Due Date: March 10th 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*
- *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. For SGD with momentum, we have ;

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

$$\begin{aligned}\Delta\boldsymbol{\theta}_t &= -\boldsymbol{v}_t \\ &= -(\alpha\boldsymbol{v}_{t-1} + \epsilon\boldsymbol{g}_t) \\ &= -\alpha\boldsymbol{v}_{t-1} - \epsilon\boldsymbol{g}_t \\ &= \alpha\Delta\boldsymbol{\theta}_{t-1} - \epsilon\boldsymbol{g}_t\end{aligned}$$

For SGD with running average,

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

Given

$$\alpha = \beta, \qquad \epsilon = \delta(1-\beta)$$

the two updates are equivalents.

2. By unrolling the average update rule, we have ;

$$\begin{aligned}\boldsymbol{v}_t &= \beta\boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t \\ &= \beta(\beta\boldsymbol{v}_{t-2} + (1-\beta)\boldsymbol{g}_{t-1}) + (1-\beta)\boldsymbol{g}_t \\ &= \beta^2\boldsymbol{v}_{t-2} + \beta(1-\beta)\boldsymbol{g}_{t-1} + (1-\beta)\boldsymbol{g}_t \\ &= \beta^2(\beta\boldsymbol{v}_{t-3} + (1-\beta)\boldsymbol{g}_{t-2}) + \beta(1-\beta)\boldsymbol{g}_{t-1} + (1-\beta)\boldsymbol{g}_t \\ &= \beta^3\boldsymbol{v}_{t-3} + \beta^2(1-\beta)\boldsymbol{g}_{t-2} + \beta(1-\beta)\boldsymbol{g}_{t-1} + (1-\beta)\boldsymbol{g}_t \\ &= \dots \\ &= \beta^t\boldsymbol{v}_0 + \sum_{i=1}^{t}\beta^{t-i}(1-\beta)\boldsymbol{g}_i \\ &= (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\boldsymbol{g}_i \quad (\text{since } v_0 = 0)\end{aligned}$$

3. Given $\boldsymbol{g}_t$ has a stationary distribution independent of t, we know from the previous subquestion that ;

$$\begin{aligned}\boldsymbol{v}_t &= (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\boldsymbol{g}_i \\ \mathbb{E}[\boldsymbol{v}_t] &= \mathbb{E}[(1-\beta)\sum_{i=1}^{t}\beta^{t-i}\boldsymbol{g}_i] \\ &= (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\mathbb{E}[\boldsymbol{g}_i] \quad (\text{linearity of expectation}) \\ &= \mathbb{E}[\boldsymbol{g}_t](1-\beta)\sum_{i=1}^{t}\beta^{t-i} \quad (\text{stationarity}) \\ &= \mathbb{E}[\boldsymbol{g}_t](1-\beta)\frac{1-\beta^t}{1-\beta} \quad (\text{sum of finite geometric series}) \\ &= \mathbb{E}[\boldsymbol{g}_t](1-\beta^t)\end{aligned}$$

Thus, we can observe that the running average is biased. If we want an unbiased estimate, we can rescale $\boldsymbol{v}_t$ by this bias, which would give us $\tilde{\boldsymbol{v}}_t = \boldsymbol{v}_t(1-\beta^t), \quad \Delta\boldsymbol{\theta}_t = -\delta\boldsymbol{v}_t$

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 2.2 satisfies

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

   where λ^{dropout} is a regularization coefficient depending on p. How does the value of p affect the regularization coefficient, λ^{dropout} ?
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 2.3 and 2.4: identify specific differences in the equations you arrive at, and discuss qualitatively what the equations tell you about the similarities and differences in the effects of weight decay and dropout (1-3 sentences).

Answer 2.

1. Given the hint, $\text{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2$, we can rearrange it as follows ; $\mathbb{E}[Z^2] = \mathbb{E}[Z]^2 + \text{Var}(Z)$

$$\begin{aligned}
\mathbb{E}_R[L(w)] &= \mathbb{E}_R[\|\boldsymbol{y} - (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w}\|^2] \\
&= \sum_{i=1}^{n} \mathbb{E}_R[(\boldsymbol{y}_i - (\boldsymbol{X}_i \odot \boldsymbol{R}_i)\boldsymbol{w})^2] \\
&= \sum_{i=1}^{n} (\mathbb{E}_R[\boldsymbol{y}_i - (\boldsymbol{X}_i \odot \boldsymbol{R}_i)\boldsymbol{w}])^2 + \text{Var}_R(\boldsymbol{y}_i - (\boldsymbol{X}_i \odot \boldsymbol{R}_i)\boldsymbol{w}) \\
&= \sum_{i=1}^{n} (\boldsymbol{y}_i - \mathbb{E}_R[\boldsymbol{R}_i] \odot \boldsymbol{X}_i)\boldsymbol{w})^2 + \text{Var}_R(\boldsymbol{y}_i - (\boldsymbol{X}_i \odot \boldsymbol{R}_i)\boldsymbol{w})
\end{aligned}$$

Deriving the first block and keeping in mind that the expected value of a Bernoulli random variable R is $\mathbb{E}(R) = \Pr(R=1)\cdot 1 + \Pr(R=0)\cdot 0 = p\cdot 1 + q\cdot 0 = p$ we get ;

$$\begin{aligned}
&\sum_{i=1}^{n} (\boldsymbol{y}_i - \mathbb{E}_R[\boldsymbol{R}_i] \odot \boldsymbol{X}_i)\boldsymbol{w})^2 \\
&= \sum_{i=1}^{n} (\boldsymbol{y}_i - p\boldsymbol{X}_i\boldsymbol{w})^2 \\
&= \|\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}\|^2
\end{aligned}$$

Now doing the same thing for the second block with $\mathbb{E}[R^2] = \Pr(R=1)\cdot 1^2 + \Pr(R=0)\cdot 0^2 = p\cdot 1^2 + q\cdot 0^2 = p$ we get ;

$$
\begin{aligned}
&\mathrm{Var}_R(\boldsymbol{y}_i - (\boldsymbol{X}_i \odot \boldsymbol{R}_i)\boldsymbol{w}) \\
&= \sum_{i=1}^{n} \mathbb{E}_R[(\boldsymbol{y}_i - (\boldsymbol{X}_i \odot \boldsymbol{R}_i)\boldsymbol{w} - \mathbb{E}_R[\boldsymbol{y}_i - (\boldsymbol{X}_i \odot \boldsymbol{R}_i)\boldsymbol{w}])^2] \\
&= \sum_{i=1}^{n} \mathbb{E}_R[(\boldsymbol{y}_i - (\boldsymbol{X}_i \odot \boldsymbol{R}_i)\boldsymbol{w} - \boldsymbol{y}_i + p\boldsymbol{X}_i\boldsymbol{w})^2] \\
&= \sum_{i=1}^{n} \mathbb{E}_R[((p\boldsymbol{X}_i - \boldsymbol{R}_i \odot \boldsymbol{X}_i)\boldsymbol{w})^2] \\
&= \sum_{i=1}^{n} \mathbb{E}_R[\boldsymbol{w}^T(p\boldsymbol{X}_i^T - \boldsymbol{R}_i^T \odot \boldsymbol{X}_i^T)(p\boldsymbol{X}_i - \boldsymbol{R}_i \odot \boldsymbol{X}_i)\boldsymbol{w}] \\
&= \sum_{i=1}^{n} (\boldsymbol{w}^T(p^2\boldsymbol{X}_i^T\boldsymbol{X}_i - p\mathbb{E}_R[\boldsymbol{R}_i]\boldsymbol{X}_i^T\boldsymbol{X}_i - p\mathbb{E}_R[\boldsymbol{R}_i^T]\boldsymbol{X}_i^T\boldsymbol{X}_i + \mathbb{E}[\boldsymbol{R}^T\boldsymbol{R}](\boldsymbol{X}_i^T\boldsymbol{X}))\boldsymbol{w}) \\
&= \sum_{i=1}^{n} (\boldsymbol{w}^T(p^2\boldsymbol{X}_i^T\boldsymbol{X}_i - p^2\boldsymbol{X}_i^T\boldsymbol{X}_i - p^2\boldsymbol{X}_i^T\boldsymbol{X}_i + p(\boldsymbol{X}_i^T\boldsymbol{X}_i))\boldsymbol{w}) \\
&= \sum_{i=1}^{n} (\boldsymbol{w}^T p(1-p)(\boldsymbol{X}_i^T\boldsymbol{X}_i)\boldsymbol{w}) \\
&= p(1-p)(\boldsymbol{w}^T \,\mathrm{diag}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{w}) \\
&= p(1-p)\|\Gamma\boldsymbol{w}\|^2
\end{aligned}
$$

Combining the two together, we get ;

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

2.

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

Setting the gradient to 0, we have ;

$$
\begin{aligned}
\nabla_w \mathbb{E}[L(\boldsymbol{w})] &= \frac{\partial}{\partial \boldsymbol{w}} ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma\boldsymbol{w}||^2 \\
&= \frac{\partial}{\partial \boldsymbol{w}} (\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}) + p(1-p)(\Gamma\boldsymbol{w})^T(\Gamma\boldsymbol{w}) \\
&= \frac{\partial}{\partial \boldsymbol{w}} (\boldsymbol{y}^T - p\boldsymbol{w}^T\boldsymbol{X}^T)(\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}) + p(1-p)(\boldsymbol{w}^T\Gamma^T\Gamma\boldsymbol{w}) \\
&= \frac{\partial}{\partial \boldsymbol{w}} \boldsymbol{y}^T\boldsymbol{y} - p\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} - p\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + p^2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + p(1-p)\boldsymbol{w}^T\Gamma^2\boldsymbol{w}
\end{aligned}
$$

Using the following formulas ;

$$\frac{\partial \boldsymbol{X}\boldsymbol{w}}{\partial \boldsymbol{w}} = \boldsymbol{X}^T, \quad \frac{\partial \boldsymbol{w}^T\boldsymbol{X}\boldsymbol{w}}{\partial \boldsymbol{w}} = (\boldsymbol{X} + \boldsymbol{X}^T)\boldsymbol{w}, \quad \frac{\partial \boldsymbol{w}^T\boldsymbol{X}}{\partial \boldsymbol{w}} = \boldsymbol{X}$$

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

Now equaling this to 0 ;

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

Letting $\lambda^{\text{dropout}} = \frac{1-p}{p}$ we get ;

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

When $p = 1$, $\lambda^{\text{dropout}} = 0$ so there is no contribution from the regularization term in $L(\boldsymbol{w})$. When p is close to 0, $\lambda^{\text{dropout}} \to \infty$, thus the contribution from the regularization increases.

3.

With L_2 regularization we can write the loss as follows ;

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

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{w}}L(\boldsymbol{w}) &= \frac{\partial}{\partial \boldsymbol{w}}(||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2 + \lambda^{L_2}||\boldsymbol{w}||^2) \\
&= \frac{\partial}{\partial \boldsymbol{w}}((\boldsymbol{y}^T - \boldsymbol{w}^T\boldsymbol{X}^T)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \lambda^{L_2}\boldsymbol{w}^T\boldsymbol{w} \\
&= \frac{\partial}{\partial \boldsymbol{w}}(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + \lambda^{L_2}\boldsymbol{w}^T\boldsymbol{w} \\
&= 0 - (\boldsymbol{y}^T\boldsymbol{X})^T - \boldsymbol{X}^T\boldsymbol{y} + (\boldsymbol{X}^T\boldsymbol{X} + (\boldsymbol{X}^T\boldsymbol{X})^T)\boldsymbol{w} + 2\lambda^{L_2}\boldsymbol{w} \\
&= -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + 2\lambda^{L_2}\boldsymbol{w}
\end{aligned}$$

Equaling this to 0, we find $\boldsymbol{w}^{L_2}$ to a minimal $L(\boldsymbol{w})$;

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

4. We can observe that dropout with linear regression is somewhat equivalent to a scaled version of ridge regression. In fact, in dropout, the penalty on the weight $\boldsymbol{w}_i$ is scaled by the standard deviation of the i'th feature. As for L^2, the weight is penalized uniformly. Also, with dropout, when the probability of dropping units is higher, the regularization coefficient grows, but when the probability decreases, we can see it as an ordinary least square problem.

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) Given that $X \perp Y$, we can use some useful properties in our calculations.

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

1. b) Keeping in mind that $\mathbb{E}[h_i^{(t-1)}] = 0$, $\text{Var}(h_i^{(t-1)}) = 1$ for $1 \leq i \leq d^{(t-1)}$ and that $X \perp Y$ we have;

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

As for $\text{Var}(\boldsymbol{a}_i^{(t)})$ we have;

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

1. c) We could initialize the parameter $c = 0$, such that $\mathbb{E}[\boldsymbol{a}_i^{(t)}] = 0$ and $\text{Var}(\boldsymbol{b}_i^{(t)}) = 0$ and let $\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] = 0$ and $\text{Var}(\boldsymbol{W}_{ij}^{(t)}) = \frac{1}{d^{(t-1)}}$ so that $\text{Var}(\boldsymbol{a}_i^{(t)}) = 1$. With the Gaussian weight distribution $W_{ij}^{(t)} \sim \mathcal{N}(\mu, \sigma^2)$,

$$\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] = \mu \quad \text{Var}(\boldsymbol{W}_{ij}^{(t)}) = \sigma^2$$

Thus,

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

2. a)

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{h}_i^{(t-1)})^2] &= \int_{-\infty}^{\infty} \max\{0, \boldsymbol{a}_i^{(t-1)}\}^2 p(\boldsymbol{a}_i^{(t-1)}) d\boldsymbol{a}_i^{(t-1)} \\
&= \int_{0}^{\infty} (\boldsymbol{a}_i^{(t-1)})^2 p(\boldsymbol{a}_i^{(t-1)}) d\boldsymbol{a}_i^{(t-1)} \\
&= \frac{1}{2} \int_{-\infty}^{\infty} (\boldsymbol{a}_i^{(t-1)})^2 p(\boldsymbol{a}_i^{(t-1)}) d\boldsymbol{a}_i^{(t-1)} \quad \text{since } \boldsymbol{a}_i^{(t-1)} \text{ is symmetric} \\
&= \frac{1}{2} \mathbb{E}[(\boldsymbol{a}_i^{(t-1)})^2] \\
&= \frac{1}{2} (\mathbb{E}[(\boldsymbol{a}_i^{(t-1)})^2] - \mathbb{E}[(\boldsymbol{a}_i^{(t-1)})]^2) \\
&= \frac{1}{2} \mathrm{Var}(\boldsymbol{a}_i^{(t-1)}) \\
&= \frac{1}{2} \quad \text{since } \mathrm{Var}(\boldsymbol{a}_i^{(t-1)}) = 1
\end{aligned}$$

2. b) Based on the previous questions we know ;

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

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

We could initialize the parameters to $c = 0$ and $\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] = 0$, so that $\mathbb{E}[\boldsymbol{a}_i^{(t)}] = 0, \quad \mathrm{Var}(\boldsymbol{b}_i^{(t)}) = 0$ and simplify $\mathrm{Var}(\boldsymbol{a}_i^{(t)})$.

$$\mathrm{Var}(\boldsymbol{a}_i^{(t)}) = \frac{d^{(t-1)}}{2} \mathrm{Var}(\boldsymbol{W}_{ij}^{(t)}) = 1$$

Thus,

$$\mathrm{Var}(\boldsymbol{W}_{ij}^{(t)}) = \frac{2}{d^{(t-1)}}$$

With the Gaussian weight distribution $W_{ij}^{(t)} \sim \mathcal{N}(\mu, \sigma^2)$,

$$\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] = \mu \quad \mathrm{Var}(\boldsymbol{W}_{ij}^{(t)}) = \sigma^2$$

Thus,

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

2. c) The He Initialization has this form (ReLU version of the Glorot initialization).

2. d) This initialization scheme would work well in practice since with the He Initialization, the scale of the activations is controlled, which decreases the probability of the gradient exploding or vanishing, thus enabling the model to train better.

3. For the assumptions in (1) and given α, β for $W_{ij}^{(t)} \sim Uniform(\alpha, \beta)$ such that $\mathbb{E}[a_i^{(t)}] = 0$ and $\text{Var}(a_i^{(t)}) = 1$,

$$\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] = \frac{\alpha + \beta}{2}, \quad \text{Var}(\boldsymbol{W}_{ij}^{(t)}) = \frac{(\beta - \alpha)^2}{12}$$

Therefore,

$$c = 0, \quad \beta = -\alpha > 0, \quad \beta = \sqrt{\frac{3}{d^{(t-1)}}}$$

For the assumptions in (2) and given α, β for $W_{ij}^{(t)} \sim Uniform(\alpha, \beta)$ such that $\mathbb{E}[a_i^{(t)}] = 0$ and $\text{Var}(a_i^{(t)}) = 1$,

$$\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] = \frac{\alpha + \beta}{2}, \quad \text{Var}(\boldsymbol{W}_{ij}^{(t)}) = \frac{(\beta - \alpha)^2}{12}$$

Thus,

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

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. Batch normalization in a the case of a convolutional neural network is different than batch normalization for usual networks since filter weights are shared across the input image. In that case, it is reasonable to normalize the output in the same way so that each output value takes the mean and variance of batch * features, at different location.

With that in mind,

$$\mu_{\text{batch}} = \begin{bmatrix}(1+3+2+3+3+2+4+2+2+3+3+2)/12\\(1+2+3+2+4+4+1+2+4+3+3+2)/12\end{bmatrix}$$

$$\mu_{\text{batch}} = \begin{bmatrix}2.5\\2.5833\end{bmatrix}$$

$$\sigma^2_{\text{batch}} = \begin{bmatrix}\frac{(1-2.5)^2+(3-2.5)^2+(2-2.5)^2+(3-2.5)^2+\ldots+(2-2.5)^2}{12}\\\frac{(1-2.5833)^2+(2-2.5833)^2+(3-2.5833)^2+\ldots+(2-2.5833)^2}{12}\end{bmatrix} = \begin{bmatrix}0.5833\\1.0764\end{bmatrix}$$

$$\mu_{\text{layer}} = \left[\begin{bmatrix}(1+3+2)/3\\(1+2+3)/3\end{bmatrix}, \begin{bmatrix}(3+3+2)/3\\(2+4+4)/3\end{bmatrix}, \begin{bmatrix}(4+2+2)/3\\(1+2+4)/3\end{bmatrix}, \begin{bmatrix}(3+3+2)/3\\(3+3+2)/3\end{bmatrix}\right] = \left[\begin{bmatrix}2\\2\end{bmatrix}, \begin{bmatrix}2.6667\\3.3333\end{bmatrix}, \begin{bmatrix}2.6667\\2.3333\end{bmatrix}, \begin{bmatrix}2.6667\\2.6667\end{bmatrix}\right]$$

$$\sigma^2_{\text{layer}} = \left[\begin{bmatrix}\frac{(1-2)^2+(3-2)^2+(2-2)^2}{3}\\\frac{(1-2)^2+(2-2)^2+(3-2)^2}{3}\end{bmatrix}, \begin{bmatrix}\ldots\\\ldots\end{bmatrix}, \begin{bmatrix}\ldots\\\ldots\end{bmatrix}, \begin{bmatrix}\frac{(3-2.6667)^2+(3-2.6667)^2+(2-2.6667)^2}{3}\\\frac{(3-2.6667)^2+(3-2.6667)^2+(2-2.6667)^2}{3}\end{bmatrix}\right] = \left[\begin{bmatrix}0.6666\\0.6666\end{bmatrix}, \begin{bmatrix}0.2222\\1.8889\end{bmatrix}, \begin{bmatrix}0.8889\\1.5556\end{bmatrix}, \begin{bmatrix}0.2222\\0.3334\end{bmatrix}\right]$$

$$\mu_{\text{instance}} = \left[\begin{bmatrix}2\\2\end{bmatrix}, \begin{bmatrix}2.6667\\3.3333\end{bmatrix}, \begin{bmatrix}2.6667\\2.3333\end{bmatrix}, \begin{bmatrix}2.6667\\2.6667\end{bmatrix}\right]$$

$$\sigma^2_{\text{instance}} = \left[\begin{bmatrix}0.6666\\0.6666\end{bmatrix}, \begin{bmatrix}0.2222\\1.8889\end{bmatrix}, \begin{bmatrix}0.8889\\1.5556\end{bmatrix}, \begin{bmatrix}0.2222\\0.3334\end{bmatrix}\right]$$

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.

2. Given the following definitions and properties ; $\boldsymbol{w} = \gamma\frac{\boldsymbol{u}}{||\boldsymbol{u}||}$, $\boldsymbol{y} = \boldsymbol{u}^T\boldsymbol{x}$,$\hat{\boldsymbol{y}} = \gamma\frac{\boldsymbol{y}-\mu_y}{\sigma_y}+\beta$, $\text{Var}(x) = \text{I}$, and $\mathbb{E}[x] = 0$.

We can derive the following ;

$$\begin{aligned}\mu_y &= \mathbb{E}[y] \\ &= \mathbb{E}[\boldsymbol{u}^T\boldsymbol{x}] \\ &= \boldsymbol{u}^T\mathbb{E}[\boldsymbol{x}] \\ &= 0\end{aligned}$$

$$\begin{aligned}\text{Var}(y) &= \text{Var}(\boldsymbol{u}^T\boldsymbol{x}) \\ &= \boldsymbol{u}^T\text{Var}(x)\boldsymbol{u} \\ &= \boldsymbol{u}^T\boldsymbol{u} \\ &= ||\boldsymbol{u}||^2 \\ &= \sigma_y \\ &= ||u||\end{aligned}$$

Thus,

$$\begin{aligned}\hat{\boldsymbol{y}} &= \gamma\frac{\boldsymbol{y}-\mu_y}{\sigma_y}+\beta \\ &= \gamma\frac{\boldsymbol{u}^T\boldsymbol{x}-0}{||\boldsymbol{u}||}+\beta \\ &= (\gamma\frac{\boldsymbol{u}^T}{||\boldsymbol{u}||})\boldsymbol{x}+\beta \\ &= \boldsymbol{w}^T\boldsymbol{x}+\beta\end{aligned}$$

3. First off, $\nabla_u L(\boldsymbol{u}) = \nabla_u \boldsymbol{w}\nabla_w L = \nabla_u(\gamma\frac{\boldsymbol{u}}{||\boldsymbol{u}||})\nabla_w L = \gamma(\nabla_u\frac{\boldsymbol{u}}{||\boldsymbol{u}||})\nabla_w L$

Using the quotient rule and the previous definitions, we get:

$$\begin{aligned}\nabla_u\frac{\boldsymbol{u}}{||\boldsymbol{u}||} &= \frac{||\boldsymbol{u}||\frac{d\boldsymbol{u}}{d\boldsymbol{u}} - \boldsymbol{u}\frac{d||\boldsymbol{u}||}{u}}{||\boldsymbol{u}||^2} \\ &= \frac{||\boldsymbol{u}||\boldsymbol{I} - \boldsymbol{u}\frac{d}{d\boldsymbol{u}}((\boldsymbol{u}^T\boldsymbol{u})^{\frac{1}{2}})}{||\boldsymbol{u}||^2} \\ &= \frac{1}{||\boldsymbol{u}||^2}(||\boldsymbol{u}||\boldsymbol{I} - \boldsymbol{u}\frac{1}{2}(\boldsymbol{u}^T\boldsymbol{u})^{-\frac{1}{2}}\frac{d(\boldsymbol{u}^T\boldsymbol{u})}{du} \\ &= \frac{1}{||\boldsymbol{u}||^2}(||\boldsymbol{u}||\boldsymbol{I} - \boldsymbol{u}\frac{1}{2||\boldsymbol{u}||}2\boldsymbol{u}^T) \\ &= \frac{1}{||\boldsymbol{u}||^2}(||\boldsymbol{u}||\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^T}{||\boldsymbol{u}||}) \\ &= \frac{1}{||\boldsymbol{u}||}(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^T}{||\boldsymbol{u}||^2}) \\ &= \frac{1}{||\boldsymbol{u}||}\boldsymbol{W}^{\perp}\end{aligned}$$

With that in mind,

$$\nabla_u L(\boldsymbol{u}) = \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{u}} \nabla_w L = \frac{\gamma}{||\boldsymbol{u}||}(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^T}{||\boldsymbol{u}||^2})\nabla_w L = s\boldsymbol{W}^{\perp}\nabla_w L$$

where $s = \frac{\gamma}{||\boldsymbol{u}||}$ and $\boldsymbol{W}^{\perp} = (\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^T}{||\boldsymbol{u}||^2})$. Also, note that $\boldsymbol{W}^{\perp}\boldsymbol{u} = \boldsymbol{u} - \boldsymbol{u}\frac{\boldsymbol{u}\boldsymbol{u}^T}{||\boldsymbol{u}||^2} = \boldsymbol{u} - \boldsymbol{u} = 0$.

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. By observing the two functions, we can see that ;

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

This lets us see that $\boldsymbol{g}_t = \sigma(\boldsymbol{h}_t)$, for each $t \geq 0$. Assuming that it is also true that $\boldsymbol{g}_{t-1} = \sigma(\boldsymbol{h}_{t-1})$ and by fixing $t \geq 1$ then we can prove that $\boldsymbol{g}_t = \sigma(\boldsymbol{h}_t)$.

We have ;

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

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.

2.

$$\left\|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0}\right\| = \left\|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_{T-1}}\right\| \cdot \left\|\frac{\partial \boldsymbol{h}_{T-1}}{\partial \boldsymbol{h}_{T-2}}\right\| \cdots \left\|\frac{\partial \boldsymbol{h}_1}{\partial \boldsymbol{h}_0}\right\|$$

With the two properties given,

$$\begin{aligned}
\left\|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}}\right\| &= \left\|\boldsymbol{W}\frac{\partial \boldsymbol{h}_{t-1}}{\partial \boldsymbol{h}_{t-1}}\right\| = \|\boldsymbol{W}\sigma'\| \\
&\leq \|\boldsymbol{W}\|\,\|\sigma'\| \quad (\text{Since } ||\boldsymbol{AB}|| \leq ||\boldsymbol{A}||\,||\boldsymbol{B}||) \\
&\leq \|\boldsymbol{W}\| \cdot \gamma \quad (\text{Since } |\sigma'| \leq \gamma) \\
&= \sqrt{\lambda_1(\boldsymbol{W}^\top \boldsymbol{W})} \cdot \gamma \quad (\text{Since } ||\boldsymbol{A}|| = \sqrt{\lambda_1(\boldsymbol{A}^\top \boldsymbol{A})}\ ) \\
&\leq \sqrt{\frac{\delta^2}{\gamma^2}} \cdot \gamma \quad (\text{Since } \lambda_1(\boldsymbol{W}^\top \boldsymbol{W}) \leq \frac{\delta^2}{\gamma^2}) \\
&= |\frac{\delta}{\gamma}| \cdot \gamma \\
&= \frac{\delta}{\gamma} \cdot \gamma \quad (\text{Since } \delta > 0, \quad \gamma > 0) \\
&= \delta
\end{aligned}$$

Thus,

$$\left\|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}}\right\| \leq \delta \implies \left\|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0}\right\| \leq \delta^T \implies \left\|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0}\right\| \to 0 \quad \text{as T} \to \infty$$

3. The largest eigenvalue of the weights being larger than $\frac{\delta^2}{\gamma^2}$ is a necessary condition because otherwise the gradient of the hidden state would not become arbitrarily large. However, it is not sufficient because the product of the norms can be greater than the norm of the product.

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.

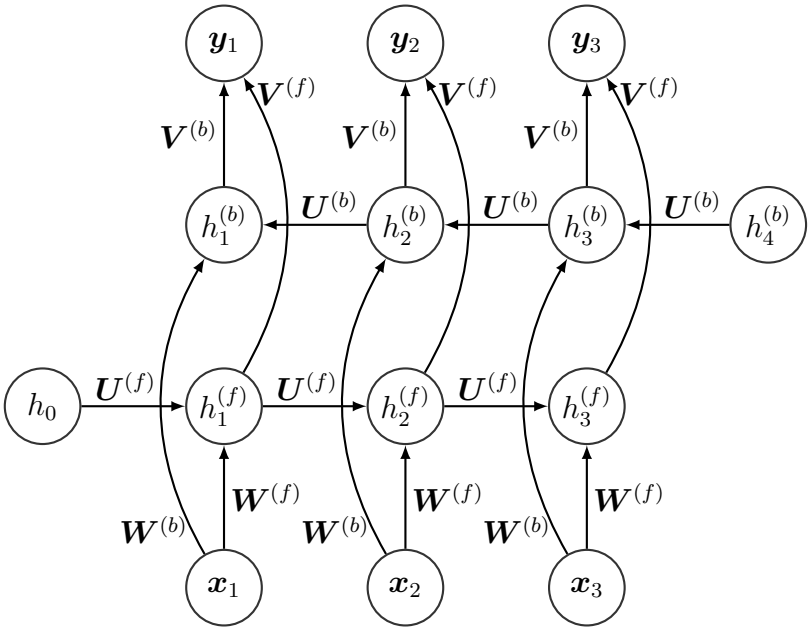

FIGURE 1 – Computational graph of the bidirectional RNN unrolled for three timesteps.

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.

2. The gradient of L_t with respect to $\boldsymbol{h}_t^{(f)}$ is

$$\nabla_{h_t^{(f)}} L_t = \left(\frac{\partial \boldsymbol{y}_t}{\partial \boldsymbol{h}_t^{(f)}}\right)^T \nabla_{y_t} L_t$$

$$\begin{aligned}
\nabla_{y_t} L_t &= \frac{\partial ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2}{\partial \boldsymbol{y}_t} \\
&= \frac{\partial (\boldsymbol{z}_t - \boldsymbol{y}_t)^T (\boldsymbol{z}_t - \boldsymbol{y}_t)}{\partial \boldsymbol{y}_t} \\
&= \frac{\partial (\boldsymbol{z}_t^T - \boldsymbol{y}_t^T)(\boldsymbol{z}_t - \boldsymbol{y}_t)}{\partial \boldsymbol{y}_t} \\
&= \frac{\partial}{\partial \boldsymbol{y}_t} (\boldsymbol{z}_t^T \boldsymbol{z}_t - \boldsymbol{z}_t^T \boldsymbol{y}_t - \boldsymbol{y}_t^T \boldsymbol{z}_t + \boldsymbol{y}_t^T \boldsymbol{y}_t) \\
&= -\boldsymbol{z}_t^T - \boldsymbol{z}_t + 2\boldsymbol{y}_t \\
&= -2(\boldsymbol{z}_t - \boldsymbol{y}_t)
\end{aligned}$$

$$\begin{aligned}
\left(\frac{\partial \boldsymbol{y}_t}{\partial \boldsymbol{h}_t^{(f)}} \right)^T &= \left(\partial \boldsymbol{h}_t^{(f)} (\boldsymbol{V}^{(f)} \boldsymbol{h}_t^{(f)} + \boldsymbol{V}^{(b)} \boldsymbol{h}_t^{(b)}) \right)^T \\
&= \boldsymbol{V}^{(f)\top}
\end{aligned}$$

Combining the two together, we get ;

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

Now for the gradient of L_t with respect to $\boldsymbol{h}_t^{(b)}$ is $\nabla_{h_t^{(b)}} L_t = \left(\frac{\partial \boldsymbol{y}_t}{\partial \boldsymbol{h}_t^{(b)}} \right)^T \nabla_{y_t} L_t$.

$$\begin{aligned}
\left(\frac{\partial \boldsymbol{y}_t}{\partial \boldsymbol{h}_t^{(b)}} \right)^T &= \left(\partial \boldsymbol{h}_t^{(b)} (\boldsymbol{V}^{(f)} \boldsymbol{h}_t^{(f)} + \boldsymbol{V}^{(b)} \boldsymbol{h}_t^{(b)}) \right)^T \\
&= \boldsymbol{V}^{(b)\top}
\end{aligned}$$

Reusing what we already computed and combining, we get ;

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

Now for $\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}}$ we have ;

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

$$\begin{aligned}
\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}} &= \frac{\partial (\sigma(\boldsymbol{W}^{(f)} \boldsymbol{x}_{t+1} + \boldsymbol{U}^{(f)} \boldsymbol{h}_t^{(f)}))}{\partial \boldsymbol{h}_t^{(f)}} \\
&= \sigma(\boldsymbol{W}^{(f)} \boldsymbol{x}_{t+1} + \boldsymbol{U}^{(f)} \boldsymbol{h}_t^{(f)})(1 - (\sigma(\boldsymbol{W}^{(f)} \boldsymbol{x}_{t+1} + \boldsymbol{U}^{(f)} \boldsymbol{h}_t^{(f)})) \cdot \boldsymbol{U}^{(f)} \\
&\quad \text{Given that} \nabla_a \sigma(a) = \sigma(a)(1 - \sigma(a)) \text{ and then applying the chain rule.} \\
&= \text{diag}(\boldsymbol{h}_{t+1}^{(f)} (1 - \boldsymbol{h}_{t+1}^{(f)})) \boldsymbol{U}^{(f)}
\end{aligned}$$

Similarly, for $\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}}$ we get ;

$$\begin{aligned}\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}} &= \frac{\partial(\sigma(\boldsymbol{W}^{(b)}\boldsymbol{x}_{t-1} + \boldsymbol{U}^{(b)}\boldsymbol{h}_t^{(b)}))}{\partial \boldsymbol{h}_t^{(b)}} \\ &= \sigma(\boldsymbol{W}^{(b)}\boldsymbol{x}_{t-1} + \boldsymbol{U}^{(b)}\boldsymbol{h}_t^{(b)})(1 - (\sigma(\boldsymbol{W}^{(b)}\boldsymbol{x}_{t-1} + \boldsymbol{U}^{(b)}\boldsymbol{h}_t^{(b)})) \cdot \boldsymbol{U}^{(b)} \\ &\quad \text{Given that} \nabla_a \sigma(a) = \sigma(a)(1 - \sigma(a)) \text{ and then applying the chain rule.} \\ &= \text{diag}(\boldsymbol{h}_{t-1}^{(b)}(1 - \boldsymbol{h}_{t-1}^{(b)}))\boldsymbol{U}^{(b)}\end{aligned}$$

3. As suggested in the question, let t denote the contribution of weight matrices when computing the recurrent hidden unit, i.e, $\boldsymbol{W}_t^{(f)}$ and $\boldsymbol{U}_t^{(f)}$ we have ;

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

Similarly, we have ;

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