Due Date: March 22nd 23:59, 2019

Name: Aditya Joshi
Matricule:

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 David Krueger, Tegan Maharaj, and Chin-Wei Huang.***

Question 1 (6-10)**.** 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 preactvations 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 either (a) a Gaussian distribution $\boldsymbol{W}_{ij}^{(t)} \sim \mathcal{N}(\mu, \sigma^2)$, or (b) a Uniform distribution $\boldsymbol{W}_{ij}^{(t)} \sim U(\alpha, \beta)$.

For both of the assumptions (1 and 2) about the distribution of the inputs to layer t listed below, and for both (a) Gaussian, and (b) Uniform sampling, design an initialization scheme that would achieve preactivations with zero-mean and unit variance at layer t, i.e.: $\mathbb{E}[\boldsymbol{a}_i^{(t)}] = 0$ and $\text{Var}(\boldsymbol{a}_i^{(t)}) = 1$, for $1 \leq i \leq d^{(t)}$.

(Hint: if $X \perp Y$, $\text{Var}(XY) = \text{Var}(X)\text{Var}(Y) + \text{Var}(X)\mathbb{E}[Y]^2 + \text{Var}(Y)\mathbb{E}[X]^2$)

1. Assume $\mathbb{E}[\boldsymbol{h}_i^{(t-1)}] = 0$ and $\text{Var}(\boldsymbol{h}_i^{(t-1)}) = 1$ for $1 \leq i \leq d^{(t-1)}$. Assume entries of $\boldsymbol{h}^{(t-1)}$ are uncorrelated (the answer should not depend on g).
   (a) Gaussian: give the values for c, μ, and σ^2 as a function of $d^{(t-1)}$.

   Consider a single entry in the preactivation $\boldsymbol{a}_i^{(t)}$. This equals

$$\begin{aligned}
\boldsymbol{a}_i^{(t)} &= \sum_{j=1}^{d^{(t-1)}} \boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)} + \boldsymbol{b}_i^{(t)} \\
\mathbb{E}[\boldsymbol{a}_i^{(t)}] &= \mathbb{E}\left[\sum_{j=1}^{d^{(t-1)}} \boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)} + \boldsymbol{b}_i^{(t)} \right] \\
\mathbb{E}[\boldsymbol{a}_i^{(t)}] &= \sum_{j=1}^{d^{(t-1)}} \mathbb{E}\big[\boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)}\big] + \mathbb{E}[\boldsymbol{b}_i^{(t)}] \\
0 &= \sum_{j=1}^{d^{(t-1)}} \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] \cdot \mathbb{E}[\boldsymbol{h}_j^{(t-1)}] + \mathbb{E}[\boldsymbol{b}_i^{(t)}] \\
0 &= \mu \cdot 0 + c \\
c &= 0
\end{aligned}$$

   Therefore we have $c = 0$ and value of μ does not matter to achieve $\mathbb{E}[\boldsymbol{a}_i^{(t)}] = 0$.

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

   Thus, if we set $\mu = 0$ we find that the variance must equal $\sigma^2 = \dfrac{1}{d^{(t-1)}}$. The initialization scheme is then $\boxed{c = 0, \mu = 0, \sigma^2 = \dfrac{1}{d^{(t-1)}}}$

(b) Uniform: give the values for c, α, and β as a function of $d^{(t-1)}$.
We consider the same equation as part (a).

$$\begin{aligned}
\boldsymbol{a}_i^{(t)} &= \sum_{j=1}^{d^{(t-1)}} \boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)} + \boldsymbol{b}_i^{(t)} \\
0 &= \sum_{j=1}^{d^{(t-1)}} \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] \cdot \mathbb{E}[\boldsymbol{h}_j^{(t-1)}] + \mathbb{E}[\boldsymbol{b}_i^{(t)}] \\
&= \sum_{j=1}^{d^{(t-1)}} \left(\frac{\alpha+\beta}{2}\right) \cdot 0 + c \\
c &= 0
\end{aligned}$$

For the variance, we have

$$\mathrm{Var}[\boldsymbol{a}_i^{(t)}] = \sum_{j=1}^{d^{(t-1)}} \mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}] \cdot \mathrm{Var}[\boldsymbol{h}_j^{(t-1)}] + \mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}] \cdot (\mathbb{E}[\boldsymbol{h}_j^{(t-1)}])^2 + \mathrm{Var}[\boldsymbol{h}_j^{(t-1)}] \cdot (\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}])^2 + 0 \tag{1}$$

$$= \sum_{j=1}^{d^{(t-1)}} \left(\frac{(\beta-\alpha)^2}{12} \cdot 1 + 0 + \frac{\alpha+\beta}{2}\right) \tag{2}$$

$$1 = d^{(t-1)} \cdot \frac{(\beta-\alpha)^2}{12} + \frac{\alpha+\beta}{2} \tag{3}$$

If we have $\beta =$ constant and $\beta =$ -constant, then $\dfrac{\alpha+\beta}{2}$ will equal 0. Then we just have to find the appropriate value of the constant. With some experimentation, we realize that if we set $\beta = \sqrt{\dfrac{3}{d^{(t-1)}}}$ and $\alpha = -\sqrt{\dfrac{3}{d^{(t-1)}}}$, we have:

$$1 = d^{(t-1)} \cdot \frac{\left(\frac{\sqrt{3}}{\sqrt{d^{(t-1)}}} + \frac{\sqrt{3}}{\sqrt{d^{(t-1)}}}\right)^2}{12} \tag{4}$$

$$\tag{5}$$

$$1 = d^{(t-1)} \cdot \frac{\left(\frac{2\sqrt{3}}{\sqrt{d^{(t-1)}}}\right)^2}{12} \tag{6}$$

$$\tag{7}$$

$$1 = \cancel{d^{(t-1)}} \cdot \frac{\cancel{12}}{\cancel{d^{(t-1)}} \cdot \cancel{12}} \tag{8}$$

Thus, we can choose the following values $\boxed{c = 0, \alpha = -\sqrt{\dfrac{3}{d^{(t-1)}}}, \beta = \sqrt{\dfrac{3}{d^{(t-1)}}}}$

2. Assume that the preactivations of the previous layer satisfy $\mathbb{E}[\boldsymbol{a}_i^{(t-1)}] = 0$, $\mathrm{Var}(\boldsymbol{a}_i^{(t-1)}) = 1$ and $\boldsymbol{a}_i^{(t-1)}$ 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) Gaussian: give the values for c, μ, and σ^2 as a function of $d^{(t-1)}$.

$$\boldsymbol{a}_i^{(t)} = \sum_{j=1}^{d^{(t-1)}} \boldsymbol{W}_{ij}^{(t)} g(\boldsymbol{a}_j^{(t-1)}) + \boldsymbol{b}_i^{(t)} \tag{9}$$

$$\mathbb{E}[\boldsymbol{a}_i^{(t)}] = \sum_{j=1}^{d^{(t-1)}} \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]\mathbb{E}[g(\boldsymbol{a}_j^{(t-1)})] + \mathbb{E}[\boldsymbol{b}_i^{(t)}] \tag{10}$$

We want to ensure that $\mathbb{E}[\boldsymbol{a}_i^{(t)}] = 0$. Note that $E[g(\boldsymbol{a}_j^{(t-1)})]$ may not equal zero. If we sample $\boldsymbol{W}_{ij}^{(t)}$ from a gaussian with zero mean, the first product will be 0 as $\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] = 0$. We can also initialize $\boldsymbol{b}_i^{(t)} = c = 0$.
Let us now compute the variance. (Note that $\mathrm{Var}[XY] = \mathbb{E}[X^2Y^2] - (\mathbb{E}[XY])^2$).
In the following steps we also assume that X and Y are independent, and thus, $E[XY] = E[X] \cdot E[Y]$.

$$\mathrm{Var}[\boldsymbol{a}_i^{(t)}] = \sum_{j=1}^{d^{(t-1)}} \mathrm{Var}[\boldsymbol{W}_{ij}^{(t)} g(\boldsymbol{a}_j^{(t-1)})] + \mathrm{Var}[\boldsymbol{b}_i^{(t)}] \tag{11}$$

$$1 = \sum_{j=1}^{d^{(t-1)}} \overbrace{\mathbb{E}[(\boldsymbol{W}_{ij}^{(t)})^2]}^{\mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}]+\overbrace{(\mathbb{E}[(\boldsymbol{W}_{ij}^{(t)})])^2}^{0}=\mathbb{E}[(\boldsymbol{W}_{ij}^{(t)})^2]} \mathbb{E}[g(\boldsymbol{a}_j^{(t-1)})^2] - (\underbrace{\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]}_{0} \cdot \mathbb{E}[g(\boldsymbol{a}_j^{(t-1)})])^2 \tag{12}$$

$$= d^{(t-1)} \cdot \underbrace{\mathrm{Var}[\boldsymbol{W}_{ij}^{(t)}]}_{\sigma^2} \cdot \mathbb{E}[g(\boldsymbol{a}_j^{(t-1)})^2] \tag{13}$$

If we compute the value of $\mathbb{E}[g(\boldsymbol{a}_j^{(t-1)})^2]$, we can find an appropriate value for σ^2.
We can compute $\mathbb{E}[g(\boldsymbol{a}_j^{(t-1)})^2]$ as follows:

$$E[g(\boldsymbol{a}_j^{(t-1)})^2] = \int_{-\infty}^{\infty} g(\boldsymbol{a}_j^{(t-1)})^2 f(\boldsymbol{a}_j^{(t-1)}) da \tag{14}$$

$$= \int_{-\infty}^{0} g(\boldsymbol{a}_j^{(t-1)})^2 f(\boldsymbol{a}_j^{(t-1)}) da + \int_{0}^{\infty} g(\boldsymbol{a}_j^{(t-1)})^2 f(\boldsymbol{a}_j^{(t-1)}) da \tag{15}$$

$$= \int_{0}^{\infty} (\boldsymbol{a}_j^{(t-1)})^2 f(\boldsymbol{a}_j^{(t-1)}) da = \frac{1}{2} \underbrace{\mathrm{Var}[\boldsymbol{a}_j^{(t-1)}]}_{1} \tag{16}$$

Therefore we have:

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

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

Thus we can choose the values $\boxed{c = 0, \mu = 0, \sigma^2 = \frac{2}{d^{(t-1)}}}$

(b) Uniform: give the values for c, α, and β as a function of $d^{(t-1)}$.

Using the same reasoning as in 10, we can have $c = 0$. The values of α and β should be $-$constant and constant so that their mean is 0. We find the exact values using the same reasoning we had in 13.

$$\text{Var}[\boldsymbol{a}_i^{(t)}] = d^{(t-1)} \cdot \underbrace{\text{Var}[\boldsymbol{W}_{ij}^{(t)}]}_{\sigma^2} \cdot \underbrace{\mathbb{E}[g(\boldsymbol{a}_j^{(t-1)})^2]}_{\frac{1}{2}\text{Var}[\boldsymbol{a}_j^{(t-1)}]} \tag{17}$$

$$1 = d^{(t-1)} \cdot \frac{(\beta - \alpha)^2}{12} \cdot \frac{1}{2} \tag{18}$$

After some experimentation like in equation 3, we can set the value of $\alpha = -\sqrt{\frac{6}{d^{(t-1)}}}$ and $\beta = \sqrt{\frac{6}{d^{(t-1)}}}$.

Thus we have $\boxed{c = 0, \alpha = -\sqrt{\frac{6}{d^{(t-1)}}}, \beta = \sqrt{\frac{6}{d^{(t-1)}}}}$

(c) What popular initialization scheme has this form ?

The **Glorot** [1] and **Kaiming He** [2] initialization schemes have this form.

(d) Why do you think this initialization would work well in practice ? Answer in 1-2 sentences.

In the case of using sigmoid or tanh activation function (as was originally analyzed in *Understanding the difficulty of training deep feedforward neural networks* [1]), the gradients of the above mentioned activation functions saturate very quickly, that is, the gradients are very close to zero beyond a certain range. This prevents any gradients from flowing backwards thus preventing any learning. Maintaining the activations around 0 with unit variance helps the flow of a gradient.

In general, we can say that maintaining the values around a certain range (close to 0) helps to keep the values (and their gradients) from exploding or vanishing.

Question 2 (4-6-4-4-3)**.** 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.

1. For squared error loss, express the loss function $L(\boldsymbol{w})$ in matrix form (in terms of $\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w}$, and $\boldsymbol{R}$).

   The dropout mask $\boldsymbol{R}$ is sampled from a Bernoulli distribution. It will have an entry of 1 if the element is to be kept (with probability p) and 0, if we drop it (with probability $(1-p)$. Thus, in matrix form, the resultant input data after applying the dropout is simply the element wise (*hadamard*) product of the two matrices ($\tilde{\boldsymbol{X}} = \boldsymbol{X} \odot \boldsymbol{R}$). Thus, the final equation for the loss is as follows.

$$L(\boldsymbol{w}) = ||\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{w}||^2 \tag{19}$$
$$= (\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{w})^T(\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{w}) \tag{20}$$

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

   Consider the loss function from equation 20,

$$\begin{aligned} L(\boldsymbol{w}) &= (\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{w})^T(\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{w}) \\ &= (\boldsymbol{y}^T - \boldsymbol{w}^T\tilde{\boldsymbol{X}}^T)(\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{w}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\tilde{\boldsymbol{X}}\boldsymbol{w} - \boldsymbol{w}^T\tilde{\boldsymbol{X}}^T\boldsymbol{y} + \boldsymbol{w}^T\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\boldsymbol{w} \end{aligned}$$

   Note that $\boldsymbol{y}^T\tilde{\boldsymbol{X}}\boldsymbol{w}$ is a scalar, and for a scalar $a^T = a$. Thus we have $(\boldsymbol{y}^T\tilde{\boldsymbol{X}}\boldsymbol{w})^T = \boldsymbol{w}^T\tilde{\boldsymbol{X}}^T\boldsymbol{y}$. Substituting this in the equation above we get (using properties of linearity of expectation):

$$L(\boldsymbol{w}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^T\tilde{\boldsymbol{X}}^T\boldsymbol{y} + \boldsymbol{w}^T\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\boldsymbol{w} \tag{21}$$
$$\mathbb{E}_{\boldsymbol{R}}[L(\boldsymbol{w})] = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^T\mathbb{E}[\tilde{\boldsymbol{X}}^T]\boldsymbol{y} + \boldsymbol{w}^T\mathbb{E}[\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}]\boldsymbol{w} \tag{22}$$
$$= \boldsymbol{y}^T\boldsymbol{y} - 2p\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{w}^T\mathbb{E}[\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}]\boldsymbol{w} \tag{23}$$

   Let us calculate the value of $\mathbb{E}[\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}]$. $\tilde{\boldsymbol{X}}^T$ and $\tilde{\boldsymbol{X}}$ are not independent. Thus, we can use the property that $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y] + \text{Cov}(X, Y)$.

$$\mathbb{E}[\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}] = \mathbb{E}[\tilde{\boldsymbol{X}}^T]\mathbb{E}[\tilde{\boldsymbol{X}}] + \text{Cov}(\tilde{\boldsymbol{X}}^T, \tilde{\boldsymbol{X}}) \tag{24}$$
$$= p^2\boldsymbol{X}^T\boldsymbol{X} + p(1-p)\Gamma \tag{25}$$

   Substituting 25 in 23,

$$= \boldsymbol{y}^T\boldsymbol{y} - 2p\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{w}^T(p^2\boldsymbol{X}^T\boldsymbol{X} + p(1-p)\Gamma)\boldsymbol{w} \tag{26}$$
$$= \boldsymbol{y}^T\boldsymbol{y} - 2p\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + p^2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^T p(1-p)\Gamma\boldsymbol{w} \tag{27}$$
$$= ||\boldsymbol{y} - 2p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma\boldsymbol{w}||^2 \tag{28}$$

3. 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} ?

Consider the equation in 27. Let us differentiate that with respect to $\boldsymbol{w}^T$.

$$\begin{aligned}
\frac{\partial L(\boldsymbol{w})}{\boldsymbol{w}^T} &= -2p\boldsymbol{X}^T\boldsymbol{y} + 2p^2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + 2\cdot\frac{p^2(1-p)}{p}\Gamma^2\boldsymbol{w} \\
0 &= -\boldsymbol{X}^T\boldsymbol{y} + p\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + \frac{p\cdot(1-p)}{p}\Gamma^2\boldsymbol{w} \\
\boldsymbol{X}^T\boldsymbol{y} &= p\boldsymbol{w}\left(\boldsymbol{X}^T\boldsymbol{X} + \frac{(1-p)}{p}\Gamma^2\right) \\
p\boldsymbol{w} &= \left(\boldsymbol{X}^T\boldsymbol{X} + \frac{(1-p)}{p}\Gamma^2\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \\
p\boldsymbol{w}^{\text{dropout}} &= \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda^{\text{dropout}}\Gamma^2\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}
\end{aligned}$$

Effect of p : Here λ^{dropout} is a regularization coefficient that depends on p. For values of p close to 1, the Γ^2 term disappears, and there is no regularization term. That is, as $p \to 1$, $\lambda^{\text{dropout}} \to 0$. As $p \to 0$, λ^{dropout} gets arbitrarily large ($\to \infty$) and the regularization increases.

4. Express the solution $\boldsymbol{w}^{L_2}$ for a linear regression problem without dropout and with L^2 regularization, with regularization coefficient λ^{L_2} in closed form.

Linear regression with L_2 regularization is often called *ridge regression*. In this form of regression, we add an additional term $||\boldsymbol{w}||^2$ to the cost function to penalize the weights from growing too large. The closed from solution for this is as follows:

$$\begin{aligned}
L(\boldsymbol{w}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \lambda^{L_2}||\boldsymbol{w}||^2 \\
&= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + \lambda^{L_2}||\boldsymbol{w}||^2 \\
\frac{\partial L(\boldsymbol{w})}{\boldsymbol{w}} &= -2\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + 2\lambda^{L_2}\boldsymbol{w} \\
\boldsymbol{X}^T\boldsymbol{y} &= \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + \lambda^{L_2}\boldsymbol{w} \\
\boldsymbol{w}^{L_2} &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda^{L_2}\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}
\end{aligned}$$

5. 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).

We see that the two equations are identical except that in the dropout case, we have $\lambda^{\text{dropout}}\Gamma$ and for L2 regularization we have $\lambda^{L_2}\boldsymbol{I}$. One difference that we observe is that the L_2 regularization is independent of the the input features $\boldsymbol{X}$. In the dropout case, it is a function of the input through the existence of Γ^2.

Question 3 (5-5-5)**.** 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).

   - SGD with momentum:
     $$\Delta\boldsymbol{\theta}_t = -\boldsymbol{v}_t \qquad \boldsymbol{v}_t = -\Delta\boldsymbol{\theta}_t \tag{29}$$
     $$\Delta\boldsymbol{\theta}_{t-1} = -\boldsymbol{v}_{t-1} \qquad \boldsymbol{v}_{t-1} = -\Delta\boldsymbol{\theta}_{t-1} \tag{30}$$

     Now let us write $\boldsymbol{v}_t$ as a function of $\Delta\boldsymbol{\theta}$.

     $$\boldsymbol{v}_t = \alpha\boldsymbol{v}_{t-1} + \epsilon\boldsymbol{g}_t \tag{31}$$
     $$-\Delta\boldsymbol{\theta}_t = -\alpha\Delta\boldsymbol{\theta}_{t-1} + \epsilon\boldsymbol{g}_t \tag{32}$$
     $$\Delta\boldsymbol{\theta}_t = \alpha\Delta\boldsymbol{\theta}_{t-1} - \epsilon\boldsymbol{g}_t \tag{33}$$

   - SGD with running average of $\boldsymbol{g}_t$:
     $$\Delta\boldsymbol{\theta}_t = -\delta\boldsymbol{v}_t \qquad \boldsymbol{v}_t = \frac{-\Delta\boldsymbol{\theta}_t}{\delta} \tag{34}$$
     $$\Delta\boldsymbol{\theta}_{t-1} = -\delta\boldsymbol{v}_{t-1} \quad \boldsymbol{v}_{t-1} = \frac{-\Delta\boldsymbol{\theta}_{t-1}}{\delta} \tag{35}$$

     Now let us write $\boldsymbol{v}_t$ as a function of $\Delta\boldsymbol{\theta}$.

     $$\boldsymbol{v}_t = \beta\boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t \tag{36}$$
     $$\frac{-\Delta\boldsymbol{\theta}_t}{\delta} = \beta\frac{-\Delta\boldsymbol{\theta}_{t-1}}{\delta} + (1-\beta)\boldsymbol{g}_t \tag{37}$$
     $$\Delta\boldsymbol{\theta}_t = \beta\Delta\boldsymbol{\theta}_{t-1} - \delta\cdot(1-\beta)\boldsymbol{g}_t \tag{38}$$

     We can see the equivalence in equations 33 and 38 with the relations between hyper-parameters as follows:

     $$\alpha = \beta$$
     $$\epsilon = \delta\cdot(1-\beta)$$

2. Unroll the running average update rule, i.e. express $\boldsymbol{v}_t$ as a linear combination of $\boldsymbol{g}_i$'s ($1 \leq i \leq t$). The running average update is given as:

$$\boldsymbol{v}_t = \beta \boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t \tag{39}$$
$$\boldsymbol{v}_{t-1} = \beta \boldsymbol{v}_{t-2} + (1-\beta)\boldsymbol{g}_{t-1} \tag{40}$$
$$\vdots \tag{41}$$
$$\boldsymbol{v}_1 = (1-\beta)\boldsymbol{g}_1 \tag{42}$$

We can unroll the steps as follows:

$$\begin{aligned}
\boldsymbol{v}_t &= \beta^2 \boldsymbol{v}_{t-2} + \beta \cdot (1-\beta)\boldsymbol{g}_{t-1} + (1-\beta)\boldsymbol{g}_t \\
\boldsymbol{v}_t &= \beta^2 \Big(\beta \boldsymbol{v}_{t-3} + (1-\beta)\boldsymbol{g}_{t-2} \Big) + \beta \cdot (1-\beta)\boldsymbol{g}_{t-1} + (1-\beta)\boldsymbol{g}_t \\
\boldsymbol{v}_t &= \beta^3 \boldsymbol{v}_{t-3} + \beta^2 \cdot (1-\beta)\boldsymbol{g}_{t-2} + \beta \cdot (1-\beta)\boldsymbol{g}_{t-1} + (1-\beta)\boldsymbol{g}_t \\
&\vdots \\
\boldsymbol{v}_t &= (1-\beta) \sum_{i=0}^{t-1} \beta^i \boldsymbol{g}_{t-i}
\end{aligned}$$

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 $\boldsymbol{v}_t$.
From the above equation we have that:

$$\begin{aligned}
\boldsymbol{v}_t &= (1-\beta) \sum_{i=0}^{t-1} \beta^i \boldsymbol{g}_{t-i} \\
\mathbb{E}[\boldsymbol{v}_t] &= (1-\beta) \sum_{i=0}^{t-1} \underbrace{\mathbb{E}[\boldsymbol{g}_{t-i}]}_{\text{stationary}} \beta^i \\
&= (1-\beta) \cdot \mathbb{E}[\boldsymbol{g}_t] \cdot \underbrace{\sum_{i=0}^{t-1} \beta^i}_{\text{geometric series}} \\
&= (1-\beta) \cdot \mathbb{E}[\boldsymbol{g}_t] \cdot \frac{1-\beta^t}{1-\beta} \\
&= \mathbb{E}[\boldsymbol{g}_t] \cdot (1-\beta^t)
\end{aligned}$$

Thus $\mathbb{E}[\boldsymbol{v}_t] \neq \mathbb{E}[\boldsymbol{g}_t]$ and we can eliminate this bias by rescaling $\boldsymbol{v}_t$ by a value of $(1-\beta^t)$. That is,

$$\boldsymbol{v}_t^{\text{new}} = \frac{\boldsymbol{v}_t^{\text{old}}}{(1-\beta^t)}$$

Question 4 (5-5-5)**.** This question is about weight normalization. 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 $\boldsymbol{w}$.

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

   The equation $\hat{y} = \gamma \cdot \dfrac{y - \mu_y}{\sigma_y} + \beta$ can be written as follows:

$$\begin{aligned}
\hat{y} &= \gamma \cdot \frac{y - \mathbb{E}[y]}{\sqrt{\text{Var}[y]}} + \beta \\
&= \gamma \cdot \frac{y - \mathbb{E}[\boldsymbol{u}^T \boldsymbol{x}]}{\sqrt{\text{Var}[\boldsymbol{u}^T \boldsymbol{x}]}} + \beta \\
&= \gamma \cdot \frac{y - \boldsymbol{u}^T \overbrace{\mathbb{E}[\boldsymbol{x}]}^{0}}{\sqrt{||\boldsymbol{u}||^2 \underbrace{\text{Var}[\boldsymbol{x}]}_{\boldsymbol{I}}}} + \beta \\
&= \gamma \cdot \frac{\boldsymbol{u}^T \boldsymbol{x}}{\sqrt{||\boldsymbol{u}||^2}} + \beta \\
&= \underbrace{\gamma \cdot \frac{\boldsymbol{u}}{||\boldsymbol{u}||}}_{\boldsymbol{w}} \boldsymbol{x} + \beta \\
\hat{y} &= \boldsymbol{w}^T \boldsymbol{x} + \beta
\end{aligned}$$

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

   Using the derivative chain rule, we can express the gradient $\nabla_{\boldsymbol{u}} L$ as follows:

$$\nabla_{\boldsymbol{u}} L = \nabla_{\boldsymbol{u}} \boldsymbol{w} \nabla_{\boldsymbol{w}} L \tag{43}$$

   Let us find the value of $\nabla_{\boldsymbol{u}} \boldsymbol{w}$ by differentiating each component of the vector valued function.

$$\underline{\boldsymbol{i} = \boldsymbol{j}} \qquad\qquad \underline{\boldsymbol{i} \neq \boldsymbol{j}} \tag{44}$$

$$\tag{45}$$

$$\frac{\partial \boldsymbol{w}_i}{\partial \boldsymbol{u}_i} = \gamma \cdot \frac{\partial}{\partial \boldsymbol{u}_i}\left(\frac{\boldsymbol{u}_i}{||\boldsymbol{u}||^2}\right) \qquad \frac{\partial \boldsymbol{w}_i}{\partial \boldsymbol{u}_j} = \gamma \cdot \frac{\partial}{\partial \boldsymbol{u}_j}\left(\frac{\boldsymbol{u}_j}{||\boldsymbol{u}||^2}\right) \tag{46}$$

$$= \frac{\gamma}{||\boldsymbol{u}||} \cdot \frac{1 - \boldsymbol{u}_i^2}{||\boldsymbol{u}||^2} \qquad = \frac{\gamma}{||\boldsymbol{u}||} \cdot \frac{0 - \boldsymbol{u}_i \boldsymbol{u}_j}{||\boldsymbol{u}||^2} \tag{47}$$

$$\tag{48}$$

$$\nabla_{\boldsymbol{u}} L = \overbrace{\frac{\gamma}{||\boldsymbol{u}||}}^{\text{scalar}} \cdot \left(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^T}{||\boldsymbol{u}||^2}\right) \tag{49}$$

$$= s\boldsymbol{W}^{\perp} \tag{50}$$

   Substituting 50 in 43 we get:

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

   where $s = \frac{\gamma}{||\boldsymbol{u}||}$ and $\boldsymbol{W}^{\perp} = \left(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^T}{||\boldsymbol{u}||^2}\right)$

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.

3. Figure 1 shows the norm of $\boldsymbol{u}$ as a function of number of updates made to a two-layer MLP using gradient descent. Different curves correspond to models trained with different log-learning rate. Explain why (1) the norm is increasing, and (2) why larger learning rate corresponds to faster growth.
   (Hint: Use the Pythagorean theorem and the fact that $\boldsymbol{W}^{\perp}\boldsymbol{u} = 0$ from question 4.2).

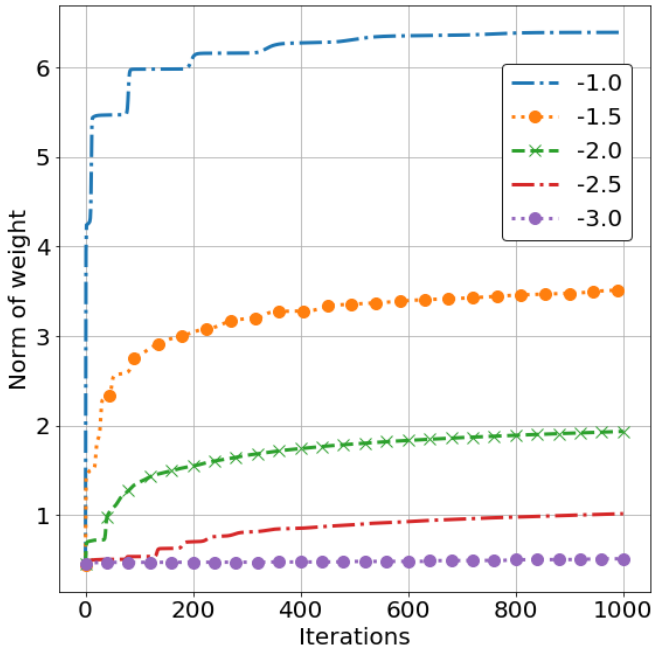

FIGURE 1 – Norm of parameters with different learning rate.

For a gradient descent update we have that:

$$\begin{aligned}\boldsymbol{u}_{t+1} &= \boldsymbol{u}_t - \eta \cdot \nabla_{\boldsymbol{u}} L \\ &= \boldsymbol{u}_t - \eta \cdot s\boldsymbol{W}^{\perp}\nabla_{\boldsymbol{w}} L\end{aligned}$$

The matrix $\boldsymbol{W}^{\perp}$ projects the gradient of $\boldsymbol{w}$ and thus gradient of $\boldsymbol{u}$ in a direction orthogonal to $\boldsymbol{u}$. Thus, by pythagoras theorem, we have that:

$$||\boldsymbol{u}_{i+1}|| > ||\boldsymbol{u}_t - \eta \cdot s\boldsymbol{W}^{\perp}\nabla_{\boldsymbol{w}} L||$$

Thus, the updated vector will have a norm that is greater than the initial value. This explains why the norm increases.

Secondly, the norm of the vector $\boldsymbol{u}_{t+1}$ grows proportional to the learning rate η. This implies that for larger values of η, for the inequality to hold, the norm of the updated vector grows faster. This explains why the larger learning rate corresponds to faster growth.

Question 5 (5-5-5)**.** 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 $\boldsymbol{h}_t$).
   Let us assume that $\sigma(\boldsymbol{h}_0)$ and $\boldsymbol{g}_0$ are initialized to the $\mathbf{0}$ vector. Thus we have that:

$$\begin{aligned}
\boldsymbol{g}_t &= \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}) \\
\boldsymbol{g}_1 &= \sigma(\boldsymbol{W}\boldsymbol{g}_0 + \underbrace{\boldsymbol{U}\boldsymbol{x}_1 + \boldsymbol{b}}_{\boldsymbol{h}_1 - \boldsymbol{W}\sigma(\boldsymbol{h}_0)}) \\
\boldsymbol{g}_1 &= \sigma(\underbrace{\boldsymbol{W}\boldsymbol{g}_0}_{0} + \boldsymbol{h}_1 - \underbrace{\boldsymbol{W}\sigma(\boldsymbol{h}_0)}_{0}) \\
\boldsymbol{g}_1 &= \sigma(\boldsymbol{h}_1)
\end{aligned}$$

   For the general induction case, let us assume that $\boldsymbol{g}_t = \sigma(\boldsymbol{h}_t)$. Now we have:

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

   Thus, we have the base case that $\boldsymbol{g}_0 = \sigma(\boldsymbol{h}_0)$ and have proved the induction step above. Therefore in general we can say that $\boxed{\boldsymbol{g}_t = \sigma(\boldsymbol{h}_t)}$

*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. $|\sigma'(x)| \leq \gamma$ 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 upper-bounded by $\frac{\delta^2}{\gamma^2}$ for some $0 \leq \delta < 1$, gradients of the hidden state will vanish over time, i.e.

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

Use the following properties of the L_2 operator norm

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

By the multivariate chain rule, we have that:

$$\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0} = \frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_{T-1}} \cdot \frac{\partial \boldsymbol{h}_{T-1}}{\partial \boldsymbol{h}_{T-2}} \cdot \frac{\partial \boldsymbol{h}_{T-2}}{\partial \boldsymbol{h}_{T-3}} \cdot \cdots \cdot \frac{\partial \boldsymbol{h}_1}{\partial \boldsymbol{h}_0} = \prod_{i=0}^{T-1} \frac{\partial \boldsymbol{h}_{i+1}}{\partial \boldsymbol{h}_i} \tag{51}$$

Let us consider one term in the product $\frac{\partial \boldsymbol{h}_{i+1}}{\partial \boldsymbol{h}_i}$. This is equal to:

$$\frac{\partial \boldsymbol{h}_{i+1}}{\partial \boldsymbol{h}_i} = \frac{\partial \boldsymbol{h}_{i+1}}{\partial \boldsymbol{a}_{i+1}} \cdot \frac{\partial \boldsymbol{a}_{i+1}}{\partial \boldsymbol{h}_i}$$

Now we have that:

$$\boldsymbol{h}_{i+1} = \sigma(\boldsymbol{a}_{i+1}) \qquad \therefore \frac{\partial \boldsymbol{h}_{i+1}}{\partial \boldsymbol{a}_{i+1}} = \sigma'(\boldsymbol{a}_{i+1})$$

$$\boldsymbol{a}_{i+1} = \boldsymbol{W}\boldsymbol{h}_i + \boldsymbol{U}\boldsymbol{x}_{i+1} + \boldsymbol{b} \qquad \therefore \frac{\partial \boldsymbol{a}_{i+1}}{\partial \boldsymbol{h}_i} = \boldsymbol{W}$$

Thus we have:

$$\frac{\partial \boldsymbol{h}_{i+1}}{\partial \boldsymbol{h}_i} = \sigma'(\boldsymbol{a}_{i+1})\boldsymbol{W} \tag{52}$$

$$\left|\left|\frac{\partial \boldsymbol{h}_{i+1}}{\partial \boldsymbol{h}_i}\right|\right| = \left|\left|\sigma'(\boldsymbol{a}_{i+1})\boldsymbol{W}\right|\right| \tag{53}$$

$$\leq \left|\left|\sigma'(\boldsymbol{a}_{i+1})\right|\right| \cdot \left|\left|\boldsymbol{W}\right|\right| \tag{54}$$

$$\leq c \cdot \gamma \cdot \sqrt{\lambda_1 \boldsymbol{W}^\top \boldsymbol{W}} \tag{55}$$

$$\leq c \cdot \gamma \cdot \frac{\delta}{\gamma} = c \cdot \delta \tag{56}$$

Substituting 56 in 51 we have:

$$\left|\left|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0}\right|\right| = \prod_{i=0}^{T-1} c \cdot \delta = c \cdot \delta^{T-1}$$

Thus we see that as $T \to \infty$, the term δ^{T-1} where $0 \leq \delta < 1$ will go towards zero, and as a consequence, the gradients of the hidden state will vanish over time.

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.

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).

   The inequalities above would not hold if $\sqrt{\lambda_1 \boldsymbol{W}^\top \boldsymbol{W}} > \frac{\delta^2}{\gamma^2}$. The gradient *could* explode. It is a *necessary* but not *sufficient* condition for the gradient to explode. It is a *necessary* condition because we have already proved that if $\sqrt{\lambda_1 \boldsymbol{W}^T \boldsymbol{W}} < \frac{\delta^2}{\lambda^2}$, then the gradients vanish.

Question 6 (6-12)**.** Denote by σ the logistic sigmoid function. 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.

1. Draw the computational graph for this RNN, unrolled for 3 time steps (from $t = 1$ to $t = 3$) Include and label the initial hidden states for both the forward and backward RNNs, $h_0^{(f)}$ and $h_4^{(b)}$ respectively. You may draw this by hand ; you may also use a computer rendering package such as TikZ, but you are not required to do so. Label each node and edge with the corresponding hidden unit or weight.

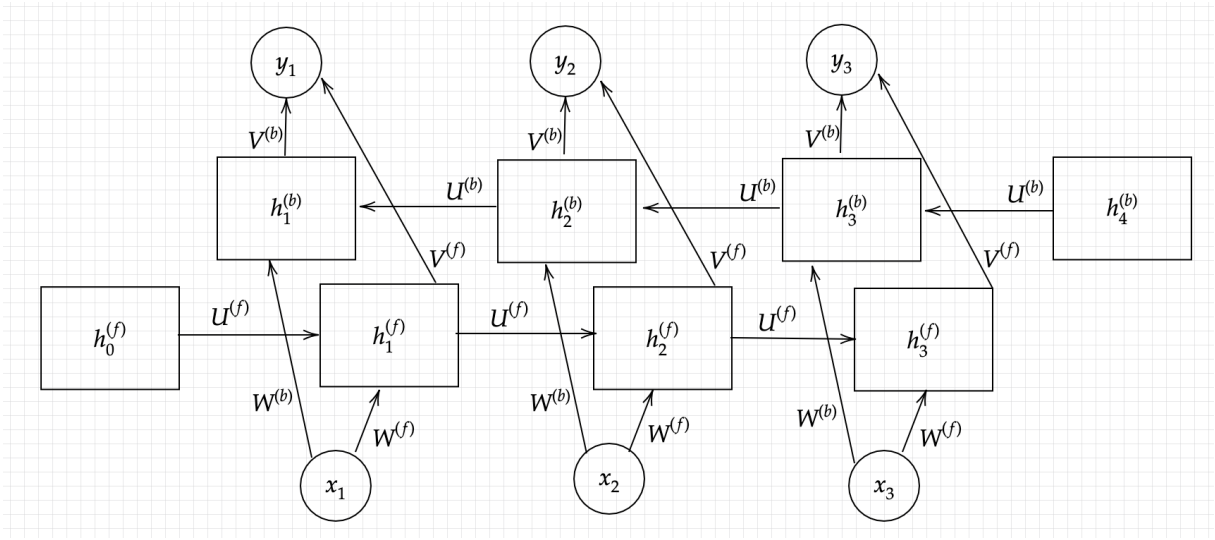

FIGURE 2 – Bidirectional RNN

*2. 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 $L_t = ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2$. 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 $\nabla_{\boldsymbol{h}_{t-1}^{(b)}} L$ respectively). Then derive $\nabla_{\boldsymbol{W}^{(f)}} L$ and $\nabla_{\boldsymbol{U}^{(b)}} L$.

Let us first find the gradient $\nabla_{\boldsymbol{h}_t^{(f)}}$. In general $\nabla_{\boldsymbol{h}_t} L$ does not depend on the hidden states of previous time steps. By the multivariate chain rule, we have the following.

$$\nabla_{\boldsymbol{h}_3^{(f)}} L = \nabla_{\boldsymbol{h}_3^{(f)}} L_3$$
$$\nabla_{\boldsymbol{h}_3^{(f)}} L = \frac{\partial L_3}{\partial \boldsymbol{y}_3} \cdot \frac{\partial \boldsymbol{y}_3}{\partial \boldsymbol{h}_3^{(f)}}$$

$$\nabla_{\boldsymbol{h}_2^{(f)}} L = \nabla_{\boldsymbol{h}_2^{(f)}} L_2 + \nabla_{\boldsymbol{h}_2^{(f)}} L_3$$
$$\nabla_{\boldsymbol{h}_2^{(f)}} L = \frac{\partial L_2}{\partial \boldsymbol{y}_2} \cdot \frac{\partial \boldsymbol{y}_2}{\partial \boldsymbol{h}_2^{(f)}} + \underbrace{\frac{\partial L_3}{\partial \boldsymbol{y}_3} \cdot \frac{\partial \boldsymbol{y}_3}{\partial \boldsymbol{h}_3^{(f)}}}_{\nabla_{\boldsymbol{h}_3^{(f)}} L} \cdot \frac{\boldsymbol{h}_3^{(f)}}{\boldsymbol{h}_2^{(f)}}$$

$$\nabla_{\boldsymbol{h}_1^{(f)}} L = \nabla_{\boldsymbol{h}_1^{(f)}} L_1 + \nabla_{\boldsymbol{h}_1^{(f)}} L_2 + \nabla_{\boldsymbol{h}_1^{(f)}} L_3$$
$$\nabla_{\boldsymbol{h}_1^{(f)}} L = \frac{\partial L_1}{\partial \boldsymbol{y}_1} \cdot \frac{\partial \boldsymbol{y}_1}{\partial \boldsymbol{h}_1^{(f)}} + \frac{\partial L_2}{\partial \boldsymbol{y}_2} \cdot \frac{\partial \boldsymbol{y}_2}{\partial \boldsymbol{h}_2^{(f)}} \cdot \frac{\partial \boldsymbol{h}_2^{(f)}}{\partial \boldsymbol{h}_1^{(f)}} + \frac{\partial L_3}{\partial \boldsymbol{y}_3} \cdot \frac{\partial \boldsymbol{y}_3}{\partial \boldsymbol{h}_3^{(f)}} \cdot \frac{\partial \boldsymbol{h}_3^{(f)}}{\partial \boldsymbol{h}_2^{(f)}} \cdot \frac{\partial \boldsymbol{h}_2^{(f)}}{\partial \boldsymbol{h}_1^{(f)}}$$
$$= \frac{\partial L_1}{\partial \boldsymbol{y}_1} \cdot \frac{\partial \boldsymbol{y}_1}{\partial \boldsymbol{h}_1^{(f)}} + \frac{\partial \boldsymbol{h}_2^{(f)}}{\partial \boldsymbol{h}_1^{(f)}} \underbrace{\left(\frac{\partial L_2}{\partial \boldsymbol{y}_2} \cdot \frac{\partial \boldsymbol{y}_2}{\partial \boldsymbol{h}_2^{(f)}} + \frac{\partial L_3}{\partial \boldsymbol{y}_3} \cdot \frac{\partial \boldsymbol{y}_3}{\partial \boldsymbol{h}_3^{(f)}} \cdot \frac{\partial \boldsymbol{h}_3^{(f)}}{\partial \boldsymbol{h}_2^{(f)}} \right)}_{\nabla_{\boldsymbol{h}_2^{(f)}} L}$$

From this pattern, we can infer the following equation:

$$\nabla_{\boldsymbol{h}_t^{(f)}} L = \nabla_{\boldsymbol{h}_t^{(f)}} L_t + \nabla_{\boldsymbol{h}_t^{(f)}} \boldsymbol{h}_{t+1}^{(f)} \cdot \nabla_{\boldsymbol{h}_{t+1}^{(f)}} L \tag{57}$$

Now let us calculate the actual gradients. The first term $\nabla_{\boldsymbol{h}_t^{(f)}} L_t$ can be decomposed as follows:

$$\nabla_{\boldsymbol{h}_t^{(f)}} L_t = \nabla_{\boldsymbol{y}_t} L \cdot \nabla_{\boldsymbol{h}_t^{(f)}} \boldsymbol{y}_t \tag{58}$$
$$= -2(\boldsymbol{z}_t - \boldsymbol{y}_t) \cdot \boldsymbol{V}^{(f)} \tag{59}$$

We can work out the first part of the second term $\nabla_{\boldsymbol{h}_t^{(f)}} \boldsymbol{h}_{t+1}^{(f)}$ as follows (using the fact that $\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x))$

$$\nabla_{\boldsymbol{h}_t^{(f)}} \boldsymbol{h}_{t+1}^{(f)} = \boldsymbol{h}_{t+1}^{(f)} (1 - \boldsymbol{h}_{t+1}^{(f)}) \cdot \boldsymbol{U}^{(f)} \tag{60}$$

Finally combining 59 and 60 in 57 we have:

$$\boxed{\nabla_{\boldsymbol{h}_t^{(f)}} L = -2(\boldsymbol{z}_t - \boldsymbol{y}_t) \cdot \boldsymbol{V}^{(f)} + \boldsymbol{h}_{t+1}^{(f)} (1 - \boldsymbol{h}_{t+1}^{(f)}) \cdot \boldsymbol{U}^{(f)} \cdot \nabla_{\boldsymbol{h}_{t+1}^{(f)}} L} \tag{61}$$

Similarly, we can derive the values for $\nabla_{\boldsymbol{h}_t^{(b)}} L$:

$$\nabla_{\boldsymbol{h}_t^{(b)}} L = \nabla_{\boldsymbol{h}_t^{(b)}} L_t + \nabla_{\boldsymbol{h}_t^{(b)}} \boldsymbol{h}_{t-1}^{(b)} \cdot \nabla_{\boldsymbol{h}_{t-1}^{(b)}} L \tag{62}$$

$$= \boxed{-2(\boldsymbol{z}_t - \boldsymbol{y}_t) \cdot \boldsymbol{V}^{(b)} + \boldsymbol{h}_t^{(b)}(1 - \boldsymbol{h}_t^{(b)}) \cdot \boldsymbol{U}^{(b)} \cdot \nabla_{\boldsymbol{h}_{t-1}^{(b)}} L} \tag{63}$$

Now we find the gradient with respect to the parameter $\nabla_{\boldsymbol{W}^{(f)}} L$. This equals:

$$\nabla_{\boldsymbol{W}^{(f)}} L = \sum_t \underbrace{\nabla_{\boldsymbol{h}_t^{(f)}} L}_{\text{calculated in 61}} \cdot \nabla_{\boldsymbol{W}^{(f)}} \boldsymbol{h}_t^{(f)}$$

Note that we have calculated the quantity $\nabla_{\boldsymbol{h}_t^{(f)}} L$ in 61. The second quantity $\nabla_{\boldsymbol{W}^{(f)}} \boldsymbol{h}_t^{(f)}$ is the derivative of a vector with respect to a matrix which is a tensor. In the following derivation we ease the notation for simplicity.

$$\nabla_{\boldsymbol{W}^{(f)}} L = \boxed{\sum_t \left(\nabla_{\boldsymbol{h}_t^{(f)}} L = -2(\boldsymbol{z}_t - \boldsymbol{y}_t) \cdot \boldsymbol{V}^{(f)} + \boldsymbol{h}_{t+1}^{(f)}(1 - \boldsymbol{h}_{t+1}^{(f)}) \cdot \boldsymbol{U}^{(f)} \cdot \nabla_{\boldsymbol{h}_{t+1}^{(f)}} L \right) \cdot \left(\boldsymbol{h}_t^{(f)}(1 - \boldsymbol{h}_t^{(f)}) \cdot \boldsymbol{x}_t \right)}$$

Similary we find the gradient with respect to the parameter $\nabla_{\boldsymbol{U}^{(b)}} L$. This equals:

$$\nabla_{\boldsymbol{U}^{(b)}} L = \sum_t \nabla_{\boldsymbol{h}_t^{(b)}} L \cdot \nabla_{\boldsymbol{U}^{(b)}} \boldsymbol{h}_t^{(b)}$$

We have calculated the first quantity $\nabla_{\boldsymbol{h}_t^{(b)}} L$ in 63. The second quantity works out to be $\boldsymbol{h}_t^{(b)}(1 - \boldsymbol{h}_t^{(b)}) \cdot \boldsymbol{h}_{t+1}^{(b)}$ using the fact that $\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x))$. Thus we have:

$$\nabla_{\boldsymbol{U}^{(b)}} L = \boxed{\sum_t \left(-2(\boldsymbol{z}_t - \boldsymbol{y}_t) \cdot \boldsymbol{V}^{(b)} + \boldsymbol{h}_t^{(b)}(1 - \boldsymbol{h}_t^{(b)}) \cdot \boldsymbol{U}^{(b)} \cdot \nabla_{\boldsymbol{h}_{t-1}^{(b)}} L \right) \cdot \left(\boldsymbol{h}_t^{(b)}(1 - \boldsymbol{h}_t^{(b)}) \cdot \boldsymbol{h}_{t+1}^{(b)} \right)}$$

Références

[1] X. Glorot and Y. Bengio, *Understanding the difficulty of training deep feedforward neural networks*, in Proceedings of the thirteenth international conference on artificial intelligence and statistics, 2010, pp. 249–256.

[2] K. He, X. Zhang, S. Ren, and J. Sun, *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*, CoRR, abs/1502.01852 (2015).