Due Date: March 22nd 23:59, 2019

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 preactivations 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 $\mathrm{Var}(\boldsymbol{a}_i^{(t)}) = 1$, for $1 \leq i \leq d^{(t)}$.

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

1. Assume $\mathbb{E}[\boldsymbol{h}_i^{(t-1)}] = 0$ and $\mathrm{Var}(\boldsymbol{h}_i^{(t-1)}) = 1$ 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 $d^{(t-1)}$.
   (b) Uniform: give the values for c, α, and β as a function of $d^{(t-1)}$.
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 $d^{(t-1)}$.
   (b) Uniform: give the values for c, α, and β as a function of $d^{(t-1)}$.
   (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.

Answer 1.

1. For $1 \leq i \leq d^{(t)}, 1 \leq j \leq d^{(t-1)}$, $\boldsymbol{W}_{ij}^{(t)}$ from Gaussian or Uniform distributions means $\boldsymbol{W}_{ij}^{(t)}$ and $\boldsymbol{h}_j^{(t-1)}$ are independent: $\boldsymbol{W}_{ij}^{(t)} \perp \boldsymbol{h}_j^{(t-1)}$.
   Given $\mathbb{E}[h_i^{(t-1)}] = 0$ and $\mathrm{Var}(\boldsymbol{h}_i^{(t-1)}) = 1$, for both distributions:

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

   and,

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

   One possible parameters initialize scheme is to let $c = 0$, such that $\mathbb{E}[\boldsymbol{a}_i^{(t)}] = 0$ and $\mathrm{Var}(\boldsymbol{b}_i^{(t)}) = 0$, and to let $\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] = 0$ and $\mathrm{Var}(\boldsymbol{W}_{ij}^{(t)}) = \frac{1}{d^{(t-1)}}$ such that $\mathrm{Var}(\boldsymbol{a}_i^{(t)}) = 1$

   (a) For a Gaussian distribution: $\boldsymbol{W}_{ij}^{(t)} \sim \mathcal{N}(\mu, \sigma^2)$,

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

   Therefore, the scheme is equivalent to:

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

   (b) For a Uniform distribution: $\boldsymbol{W}_{ij}^{(t)} \sim U(\alpha, \beta)$,

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

   Therefore, the scheme is equivalent to:

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

2. Given $\mathbb{E}[\boldsymbol{a}_i^{(t-1)}] = 0, \mathrm{Var}(\boldsymbol{a}_i^{(t-1)}) = 1$ and $\boldsymbol{a}_i^{(t-1)}$ has a symmetric distribution for $1 \leq i \leq d^{(t-1)}$, we have:

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{h}_i^{(t-1)})^2] &= \int_{-\infty}^{\infty} \max(0, \boldsymbol{a}_i^{(t-1)})^2 \; p(x)dx \\
&= \int_{0}^{\infty} (\boldsymbol{a}_i^{(t-1)})^2 \; p(x)dx \\
&= \frac{1}{2}\int_{-\infty}^{\infty} \big(\boldsymbol{a}_i^{(t-1)} - \mathbb{E}[\boldsymbol{a}_i^{(t-1)}]\big)^2 p(x)dx \\
&= \frac{1}{2}\mathrm{Var}(\boldsymbol{a}_i^{(t-1)}) \\
&= \frac{1}{2}
\end{aligned}$$

From previous question, we know:

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

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

One possible parameters initialize scheme is to let $c = 0$ and $\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] = 0$, such that $\mathbb{E}[\boldsymbol{a}_i^{(t)}] = 0, \mathrm{Var}(\boldsymbol{b}_i^{(t)}) = 0$ and simply $\mathrm{Var}(\boldsymbol{a}_i^{(t)})$ to:

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

Therefore, we need

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

(a) For a Gaussian distribution: $\boldsymbol{W}_{ij}^{(t)} \sim \mathcal{N}(\mu, \sigma^2)$,

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

Therefore, the scheme is equivalent to:

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

(b) For a Uniform distribution: $\boldsymbol{W}_{ij}^{(t)} \sim U(\alpha, \beta)$,

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

Therefore, the scheme is equivalent to:

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

(c) He Normal (He-et-al) Initialization shceme has this form.

(d) By using this initialization scheme, the expectation and variance of the output of each layer will be very close to 0 and 1, respectively ; which maintains a certain level of gradient for each layer, decreases the possibility of the gradient exploding or vanishing, and therefore gets a better trained model.

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 $\boldsymbol{R}_{ij} \sim \mathrm{Bern}(p)$ 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 $\boldsymbol{R}$).
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 $\mathbb{E}[L(\boldsymbol{w})] = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma\boldsymbol{w}||^2$.
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} ?

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.
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).

Answer 2.

1. $L(\boldsymbol{w}) = ||\boldsymbol{y} - (\boldsymbol{R} \odot \boldsymbol{X})\boldsymbol{w}||^2$, where $\odot$ is an element-wise multiplication.
2.

$$\begin{aligned}
\mathbb{E}[L(\boldsymbol{w})] &= \mathbb{E}[||\boldsymbol{y} - (\boldsymbol{R}\odot\boldsymbol{X})\boldsymbol{w}||^2] \\
&= \mathbb{E}[(\boldsymbol{y} - (\boldsymbol{R}\odot\boldsymbol{X})\boldsymbol{w})^\top(\boldsymbol{y} - (\boldsymbol{R}\odot\boldsymbol{X})\boldsymbol{w})] \\
&= \mathbb{E}[\boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{w}^\top(\boldsymbol{R}\odot\boldsymbol{X})^\top\boldsymbol{y} + \boldsymbol{w}^\top(\boldsymbol{R}\odot\boldsymbol{X})^\top(\boldsymbol{R}\odot\boldsymbol{X})\boldsymbol{w}] \\
&= \boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{w}^\top\mathbb{E}[(\boldsymbol{R}\odot\boldsymbol{X})]^\top\boldsymbol{y} + \boldsymbol{w}^\top\mathbb{E}[(\boldsymbol{R}\odot\boldsymbol{X})^\top(\boldsymbol{R}\odot\boldsymbol{X})]\boldsymbol{w}
\end{aligned}$$

For the second item,

$$\big(\mathbb{E}[\boldsymbol{R}\odot\boldsymbol{X}]\big)_{ij}=\mathbb{E}[(\boldsymbol{R}\odot\boldsymbol{X})_{ij}]=p\boldsymbol{X}_{ij}$$

For the last item,

$$\begin{aligned}
&\mathbb{E}\big[\big((\boldsymbol{R}\odot\boldsymbol{X})^\top(\boldsymbol{R}\odot\boldsymbol{X})\big)_{ij}\big]\\
=&\mathbb{E}\Big[\sum_{k=1}^{N}(\boldsymbol{R}\odot\boldsymbol{X})_{ki}(\boldsymbol{R}\odot\boldsymbol{X})_{kj}\Big]\\
=&\sum_{k=1}^{d}\mathbb{E}\big[\boldsymbol{R}_{ki}\boldsymbol{R}_{kj}\big]\boldsymbol{X}_{ki}\boldsymbol{X}_{kj}\\
=&\begin{cases}p^2(\boldsymbol{X}^T\boldsymbol{X})_{ij} & \text{if } i\neq j\\ \sum_{k=1}^{d}\mathbb{E}[\boldsymbol{R}_{ki}^2]\boldsymbol{X}_{ki}^2=p(\boldsymbol{X}^\top\boldsymbol{X})_{ii} & \text{if } i=j\end{cases}
\end{aligned}$$

With the two items re-written, we have:

$$\begin{aligned}
\mathbb{E}[L(\boldsymbol{w})]&=\boldsymbol{y}^\top\boldsymbol{y}-2\boldsymbol{w}^\top\mathbb{E}[(\boldsymbol{R}\odot\boldsymbol{X})]^\top\boldsymbol{y}+\boldsymbol{w}^\top\mathbb{E}[(\boldsymbol{R}\odot\boldsymbol{X})^\top(\boldsymbol{R}\odot\boldsymbol{X})]\boldsymbol{w}\\
&=\boldsymbol{y}^\top\boldsymbol{y}-2p\boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{y}+\boldsymbol{w}^\top\mathbb{E}[(\boldsymbol{R}\odot\boldsymbol{X})^\top(\boldsymbol{R}\odot\boldsymbol{X})]\boldsymbol{w}\\
&=\boldsymbol{y}^\top\boldsymbol{y}-2p\boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{y}+p^2\boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}-p^2\boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}+\boldsymbol{w}^\top\mathbb{E}[(\boldsymbol{R}\odot\boldsymbol{X})^\top(\boldsymbol{R}\odot\boldsymbol{X})]\boldsymbol{w}\\
&=||\boldsymbol{y}-p\boldsymbol{X}\boldsymbol{w}||^2+\boldsymbol{w}^\top\Big(\mathbb{E}[(\boldsymbol{R}\odot\boldsymbol{X})^\top(\boldsymbol{R}\odot\boldsymbol{X})]-p^2\boldsymbol{X}^\top\boldsymbol{X}\Big)\boldsymbol{w}\\
&=||\boldsymbol{y}-p\boldsymbol{X}\boldsymbol{w}||^2+\boldsymbol{w}^\top\big(p(1-p)\mathrm{diag}([(\boldsymbol{X}^\top\boldsymbol{X})_{11},(\boldsymbol{X}^\top\boldsymbol{X})_{22},\cdots,(\boldsymbol{X}^\top\boldsymbol{X})_{dd}]^\top)\big)\boldsymbol{w}\\
&=||\boldsymbol{y}-p\boldsymbol{X}\boldsymbol{w}||^2+(p(1-p)\boldsymbol{w}^\top\Gamma^\top\Gamma\boldsymbol{w}\\
&=||\boldsymbol{y}-p\boldsymbol{X}\boldsymbol{w}||^2+(p(1-p)(\Gamma\boldsymbol{w})^\top\Gamma\boldsymbol{w}\\
&=||\boldsymbol{y}-p\boldsymbol{X}\boldsymbol{w}||^2+p(1-p)||\Gamma\boldsymbol{w}||^2
\end{aligned}$$

3. For a matrix $\boldsymbol{A}\in\mathbb{R}^{d\times d}$ and a vector $\boldsymbol{w}\in\mathbb{R}^{d\times 1}$, formulas of matrix derivatives (denominator layout):

$$\frac{\partial\boldsymbol{A}\boldsymbol{w}}{\partial\boldsymbol{w}}=\boldsymbol{A}^\top,\qquad\frac{\partial\boldsymbol{w}^\top A}{\partial\boldsymbol{w}}=\boldsymbol{A},\qquad\frac{\partial\boldsymbol{w}^\top\boldsymbol{A}\boldsymbol{w}}{\partial\boldsymbol{w}}=(\boldsymbol{A}+\boldsymbol{A}^\top)\boldsymbol{w}$$

are used for the following calculation:

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

Let $\frac{\partial(\mathbb{E}[L(\boldsymbol{w})])}{\partial \boldsymbol{w}} = 0$, we can find $\boldsymbol{w}^{\text{dropout}}$ to a minimal $L(\boldsymbol{w})$:

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

Obviously, $\lambda^{\text{dropout}} = \frac{1-p}{p}$. When p is close to 0, λ^{dropout} grows to ∞ ; when p is close to 1, λ^{dropout} decreases to 0.

4. When use L_2 regularization, the loss function can be written as:

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

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

Let $\frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}} = 0$, we find $\boldsymbol{w}^{L_2}$ to a minimal $L(\boldsymbol{w})$:

$$\boldsymbol{w}^{L_2} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{L_2})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

5. Dropout with linear regression is equivalent to ridge regression with a particular form for Γ. This form essentially scales the weight cost for weight $\boldsymbol{w}_i$ by the standard deviation of the i_{th} dimension of the data. Both dropout and weight decay can be considered a kind of regularization method in training a network, and the theoretical solution of the optimal weights from two methods have similar structures, indicating regularization method can be considered a special case of dropout.

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).

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 3.

1.
- for SGD with momentum,

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

- for SGD with running average of $\boldsymbol{g}_t$,

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

Given:

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

the two update rules are equivalent.

2.

$$\begin{aligned}\Delta\boldsymbol{\theta}_t &= \beta\Delta\boldsymbol{\theta}_{t-1} - \delta(1-\beta)\boldsymbol{g}_t \\ &= \beta(\beta\Delta\boldsymbol{\theta}_{t-2} - \delta(1-\beta)\boldsymbol{g}_{t-1}) - \delta(1-\beta)\boldsymbol{g}_t \\ &= \beta^2\Delta\boldsymbol{\theta}_{t-1} - \delta\beta(1-\beta)\boldsymbol{g}_{t-1} - \delta(1-\beta)\boldsymbol{g}_t \\ &\cdots \\ &= \beta^t\Delta\boldsymbol{\theta}_0 - \delta(1-\beta)\big(\beta^{t-1}\boldsymbol{g}_1 + \beta^{t-2}\boldsymbol{g}_2 + \cdots + \beta\boldsymbol{g}_{t-1} + \beta^0\boldsymbol{g}_t\big) \\ &= \beta^t(-\delta\boldsymbol{v}_0) - \delta(1-\beta)\sum_{i=1}^{t}\beta^{t-i}\boldsymbol{g}_i \\ &= -\delta(1-\beta)\sum_{i=1}^{t}\beta^{t-i}\boldsymbol{g}_i\end{aligned}$$

$$\boldsymbol{v}_t = -\frac{\Delta\boldsymbol{\theta}_t}{\delta} = (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\boldsymbol{g}_i$$

3. Given $\boldsymbol{g}_t$ has a stationary distribution independent of t,

$$\begin{aligned}\mathbb{E}[\boldsymbol{v}_t] &= \mathbb{E}[\frac{\Delta\boldsymbol{\theta}_t}{-\delta}] \\ &= \mathbb{E}[(1-\beta)\sum_{i=1}^{t}\beta^{t-i}\boldsymbol{g}_i] \\ &= \mathbb{E}[\boldsymbol{g}_t](1-\beta)\sum_{i=1}^{t}\beta^{t-i} \\ &= (1-\beta)\frac{1-\beta^t}{1-\beta}\mathbb{E}[\boldsymbol{g}_t] \\ &= (1-\beta^t)\mathbb{E}[\boldsymbol{g}_t]\end{aligned}$$

By importing an exponential weighted mean factor $\frac{1}{1-\beta^t}$; that is, let $\tilde{\boldsymbol{v}}_t = \frac{1}{1-\beta^t}\boldsymbol{v}_t$, then:

$$\mathbb{E}[\tilde{\boldsymbol{v}}_t] = \mathbb{E}[\boldsymbol{g}_t]$$

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 ($\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{I}$) and centered at 0 ($\mathbb{E}[\boldsymbol{x}] = \boldsymbol{0}$). Show that $\hat{y} = \boldsymbol{w}^\top\boldsymbol{x}+\beta$.
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] $\boldsymbol{W}^\perp\boldsymbol{u} = \boldsymbol{0}$.
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).

Answer 4.

1. Given $\boldsymbol{w} := \gamma\frac{\boldsymbol{u}}{||\boldsymbol{u}||}$, $y = \boldsymbol{u}^\top\boldsymbol{x}$, $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{I}$, and $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{0}$, we have:

$$\mu_y = \mathbb{E}[y] = \mathbb{E}[\boldsymbol{u}^\top\boldsymbol{x}] = \mathbb{E}\big[\sum_i \boldsymbol{u}_i\boldsymbol{x}_i\big] = \sum_i \mathbb{E}[\boldsymbol{u}_i\boldsymbol{x}_i] = \sum_i \boldsymbol{u}_i\mathbb{E}[\boldsymbol{x}_i] = 0$$

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.

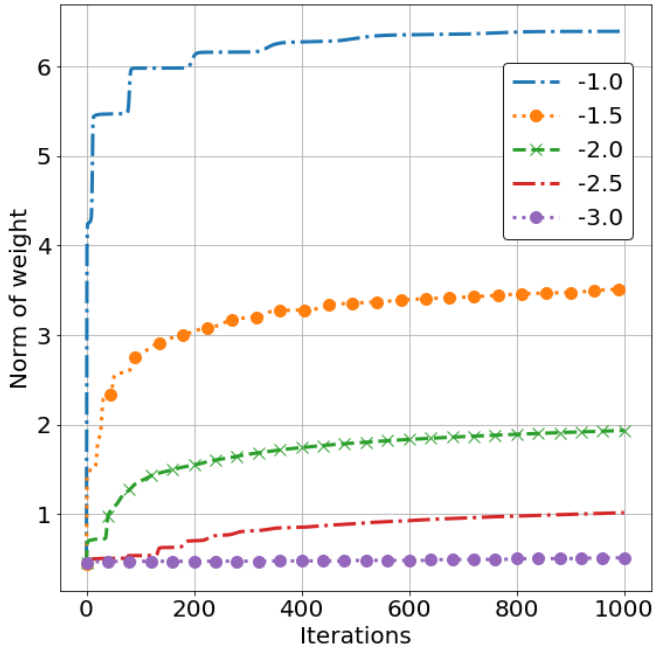


FIGURE 1 – Norm of parameters with different learning rate.

$$
\begin{aligned}
\gamma \cdot \frac{y-\mu_y}{\sigma_y} &= \gamma \frac{\boldsymbol{u}^\top \boldsymbol{x} - 0}{\sqrt{\mathrm{Var}(\boldsymbol{u}^\top \boldsymbol{x})}} \\
&= \gamma \frac{\boldsymbol{u}^\top \boldsymbol{x}}{\sqrt{\mathrm{Var}(\sum_i \boldsymbol{u}_i \boldsymbol{x}_i)}} \\
&= \gamma \frac{\boldsymbol{u}^\top \boldsymbol{x}}{\sqrt{\sum_i \boldsymbol{u}_i^2 \mathrm{Var}(\boldsymbol{x}_i)}} \qquad (\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{I}, \quad \mathrm{Cov}(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0 \text{ if } i \neq j) \\
&= \gamma \frac{\boldsymbol{u}^\top \boldsymbol{x}}{\sqrt{\sum_i \boldsymbol{u}_i^2}} \\
&= \gamma \frac{\boldsymbol{u}^\top \boldsymbol{x}}{||\boldsymbol{u}||} \\
&= \boldsymbol{w}^\top \boldsymbol{x}
\end{aligned}
$$

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

2. First,

$$
\begin{aligned}
\frac{\partial ||\boldsymbol{u}||}{\partial \boldsymbol{u}} &= \frac{\partial (\boldsymbol{u}^\top \boldsymbol{u})^{1/2}}{\partial \boldsymbol{u}} \\
&= \frac{1}{2} (\boldsymbol{u}^\top \boldsymbol{u})^{-\frac{1}{2}} \frac{\partial \boldsymbol{u}^\top \boldsymbol{u}}{\partial \boldsymbol{u}} \\
&= \frac{1}{2} \frac{2\boldsymbol{u}^\top}{||\boldsymbol{u}||} \\
&= \frac{\boldsymbol{u}^\top}{||\boldsymbol{u}||}
\end{aligned}
$$

Given $\boldsymbol{W}^{\perp} = \left(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^{\top}}{||\boldsymbol{u}||^2}\right)$, let $s = \frac{\gamma}{||\boldsymbol{u}||}$, then:

$$\begin{aligned}
\nabla_{\boldsymbol{u}} L &= \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{u}} \nabla_{\boldsymbol{w}} L \\
&= \gamma \frac{\partial}{\partial \boldsymbol{u}} (\frac{\boldsymbol{u}}{||\boldsymbol{u}||}) \nabla_{\boldsymbol{w}} L \\
&= \gamma \frac{||\boldsymbol{u}|| \boldsymbol{I} - \boldsymbol{u} \partial ||\boldsymbol{u}|| / \partial \boldsymbol{u}}{||\boldsymbol{u}||^2} \nabla_{\boldsymbol{w}} L \\
&= \gamma \left(\frac{\boldsymbol{I}}{||\boldsymbol{u}||} - \frac{\boldsymbol{u}}{||\boldsymbol{u}||^2} \frac{\partial ||\boldsymbol{u}||}{\partial \boldsymbol{u}} \right) \nabla_{\boldsymbol{w}} L \\
&= \gamma \left(\frac{\boldsymbol{I}}{||\boldsymbol{u}||} - \frac{\boldsymbol{u}}{||\boldsymbol{u}||^2} \frac{\boldsymbol{u}^{\top}}{||\boldsymbol{u}||} \right) \nabla_{\boldsymbol{w}} L \\
&= \frac{\gamma}{||\boldsymbol{u}||} (\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^{\top}}{||\boldsymbol{u}||^2}) \nabla_{\boldsymbol{w}} L \\
&= s \boldsymbol{W}^{\perp} \nabla_{\boldsymbol{w}} L
\end{aligned}$$

3. First,

$$\begin{aligned}
\boldsymbol{u}^{\top} \nabla_{\boldsymbol{u}} L &= \boldsymbol{u}^{\top} s \boldsymbol{W}^{\perp} \nabla_{\boldsymbol{w}} L \\
&= s \boldsymbol{u}^{\top} \left(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^{\top}}{||\boldsymbol{u}||^2} \right) \\
&= s \left(\boldsymbol{u}^{\top} - \frac{\boldsymbol{u}^{\top} \boldsymbol{u} \boldsymbol{u}^{\top}}{||\boldsymbol{u}||^2} \right) \nabla_{\boldsymbol{w}} L \\
&= s (\boldsymbol{u}^{\top} - \boldsymbol{u}^{\top}) \nabla_{\boldsymbol{w}} L \\
&= 0
\end{aligned}$$

which also means

$$(\nabla_{\boldsymbol{u}} L)^{\top} \boldsymbol{u} = (\boldsymbol{u}^{\top} \nabla_{\boldsymbol{u}} L)^{\top} = 0$$

Let $\boldsymbol{u}'$ be the one time step updated $\boldsymbol{u}$, that is:

$$\boldsymbol{u}' = \boldsymbol{u} - \alpha \nabla_{\boldsymbol{u}} L$$

then,

$$\begin{aligned}
||\boldsymbol{u}'||^2 - ||\boldsymbol{u}||^2 &= (\boldsymbol{u} - \alpha \nabla_{\boldsymbol{u}} L)^{\top} (\boldsymbol{u} - \alpha \nabla_{\boldsymbol{u}} L) - \boldsymbol{u}^{\top} \boldsymbol{u} \\
&= \boldsymbol{u}^{\top} \boldsymbol{u} - \boldsymbol{u}^{\top} \alpha \nabla_{\boldsymbol{u}} L - \alpha (\nabla_{\boldsymbol{u}} L)^{\top} \boldsymbol{u} + \alpha^2 (\nabla_{\boldsymbol{u}} L)^{\top} \nabla_{\boldsymbol{u}} L - \boldsymbol{u}^{\top} \boldsymbol{u} \\
&= \alpha^2 (\nabla_{\boldsymbol{u}} L)^{\top} \nabla_{\boldsymbol{u}} L - \alpha \left(\boldsymbol{u}^{\top} \nabla_{\boldsymbol{u}} L + (\nabla_{\boldsymbol{u}} L)^{\top} \boldsymbol{u} \right) \\
&= \alpha^2 ||\nabla_{\boldsymbol{u}} L||^2 + \alpha (0 + 0) \\
&= ||\alpha \nabla_{\boldsymbol{u}} L||^2
\end{aligned}$$

This indicates that $\Delta \boldsymbol{u} = -\alpha \nabla_u L$($\alpha$ is learning rate) is always orthonormal to $\boldsymbol{u}$. $||\boldsymbol{u}'||, ||\boldsymbol{u}||$, and $||\Delta \boldsymbol{u}||$ can form a right triangle. According to the Pythagorean theorem, $||\boldsymbol{u}'||$ is the longest. That is, if $||\Delta \boldsymbol{u}|| > 0$, the norm of updated weight $\boldsymbol{u}'$ satisfies:

$$||\boldsymbol{u}'|| = ||\boldsymbol{u} + \Delta \boldsymbol{u}|| > ||\boldsymbol{u}||$$

This explains that the norm of weight is always increasing.

Larger learning rate corresponds to larger $||\Delta \boldsymbol{u}||$ in early iteration, and therefore leading to a larger $||\boldsymbol{u}'||$. This explains larger learning rate corresponds to faster growth of $||\boldsymbol{u}||$ in early several iterations.

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 $\boldsymbol{h}_t$). 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. $|\sigma'(x)| \leq \gamma$ 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.

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

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. I will first make a guess, then prove it by mathematical induction.
   Assume existing a function σ^{-1} such that:

$$\sigma^{-1}(\sigma(x)) = x$$

from

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

we can derive:

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

take $\boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}$ into $\boldsymbol{h}_t$, we have:

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

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.

This implies:

$$\boldsymbol{g}_t = \sigma(\boldsymbol{h}_t)$$

Prove: At tims step $t = 0$, $\boldsymbol{g}_0$ and $\boldsymbol{h}_0$ are initialized to 0.
At time step $t = 1$, based on the formulas give for $\boldsymbol{g}_t$ and $\boldsymbol{h}_t$, we have:

$$\begin{aligned}
\boldsymbol{h}_1 &= \boldsymbol{W}\sigma(\boldsymbol{h}_0) + \boldsymbol{U}\boldsymbol{x}_1 + \boldsymbol{b} = \boldsymbol{U}\boldsymbol{x}_1 + \boldsymbol{b} \\
\boldsymbol{g}_1 &= \sigma(\boldsymbol{W}\boldsymbol{g}_0 + \boldsymbol{U}\boldsymbol{x}_1 + \boldsymbol{b}) = \sigma(\boldsymbol{U}\boldsymbol{x}_1 + \boldsymbol{b}) \\
&= \sigma(\boldsymbol{h}_1)
\end{aligned}$$

which means $\boldsymbol{g}_t = \sigma(\boldsymbol{h}_t)$ holds for $t = 1$.
Assume for the time $t-1$, $\boldsymbol{g}_{t-1} = \sigma(\boldsymbol{h}_{t-1})$ holds, then for the time step t:

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

which means $\boldsymbol{g}_t = \sigma(\boldsymbol{h}_t)$ holds as well for time step t.

2.

$$\begin{aligned}
\left|\left|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0}\right|\right| &= \left|\left|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_{T-1}}\frac{\partial \boldsymbol{h}_{T-1}}{\partial \boldsymbol{h}_{T-2}}\cdots\frac{\partial \boldsymbol{h}_1}{\partial \boldsymbol{h}_0}\right|\right| \\
&= ||\boldsymbol{W}^\top\sigma'(\boldsymbol{h}_{T-1})\ \boldsymbol{W}^\top\sigma'(\boldsymbol{h}_{T-1})\ \cdots\ \boldsymbol{W}^\top\sigma'(\boldsymbol{h}_0)|| \\
&= ||\prod_{i=0}^{T-1}\boldsymbol{W}^\top\sigma'(\boldsymbol{h}_i)|| \\
&\leq \prod_{i=0}^{T-1}||\boldsymbol{W}^\top\sigma'(\boldsymbol{h}_i)|| \\
&\leq ||\boldsymbol{W}^\top||^T\prod_{i=0}^{T-1}||\sigma'(\boldsymbol{h}_i)|| \\
&= ||\boldsymbol{W}||^T\prod_{i=0}^{T-1}||\sigma'(\boldsymbol{h}_i)|| \\
&\leq \left(\sqrt{\lambda_1(\boldsymbol{W}^\top\boldsymbol{W})}\right)^T\gamma^T \\
&\leq \left(\frac{\delta}{\gamma}\right)^T\gamma^T \\
&= \delta^T
\end{aligned}$$

Since $0 \leq \delta < 1$, as $T \to \infty$, the norm of the gradient of the hidden states will be 0, meaning the gradient of the hidden states will vanish.

3. The condition of the largest eigenvalue of the weights being larger than $\frac{\delta^2}{\gamma^2}$ is a necessary but not sufficient for the gradient to explode. There are some other factors which affect the gradient, including the relation of the directions of matrices $\boldsymbol{W}$ and $\sigma'(\boldsymbol{h}_i)$(for $0 \leq i \leq T-1$).

Question 6 (6-12)**.** Denote by σ the logistic sigmoid function. Consider the following Bidirectional RNN:

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

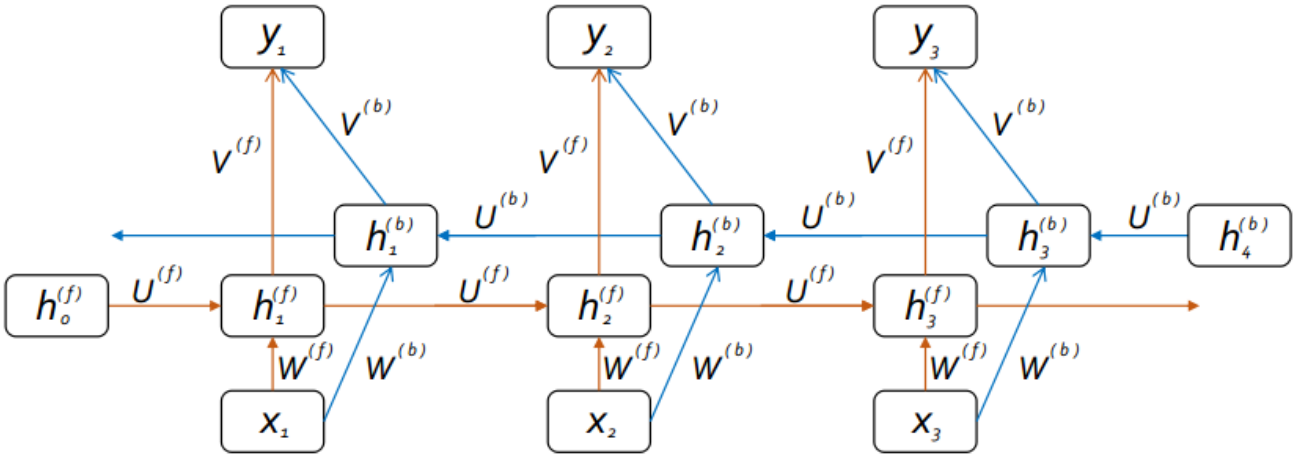
where the superscripts f and b correspond to the forward and backward RNNs respectively.

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.

*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 $\nabla_{\boldsymbol{U}^{(b)}} L$.

Answer 6.

1. The computational graph is as follow:



2. Note: denominator display of matrix derivative is used for the following derivation.
   First we have:

$$\frac{\partial L}{\partial L_t} = 1$$

then,

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

At the final time step τ, $\boldsymbol{h}_\tau^{(f)}$ only has $\boldsymbol{y}_\tau$ as a descendent, its gradient is:

$$\nabla_{\boldsymbol{h}_\tau^{(f)}} L = \boldsymbol{V}^{(f)\top} \nabla_{\boldsymbol{y}_\tau} L = 2\boldsymbol{V}^{(f)\top}(\boldsymbol{y}_\tau - \boldsymbol{z}_\tau)$$

Similarly, at the first time step ($t = 1$), $\boldsymbol{h}_1^{(b)}$ only has $\boldsymbol{y}_1$ as a descendent, its gradient is:

$$\nabla_{\boldsymbol{h}_1^{(b)}} L = \boldsymbol{V}^{(b)\top} \nabla_{\boldsymbol{y}_1} L = 2\boldsymbol{V}^{(b)\top}(\boldsymbol{y}_1 - \boldsymbol{z}_1)$$

For the time step t ($1 < t < \tau$), the gradients $\nabla_{\boldsymbol{h}_t^{(f)}} L$ and $\nabla_{\boldsymbol{h}_t^{(b)}} L$ are composed of two parts: one from the next or previous time step, and the other from the output $\boldsymbol{y}_t$.
So, we have:

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

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

Introduce dummy variables $\boldsymbol{W}_t^{(f)}$ and $\boldsymbol{U}_t^{(b)}$ that are defined to be copies of $\boldsymbol{W}^{(f)}$ and $\boldsymbol{U}^{(b)}$ respectively, but with each $\boldsymbol{W}_t^{(f)}$ and $\boldsymbol{U}_t^{(b)}$ used only at time step t.

$$\begin{aligned}\nabla_{\boldsymbol{W}^{(f)}} L &= \sum_t \nabla_{\boldsymbol{W}_t^{(f)}} L \\ &= \sum_t \Big(\frac{\partial \boldsymbol{h}_t^{(f)}}{\partial \boldsymbol{W}_t^{(f)}}\Big)^\top (\nabla_{\boldsymbol{h}_t^{(f)}} L) \\ &= \sum_t \mathrm{diag}\Big(\boldsymbol{h}_t^{(f)} - (\boldsymbol{h}_t^{(f)})^2\Big)(\nabla_{\boldsymbol{h}_t^{(f)}} L)\boldsymbol{x}_t^\top\end{aligned}$$

$$
\begin{aligned}
\nabla_{\boldsymbol{U}^{(b)}} L &= \sum_t \nabla_{\boldsymbol{U}_t^{(b)}} L \\
&= \sum_t \Big(\frac{\partial \boldsymbol{h}_t^{(b)}}{\partial \boldsymbol{U}_t^{(b)}} \Big)^\top (\nabla_{\boldsymbol{h}_t^{(b)}} L) \\
&= \sum_t \operatorname{diag}\Big(\boldsymbol{h}_t^{(b)} - (\boldsymbol{h}_t^{(b)})^2 \Big) (\nabla_{\boldsymbol{h}_t^{(b)}} L) \boldsymbol{h}_{t+1}^\top
\end{aligned}
$$