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. (a) SGD with momentum:

$$\boldsymbol{v}_t = \alpha \boldsymbol{v}_{t-1} + \epsilon \boldsymbol{g}_t \qquad \Delta\boldsymbol{\theta}_t = -\boldsymbol{v}_t$$

Replacing $\boldsymbol{v}_t$ with $-\Delta\boldsymbol{\theta}_t$ and $\boldsymbol{v}_{t-1}$ with $-\Delta\boldsymbol{\theta}_{t-1}$

$$\Delta\boldsymbol{\theta}_t = \alpha\Delta\boldsymbol{\theta}_{t-1} - \epsilon \boldsymbol{g}_t$$

(b) SGD with running average of $\boldsymbol{g}_t$:

$$\boldsymbol{v}_t = \beta \boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t \qquad \Delta\boldsymbol{\theta}_t = -\delta\boldsymbol{v}_t$$

Replacing $\boldsymbol{v}_t$ with $\frac{-\Delta\boldsymbol{\theta}_t}{\delta}$ and $\boldsymbol{v}_{t-1}$ with $\frac{-\Delta\boldsymbol{\theta}_{t-1}}{\delta}$

$$\Delta\boldsymbol{\theta}_t = \beta\Delta\boldsymbol{\theta}_{t-1} - \delta(1-\beta)\boldsymbol{g}_t$$

The above two equations for update rules are equivalent for the following relation.

$$\alpha = \beta \quad \epsilon = \delta(1-\beta)\boldsymbol{g}_t$$

2. SGD with running average of $\boldsymbol{g}_t$:

$$\boldsymbol{v}_t = \beta \boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t$$
$$\boldsymbol{v}_{t-1} = \beta \boldsymbol{v}_{t-2} + (1-\beta)\boldsymbol{g}_{t-1}$$
$$\boldsymbol{v}_1 = \beta \boldsymbol{v}_0 + (1-\beta)\boldsymbol{g}_0 \quad \text{where } \boldsymbol{v}_0 = 0$$

Unrolling the steps till few timestamps

$$\boldsymbol{v}_t = \beta(\beta((\beta\boldsymbol{v}_{t-3} + (1-\beta)\boldsymbol{g}_{t-2}) + (1-\beta)\boldsymbol{g}_{t-1}) + (1-\beta)\boldsymbol{g}_{t-1}) + (1-\beta)\boldsymbol{g}_t$$
$$\boldsymbol{v}_t = \beta^3\boldsymbol{v}_{t-3} + \beta^2(1-\beta)\boldsymbol{g}_{t-2} + \beta^1(1-\beta)\boldsymbol{g}_{t-1} + \beta^0(1-\beta)\boldsymbol{g}_t$$

If we unroll till $\boldsymbol{v}_1$, then the equation becomes independent of $\boldsymbol{v}$. Hence we can write $\boldsymbol{v}_t$ as a function of $\boldsymbol{g}_i's$:

$$\boldsymbol{v}_t = (1-\beta)\sum_{i=0}^{t-1}\beta^i\boldsymbol{g}_{t-i}$$

3. SGD with running average of $\boldsymbol{g}_t$:

$$\boldsymbol{v}_t = (1-\beta)\sum_{i=0}^{t-1}\beta^i\boldsymbol{g}_{t-i}$$
$$\mathbb{E}[\boldsymbol{v}_t] = (1-\beta)\sum_{i=0}^{t-1}\beta^i\mathbb{E}[\boldsymbol{g}_{t-i}]$$

Since $\boldsymbol{g}_t$ is stationary and independent of t

$$\mathbb{E}[\boldsymbol{v}_t] = (1-\beta)\mathbb{E}[\boldsymbol{g}_t]\sum_{i=0}^{t-1}\beta^i$$

Expanding geometric series

$$\mathbb{E}[\boldsymbol{v}_t] = (1-\beta)\mathbb{E}[\boldsymbol{g}_t]\frac{1-\beta^i}{1-\beta}$$
$$\mathbb{E}[\boldsymbol{v}_t] = \mathbb{E}[\boldsymbol{g}_t](1-\beta^i)$$

Therefore, $\mathbb{E}[\boldsymbol{v}_t] = \mathbb{E}[\boldsymbol{g}_t]$ and they can be equal if $\mathbb{E}[\boldsymbol{v}_t]$ is scaled by $(1-\beta)$ i.e. $\boldsymbol{v}_t \leftarrow (\boldsymbol{v}_t)/(1-\beta^t)$

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.

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

$$E[L(\boldsymbol{w})] = \sum_{i=1}^{n} E[(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{R}_i \boldsymbol{w})^2]$$

Using $E[Z^2] = E[Z]^2 + Var(Z)$

$$E[L(\boldsymbol{w})] = \sum_{i=1}^{n} E[\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{R}_i \boldsymbol{w}]^2 + Var(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{R}_i \boldsymbol{w})$$

Solving first half of equation. Using $E[R] = p$

$$\sum_{i=1}^{n} E[\boldsymbol{y}_i - (\boldsymbol{X}_i \boldsymbol{R}_i)\boldsymbol{w}]^2 = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||$$

Solving second half of equation. Using $Var(Z) = E[Z^2] - E[Z]^2$

$$\sum_{i=1}^{n} Var(\boldsymbol{y}_i - (\boldsymbol{X}_i \boldsymbol{R}_i)\boldsymbol{w}) = \sum_{i=1}^{n} E[(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{R}_i \boldsymbol{w})^2] - E[\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{R}_i \boldsymbol{w}]^2$$

Using Mean and Variance of Bernoulli distribution. $E[R] = p$, $Var(R) = p(1-p)$

$$= \sum_{i=1}^{n} \boldsymbol{w}^\top p(1-p)(\boldsymbol{X}^\top \boldsymbol{X})_{ii} \boldsymbol{w}$$

Since $\Gamma_{ii} = (\boldsymbol{X}^\top \boldsymbol{X})_{ii}^{1/2}$

$$= p(1-p)||\Gamma \boldsymbol{w}||^2$$

Adding both the parts

$$E[L(\boldsymbol{w})] = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}|| + p(1-p)||\Gamma \boldsymbol{w}||^2$$

2. Expectation of Loss function derived in the previous question

$$\mathbb{E}[L(\boldsymbol{w})] = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma\boldsymbol{w}||^2$$

Differentiating w.r.t $\boldsymbol{W}$ and equating it to zero.

$$\frac{\partial\mathbb{E}[L(\boldsymbol{w})]}{\partial\boldsymbol{w}} = 2(\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w})(-p\boldsymbol{X}) + 2p(1-p)(\Gamma^2\boldsymbol{w})$$

$$\begin{aligned} 0 &= -2p\boldsymbol{X}^\top\boldsymbol{y} + 2p^2\boldsymbol{X}^\top\boldsymbol{X} + 2p(1-p)\Gamma^2\boldsymbol{w} \\ p\boldsymbol{X}^\top\boldsymbol{y} &= p^2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + p(1-p)\Gamma^2\boldsymbol{w} \\ \boldsymbol{X}^\top\boldsymbol{y} &= p\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + (1-p)\Gamma^2\boldsymbol{w} \\ \boldsymbol{X}^\top\boldsymbol{y} &= p\boldsymbol{w}(\boldsymbol{X}^\top\boldsymbol{X} + \frac{1-p}{p}\Gamma^2) \end{aligned}$$

Let $\gamma^{dropout} = \frac{1-p}{p}$

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

The values of $\gamma^{dropout}$ will increase with decrease in the probability p whose value is between 0 and 1. Lowering the probability will drop more units and increase $\gamma^{dropout}$

3. Ridge Regression: Linear Regression with L2 regularization

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

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

Differentiating w.r.t $\boldsymbol{W}$ and equating it to zero.

$$\frac{\partial L(\boldsymbol{w})}{\partial\boldsymbol{w}} = 0 - 2\boldsymbol{X}^\top\boldsymbol{y} + 2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + 2\lambda^{L_2}\boldsymbol{w}$$

$$\begin{aligned} 0 &= -\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + \lambda^{L_2}\boldsymbol{w} \\ \boldsymbol{X}^\top\boldsymbol{y} &= \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + \lambda^{L_2}\boldsymbol{w} \\ \boldsymbol{X}^\top\boldsymbol{y} &= \boldsymbol{w}(\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{L_2}\boldsymbol{I}) \\ \boldsymbol{w} &= (\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{L_2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \end{aligned}$$

..

4. The difference between the equations in 2.2 and 2.3 is of the term $\lambda^{dropout}\Gamma^2$ and $\lambda^{L_2}\boldsymbol{I}$. As we can see that the weights in the case of dropouts also depends on the input itself in the form of $\Gamma^2 = \boldsymbol{X}^\top\boldsymbol{X}$ whereas the weight in the L_2 regularization is independent of input.
Also, the weights are reduced uniformly in L_2 regularization unlike the dropout layer where it can be controlled by changing the probability i.e. lower probability will drop more units and the $\lambda^{dropout}$ will increase. The $\lambda^{dropout}$ and λ^{L_2} are related and one is a factor of another.

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:

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

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 $\mathrm{Var}(a_i^{(t)}) = 1$, $1 \leq i \leq d^{(t)}$) for the assumptions about either the activations or pre-activations of layer $t-1$ listed below. Note we are not 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 $\mathrm{Var}(h_i^{(t-1)}) = 1$ for $1 \leq i \leq d^{(t-1)}$. Also, assume entries of $\boldsymbol{h}^{(t-1)}$ are uncorrelated (the answer should not depend on g).
   (a) Show $\mathrm{Var}(XY) = \mathrm{Var}(X)\mathrm{Var}(Y) + \mathrm{Var}(X)\mathbb{E}[Y]^2 + \mathrm{Var}(Y)\mathbb{E}[X]^2$ when $X \perp Y$
   (b) Write $\mathbb{E}[a_i^{(t)}]$ and $\mathrm{Var}(a_i^{(t)})$ in terms of $c, \mu, \sigma^2, \mathrm{Var}(h_i^{(t-1)}), \mathbb{E}[h_i^{(t-1)}]$.
   (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 $\mathrm{Var}(a_i^{(t)}) = 1$ for $1 \leq i \leq d^{(t)}$.
2. Now assume that the pre-activations of the previous layer satisfy $\mathbb{E}[a_i^{(t-1)}] = 0$, $\mathrm{Var}(a_i^{(t-1)}) = 1$ and $a_i^{(t-1)}$ has a symmetric distribution for $1 \leq i \leq d^{(t-1)}$. Assume entries of $\boldsymbol{a}^{(t-1)}$ are uncorrelated. Consider the case of ReLU activation: $g(x) = \max\{0, x\}$.
   (a) 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 $\mathrm{Var}(a_i^{(t)}) = 1$ for $1 \leq i \leq d^{(t)}$.
   (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 $\mathrm{Var}(a_i^{(t)}) = 1$.

Answer 3. .

1. (a)

$$Var(XY) = E[(XY)^2] - E[(XY)]^2$$
$$Var(XY) = E[X^2Y^2] - E[(XY)]^2$$

Since $X \perp Y$

$$Var(XY) = E[X^2]E[Y^2] - E[X]^2E[Y]^2$$
$$Var(XY) = \left(Var(X) + E[X]^2\right)\left(Var(Y) + E[Y]^2\right) - E[X]^2E[Y]^2$$
$$Var(XY) = Var(X)Var(Y) + Var(X)E[Y]^2 + Var(Y)E[X]^2 + E[X]^2E[Y]^2 - E[X]^2E[Y]^2$$
$$Var(XY) = Var(X)Var(Y) + Var(X)E[Y]^2 + Var(Y)E[X]^2$$

(b)

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

$$a_i^{(t)} = \sum_{j=1}^{d^{t-1}} \left(\boldsymbol{W}_{ij}^{(t)}\boldsymbol{h}_j^{(t-1)}\right) + b_i^{(t)}$$

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

Since these are independent variables

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

$E[\boldsymbol{W}_{ij}^{(t)}] = \mu$ and $E[b_i^{(t)}] = c$

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

Here, $E[\boldsymbol{h}_j^{(t-1)}] = 0$

$$\therefore E[a_i^{(t)}] = c$$

$$Var(a_i^{(t)}) = Var(\sum_{j=1}^{d^{t-1}} \boldsymbol{W}_{ij}^{(t)}\boldsymbol{h}_j^{(t-1)} + b_i^{(t)})$$

Using $Var(X+Y) = Var(X) + Var(Y)$

$$Var(a_i^{(t)}) = \sum_{j=1}^{d^{t-1}} Var(\boldsymbol{W}_{ij}^{(t)}\boldsymbol{h}_j^{(t-1)}) + Var(b^{(t)})$$

Using $\mathrm{Var}(XY) = \mathrm{Var}(X)\mathrm{Var}(Y) + \mathrm{Var}(X)\mathbb{E}[Y]^2 + \mathrm{Var}(Y)\mathbb{E}[X]^2$ when $X \perp Y$

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

Putting values $E[\boldsymbol{W}^{(t)}] = \mu$, $E[h^{(t-1)}] = 0$, $Var(\boldsymbol{W}^{(t)}) = \sigma^2$, $Var(h^{(t-1)}) = 1$, $Var(b^{(t)}) = 0$ in equation.

$$Var(a^{(t)}) = \sum_{j=1}^{d^{t-1}} \left(\sigma^2 \cdot Var(h^{(t-1)}) + \sigma^2 \cdot E(h^{(t-1)})^2 + 1 \cdot \mu^2\right) + 0$$

$$Var(a^{(t)}) = \sum_{j=1}^{d^{t-1}} \left(\sigma^2 \cdot 1 + \sigma^2 \cdot 0 + 1 \cdot \mu^2\right) + 0$$

$$Var(a^{(t)}) = \sum_{j=1}^{d^{t-1}} \left(\sigma^2 \cdot 1 + \sigma^2 \cdot 0 + 1 \cdot \mu^2\right)$$

$$Var(a^{(t)}) = d^{(t-1)}(\sigma^2 + \mu^2)$$

(c) Using equations solved in the above question.

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

For $E[a_i^{(t)}] = 0$

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

$$c = 0$$

$$Var(a^{(t)}) = \sum_{j=1}^{d^{t-1}} \left(\sigma^2 \cdot Var(h^{(t-1)}) + \sigma^2 \cdot E[h^{(t-1)}]^2 + 1 \cdot \mu^2\right) + 0$$

For $Var(a_i^{(t-1)}) = 1$

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

$\therefore$ for $E[a_i^{(t)}] = 0$ and $Var(a_i^{(t)}) = 1 \rightarrow c = 0, \mu = 0$ and $\sigma^2 = 1/d^{(t-1)}$

2. (a)

$$h_i^{(t)} = g(a_i^{(t)})$$
$$h_i^{(t)} = max(0, a_i^{(t)})$$
$$E[(h_i^{(t-1)})^2] = E[max(a_i^{(t-1)})^2]$$
$$E[(h_i^{(t-1)})^2] = \int_{-\infty}^{\infty} max(0, a_i^{(t-1)})^2 p(a_i^{(t-1)}) da_i^{(t-1)}]$$

We can write this half the integral over real domain i.e. greater than 0

$$E[(h_i^{(t-1)})^2] = \frac{1}{2}\int_{-\infty}^{\infty} a_i^{(t-1)})^2 p(a_i^{(t-1)}) da_i^{(t-1)}]$$

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

Subtracting with 0 as $E[(a_i^{(t-1)})] = 0$

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

$$E[(h_i^{(t-1)})^2] = \frac{1}{2}Var(a_i^{(t-1)})$$

Using $Var(a_i^{(t-1)}) = 1$

$$E[(h_i^{(t-1)})^2] = \frac{1}{2}$$

(b)

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

For $E[a_i^{(t)}] = 0$

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

$$Var(a_i^{(t)}) = \sum_{j=1}^{d^{t-1}} \Big(Var(\boldsymbol{W}_{ij}^{(t)})Var(\boldsymbol{h}_j^{(t-1)}) + Var(\boldsymbol{W}_{ij}^{(t)})E[\boldsymbol{h}_j^{(t-1)}]^2 + Var(\boldsymbol{h}_j^{(t-1)})E[\boldsymbol{W}_{ij}^{(t)}]^2\Big) + Var(b_i^{(t)})$$

For $Var(a^{(t)}) = 1$

$$1 = \sum_{j=1}^{d^{t-1}} \Big(Var(\boldsymbol{W}_{ij}^{(t)})Var(\boldsymbol{h}_j^{(t-1)}) + Var(\boldsymbol{W}_{ij}^{(t)})E[\boldsymbol{h}_j^{(t-1)}]^2 + Var(\boldsymbol{h}_j^{(t-1)})E[\boldsymbol{W}_{ij}^{(t)}]^2\Big) + Var(b_i^{(t)})$$

Since $\mu = 0 \rightarrow E[\boldsymbol{W}_{ij}^{(t)}]^2 = 0$. Also $Var(b_i^{(t)}) = 0$

$$1 = \sum_{j=1}^{d^{t-1}} \Big(Var(\boldsymbol{W}_{ij}^{(t)})Var(\boldsymbol{h}_j^{(t-1)}) + Var(\boldsymbol{W}_{ij}^{(t)})E[\boldsymbol{h}_j^{(t-1)}]^2\Big)$$

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

Using $E[Z^2] = Var(Z) + E[Z]^2$

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

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

Using $E[(\boldsymbol{h}_j^{(t-1)})^2] = 1/2$ from previous question

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

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

$\therefore$ for $E[a_i^{(t)}] = 0$ and $Var(a^{(t)}) = 1 \rightarrow c = 0, \mu = 0, \sigma^2 = \frac{2}{d^{(t-1)}}$

(c) The above initialization is He Normal Initialization

(d) In this method, the weights are initialized keeping in mind the size of the previous layer which helps in attaining a global minimum of the cost function faster and more efficiently.The weights are still random but differ in range depending on the size of the previous layer of neurons. This provides a controlled initialization hence the faster and more efficient gradient descent.

3. For $W_{ij}^{(t)} \sim Uniform(\alpha,\beta)$, $E[W_{ij}^{(t)}] = \frac{\alpha+\beta}{2}, Var(W_{ij}^{(t)} = \frac{(\beta-\alpha)^2}{12}$
   (a) For question 1

$$\mathbb{E}[a_i^{(t)}] = \sum_{j=1}^{d^{(t-1)}} E[W_{ij}^t]E[h_j^{(t-1)} + E[b_i^t]$$

$$0 = \sum_{j=1}^{d^{(t-1)}} E[W_{ij}^t]E[h_j^{(t-1)}] + E[b_i^t]$$

$$0 = d^{(t-1)} \cdot \frac{\alpha+\beta}{2} \cdot 0 + c$$

$$c = 0$$

$$Var(a_i^{(t)}) = \sum_{j=1}^{d^{(t-1)}} \left(Var(W_{ij}^t) + E[W_{ij}^t]^2 \right)$$

$$1 = d^{(t-1)} \left(Var(W_{ij}^t) + E[W_{ij}^t]^2 \right)$$

$$1 = d^{(t-1)} \left(\frac{(\beta-\alpha)^2}{12} + (\frac{\alpha+\beta}{2})^2 \right)$$

$$1 = d^{(t-1)} \left(\frac{\beta^2+\alpha^2+\alpha\beta}{3} \right)$$

$$\frac{3}{d^{(t-1)}} = (\beta^2+\alpha^2+\alpha\beta)$$

Here we can put $\alpha = -\sqrt{\frac{3}{d^{(t-1)}}}$ and $\beta = \sqrt{\frac{3}{d^{(t-1)}}}$ as one of the solution of the equation.

   (b) For question 2

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

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

$$\alpha = -\beta, c = 0$$

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

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

$$1 = d^{t-1} \frac{(\beta-\alpha)^2}{12} \frac{1}{2}$$

Here we can put $\alpha = -\sqrt{\frac{6}{d^{(t-1)}}}$ and $\beta = \sqrt{\frac{6}{d^{(t-1)}}}$ since $\alpha = -\beta$ for $E[a_i^{(t)}] = 0$. This is called He Uniform Initialization.

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 ($\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{I}$) 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] $\boldsymbol{W}^\perp \boldsymbol{u} = \boldsymbol{0}$.

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.

Answer 4. .

1. **Batch Norm:** Though we normalize each neuron/unit, but in case of convolutions, the weights are shared across inputs. Therefore, we apply batch normalization using mean and variance per feature map across entire batch.

$$channel : 1 : \mu = 30/12 = 2.50, \sigma^2 = 0.58$$
$$channel : 2 : \mu = 31/12 = 2.58, \sigma^2 = 1.07$$
$$\boldsymbol{\mu}_{batch} = \begin{bmatrix}2.50, 2.58\end{bmatrix}$$
$$\boldsymbol{\sigma}^2_{batch} = \begin{bmatrix}0.58, 1.07\end{bmatrix}$$

Layer Norm: Normalises all the activations of a single layer for each given example in a batch independently rather than across a batch like batch norm.

$$example : 1 : \mu = 12/06 = 2.00, \sigma^2 = 0.66$$
$$example : 2 : \mu = 18/06 = 3.00, \sigma^2 = 0.66$$
$$example : 3 : \mu = 15/06 = 2.50, \sigma^2 = 1.25$$
$$example : 4 : \mu = 16/06 = 2.66, \sigma^2 = 0.22$$
$$\boldsymbol{\mu}_{layer} = \begin{bmatrix}2.00, 3.00, 2.50, 2.66\end{bmatrix}$$
$$\boldsymbol{\sigma}^2_{layer} = \begin{bmatrix}0.66, 0.66, 1.25, 0.22\end{bmatrix}$$

Instance Norm: is similar to layer normalization but it computes the mean/standard deviation and normalize across each channel in each training example.

$$example : 1 : channel : 1 : u = 06/03 = 2.00, \sigma^2 = 0.66$$
$$example : 1 : channel : 2 : \mu = 06/03 = 2.00, \sigma^2 = 0.66$$
$$example : 2 : channel : 1 : \mu = 08/03 = 2.66, \sigma^2 = 0.22$$
$$example : 2 : channel : 2 : \mu = 10/03 = 3.33, \sigma^2 = 0.88$$
$$example : 3 : channel : 1 : \mu = 08/03 = 2.66, \sigma^2 = 0.88$$
$$example : 3 : channel : 2 : \mu = 07/03 = 2.33, \sigma^2 = 1.55$$
$$example : 4 : channel : 1 : \mu = 08/03 = 2.66, \sigma^2 = 0.22$$
$$example : 4 : channel : 2 : \mu = 08/03 = 2.66, \sigma^2 = 0.22$$
$$\boldsymbol{\mu}_{instance} = \Big[\begin{bmatrix}2.00, 2.00\end{bmatrix}, \begin{bmatrix}2.66, 3.33\end{bmatrix}, \begin{bmatrix}2.66, 2.33\end{bmatrix}, \begin{bmatrix}2.66, 2.66\end{bmatrix}\Big]$$
$$\boldsymbol{\sigma}^2_{instance} = \Big[\begin{bmatrix}0.66, 0.66\end{bmatrix}, \begin{bmatrix}0.22, 0.88\end{bmatrix}, \begin{bmatrix}0.88, 1.55\end{bmatrix}, \begin{bmatrix}0.22, 0.22\end{bmatrix}\Big]$$

2.

$$\hat{y} = \gamma \cdot \frac{y - \mu_y}{\sigma_y} + \beta$$

We can write μ_y as $\mathbb{E}[y] = \mathbb{E}[\boldsymbol{u}^\top \boldsymbol{x}] = \boldsymbol{u}^\top \mathbb{E}[\boldsymbol{x}]$. Since $\mathbb{E}[\boldsymbol{x}] = 0]$ and placing it in the above equation gives $\mathbb{E}[y] = 0$

$$\hat{y} = \gamma \cdot \frac{y - (\mathbb{E}[y] = 0)}{\sqrt{Var(y)}} + \beta$$

We can write $Var(y) = Var(\boldsymbol{u}^\top x)$ and $Var(\boldsymbol{x}) = \boldsymbol{I}$

$$\hat{y} = \gamma \cdot \frac{\boldsymbol{u}^\top \boldsymbol{x}}{\sqrt{||\boldsymbol{u}||^2 \boldsymbol{I}}} + \beta$$

$$\hat{y} = \gamma \cdot \frac{\boldsymbol{u}^\top \boldsymbol{x}}{||\boldsymbol{u}||} + \beta$$

Replacing $\gamma \cdot \frac{\boldsymbol{u}}{||\boldsymbol{u}||}$ with w

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

3.

$$\nabla_u L = \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{u}} \nabla_{\boldsymbol{w}} L$$

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

Using quotient rule, and that gradient of norm is unit vector i.e. $\nabla_u ||\boldsymbol{u}|| = \boldsymbol{u}/||\boldsymbol{u}||$

$$\frac{\partial \boldsymbol{w}}{\partial \boldsymbol{u}} = \frac{||\boldsymbol{u}||\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^\top}{||\boldsymbol{u}||}}{||\boldsymbol{u}^2||}$$

$$\frac{\partial \boldsymbol{w}}{\partial \boldsymbol{u}} = \frac{\gamma}{||\boldsymbol{u}||} \left(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^\top}{||\boldsymbol{u}^2||} \right)$$

$$\nabla_u L = \frac{\gamma}{||\boldsymbol{u}||} \left(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^\top}{||\boldsymbol{u}^2||} \right) \nabla_{\boldsymbol{w}} L$$

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

$$\nabla_u L = s \boldsymbol{W}^\perp \nabla_{\boldsymbol{w}} L$$

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).

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.

Answer 5. .

1. Assuming $\boldsymbol{g}_0 = \sigma(\boldsymbol{h}_0) = 0$

$$\boldsymbol{g}_t = \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$$
$$\boldsymbol{g}_1 = \sigma(\boldsymbol{W}\boldsymbol{g}_0 + \boldsymbol{U}\boldsymbol{x}_1 + \boldsymbol{b})$$
$$\boldsymbol{h}_1 = \boldsymbol{W}\sigma(\boldsymbol{h}_0) + \boldsymbol{U}\boldsymbol{x}_1 + \boldsymbol{b}$$
$$\boldsymbol{U}\boldsymbol{x}_1 + \boldsymbol{b} = \boldsymbol{h}_1 - \boldsymbol{W}\sigma(\boldsymbol{h}_0)$$

Using this value in the above equation.

$$\boldsymbol{g}_1 = \sigma(\boldsymbol{W}\boldsymbol{g}_0 + \Big(\boldsymbol{h}_1 - \boldsymbol{W}\sigma(\boldsymbol{h}_0)\Big))$$
$$\boldsymbol{g}_1 = \sigma(\boldsymbol{W}\cdot 0 + \Big(\boldsymbol{h}_1 - \boldsymbol{W}\cdot 0)\Big))$$
$$\boldsymbol{g}_1 = \sigma(\boldsymbol{h}_1)$$

Using induction case, we can assume that $\boldsymbol{g}_{t-1} = \sigma(\boldsymbol{h}_{t-1})$

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

Using this value in the above equation.

$$\boldsymbol{g}_t = \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \Big(\boldsymbol{h}_t - \boldsymbol{W}\sigma(\boldsymbol{h}_{t-1})\Big))$$
$$\boldsymbol{g}_t = \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \Big(\boldsymbol{h}_t - \boldsymbol{W}\boldsymbol{g}_{t-1}\Big))$$
$$\boldsymbol{g}_t = \sigma(\boldsymbol{h}_t)$$

2. Using multi-variate chain rule

$$\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0} = \frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_{T-1}} \cdot \frac{\partial \boldsymbol{h}_{T-1}}{\partial \boldsymbol{h}_{T-2}} \cdots \frac{\partial \boldsymbol{h}_1}{\partial \boldsymbol{h}_0} = \prod_{i=1}^{T} \frac{\partial \boldsymbol{h}_i}{\partial \boldsymbol{h}_{i-1}}$$

Considering one at a time

$$\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} = \boldsymbol{W} \cdot \frac{\partial \sigma(\boldsymbol{h}_{t-1})}{\partial \boldsymbol{h}_{t-1}}$$

$$\left|\left|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}}\right|\right| = \left|\left|\boldsymbol{W} \cdot \frac{\partial \sigma(\boldsymbol{h}_{t-1})}{\partial \boldsymbol{h}_{t-1}}\right|\right|$$

Use the following properties of the L_2 operator norm

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

$$\left|\left|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}}\right|\right| \leq ||\boldsymbol{W}|| \cdot \left|\left|\frac{\partial \sigma(\boldsymbol{h}_{t-1})}{\partial \boldsymbol{h}_{t-1}}\right|\right|$$

$$\left|\left|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}}\right|\right| \leq \sqrt{\lambda_1(\boldsymbol{W}^\top \boldsymbol{W})} \cdot \gamma$$

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

The δ is bounded by [0,1]

$$\left|\left|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0}\right|\right| \leq \prod_{i=1}^{T} \left|\left|.\frac{\partial \boldsymbol{h}_i}{\partial \boldsymbol{h}_{i-1}}\right|\right| \leq \delta^T$$

As the $T \to \infty, \delta^T \to 0$. Therefore the gradients of the hidden state will vanish over time.

$$\left|\left|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0}\right|\right| \to 0 \text{ as } T \to \infty$$

..

3. If the inequalities are reversed i.e. $\lambda_1(\boldsymbol{W}^\top \boldsymbol{W}) > \frac{\delta^2}{\gamma^2}$ then the gradients would in fact explode instead of vanish. It is a necessary but not sufficient condition for the gradient to explode. The product of the norms can be greater than norm of product, so we cant say that it is a sufficient condition.

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

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.

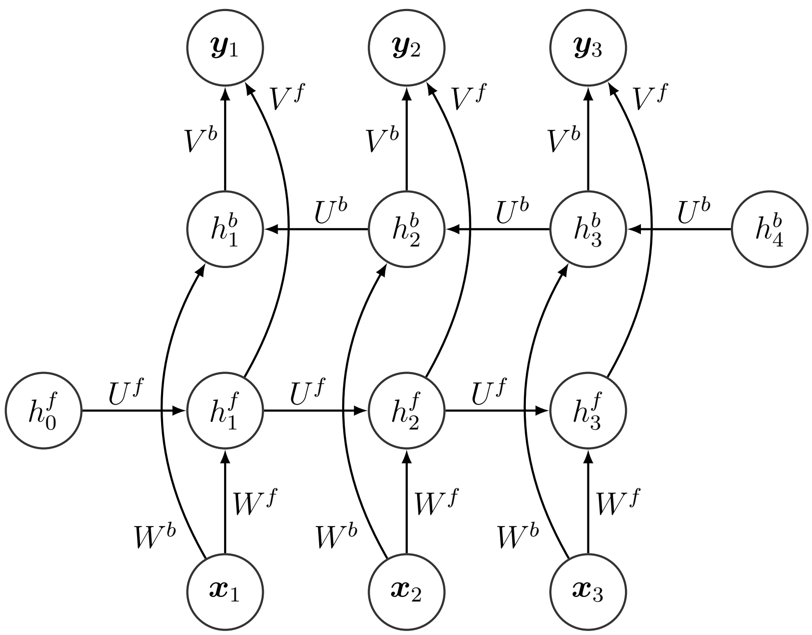

FIGURE 1 – Computational graph of the bidirectional RNN unrolled for three timesteps.

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. Labelled in the figure.
2.

$$
\begin{aligned}
\boldsymbol{h}_t^{(f)} &= \sigma(\boldsymbol{W}^{(f)}\boldsymbol{x}_t + \boldsymbol{U}^{(f)}\boldsymbol{h}_{t-1}^{(f)}) \\
\boldsymbol{h}_t^{(b)} &= \sigma(\boldsymbol{W}^{(b)}\boldsymbol{x}_t + \boldsymbol{U}^{(b)}\boldsymbol{h}_{t+1}^{(b)}) \\
\boldsymbol{y}_t &= \boldsymbol{V}^{(f)}\boldsymbol{h}_t^{(f)} + \boldsymbol{V}^{(b)}\boldsymbol{h}_t^{(b)} \\
L_t &= ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2 \\
\nabla_{\boldsymbol{h}_t^{(f)}} L_t &= \nabla_{\boldsymbol{y}_t} L_t . \nabla_{\boldsymbol{h}_t^{(f)}} \boldsymbol{y}_t \\
&= -2(\boldsymbol{z}_t - \boldsymbol{y}_t)\boldsymbol{V}^{(f)} \\
\nabla_{\boldsymbol{h}_t^{(b)}} L_t &= \nabla_{\boldsymbol{y}_t} L_t . \nabla_{\boldsymbol{h}_t^{(b)}} \boldsymbol{y}_t \\
&= -2(\boldsymbol{z}_t - \boldsymbol{y}_t)\boldsymbol{V}^{(b)} \\
\text{Using } \sigma'(x) &= \sigma'(x)(1-\sigma'(x)) \text{ in the below equations} \\
\boldsymbol{h}_{t+1}^{(f)} &= \sigma(\boldsymbol{W}^{(f)}\boldsymbol{x}_t + \boldsymbol{U}^{(f)}\boldsymbol{h}_t^{(f)}) \\
\boldsymbol{h}_{t-1}^{(b)} &= \sigma(\boldsymbol{W}^{(b)}\boldsymbol{x}_t + \boldsymbol{U}^{(b)}\boldsymbol{h}_t^{(b)}) \\
\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}} &= \boxed{diag\big(\boldsymbol{h}_{t+1}^{(f)}(1-\boldsymbol{h}_{t+1}^{(f)})\big)U^{(f)}} \\
\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}} &= \boxed{diag\big(\boldsymbol{h}_{t-1}^{(b)}(1-\boldsymbol{h}_{t-1}^{(b)})\big)U^{(b)}}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\nabla_{\boldsymbol{h}_t^{(f)}} L &= \nabla_{\boldsymbol{h}_t^{(f)}} L_t + \left(\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}}\right)^\top \nabla_{\boldsymbol{h}_{t+1}^{(f)}} L \\
&= \boxed{-2(\boldsymbol{z}_t - \boldsymbol{y}_t)\boldsymbol{V}^{(f)} + \Big(diag\big(\boldsymbol{h}_{t+1}^{(f)}(1-\boldsymbol{h}_{t+1}^{(f)})\big)U^{(f)}\Big)^\top \nabla_{\boldsymbol{h}_{t+1}^{(f)}} L}
\end{aligned}
$$

$$
\begin{aligned}
\nabla_{\boldsymbol{h}_t^{(b)}} L &= \nabla_{\boldsymbol{h}_t^{(b)}} L_t + \left(\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}}\right)^\top \nabla_{\boldsymbol{h}_{t-1}^{(b)}} L \\
&= \boxed{-2(\boldsymbol{z}_t - \boldsymbol{y}_t)\boldsymbol{V}^{(b)} + \Big(diag\big(\boldsymbol{h}_{t-1}^{(b)}(1-\boldsymbol{h}_{t-1}^{(b)})\big)U^{(b)}\Big)^\top \nabla_{\boldsymbol{h}_{t-1}^{(b)}} L}
\end{aligned}
$$

3.

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

$$\nabla_{\boldsymbol{W}^{(f)}} L = \sum_t \nabla_{\boldsymbol{W}_{(t)}^{(f)}} L$$

Here, $\nabla_{\boldsymbol{W}^{(f)}} L$ takes into account the consideration of $\boldsymbol{W}$ to function L due to all edges in computational graph. To resolve this ambiguity, we introduce dummy variables $\boldsymbol{W}_{(t)}$ that are defined to be the copies of $\boldsymbol{W}$ but with each $\boldsymbol{W}_{(t)}$ used to denote contribution at timestep t. So, used $\nabla_{\boldsymbol{W}_{(t)}}$ to denote contribution of weight at time step t to the gradient.

$$\frac{\partial L}{\partial \boldsymbol{W}_{(t)}^{(f)}} = \frac{\partial L}{\partial \boldsymbol{h}_{(t)}^{(f)}} \cdot \frac{\partial \boldsymbol{h}_{(t)}^{(f)}}{\partial \boldsymbol{W}_{(t)}^{(f)}}$$

$$\nabla_{\boldsymbol{W}_{(t)}^{(f)}} L = diag(\boldsymbol{h}_t^f(1-\boldsymbol{h}_t^f))\big(\nabla_{\boldsymbol{h}_t^{(f)}} L\big)\boldsymbol{x}_t^\top$$

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

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

$$\nabla_{\boldsymbol{U}^{(b)}} L = \sum_t \nabla_{\boldsymbol{U}_{(t)}^{(b)}} L$$

Similarly, using $\nabla_{\boldsymbol{U}_{(t)}}$ to denote contribution of weight at time step t to the gradient.

$$\frac{\partial L}{\partial \boldsymbol{U}_{(t)}^{(b)}} = \frac{\partial L}{\partial \boldsymbol{h}_{(t)}^{(b)}} \cdot \frac{\partial \boldsymbol{h}_{(t)}^{(b)}}{\partial \boldsymbol{U}_{(t)}^{(b)}}$$

$$\nabla_{\boldsymbol{U}_{(t)}^{(b)}\boldsymbol{h}_t^{(b)}} = diag\big(\boldsymbol{h}_t^b(1-\boldsymbol{h}_t^b)\big)\big(\nabla_{\boldsymbol{h}_t^{(b)}} L\big)\big(\boldsymbol{h}_{t+1}^{(b)}\big)^\top$$

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