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 th 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 preacti

vations 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 $1 \le i \le d^{(t)}$. Note we are not asking for a general formula ; you just need to provide **one setting** that meets these criteria (there are many possiblities).

(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 $1 \le i \le d^{(t-1)}$. Assume entries of $\boldsymbol{h}^{(t-1)}$ are uncorrelated (the answer should not depend on g).
   (a) Gaussian: give a value for c, μ, and σ^2 as a function of $d^{(t-1)}$.
   (b) Uniform: give a value 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 $1 \le i \le d^{(t-1)}$. Assume entries of $\boldsymbol{a}^{(t-1)}$ are uncorrelated. Consider the case of ReLU activation: $g(x) = \max\{0, x\}$.
   (a) Gaussian: give a value for c, μ, and σ^2 as a function of $d^{(t-1)}$.
   (b) Uniform: give a value 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. In both questions let $c = 0$, $\mu = 0$ and $\alpha = -\beta$ where $\beta > 0$ such that $\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] = 0$ and thus the first moment condition $\mathbb{E}[\boldsymbol{a}_i^{(t)}] = 0$ is satisfied. This also means for $j \neq k$, if we write $a_{ij} = \boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)}$,

$$\mathrm{Cov}(\boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)}, \boldsymbol{W}_{ik}^{(t)} \boldsymbol{h}_k^{(t-1)}) = \mathbb{E}[(a_{ij} - \mathbb{E}[a_{ij}])(a_{ik} - \mathbb{E}[a_{ik}])] = \mathbb{E}[a_{ij} a_{ik}] = 0$$

where we use the independence and zero mean of $\boldsymbol{W}_{ij}^{(t)}$ and $\boldsymbol{W}_{ik}^{(t)}$:

$$\mathbb{E}[a_{ij}] = \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]\mathbb{E}[\boldsymbol{h}_j^{(t-1)}] = 0 \qquad \mathbb{E}[a_{ij} a_{ik}] = \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]\mathbb{E}[\boldsymbol{W}_{ik}^{(t)}]\mathbb{E}[\boldsymbol{h}_j^{(t-1)} \boldsymbol{h}_k^{(t-1)}] = 0$$

Note that the variance of $U(-b, b)$ is $\frac{(b-(-b))^2}{12} = \frac{b^2}{3}$.

1. Now the second moment condition:

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

   where the second equality is because $\mathrm{Cov}(\boldsymbol{W}_{ij}^{(t)} \boldsymbol{h}_j^{(t-1)}, \boldsymbol{W}_{ik}^{(t)} \boldsymbol{h}_k^{(t-1)}) = 0$ for $j \neq k$ (uncorrelated), and the third equality is due to $\boldsymbol{W}_{ij}^{(t)} \perp \boldsymbol{h}_j^{(t-1)}$. Since $\boldsymbol{W}_{ij}^{(t)}$'s are i.i.d., $\mathrm{Var}(\boldsymbol{W}_{ij}^{(t)}) = \frac{1}{d^{(t-1)}}$. Hence, to satisfy the second-moment constraint, we would need $\sigma^2 = \frac{1}{d^{(t-1)}}$ or $\beta = \sqrt{\frac{3}{d^{(t-1)}}}$.

2. Let $A = \{\boldsymbol{a}_i^{(t-1)} \geq 0\}$, and $\delta_A \in \{0, 1\}$ be the indicator function that is equal to 1 iff A is satisfied. First, let's observe that

$$\mathbb{E}[(\boldsymbol{h}_i^{(t-1)})^2] = \mathbb{E}[g(\boldsymbol{a}_i^{(t-1)})^2] = \mathbb{E}[g(\boldsymbol{a}_i^{(t-1)})^2 \delta_A] = \mathbb{E}[\boldsymbol{a}_i^{(t-1)2} \delta_A] = \frac{1}{2}\mathbb{E}[\boldsymbol{a}_i^{(t-1)2}] = \frac{1}{2}$$

   The second last equality is due to the symmetric distribution of $\boldsymbol{a}_i^{(t-1)}$ and the symmetric nature of the function $x \mapsto x^2$.

   Now the second moment condition:

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

Likewise, to have $\text{Var}(\boldsymbol{W}_{ij}^{(t)}) = \frac{2}{d^{(t-1)}}$, we would need $\sigma^2 = \frac{2}{d^{(t-1)}}$ or $\beta = \sqrt{\frac{6}{d^{(t-1)}}}$.

This is known as the He Initialization (which is the ReLU version of Glorot init). If the parameters are initialized this way, roughly half of the neurons are activated, and the scale of the activations is controlled (i.e. it will not grow as the dimensionality of the input increases).

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 $\boldsymbol{R}$).
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. When a particular dropout mask $\boldsymbol{R}$ is applied, we are dealing with the matrix $\boldsymbol{X}' = \boldsymbol{X} \odot \boldsymbol{R}$. The loss function is
$$L(\boldsymbol{w}) = ||\boldsymbol{y} - (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w}||^2 = ||\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{w}||^2$$
2. Let $\boldsymbol{x}_i$ and $\boldsymbol{r}_i$ be the i'th rows of $\boldsymbol{X}$ and $\boldsymbol{R}$. Note that $\mathbb{E}[\boldsymbol{r}_{ij}] = p$ and $\text{Var}(\boldsymbol{r}_{ij}) = p(1-p)$. Since the quadratic term is a sum of squared error, using the relation $\mathbb{E}[Z^2] = \text{Var}(Z) + \mathbb{E}[Z]^2$ and the fact that $\boldsymbol{r}_{ij}$ is i.i.d., we have

$$\begin{aligned}
\mathbb{E}[L(\boldsymbol{w})] &= \sum_{i=1}^{n} \mathbb{E}[(\boldsymbol{y}_i - \boldsymbol{w}^\top(\boldsymbol{x}_i \odot \boldsymbol{r}_i))^2] \\
&= \sum_{i=1}^{n} \mathbb{E}[\boldsymbol{y}_i - \boldsymbol{w}^\top(\boldsymbol{x}_i \odot \boldsymbol{r}_i)]^2 + \text{Var}(\boldsymbol{y}_i - \boldsymbol{w}^\top(\boldsymbol{x}_i \odot \boldsymbol{r}_i)) \\
&= \sum_{i=1}^{n} (\boldsymbol{y}_i - p\boldsymbol{w}^\top \boldsymbol{x}_i)^2 + p(1-p) \sum_{j=1}^{d} (\boldsymbol{w}_j \boldsymbol{x}_{ij})^2 \\
&= ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma \boldsymbol{w}||^2
\end{aligned}$$

3. Setting the gradient of the above equation to zero, we have

$$\nabla_{\boldsymbol{w}}\mathbb{E}[L(\boldsymbol{w})] = -p\boldsymbol{X}^\top(\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}) + p(1-p)\Gamma^2\boldsymbol{w} = 0$$

After rearrangement,

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

Letting $\lambda^{\text{dropout}} = \frac{1-p}{p}$ and inverting the matrix on the left yield

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

4. The solution to L^2 regularized linear regression problem is

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

5. In L^2 the weight is penalized uniformly, whereas in the case of dropout, the penalty on weight parameter $\boldsymbol{w}_i$ is scaled by the standard deviation of the i'th feature. When the probability of dropping out the units is higher (lower p), the regularization coefficient grows ; when less dropout is applied (larger p), it gets back to the ordinary least square problem, i.e. we can view dropout as scaled L^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).
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,

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

   For SGD with running average,

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

   To equalize the two, set $(\alpha, \epsilon) = (\beta, \delta(1-\beta))$, and reversely $(\beta, \delta) = (\alpha, \frac{\epsilon}{1-\alpha})$.
2. For SGD with running average we have:

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

3. Taking the expectation of the running average yields

$$\begin{aligned}\mathbb{E}[\boldsymbol{v}_t] &= \mathbb{E}\left[(1-\beta)\sum_{i=1}^{t}\beta^{t-i}\boldsymbol{g}_i\right] \\ &= (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\mathbb{E}[\boldsymbol{g}_i] && \text{(Linearity)} \\ &= \mathbb{E}[\boldsymbol{g}_t](1-\beta)\sum_{i=1}^{t}\beta^{t-i} && \text{(Stationarity)} \\ &= \mathbb{E}[\boldsymbol{g}_t](1-\beta^t) && \text{(Telescoping sum)}\end{aligned}$$

   One can use $\frac{\boldsymbol{v}_t}{1-\beta^t}$ instead as an unbiased estimate.

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

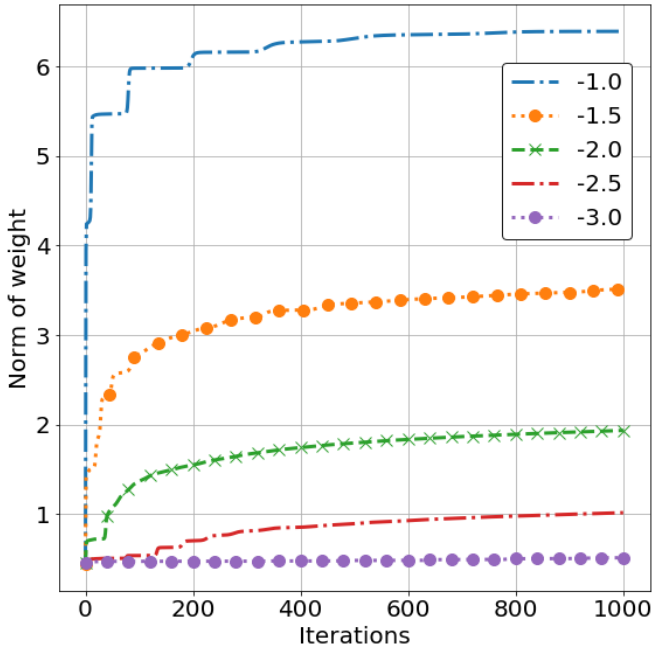

FIGURE 1 – Norm of parameters with different learning rate.

Answer 4.

1. Since the data is centered and whitened, $\mu_y = 0$ and $\sigma_y = \text{Var}(\boldsymbol{u}^\top \boldsymbol{x})^{0.5} = \mathbb{E}[\boldsymbol{u}^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{u}]^{0.5} = (\boldsymbol{u}^\top \boldsymbol{u})^{0.5} = ||\boldsymbol{u}||$. Thus,
$$\hat{y} = \gamma \cdot \frac{\boldsymbol{u}^\top \boldsymbol{x}}{||\boldsymbol{u}||} + \beta = \boldsymbol{w}^\top \boldsymbol{x} + \beta$$

2. The gradient of the norm is the unit vector itself $\nabla_{\boldsymbol{u}} ||\boldsymbol{u}|| = \frac{\boldsymbol{u}}{||\boldsymbol{u}||}$. By the quotient rule, $\frac{\partial}{\partial \boldsymbol{u}} \frac{\boldsymbol{u}}{||\boldsymbol{u}||} = \frac{||\boldsymbol{u}||\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top / ||\boldsymbol{u}||}{||\boldsymbol{u}||^2} = \frac{1}{||\boldsymbol{u}||}\left(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^\top}{||\boldsymbol{u}||^2}\right)$. Now
$$\nabla_{\boldsymbol{u}} L(\boldsymbol{u}) = \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{u}} \nabla_{\boldsymbol{w}} L = \frac{\gamma}{||\boldsymbol{u}||}\left(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^\top}{||\boldsymbol{u}||^2}\right) \nabla_{\boldsymbol{w}} L = s\boldsymbol{W}^* \nabla_{\boldsymbol{w}} L$$
where $s = \frac{\gamma}{||\boldsymbol{u}||}$ and $\boldsymbol{W}^\perp = \left(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^\top}{||\boldsymbol{u}||^2}\right)$. Also $\boldsymbol{W}^\perp \boldsymbol{u} = \boldsymbol{u} - \boldsymbol{u}\frac{\boldsymbol{u}^\top \boldsymbol{u}}{||\boldsymbol{u}||^2} = \boldsymbol{u} - \boldsymbol{u} = \boldsymbol{0}$.

3. The gradient descent update is defined as $\boldsymbol{u}_{t+1} = \boldsymbol{u}_t - \alpha \nabla_{\boldsymbol{u}_t} L$, where $\alpha > 0$ is the learning rate. From the last question, we know that $\nabla_{\boldsymbol{u}_t} L$ is perpendicular to $\boldsymbol{u}_t$. Letting $c = ||\alpha \nabla_{\boldsymbol{u}_t} L|| / ||\boldsymbol{u}_t||$, by Pythagorean theorem, we have
$$||\boldsymbol{u}_{t+1}|| = \sqrt{||\boldsymbol{u}_t||^2 + ||\alpha \nabla_{\boldsymbol{u}_t} L||^2} = ||\boldsymbol{u}_t|| \sqrt{1 + c^2}$$
which is strictly larger than $||\boldsymbol{u}_t||$ as long as $\alpha > 0$ and $\nabla_{\boldsymbol{u}_t} L$ has non-zero entries, hence the monotonic growth of $||\boldsymbol{u}||$. Furthermore, since $c \propto \frac{\alpha}{||\boldsymbol{u}_t||^2}$, the larger α is relative to $||\boldsymbol{u}_t||$, the faster the norm grows.

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.

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. $|\sigma'| \leq \gamma$ 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. Let $\boldsymbol{g}_t := \sigma(\boldsymbol{h}_t)$. Assume it is also true that $\boldsymbol{g}_{t-1} = \sigma(\boldsymbol{h}_{t-1})$. Then

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

2. For consecutive units, the Jacobian is

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

   Recall the following properties of 2-norm:

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

   from which we have

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

   which means the 2-norm is bounded by some $\delta \in [0, 1)$. Applying the sub-multiplicativity T times gives

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

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. By contraposition, gradients of the hidden state would not become arbitrarily large if the largest eigenvalue of weights is not larger than $\frac{\delta^2}{\gamma^2}$. Thus it is a necessary condition for gradient explosion. It is not sufficient: the product of the norms can be greater than the norm of the product. This can happen if the hidden state is orthogonal to the largest eigenvector of $\boldsymbol{W}$.

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. See Figure 2.

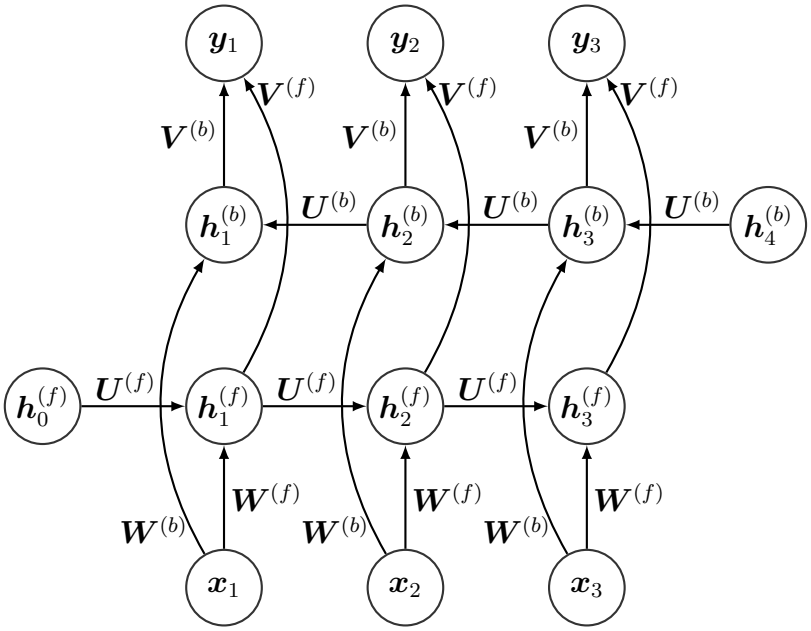

FIGURE 2 – Computational graph of the bidirectional RNN unrolled for three timesteps.

2. The gradients of L_t wrt $\boldsymbol{h}_t^{(f)}$ and $\boldsymbol{h}_t^{(b)}$ are

$$\nabla_{\boldsymbol{h}_t^{(f)}} L_t = \left(\frac{\partial \boldsymbol{y}_t}{\partial \boldsymbol{h}_t^{(f)}}\right)^\top \nabla_{\boldsymbol{y}_t} L_t = -2\boldsymbol{V}^{(f)\top} \cdot (\boldsymbol{z}_t - \boldsymbol{y}_t)$$

$$\nabla_{\boldsymbol{h}_t^{(b)}} L_t = \left(\frac{\partial \boldsymbol{y}_t}{\partial \boldsymbol{h}_t^{(b)}}\right)^\top \nabla_{\boldsymbol{y}_t} L_t = -2\boldsymbol{V}^{(b)\top} \cdot (\boldsymbol{z}_t - \boldsymbol{y}_t)$$

The Jacobian matrices of the recurrent units are

$$\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}} = \text{diag}(\boldsymbol{h}_{t+1}^{(f)}(1 - \boldsymbol{h}_{t+1}^{(f)}))\boldsymbol{U}^{(f)}$$

$$\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}} = \text{diag}(\boldsymbol{h}_{t-1}^{(b)}(1 - \boldsymbol{h}_{t-1}^{(b)}))\boldsymbol{U}^{(b)}$$

Combining the above, we can express the gradients of L wrt $\boldsymbol{h}_t^{(f)}$ and $\boldsymbol{h}_t^{(b)}$ using total derivative:

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

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

Let the subscript t denote the contribution of the weight matrices when computing the recurrent unit indexed by t, e.g. $\boldsymbol{W}_t^{(f)}$ and $\boldsymbol{U}_t^{(b)}$. Now, the gradients wrt the i'th row of $\boldsymbol{W}_t^{(f)}$ and the i'th row of $\boldsymbol{U}_t^{(b)}$ are

$$\nabla_{(\boldsymbol{W}_t^{(f)})_i} L = \left(\frac{\partial \boldsymbol{h}_t^{(f)}}{\partial (\boldsymbol{W}_t^{(f)})_i}\right)^\top \nabla_{\boldsymbol{h}_t^{(f)}} L = \left(\text{diag}(\boldsymbol{h}_t^{(f)}(1 - \boldsymbol{h}_t^{(f)}))(\boldsymbol{e}_i \boldsymbol{x}_t^\top)\right)^\top \nabla_{\boldsymbol{h}_t^{(f)}} L$$

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

where $\boldsymbol{e}_i$ is a one-hot vector with its i'th entry being 1. In total, in matrix form,

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

$$\nabla_{\boldsymbol{U}^{(b)}} L = \sum_t \nabla_{\boldsymbol{U}_t^{(f)}} L = \sum_t \text{diag}\left(\boldsymbol{h}_t^{(b)}(1 - \boldsymbol{h}_t^{(b)})\right)\left(\nabla_{\boldsymbol{h}_t^{(b)}} L\right)\boldsymbol{h}_{t+1}^\top$$