}^{t} \beta^{t-i} \mathbb{E}[\boldsymbol{g}_t] \\
&= (1-\beta)\mathbb{E}[\boldsymbol{g}_t]\Sigma_{i=1}^{t} \beta^{t-i} \\
&= (1-\beta)\mathbb{E}[\boldsymbol{g}_t]\Sigma_{i=0}^{t} \beta^{t-1} \xrightarrow{Telescoping-Sum} \\
&= (1-\beta)\mathbb{E}[\boldsymbol{g}_t]\frac{1-\beta^t}{1-\beta} \xrightarrow{(1-\beta)cancelled-out} \\
&= \mathbb{E}[\boldsymbol{g}_t](1-\beta^t)
\end{aligned} \tag{2}$$

We showed that the running average is biased. If we rescale $\boldsymbol{v}_t$ by $\frac{1}{1-\beta^t}$ then the bias will be removed and we will have an unbiased estimation of running average.

Question 2 (7-5-5-3)**.** 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 $\boldsymbol{R}_{ij} \sim \text{Bern}(p)$ 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 $\Gamma_{ii} = (\boldsymbol{X}^\top \boldsymbol{X})_{ii}^{1/2}$. Show that the *expectation (over $\boldsymbol{R}$)* of the loss function can be rewritten as $\mathbb{E}[L(\boldsymbol{w})] = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma\boldsymbol{w}||^2$. *Hint: 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. Show that the solution $\boldsymbol{w}^{\text{dropout}}$ that minimizes the expected loss from question 2.1 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} ?
3. Express the loss function for a linear regression problem without dropout and with L^2 regularization, with regularization coefficient λ^{L_2}. Derive its closed form solution $\boldsymbol{w}^{L_2}$.
4. Compare the results of 2.2 and 2.3: 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. Since we are trying to find the expectation over a squared term we can re-write this formula $\text{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2$ in the following form $\mathbb{E}[Z^2] = \mathbb{E}[Z]^2 + \text{Var}(Z)$. $L(\boldsymbol{w})$ have a squared form so let $Z = L(\boldsymbol{w}) = ||\boldsymbol{y} - (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w}||^2$.
$\boldsymbol{R}_{ij}$ is a Bernoulli variable so by definition we know $\mathbb{E}[r_{ij}] = p$, and $Var(r_{ij}) = (1-p)(0-p)^2 + p(1-p)^2 = p^2(1-p) + p(1-p-p+p^2) = p - p^2 = p(1-p)$.

$$\begin{aligned}
\mathbb{E}[L(\boldsymbol{w})] &= \mathbb{E}[||\boldsymbol{y}_i - (\boldsymbol{x}_i \odot \boldsymbol{r}_i)\boldsymbol{w}||^2] = \Sigma_{i=1}^n \mathbb{E}[(\boldsymbol{y}_{i-(\boldsymbol{x}_i \odot \boldsymbol{r}_i)\boldsymbol{w})^2]} \\
&= \Sigma_{i=1}^n \mathbb{E}[(\boldsymbol{y}_i - (\boldsymbol{x}_i \odot \boldsymbol{r}_i)\boldsymbol{w})]^2 + Var(\boldsymbol{y}_i - (\boldsymbol{x}_i \odot \boldsymbol{r}_i)\boldsymbol{w}) \\
&= \Sigma_{i=1}^n (\boldsymbol{y}_i - (\boldsymbol{x}_i \odot \mathbb{E}[\boldsymbol{r}_i])\boldsymbol{w})^2 + Var(\boldsymbol{y}_i - (\boldsymbol{x}_i \odot \boldsymbol{r}_i)\boldsymbol{w}) \xrightarrow{Bernoulli-Variable} \\
&= \Sigma_{i=1}^n (\boldsymbol{y}_i - p\boldsymbol{x}_i\boldsymbol{w})^2 + \Sigma_{i=1}^n \boldsymbol{w}^T p(1-p)(\boldsymbol{X}_i^T \boldsymbol{X}_i)\boldsymbol{w} \\
&= ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)diag(\boldsymbol{X}^T\boldsymbol{X})w \\
&= ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma\boldsymbol{w}||^2
\end{aligned} \tag{3}$$

2. How does p affects the regularization coefficient, $\lambda^{dropout}$? In order to find the $\boldsymbol{w}^{dropout}$ which minimize expected loss $\mathbb{E}[L(\boldsymbol{w})] = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma\boldsymbol{w}||^2$ we need to set the gradient of $\mathbb{E}[L(\boldsymbol{w})]$ to zero:

$$\begin{aligned}
\frac{d\mathbb{E}[L(\boldsymbol{w})}{d\boldsymbol{w}} &= -2(-p\boldsymbol{X}^T)(\boldsymbol{y} - pX\boldsymbol{x}^{dropout}) + 2p(1-p)\Gamma^T(\Gamma\boldsymbol{w}^{dropout}) \\
&= -2p\boldsymbol{X}^T\boldsymbol{y} + 2p^2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}^{dropout} + 2p(1-p)(\Gamma^T\Gamma)\boldsymbol{w}^{dropout} = 0 \xrightarrow{Note-that:} \\
&(\Gamma^T\Gamma = (diag(\boldsymbol{X}^T\boldsymbol{X}))^T(diag(\boldsymbol{X}^T\boldsymbol{X})^{0.5}) = diag(\boldsymbol{X}^T\boldsymbol{X})^{0.5}diag(\boldsymbol{X}^T\boldsymbol{X})^{0.5} = \Gamma^2) \\
&= -p\boldsymbol{X}^T\boldsymbol{y} + p^2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}^{dropout} + p(1-p)(\Gamma^T\Gamma)\boldsymbol{w}^{dropout} \\
&= p^2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}^{dropout} + p(1-p)\Gamma^2\boldsymbol{w}dropout = p\boldsymbol{X}^T\boldsymbol{y} \xrightarrow{divide-by-p} \\
&(\boldsymbol{X}^T\boldsymbol{X} + (\frac{1}{p} - 1)\Gamma^2)p\boldsymbol{w}^{dropout} = \boldsymbol{X}^T\boldsymbol{y} \longrightarrow \\
&p\boldsymbol{w}^{dropout} = (\boldsymbol{X}^T\boldsymbol{X} + (\frac{1}{p} - 1)\Gamma^2)^{-1}\boldsymbol{X}^T\boldsymbol{y}
\end{aligned} \tag{4}$$

Until now we showed that the solution $\boldsymbol{w}^{dropout}$ that minimizes expected loss satisfies the form mentioned in the problem. $\lambda^{dropout}$ is equal to $(\frac{1}{p} - 1)$.

How does the value of p affect the regularization coefficient, $\lambda^{dropout}$? If we set p to 1 then the value of $\lambda^{dropout}$ will be zero which means $L(\boldsymbol{w})$ have no regularization(it turned regularization off). If p is small and goes to zero then the value of $(\frac{1}{p} - 1)$ goes to infinity and the contribution from dropout increases.

3. The loss function is $L(\boldsymbol{w}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2 + \lambda^{L_2}||\boldsymbol{w}||^2$. In order to derive its closed form solution we need to take the derivative with respect to w and set it to zero:

$$\begin{aligned}
&\frac{dL(\boldsymbol{w})}{d\boldsymbol{w}} = (-2X^T)(\boldsymbol{y} - X\boldsymbol{w}^{L_2}) + 2\lambda^{L_2}\boldsymbol{w}^{L_2} = -X^T\boldsymbol{y} + X^TX\boldsymbol{w}^{L_2} + \lambda^{L_2}\boldsymbol{w}^{L_2} = 0 \xrightarrow{factore-\boldsymbol{w}^{L_2}} \\
&\boldsymbol{w}^{L_2}(X^TX + \lambda^{L_2}) = X^T\boldsymbol{y} \longrightarrow \boldsymbol{w}^{L_2} = (X^TX + \lambda^{L_2})^{-1}X^T\boldsymbol{y}
\end{aligned} \tag{5}$$

4. In part 2.2, the regularization term $\lambda^{dropout}$ is scaled by Γ^2 which is the standard deviation of each feature i. If p is equal to 1, the dropout is off and we will only have least square and when p is small then the regularization term increases. In part 2.3. λ^{L_2} is scaled by I and are penalized uniformly. One's might think of dropout as scaled L^2.

Question 3 (6-10-2)**.** 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:

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

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 $W_{ij}^{(t)} \sim \mathcal{N}(\mu, \sigma^2)$.

Your task is to 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 $\text{Var}(a_i^{(t)}) = 1$, $1 \le i \le d^{(t)}$) for the assumptions about either the activations or pre-activations of layer $t-1$ listed below. Note we are not asking for a general formula ; you just need to provide one setting that meets these criteria (there are many possiblities).

1. First 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 $1 \le i \le d^{(t-1)}$. Also, assume entries of $\boldsymbol{h}^{(t-1)}$ are uncorrelated (the answer should not depend on g).
   (a) 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) 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) 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 $1 \le i \le d^{(t)}$.
2. Now 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 $1 \le i \le d^{(t-1)}$. Assume entries of $\boldsymbol{a}^{(t-1)}$ are uncorrelated. Consider the case of ReLU activation: $g(x) = \max\{0, x\}$.
   (a) Derive $\mathbb{E}[(h_i^{(t-1)})^2]$
   (b) Using the result from (a), 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 $1 \le i \le d^{(t)}$.
   (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.
3. For both assumptions (1,2) give values α, β for $W_{ij}^{(t)} \sim Uniform(\alpha, \beta)$ such that $\mathbb{E}[a_i^{(t)}] = 0$ and $\text{Var}(a_i^{(t)}) = 1$.

Answer 3. 1. (a) X and Y are independent then $X|Y = X$, $Y|X = Y$.

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

(b) Write $\mathbb{E}[a_i^{(t)}]$ and $\text{Var}(a_i^{(t)})$ in terms of $c, \mu, \sigma^2, \text{Var}(h_i^{(t-1)}), \mathbb{E}[h_i^{(t-1)}]$.

$$\begin{aligned} E[a_i^{(t)}] &= E[W_i^{(t)} h^{(t-1)} + b_i^{(t)}] \xrightarrow{E[X+Y]=E[x]+E[Y]} \\ &= E[W_i(t)h^{(t-1)}] + E[b_i(t)] \xrightarrow{if(X\perp Y)then(E[XY]=E[X]E[Y]}) \\ &= E[W_i^{(t)}]E[h^{(t-1)}] + E[b_i^{(t)}] \xrightarrow{E[h^{(t-1)}]=0,(1\leq i\leq d^{(t-1)})} \\ &= E[W_i^{(t)}]\cancelto{0}{E[h^{(t-1)}]} + E[b_i^{(t)}] \xrightarrow{(b_i^{(t)}=c)so:(E[b_i^{(t)}]=c)} \\ &= E[b_i^{(t)}] \xrightarrow{E[a_i^{(t)}]=0} \\ &= c = 0 \end{aligned} \tag{7}$$

$$\begin{aligned} Var[a_i^{(t)}] &= Var(W_i^{(t)} h^{(t-1)} + b_i^{(t)}) \xrightarrow{Var(X+Y)=Var(x)+Var(Y)} \\ &= Var(W_i(t)h^{(t-1)}) + Var(b_i(t)) \xrightarrow{part-a} \\ &= Var(W_i^{(t)})Var(h^{(t-1)}) + Var(W_i^{(t)})E[h^{(t-1)}]^2 + E[W_i^{(t)}]^2 Var(h^{(t-1)}) + Var(b_i^{(t)}) \xrightarrow{problemdef.} \\ &= Var(W_i^{(t)})\cancelto{1_{d^{(t-1)}}}{Var(h^{(t-1)})} + Var(W_i^{(t)})\cancelto{0}{E[h^{(t-1)}]^2} + E[W_i^{(t)}]^2\cancelto{1_{d^{(t-1)}}}{Var(h^{(t-1)})} + \cancelto{0}{Var(b_i^{(t)})} \\ &= Var(W_i^{(t)})1_{d^{(t-1)}} + E[W_i^{(t)}]^2 1_{d^{(t-1)}} \\ &= 1 \end{aligned} \tag{8}$$

We need to write the above mentioned formula in terms of $c, \mu, \sigma^2, \text{Var}(h_i^{(t-1)}), \mathbb{E}[h_i^{(t-1)}]$. We already find out that $E[a_i^{(t)}] = c = 0$ and we also have:
$Var[a_i^{(t)}] = Var(W_i^{(t)})1_{d^{(t-1)}} + E[W_i^{(t)}]^2 1_{d^{(t-1)}}$.
From problem definition we can derive $E[W_i^{(t)}] = \mu 1^T_{d^{(t-1)}}$, $Var(W_i^{(t)}) = \sigma^2 1^T_{d^{(t-1)}}$, and from $E[a_i^{(t)}] = 0$ we have $E[W_i^{(t)}] = 0$ so $\mu = 0$
Now, we need to rewrite $Var[a_i^{(t)}]$ formula.

$$\begin{aligned} Var[a_i^{(t)}] &= Var(W_i^{(t)})1_{d^{(t-1)}} + E[W_i^{(t)}]^2 1_{d^{(t-1)}} \\ &= \sigma^2 1^T_{d^{(t-1)}} 1_{d^{(t-1)}} + \mu^2 1^T_{d^{(t-1)}} 1_{d^{(t-1)}} \\ &= (\sigma^2 + \mu^2) d^{(t-1)} \xrightarrow{problemdef.} \\ &= \sigma^2 d^{(t-1)} = 1 \end{aligned} \tag{9}$$

Which means $\sigma^2 = \frac{1}{d^{(t-1)}}$.

(c) I take these properties into account in part b after deriving the formula.
Here is all I find out:

$$c = \mu = 0$$

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

2. (a) Derive $\mathbb{E}[(h_i^{(t-1)})^2]$:

$$\begin{aligned}
\mathbb{E}[(h_i^{(t-1)})^2] &= \mathbb{E}[g(a_i^{(t-1)})^2] \\
&= \int_{-\infty}^{+\infty} (max(0, a_i^{t-1}))^2 f(a_i^{t-1}) da_i^{t-1} = \int_{0}^{+\infty} (a_i^{t-1})^2 f(a_i^{t-1}) da_i^{t-1} \xrightarrow{symmetric} \\
&= \frac{1}{2} \int_{-\infty}^{+\infty} (a_i^{t-1})^2 f(a_i^{t-1}) da_i^{t-1} = \frac{1}{2} \mathbb{E}[(a_i^{(t-1)})^2] \\
&= \frac{1}{2} (\mathrm{Var}(a_i^{(t-1)}) + (E[a_i^{(t-1)}])^2) \xrightarrow{\mathbb{E}[a_i^{(t-1)}]=0} \\
&= \frac{1}{2} (\mathrm{Var}(a_i^{(t-1)}) + \cancelto{0}{(E[a_i^{(t-1)}])^2}) = \frac{1}{2} \mathrm{Var}(a_i^{(t-1)}) \xrightarrow{\mathrm{Var}(a_i^{(t-1)})=1} \\
&= \frac{1}{2} \cancelto{1}{\mathrm{Var}(a_i^{(t-1)})} = \frac{1}{2}
\end{aligned} \tag{10}$$

(b) Now we need to find out the values for c, μ, and σ^2 as a function of $d^{(t-1)}$.

$$\begin{aligned}
\mathbb{E}[a_i^{(t)}] &= \mathbb{E}[W_i^{(t)} h_i^{(t-1)} + b_i^{(t)}] \xrightarrow{\mathbb{E}[X+Y]=E[x]+E[Y]} \\
&= \mathbb{E}[W_i(t) h_i^{(t-1)}] + \mathbb{E}[b_i(t)] \xrightarrow{if(X \perp Y) then(\mathbb{E}[XY]=\mathbb{E}[X]\mathbb{E}[Y]}) \\
&= \mathbb{E}[W_i^{(t)}] \mathbb{E}[h_i^{(t-1)}] + \mathbb{E}[b_i^{(t)}] \xrightarrow{(b_i^{(t)}=c) so: (\mathbb{E}[b_i^{(t)}]=c)} \\
&= \mathbb{E}[W_i^{(t)}] \mathbb{E}[h_i^{(t-1)}] + c = 0 \quad (*) \\
\mathrm{Var}(h_i^{(t-1)}) &= \frac{1}{2} - \mathbb{E}[h_i^{(t-1)}]^2 \longrightarrow \mathrm{Var}(h^{(t-1)}) = \frac{1}{2} 1_{d^{t-1}} - \mathbb{E}[h^{(t-1)}]^2
\end{aligned} \tag{11}$$

$$\begin{aligned}
Var[a_i^{(t)}] &= \mathrm{Var}(W_i^{(t)} h_i^{(t-1)} + b_i^{(t)}) \xrightarrow{Var(X+Y)=Var(x)+Var(Y)} \\
&= \mathrm{Var}(W_i(t) h_i^{(t-1)}) + \mathrm{Var}(b_i(t)) \xrightarrow{part-a} \\
&= \mathrm{Var}(W_i^{(t)}) (\frac{1}{2} 1_{d^{t-1}} - \mathbb{E}[h^{(t-1)}]^2) + \mathrm{Var}(W_i^{(t)}) \mathbb{E}[h_i^{(t-1)}]^2 \\
&\quad + \mathbb{E}[W_i^{(t)}]^2 (\frac{1}{2} 1_{d^{t-1}} - \mathbb{E}[h^{(t-1)}]^2) + \mathrm{Var}(b_i^{(t)}) \xrightarrow{problemdef.} \\
&= \frac{1}{2} \mathrm{Var}(W_i^{(t)}) 1_{d^{t-1}} - \mathrm{Var}(W_i^{(t)}) \mathbb{E}[h^{(t-1)}]^2 + \mathrm{Var}(W_i^{(t)}) \mathbb{E}[h_i^{(t-1)}]^2 \\
&\quad + \frac{1}{2} \mathbb{E}[W_i^{(t)}]^2 1_{d^{t-1}} - \mathbb{E}[W_i^{(t)}]^2 \mathbb{E}[h^{(t-1)}]^2) + \cancelto{0}{\mathrm{Var}(b_i^{(t)})} \\
&= \frac{1}{2} \mathrm{Var}(W_i^{(t)}) 1_{d^{(t-1)}} + \frac{1}{2} \mathbb{E}[W_i^{(t)}]^2 1_{d^{(t-1)}} - \mathbb{E}[W_i^{(t)}]^2 \mathbb{E}[h^{(t-1)}]^2) \xrightarrow{from(*)} \\
&= \frac{1}{2} \mathrm{Var}(W_i^{(t)}) 1_{d^{(t-1)}} + \frac{1}{2} \mathbb{E}[W_i^{(t)}]^2 1_{d^{(t-1)}} - \cancel{\mathbb{E}[W_i^{(t)}]^2} (\frac{-c}{\cancel{\mathbb{E}[W_i^{(t)}]} d^{(t-1)}})^2 d^{(t-1)} \\
&= \frac{1}{2} \mathrm{Var}(W_i^{(t)}) 1_{d^{(t-1)}} + \frac{1}{2} \mathbb{E}[W_i^{(t)}]^2 1_{d^{(t-1)}} - \frac{-c^2}{d^{(t-1)}} \\
&= 1
\end{aligned} \tag{12}$$

Lets assume that $Var(W_i^{(t)}) = \sigma^2$ and $E[W_i^{(t)}] = \mu$ since W_{ij} sampled form Gaussian

distribution.

$$\begin{aligned} Var[a_i^{(t)}] &= \frac{1}{2}\text{Var}(W_i^{(t)})1_{d^{(t-1)}} + \frac{1}{2}\mathbb{E}[W_i^{(t)}]^2 1_{d^{(t-1)}} - \frac{-c^2}{d^{(t-1)}} \\ &= \frac{1}{2}\sigma^2 1_{d^{(t-1)}}^T 1_{d^{(t-1)}} + \frac{1}{2}\mu^2 1_{d^{(t-1)}}^T 1_{d^{(t-1)}} - \frac{-c^2}{d^{(t-1)}} \\ &= \frac{1}{2}((\sigma^2+\mu^2)(d^{(t-1)})) - \frac{-c^2}{d^{(t-1)}} = 1 \end{aligned} \tag{13}$$

One solution is when we set $c = 0, and \quad \mu = 0$ which results $\sigma^2 = \frac{2}{d^{(t-1)}}$.

(c) The He initialization also have this form.

(d) Using this normalization method, the scale of the output activation is controlled and it is great because it will give us unit variance pre-activations. The ReLU turns approximately half of the neurons (the negative ones) into zeros and it help to control the scale of the activations.

3. For part 1 we know $Var(W_i^{(t)}) = \frac{(\beta-\alpha)^2}{12}$ and $E[W_i^{(t)}] = \frac{\alpha+\beta}{2}$ since $W_{ij}$is sampled form Uniform distribution. We also know that $\boldsymbol{W}_i^{(t)}$'s are independent and identically distributed random variables, $\text{Var}(\boldsymbol{W}_i^{(t)}) = \frac{1}{d^{(t-1)}}$.
We can write:

$$\mathbb{E}[a_i^t] = c = 0$$

$$\begin{aligned} Var[a_i^{(t)}] &= \text{Var}(\boldsymbol{W}_i^{(t)}) + \mathbb{E}[\boldsymbol{W}_i^{(t)}]^2 \\ &= \frac{(\beta-\alpha)^2}{12} d^{(t-1)} + (\frac{\alpha+\beta}{2})^2 d^{(t-1)} \\ &= \frac{4(\alpha^2+\beta^2-\alpha\beta)}{12} = \frac{(\alpha-\beta)(\alpha+\beta)}{3} d^{(t-1)} = 1 \end{aligned} \tag{14}$$

$\alpha = -\sqrt{\frac{3}{d^{(t-1)}}} \quad \beta = \sqrt{\frac{3}{d^{(t-1)}}}$ is a possible solution for the equation above having the constraints in mind. This seems to be Glorot Uniform initialization.
For part 2 we know $Var(W_i^{(t)}) = \frac{(\beta-\alpha)^2}{12}$ and $E[W_i^{(t)}] = \frac{\alpha+\beta}{2}$ since $W_{ij}$is sampled form Uniform distribution.
We can write:

$$\mathbb{E}[a_i^t] = c = 0$$

$$\begin{aligned} Var[a_i^{(t)}] &= \frac{1}{2}\text{Var}(\boldsymbol{W}_i^{(t)}) + \frac{1}{2}\mathbb{E}[\boldsymbol{W}_i^{(t)}]^2 \\ &= \frac{1}{2}\frac{(\beta-\alpha)^2}{12} d^{(t-1)} + \frac{1}{2}(\frac{\alpha+\beta}{2})^2 d^{(t-1)} \\ &= \frac{4(\alpha^2+\beta^2-\alpha\beta)}{24} = \frac{(\alpha-\beta)(\alpha+\beta)}{6} d^{(t-1)} = 1 \end{aligned} \tag{15}$$

$\alpha = -\sqrt{\frac{6}{d^{(t-1)}}} \quad \beta = \sqrt{\frac{6}{d^{(t-1)}}}$ is a possible solution for the equation above having the constraints in mind. This is He Uniform initialization.
As far as I know one can solve this problem differently, I just wrote one possible answer for the questions in hand.

Question 4 (4-6-6)**.** This question is about normalization techniques.

1. Batch normalization, layer normalization and instance normalization all involve calculating the mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma^2}$ with respect to different subsets of the tensor dimensions. Given the following 3D tensor, calculate the corresponding mean and variance tensors for each normalization technique: $\boldsymbol{\mu}_{batch}$, $\boldsymbol{\mu}_{layer}$, $\boldsymbol{\mu}_{instance}$, $\boldsymbol{\sigma}^2_{batch}$, $\boldsymbol{\sigma}^2_{layer}$, and $\boldsymbol{\sigma}^2_{instance}$.

$$\left[\begin{bmatrix}1,3,2\\1,2,3\end{bmatrix},\begin{bmatrix}3,3,2\\2,4,4\end{bmatrix},\begin{bmatrix}4,2,2\\1,2,4\end{bmatrix},\begin{bmatrix}3,3,2\\3,3,2\end{bmatrix}\right]$$

   The size of this tensor is 4 x 2 x 3 which corresponds to the batch size, number of channels, and number of features respectively.

2. For the next two subquestions, 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 $\boldsymbol{w}$.

   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 $\hat{y} = \boldsymbol{w}^\top \boldsymbol{x} + \beta$.

3. 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] $\boldsymbol{W}^\perp \boldsymbol{u} = \boldsymbol{0}$.

Answer 4. 1. (a) **Batch Normalization:** Batch normalization normalizes activations in a network across the mini-batch. The mean and variance for this technique will be of size *number of channels* $\times$ *number of features*.Using the formula $\mu_{batch} = \frac{1}{batch-size}\Sigma_i^{batch-size}\boldsymbol{x}_i$ we have:

$$\mu_{batch} = \begin{bmatrix}11/4, 11/4, 8/4\\7/4, 11/4, 13/4\end{bmatrix} = \begin{bmatrix}2.75, 2.75, 2\\1.75, 2.75, 3.25\end{bmatrix}$$

We can calculate the variance using:

$$\sigma^2_{batch} = \frac{1}{batch-size}\Sigma_i^{batch-size}(\boldsymbol{x}_i - \mu_{batch})^2 = \begin{bmatrix}1.1875, 1.1875, 0\\0.6875, 0.6875, 0.6875\end{bmatrix}$$

(b) **Layer Normalization:** Layer normalization normalizes input across the features. Each of our training examples have 3 features so the mean and variance for this technique will be of size *batch size* $\times$ *number of features*. Using the formula $\mu_{layer} = \frac{1}{feature-size}\Sigma_j^{feature-size}\boldsymbol{x}_{ij}$ we have:

$$\mu_{layer} = \begin{bmatrix}1, 2.5, 2.5\\2.5, 3.5, 3\\2.5, 2, 3\\3, 3, 2\end{bmatrix}$$

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.

$$\sigma^2_{layer} = \begin{bmatrix} 0, 0.25, 0.25 \\ 0.25, 0.25, 1 \\ 2.25, 0, 1 \\ 0, 0, 0 \end{bmatrix}$$

(c) **Instance Normalization:** Instance normalization normalizes across each channel in each training example in a minibatch instead of normalizing across input features in a training example. Each of our training examples have 2 channels so the mean and variance for this technique will be of shape *batch size* $\times$ *number of channels*. Using the formula in the slides we have:

$$\mu_{instance} = \frac{1}{HW}\Sigma^W_{l=1}\Sigma^H_{m=1}x_{tilm} = \begin{bmatrix} 2, 2 \\ 8/3, 10/3 \\ 8/3, 8/3 \\ 8/3, 8/3 \end{bmatrix} = \begin{bmatrix} 2, 2 \\ 2.66, 3.33 \\ 2.66, 2.66 \\ 2.66, 2.66 \end{bmatrix}$$

We can calculate the variance using:

$$\sigma^2_{instance} = \frac{1}{HW}\Sigma^W_{l=1}\Sigma^H_{m=1}(x_{tilm} - \mu_{ti})^2 = \begin{bmatrix} 0.66, 0.66 \\ 0.22, 0.88 \\ 0.88, 1.55 \\ 0.22, 0.22 \end{bmatrix}$$

2. We know $\boldsymbol{w} := \gamma\frac{\boldsymbol{u}}{||\boldsymbol{u}||}$ where γ is scalar parameter controlling the magnitude and $\boldsymbol{u}$ is a vector controlling the direction of $\boldsymbol{w}$. We also know that $\hat{y} = \gamma \cdot \frac{y-\mu_y}{\sigma_y} + \beta$ where $y = \boldsymbol{u}^\top\boldsymbol{x}$, and for vector x we know that ($\text{Var}(\boldsymbol{x}) = \boldsymbol{I}$), and ($\mathbb{E}[\boldsymbol{x}] = \boldsymbol{0}$). Since x is centered at 0 ($\mathbb{E}[\boldsymbol{x}] = \boldsymbol{0}$), $\mu_y = \mathbb{E}[\boldsymbol{y}] = \mathbb{E}[u^T\boldsymbol{x}] = u^T\mathbb{E}[\boldsymbol{x}] = 0$, and since $\boldsymbol{x}$ is whitened ($\text{Var}(\boldsymbol{x}) = \boldsymbol{I}$) we have $\sigma_y = Var(y)^{1/2} = Var(u^T\boldsymbol{x})^{1/2} = (u^TVar(\boldsymbol{x})u)^{1/2} = (u^T\boldsymbol{I}u)^{1/2} = (u^Tu)^{1/2} = ||\boldsymbol{u}||$. So we can write $\hat{y} = \gamma \cdot \frac{y-\mu_y}{\sigma_y} + \beta = \gamma \cdot \frac{u^T\boldsymbol{x}-0}{||\boldsymbol{u}||} + \beta = \gamma \cdot \frac{u^T\boldsymbol{x}}{||\boldsymbol{u}||} + \beta = \boldsymbol{w}^\top\boldsymbol{x} + \beta$

3. Our objective is to show that: $\nabla_{\boldsymbol{u}}L = s\boldsymbol{W}^\perp\nabla_{\boldsymbol{w}}L$ for some s.

$$\begin{aligned}
&\nabla_{\boldsymbol{u}}L = \nabla_{\boldsymbol{u}}w.\nabla_{\boldsymbol{w}}L = \nabla_{\boldsymbol{u}}(\gamma\frac{u}{||u||}).\nabla_{\boldsymbol{w}}L \xrightarrow{quotient-rule} \\
&\nabla_{\boldsymbol{u}}\frac{u}{||u||} = \frac{||u||I - u\frac{d}{du}(u^Tu)^{0.5}}{||u||^2} = \frac{||u||I - \frac{1}{2}u.(u^Tu)^{-0.5}\frac{d(u^Tu)}{du}}{||u||^2} \\
&\qquad = \frac{||u||I - u\frac{1}{2||u||}(2u^T)}{||u||^2} = \frac{||u||I - \frac{uu^T}{||u||}}{||u||^2} = \frac{I - \frac{uu^T}{||u||^2}}{||u||} = \frac{1}{||u||}\boldsymbol{W}^\perp \\
&\nabla_{\boldsymbol{u}}L = \frac{\gamma}{||u||}(I - \frac{uu^T}{||u||^2})\nabla_{\boldsymbol{w}}L = \frac{\gamma}{||u||}\boldsymbol{W}^\perp\nabla_{\boldsymbol{w}}L = s\boldsymbol{W}^\perp\nabla_{\boldsymbol{w}}L, \quad where \quad s = \frac{\gamma}{||u||}
\end{aligned} \tag{16}$$

Question 5 (4-6-4)**.** 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 [2] of matrix $\boldsymbol{A}$ ($||\boldsymbol{A}|| := \max_{\boldsymbol{x}:||\boldsymbol{x}||=1} ||\boldsymbol{A}\boldsymbol{x}||$). Assume $\sigma(x)$ has bounded derivative, i.e. $|\sigma'| \leq \gamma$ 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 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. $\boldsymbol{g}_t = \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$ and $\boldsymbol{h}_t = \boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}$ are given.
Initial point: $\boldsymbol{g}_0 = \sigma(\boldsymbol{h}_0)$ satisfies , assuming that it is also true for $\boldsymbol{g}_{t-1} = \sigma(\boldsymbol{h}_{t-1})$ then we should prove that $\boldsymbol{g}_t = \sigma(\boldsymbol{h}_t)$.
We can write:
$\boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b} = \boldsymbol{h}_t - \boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) \longrightarrow \boldsymbol{g}_t = \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$
$= \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{h}_t - \boldsymbol{W}\sigma(\boldsymbol{h}_{t-1})) = \sigma(\boldsymbol{h}_t + \boldsymbol{W}(\boldsymbol{g}_{t-1} - \sigma(\boldsymbol{h}_{t-1})))$
Then $\boldsymbol{g}_t = \sigma(\boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}) = \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$.

2.
$$\frac{\partial h_T}{\partial h_0} = \frac{\partial h_T}{\partial h_{T-1}} \frac{\partial h_{T-1}}{\partial h_{T-2}} \cdots \frac{\partial h_t}{\partial h_{t-1}} \cdots \frac{\partial h_1}{\partial h_0}$$

$\frac{\partial h_t}{\partial h_{t-1}} = \frac{\partial(\boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})}{\partial h_{t-1}} = \boldsymbol{W}\frac{\partial \sigma(\boldsymbol{h}_{t-1}}{\partial \boldsymbol{h}_{t-1}} = \boldsymbol{W}\sigma'$
Recall the properties of the L_2 operator norm:

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

From the equations above we have:

$$\left|\left|\frac{\partial h_t}{\partial h_{t-1}}\right|\right| = \left|\left|\boldsymbol{W}\sigma'\right|\right| = \left|\left|\boldsymbol{W}\right|\right|\gamma = \sqrt{\lambda_1(\boldsymbol{W}^\top \boldsymbol{W})}\gamma \leq \sqrt{\frac{\delta^2}{\gamma^2}}\gamma = \left|\frac{\delta}{\gamma}\right| \cdot \gamma = \delta$$

2. 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.

Until now, we showed that 2-norm is bounded by δ. Now we need to calculate it for the

$$\left|\left|\frac{\partial h_T}{\partial h_0}\right|\right| \leq \Pi_{t=1}^{T}\left|\left|\frac{\partial h_t}{\partial h_{t-1}}\right|\right| \leq \delta^T$$

We know $0 \leq \delta < 1$ so if $\boldsymbol{T}$ goes to ∞ the value of δ^T will goes to zero and the gradiets will vanish.

3. We know that the largest eigenvalue of weights is not larger than $\frac{\delta^2}{\gamma^2}$. And because of that the gradient of hidden states cannot explode. So it is necessary condition for gradient explosion. But it cannot always prevent explosion. One possible case where it is not sufficient is when the largest eigenvalue is orthogonal to hidden state.

Question 6 (4-8-8)**.** 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 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 $L_t = ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2$.

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

1. First, complete the following computational graph for this RNN, unrolled for 3 time steps (from $t = 1$ to $t = 3$). Label 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.

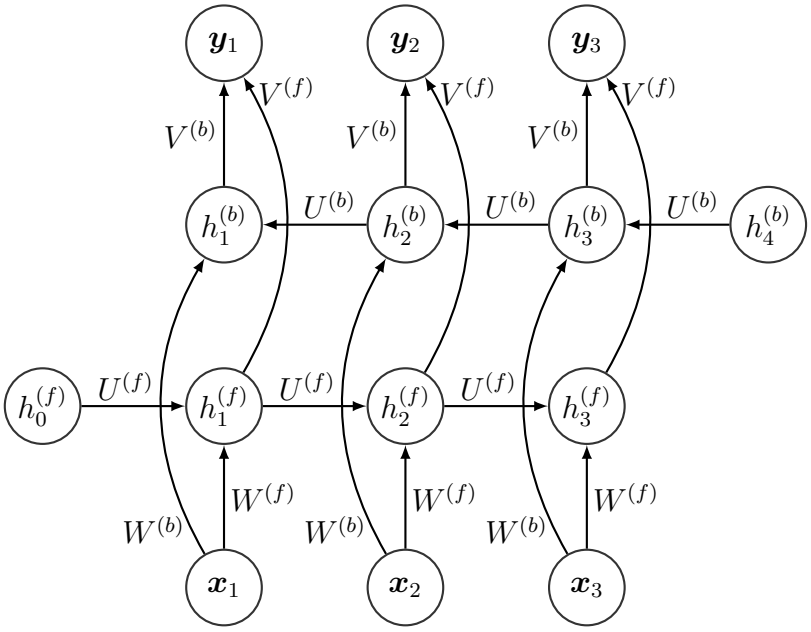

FIGURE 1 – Computational graph of the bidirectional RNN unrolled for three timesteps.

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:

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

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

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. Now 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.
Hint: It might be useful to consider the contribution of the weight matrices when computing the recurrent hidden unit at a particular time t and how those contributions might be aggregated.

Answer 6. 1. The graph is completed.

2. Given $L = \sum_t L_t$ where $L_t = ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2$ and

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

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

we need to find out the values 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 $\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}}$. Lets start with $\nabla_{\boldsymbol{h}_t^{(f)}} L_t$:

- $\nabla_{\boldsymbol{h}_t^{(f)}} L_t = \nabla_{\boldsymbol{h}_t^{(f)}} ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2 = 2||\boldsymbol{z}_t - \boldsymbol{y}_t||_2 \nabla_{h_t^{(f)}} (\boldsymbol{z}_t - (\boldsymbol{V}^{(f)} \boldsymbol{h}_t^{(f)} + \boldsymbol{V}^{(b)} \boldsymbol{h}_t^{(b)}))$
$= 2||\boldsymbol{z}_t - \boldsymbol{y}_t||_2 (-\boldsymbol{V}^{(f)}) = -2\boldsymbol{V}^{(f)T} ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2 = -2\boldsymbol{V}^{(f)T} \cdot (\boldsymbol{z}_t - \boldsymbol{y}_t)$

- $\nabla_{\boldsymbol{h}_t^{(b)}} L_t = \nabla_{\boldsymbol{h}_t^{(b)}} ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2 = 2||\boldsymbol{z}_t - \boldsymbol{y}_t||_2 \nabla_{h_t^{(b)}} (\boldsymbol{z}_t - (\boldsymbol{V}^{(f)} \boldsymbol{h}_t^{(f)} + \boldsymbol{V}^{(b)} \boldsymbol{h}_t^{(b)}))$
$= 2||\boldsymbol{z}_t - \boldsymbol{y}_t||_2 (-\boldsymbol{V}^{(b)}) = -2\boldsymbol{V}^{(b)T} ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2 = -2\boldsymbol{V}^{(b)T} \cdot (\boldsymbol{z}_t - \boldsymbol{y}_t)$

- $\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}} = \nabla_{h_t^{(f)}} (\sigma(\boldsymbol{W}^{(f)} \boldsymbol{x}_{t+1} + \boldsymbol{U}^{(f)} \boldsymbol{h}_t^{(f)}))$
$= \sigma(\boldsymbol{W}^{(f)} \boldsymbol{x}_{t+1} + \boldsymbol{U}^{(f)} \boldsymbol{h}_t^{(f)})(1 - \sigma(\boldsymbol{W}^{(f)} \boldsymbol{x}_{t+1} + \boldsymbol{U}^{(f)} \boldsymbol{h}_t^{(f)})) \boldsymbol{U}^{(f)}$
$= \boldsymbol{h}_{t+1}^{(f)} (1 - \boldsymbol{h}_{t+1}^{(f)}) \boldsymbol{U}^{(f)} = diag(\boldsymbol{h}_{t+1}^{(f)} (1 - \boldsymbol{h}_{t+1}^{(f)})) \boldsymbol{U}^{(f)}$

- $\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}} = \nabla_{h_t}^{(b)} (\sigma(\boldsymbol{W}^{(b)} \boldsymbol{x}_{t-1} + \boldsymbol{U}^{(b)} \boldsymbol{h}_t^{(b)}))$
$= \sigma(\boldsymbol{W}^{(b)} \boldsymbol{x}_{t-1} + \boldsymbol{U}^{(b)} \boldsymbol{h}_t^{(b)})(1 - \sigma(\boldsymbol{W}^{(b)} \boldsymbol{x}_{t-1} + \boldsymbol{U}^{(b)} \boldsymbol{h}_t^{(b)})) . \boldsymbol{U}^{(b)}$
$= \boldsymbol{h}_{t-1}^{(b)} (1 - \boldsymbol{h}_{t-1}^{(b)}) \boldsymbol{U}^{(b)} = diag(\boldsymbol{h}_{t-1}^{(b)} (1 - \boldsymbol{h}_{t-1}^{(b)})) \boldsymbol{U}^{(b)}$

3. Derive $\nabla_{\boldsymbol{W}^{(f)}} L$ and $\nabla_{\boldsymbol{U}^{(b)}} L$ as a functions of $\nabla_{\boldsymbol{h}_t^{(f)}} L$ and $\nabla_{\boldsymbol{h}_t^{(b)}} L$:

-
$$\nabla_{\boldsymbol{W}^{(f)}} L = \Sigma_{t=1}^{dim} \frac{\partial h_t^{(f)}}{\partial \boldsymbol{W}_t^{(f)}} \nabla_{h_t^{(f)}} L$$

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

$$\nabla_{\boldsymbol{W}^{(f)}} L = \Sigma_{t=1}^{dim} \nabla_{\boldsymbol{w}_t^{(f)}} L = \boldsymbol{h}_t^{(f)} (1 - \boldsymbol{h}_t^{(f)}) \boldsymbol{x}_t \nabla_{h_t^{(f)}} L$$

$$\nabla_{\boldsymbol{W}^{(f)}} L = \Sigma_{t=1}^{dim} diag(\boldsymbol{h}_t^{(f)} (1 - \boldsymbol{h}_t^{(f)})) \cdot \nabla_{h_t^{(f)}} L \cdot \boldsymbol{x}_t^T$$

- $$\nabla_{\boldsymbol{U}^{(b)}} L = \Sigma_{t=1}^{dim} \frac{\partial h_t^{(b)}}{\partial \boldsymbol{U}_t^{(b)}} \nabla_{h_t^{(f)}} L$$

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

$$\nabla_{\boldsymbol{U}^{(b)}} L = \Sigma_{t=1}^{dim} \nabla_{\boldsymbol{U}_t^{(b)}} L = \boldsymbol{h}_t^{(b)} (1 - \boldsymbol{h}_t^{(b)}) h_{t+1}^{(T)} \nabla_{h_t^{(b)}} L$$

$$\nabla_{\boldsymbol{U}^{(b)}} L = \Sigma_{t=1}^{dim} diag(\boldsymbol{h}_t^{(b)} (1 - \boldsymbol{h}_t^{(b)})) \cdot \nabla_{h_t^{(b)}} L \cdot h_{t+1}^T$$