Due Date: March 17th 23:00, 2020

Instructions

- *For all questions, show your work!*
- *Starred questions are **hard** questions, not **bonus** questions.*
- *Please use a document preparation system such as LaTeX, unless noted otherwise.*
- *Unless noted that questions are related, assume that notation and defintions for each question are self-contained and independent*
- *All norms denote Euclidean norms unless otherwise specified.*
- *Submit your answers electronically via Gradescope.*
- *TAs for this assignment are **Jessica Thompson, Jonathan Cornford and Lluis Castrejon**.*

Question 1 (4-4-4)**.** In this question you will demonstrate that an estimate of the first moment of the gradient using an (exponential) running average is equivalent to using momentum, and is biased by a scaling factor. The goal of this question is for you to consider the relationship between different optimization schemes, and to practice noting and quantifying the effect (particularly in terms of bias/variance) of *estimating* a quantity.

Let $\boldsymbol{g}_t$ be an unbiased sample of gradient at time step t and $\Delta\boldsymbol{\theta}_t$ be the update to be made. Initialize $\boldsymbol{v}_0$ to be a vector of zeros.

1. For $t \geq 1$, consider the following update rules:
   - SGD with momentum:
   $$\boldsymbol{v}_t = \alpha\boldsymbol{v}_{t-1} + \epsilon\boldsymbol{g}_t \qquad \Delta\boldsymbol{\theta}_t = -\boldsymbol{v}_t$$
   where $\epsilon > 0$ and $\alpha \in (0, 1)$.
   - SGD with running average of $\boldsymbol{g}_t$:
   $$\boldsymbol{v}_t = \beta\boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t \qquad \Delta\boldsymbol{\theta}_t = -\delta\boldsymbol{v}_t$$
   where $\beta \in (0, 1)$ and $\delta > 0$.

   Express the two update rules recursively ($\Delta\boldsymbol{\theta}_t$ as a function of $\Delta\boldsymbol{\theta}_{t-1}$). Show that these two update rules are equivalent ; i.e. express (α, ϵ) as a function of (β, δ).

Solution

Replacing $\Delta\boldsymbol{\theta}_t$ in the definition of the SGD momentum we obtain

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

Doing the same process on the SGD wiht running average, i.e replacing $\Delta\boldsymbol{\theta}_t = -\delta\boldsymbol{v}_t$ we arrive to

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

They are equivalent if $\alpha = \beta$ and $\epsilon = \delta(1-\beta)$

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

Solution

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

So we can define the rule

$$\boldsymbol{v}_t = \beta^t \boldsymbol{v}_0 + (1-\beta)\sum_{j=t}^{0} \beta^j \boldsymbol{g}_{t-j}$$

Now we do the change of variable $t - j = i$ and assuming $\boldsymbol{v}_0 = 0$ we arrive to

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

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 $\boldsymbol{v}_t$.

Solution
The expected value of the last expression is given by

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

Considering $\mathbb{E}[\boldsymbol{g}_i]$ is independent of t and making the change of variable $t - i = j$ we obtain

$$\sum_{j=0}^{t-1} \beta^j = \frac{1-\beta^t}{1-\beta}$$

therefore

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

so such bias can be eliminated rescaling $\boldsymbol{v}_t$ as $\boldsymbol{v}_t = \frac{\boldsymbol{v}_t}{1-\beta^t}$

Question 2 (7-5-5-3)**.** The point of this question is to understand and compare the effects of different regularizers (specifically dropout and weight decay) on the weights of a network. Consider a linear regression problem with input data $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, weights $\boldsymbol{w} \in \mathbb{R}^{d \times 1}$ and targets $\boldsymbol{y} \in \mathbb{R}^{n \times 1}$. Suppose that dropout is applied to the input (with probability $1-p$ of dropping the unit i.e. setting it to 0). Let $\boldsymbol{R} \in \mathbb{R}^{n \times d}$ be the dropout mask such that $\boldsymbol{R}_{ij} \sim \text{Bern}(p)$ is sampled i.i.d. from the Bernoulli distribution.

For a squared error loss function with dropout, we then have:

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

1. Let Γ be a diagonal matrix with $\Gamma_{ii} = (\boldsymbol{X}^\top \boldsymbol{X})^{1/2}_{ii}$. Show that the *expectation (over $\boldsymbol{R}$)* of the loss function can be rewritten as $\mathbb{E}[L(\boldsymbol{w})] = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma\boldsymbol{w}||^2$. *Hint: Note we are trying to find the expectation over a squared term and use* $\mathrm{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2$.

Solution
Explicitly, the loss is done by

$$L(\boldsymbol{w}) = \sum_i \sum_j (y_i - x_{ij} R_{ij} w_j)^2$$

so using the hint

$$\mathbb{E}[L(\boldsymbol{w})] = \left(\mathbb{E}[\sum_i \sum_j y_i - x_{ij} R_{ij} w_j]\right)^2 + \mathrm{Var}\left(\sum_i \sum_j y_i - x_{ij} R_{ij} w_j\right) \tag{1}$$

Firstly, we calculate the expectation

$$\left(\mathbb{E}[\sum_i \sum_j y_i - x_{ij} R_{ij} w_j]\right)^2 = \left(\sum_i \sum_j y_i - p x_{ij} w_j\right)^2 = ||\boldsymbol{y} - p\boldsymbol{X}||^2 \tag{2}$$

where we used $\mathbb{E}[\boldsymbol{R}] = p$ because it is a Bernoulli distribution.
Now we calculate the variance of Eq.1

$$\mathrm{Var}\left(\sum_i \sum_j y_i - x_{ij} R_{ij} w_j\right) = \mathrm{Var}\left(\sum_i y_i - (\boldsymbol{X}_i \odot \boldsymbol{R}_i)\boldsymbol{w}\right)$$

By definition $\mathrm{Var}(Z) = \mathbb{E}[(Z - \mathbb{E}[Z])^2]$ and using the result from Eq.2

$$\mathrm{Var}\left(\sum_i y_i - (\boldsymbol{X}_i \odot \boldsymbol{R}_i)\boldsymbol{w}\right) = \mathbb{E}\left[\left(\sum_i y_i - (\boldsymbol{X}_i \odot \boldsymbol{R}_i)\boldsymbol{w} - \sum_i y_i - p\boldsymbol{X}_i\boldsymbol{w}\right)^2\right] \tag{3}$$

$$= \mathbb{E}\left[\left(\sum_i p\boldsymbol{X}_i\boldsymbol{w} - (\boldsymbol{X}_i \odot \boldsymbol{R}_i)\boldsymbol{w}\right)^2\right] \tag{4}$$

$$= \mathbb{E}\left[\left(\sum_i p\boldsymbol{X}_i\boldsymbol{w} - (\boldsymbol{X}_i \odot \boldsymbol{R}_i)\boldsymbol{w}\right)^T \left(\sum_i p\boldsymbol{X}_i\boldsymbol{w} - (\boldsymbol{X}_i \odot \boldsymbol{R}_i)\boldsymbol{w}\right)\right] \tag{5}$$

$$= \sum_i \boldsymbol{w}^T(-p^2\boldsymbol{X}_i^T\boldsymbol{X}_i + \boldsymbol{X}_i^T\boldsymbol{X}_i\mathbb{E}[\boldsymbol{R}_i^T\boldsymbol{R}_i])\boldsymbol{w} \tag{6}$$

Finally we define the value $\mathbb{E}[\boldsymbol{R}^T\boldsymbol{R}] = \mathrm{Var}(\boldsymbol{R}) + \mathbb{E}[\boldsymbol{R}]^2 = p(1-p) + p^2 = p$ and replacing on Eq.6, we arrive to

$$\begin{aligned}\mathrm{Var}\left(\sum_i y_i - (\boldsymbol{X}_i \odot \boldsymbol{R}_i)\boldsymbol{w}\right) &= p(1-p)\sum_i \boldsymbol{w}^T\boldsymbol{X}_i^T\boldsymbol{X}_i\boldsymbol{w} \\ &= p(1-p)\sum_i \boldsymbol{w}^T\Gamma_{ii}\boldsymbol{w} \\ &= p(1-p)||\boldsymbol{\Gamma}\boldsymbol{w}||\end{aligned}$$

Joining this result with Eq.2 we conclude that

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

2. Show that the solution $\boldsymbol{w}^{\text{dropout}}$ that minimizes the expected loss from question 2.1 satisfies

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

where λ^{dropout} is a regularization coefficient depending on p. How does the value of p affect the regularization coefficient, λ^{dropout} ?

Solution From the solution above

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

therefore

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

making the derivative to zero and doing the algebraic operations on the last equation we find that

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

Finally

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

with $\lambda^{drop} \equiv \frac{1-p}{p}$. If $p = 1$ we do not have regularization, if p is very small the regularization very high.

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

Solution In this case, the lost is defined as

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

i.e

$$L(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \lambda^{L_2}\boldsymbol{w}^T\boldsymbol{w} \Rightarrow$$
$$\frac{\partial L}{\partial \boldsymbol{w}} = -(\boldsymbol{y}^T - \boldsymbol{w}^T\boldsymbol{X}^T)\boldsymbol{X} + \lambda^{L_2}\boldsymbol{w}^T$$

Making the derivative equal to zero we obtain the condition

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

4. Compare the results of 2.2 and 2.3: identify specific differences in the equations you arrive at, and discuss qualitatively what the equations tell you about the similarities and differences in the effects of weight decay and dropout (1-3 sentences).

Solution In the standard L_2 the weights are penalized uniformly. In dropout we must use a mask so the weights are nor penalized uniformly.

Question 3 (6-10-2)**.** The goal of this question is for you to understand the reasoning behind different parameter initializations for deep networks, particularly to think about the ways that the initialization affects the activations (and therefore the gradients) of the network. Consider the following equation for the t-th layer of a deep network:

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

where $\boldsymbol{a}^{(t)}$ are the pre-activations and $\boldsymbol{h}^{(t)}$ are the activations for layer t, g is an activation function, $\boldsymbol{W}^{(t)}$ is a $d^{(t)} \times d^{(t-1)}$ matrix, and $\boldsymbol{b}^{(t)}$ is a $d^{(t)} \times 1$ bias vector. The bias is initialized as a constant vector $\boldsymbol{b}^{(t)} = [c, .., c]^\top$ for some $c \in \mathbb{R}$, and the entries of the weight matrix are initialized by sampling i.i.d. from a Gaussian distribution $W_{ij}^{(t)} \sim \mathcal{N}(\mu, \sigma^2)$.

Your task is to design an initialization scheme that would achieve a vector of **pre-activations** at layer t whose elements are zero-mean and unit variance (i.e.: $\mathbb{E}[a_i^{(t)}] = 0$ and $\text{Var}(a_i^{(t)}) = 1$, $1 \le i \le d^{(t)}$) for the assumptions about either the activations or pre-activations of layer $t-1$ listed below. Note we are not asking for a general formula ; you just need to provide one setting that meets these criteria (there are many possiblities).

1. First assume that the activations of the previous layer satisfy $\mathbb{E}[h_i^{(t-1)}] = 0$ and $\text{Var}(h_i^{(t-1)}) = 1$ for $1 \le i \le d^{(t-1)}$. Also, assume entries of $\boldsymbol{h}^{(t-1)}$ are uncorrelated (the answer should not depend on g).
   (a) Show $\text{Var}(XY) = \text{Var}(X)\text{Var}(Y) + \text{Var}(X)\mathbb{E}[Y]^2 + \text{Var}(Y)\mathbb{E}[X]^2$ when $X \perp Y$

   Solution
   By definition

   $$\text{Var}(XY) = \mathbb{E}[X^2Y^2] - \mathbb{E}[XY]^2 = \mathbb{E}[X^2]\mathbb{E}[Y^2] - (\mathbb{E}[X]\mathbb{E}[Y])^2 \tag{7}$$

   On the other hand, considering the definition of variance, we can say that

   $$\begin{aligned} \text{Var}(X)\mathbb{E}[Y]^2 &= (\mathbb{E}[X^2] - \mathbb{E}[X]^2)\mathbb{E}[Y]^2 = \mathbb{E}[X^2]\mathbb{E}[Y]^2 - \mathbb{E}[X]^2\mathbb{E}[Y]^2 \\ \text{Var}(Y)\mathbb{E}[X]^2 &= (\mathbb{E}[Y^2] - \mathbb{E}[Y]^2)\mathbb{E}[X]^2 = \mathbb{E}[Y^2]\mathbb{E}[X]^2 - \mathbb{E}[Y]^2\mathbb{E}[X]^2 \end{aligned} \tag{8}$$

   And using again the definition of variance, we have

   $$\text{Var}(X)\text{Var}(Y) = \mathbb{E}[X^2]\mathbb{E}[Y^2] - \mathbb{E}[X^2]\mathbb{E}[Y]^2 - \mathbb{E}[X]^2\mathbb{E}[Y^2] + \mathbb{E}[X]^2\mathbb{E}[Y]^2 \tag{9}$$

   Therefore adding Eq.9 with Eq.8 we obtain

   $$\text{Var}(X)\text{Var}(Y) + \text{Var}(X)\mathbb{E}[Y]^2 + \text{Var}(Y)\mathbb{E}[X]^2 = \mathbb{E}[X^2]\mathbb{E}[Y^2] - \mathbb{E}[X]^2\mathbb{E}[Y]^2 \tag{10}$$

   which is just Eq.7, therefore

   $$\text{Var}(XY) = \text{Var}(X)\text{Var}(Y) + \text{Var}(X)\mathbb{E}[Y]^2 + \text{Var}(Y)\mathbb{E}[X]^2 \tag{11}$$

   (b) Write $\mathbb{E}[a_i^{(t)}]$ and $\text{Var}(a_i^{(t)})$ in terms of $c, \mu, \sigma^2, \text{Var}(h_i^{(t-1)}), \mathbb{E}[h_i^{(t-1)}]$.
   Solution
   We know that $a_i^{(t)} = \boldsymbol{W}_i^{(t)}\boldsymbol{h}^{(t-1)} + b_i$ so

   $$\mathbb{E}[a_i^{(t)}] = \mathbb{E}[\boldsymbol{W}_i^{(t)}]\mathbb{E}[\boldsymbol{h}^{(t-1)}] + \mathbb{E}[b_i] \tag{12}$$

   $$\mathbb{E}[a_i^{(t)}] = \mu \mathbf{1}_{d^{(t-1)}}^T \mathbb{E}[\boldsymbol{h}^{(t-1)}] + c \tag{13}$$

   $$\mathbb{E}[a_i^{(t)}] = \mu d^{(t-1)} \mathbb{E}[\boldsymbol{h}_i^{(t-1)}] + c \tag{14}$$

because $W_{ij}^{(t)} \sim \mathcal{N}(\mu, \sigma^2)$ and $\mathbb{E}[b_i]$ is a constant c. Remark that each component of the matrix $\boldsymbol{W}$ has a normal distribution, so the expected value of a row $\boldsymbol{W}_i$ is $\mu \mathbf{1}_{d^{(t-1)}}^T$, where $\mathbf{1}_{d^{(t-1)}}^T$ is a row vector of ones with the dimensions of the layer $t-1$. In the same way, we assume each component of the vector $\boldsymbol{h}$ has the same distribution so $\mathbb{E}[\boldsymbol{h}] = \mathbf{1}_{d^{(t-1)}}\mathbb{E}[\boldsymbol{h}_i]$, with $\mathbf{1}_{d^{(t-1)}}$ a column vector of ones with the dimensions of the layer $t-1$. Using the Eq. 11 the variance of $a_i^{(t)}$ is

$$\mathrm{Var}(a_i^{(t)}) = \mathrm{Var}(\boldsymbol{W}_i^{(t)}\boldsymbol{h}^{(t-1)}) + \mathrm{Var}(b_i^t) \tag{15}$$

$$\mathrm{Var}(a_i^{(t)}) = \mathrm{Var}(\boldsymbol{W}_i^{(t)})\mathrm{Var}(\boldsymbol{h}^{(t-1)}) + \mathrm{Var}(\boldsymbol{W}_i^{(t)})\mathbb{E}[\boldsymbol{h}^{(t-1)}]^2 + \mathrm{Var}(\boldsymbol{h}^{(t-1)})\mathbb{E}[\boldsymbol{W}_i^{(t)}]^2 \tag{16}$$

$$\mathrm{Var}(a_i^{(t)}) = \sigma^2 d^{(t-1)}\mathrm{Var}(\boldsymbol{h}_i^{(t-1)}) + \sigma^2 d^{(t-1)}\mathbb{E}[\boldsymbol{h}_i^{(t-1)}]^2 + \mu^2 d^{(t-1)}\mathrm{Var}(\boldsymbol{h}_i^{(t-1)}) \tag{17}$$

In particular if $\mathbb{E}[\boldsymbol{h}_i^{t-1}] = 0$ and $\mathrm{Var}(\boldsymbol{h}_i^{t-1}) = 1$ we have from Eq.14 and Eq.17

$$\mathbb{E}[a_i^{(t)}] = c \tag{18}$$

and

$$\mathrm{Var}(a_i^{(t)}) = (\sigma^2 + \mu^2)(d^{(t-1)}) \tag{19}$$

(c) Give values for c, μ, and σ^2 as a function of $d^{(t-1)}$ such that $\mathbb{E}[a_i^{(t)}] = 0$ and $\mathrm{Var}(a_i^{(t)}) = 1$ for $1 \leq i \leq d^{(t)}$.

Solution
From Eq.18 $\mathbb{E}[a_i^{(t)}] = 0$ if $c = 0$ and a condition to have $\mathrm{Var}(a_i^{(t)}) = 1$ assuming $\mu = 0$ from Eq.19 is

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

2. Now assume that the pre-activations of the previous layer satisfy $\mathbb{E}[a_i^{(t-1)}] = 0$, $\mathrm{Var}(a_i^{(t-1)}) = 1$ and $a_i^{(t-1)}$ has a symmetric distribution for $1 \leq i \leq d^{(t-1)}$. Assume entries of $\boldsymbol{a}^{(t-1)}$ are uncorrelated. Consider the case of ReLU activation: $g(x) = \max\{0, x\}$.

(a) Derive $\mathbb{E}[(h_i^{(t-1)})^2]$

Solution

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

Using the definition of expectation and considering $f_p(a_i^{(t-1)})$ as the probability density of function, we have

$$\begin{aligned}
\mathbb{E}[(h_i^{(t-1)})^2] &= \int_{-\infty}^{\infty} \max\{0, a_i^{(t-1)}\}^2 f_p(a_i^{(t-1)})\, da_i^{(t-1)} \\
&= \int_{0}^{\infty} (a_i^{(t-1)})^2 f_p(a_i^{(t-1)})\, da_i^{(t-1)} \\
&= \frac{1}{2}\int_{-\infty}^{\infty} (a_i^{(t-1)})^2 f_p(a_i^{(t-1)})\, da_i^{(t-1)} \quad * \\
&= \frac{1}{2}\mathbb{E}[(a_i^{(t-1)})^2]
\end{aligned}$$

In the step $*$ we consider that the argument of the integral is an even function. By definition

$$\mathbb{E}[(a_i^{(t-1)})^2] = \left[\text{Var}(a_i^{(t-1)}) + \mathbb{E}[(a_i^{(t-1)})]^2\right] = 1$$

Therefore, we conclude that

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

(b) Using the result from (a), give values for c, μ, and σ^2 as a function of $d^{(t-1)}$ such that $\mathbb{E}[a_i^{(t)}] = 0$ and $\text{Var}(a_i^{(t)}) = 1$ for $1 \leq i \leq d^{(t)}$.

Solution

From Eq.12 we know that

$$\mathbb{E}[a_i^{(t)}] = \mathbb{E}[\boldsymbol{W}_i^{(t)}]\mathbb{E}[\boldsymbol{h}^{(t-1)}] + \mathbb{E}[b_i^{(t)}] \tag{20}$$

$$= \mathbb{E}[W_{ij}^{(t)}]\mathbf{1}_{d^{(t-1)}}^T\mathbb{E}[\boldsymbol{h}_i^{(t-1)}]\mathbf{1}_{d^{(t-1)}} + c \tag{21}$$

and considering $\mathbb{E}[a_i^{(t)}] = 0$ we arrive to

$$\mathbb{E}[\boldsymbol{h}_i^{(t-1)}] = -\frac{c}{\mathbb{E}[W_{ij}]}d^{(t-1)} \tag{22}$$

Now we will calculate

$$\text{Var}(a_i^{(t)}) = \text{Var}(\boldsymbol{W}_i^{(t)}\boldsymbol{h}^{(t-1)})$$

using the property shown in Eq.11

$$\text{Var}(a_i^{(t)}) = \text{Var}(\boldsymbol{W}_i^{(t)})\text{Var}(\boldsymbol{h}^{(t-1)}) + \text{Var}(\boldsymbol{W}_i^{(t)})\mathbb{E}[\boldsymbol{h}^{(t-1)}]^2 + \text{Var}(\boldsymbol{h}^{(t-1)})\mathbb{E}[\boldsymbol{W}_i^{(t)}]^2 \tag{23}$$

Considering that $\mathbb{E}[(\boldsymbol{h}_i^{(t-1)})^2] = \frac{1}{2}$ (item a) and considering that

$$\text{Var}[\boldsymbol{h}_i^{(t-1)}] = \mathbb{E}[(h_i^{(t-1)})^2] - \mathbb{E}[(h_i^{(t-1)})]^2$$

we obtain

$$\text{Var}[\boldsymbol{h}_i^{(t-1)}] = \frac{1}{2} - \mathbb{E}[h_i^{(t-1)}]^2$$

Replacing on Eq.23:

$$\begin{aligned}\text{Var}(a_i^{(t)}) =& \text{Var}(W_{ij}^{(t)})\left(\frac{1}{2} - \mathbb{E}[h_i^{(t-1)}]^2\right)d^{(t-1)} + \text{Var}(W_{ij}^{(t)})\mathbb{E}[h_i^{(t-1)}]^2 d^{(t-1)} \\ &+ \mathbb{E}[W_{ij}^{(t)}]^2\left(\frac{1}{2} - \mathbb{E}[h_i^{(t-1)}]^2\right)d^{(t-1)}\end{aligned}$$

and tanking into account that $\text{Var}(a_i^{(t)}) = 1$ and Eq.22 we arrive to

$$1 = \frac{1}{2}d^{(t-1)}\left[\text{Var}(W_{ij}^{(t)}) + \mathbb{E}[W_{ij}^{(t)}]^2\right] - \frac{c^2}{d^{(t-1)}} \tag{24}$$

If we have a Gaussian distribution $\text{Var}(W_{ij}^{(t)}) = \sigma^2$ and $\mathbb{E}[W_{ij}^{(t)}] = \mu$ Substituting on Eq.24 we achieve

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

Additionally, if $c = 0$, and $\mu = 0$ we obtain

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

(c) What popular initialization scheme has this form ?

Solution The scheme above is known as He normal distribution.

(d) Why do you think this initialization would work well in practice ? Answer in 1-2 sentences.
Solution
Using this scheme (which is the ReLu version of Glorot) we have more control of the number of neurons that are activated and the scale of activation.

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

Solution

Assumption 1
As $a_i^{(t)} = \boldsymbol{W}_i^{(t)} \boldsymbol{h}^{(t-1)} + b_i^{(t)}$ we have

$$\mathbb{E}[a_i^{(t)}] = \mathbb{E}[\boldsymbol{W}_i^{(t)}]\mathbb{E}[\boldsymbol{h}^{(t-1)}] + \mathbb{E}[b_i^{(t)}] \Rightarrow$$
$$\mathbb{E}[a_i^{(t)}] = c$$

Now

$$\begin{aligned}\text{Var}(a_i^{(t)}) &= \text{Var}(\boldsymbol{W}_i^{(t)})\text{Var}(\boldsymbol{h}^{(t-1)}) + \text{Var}(\boldsymbol{W}_i^{(t)})\mathbb{E}[\boldsymbol{h}^{(t-1)}]^2 + \text{Var}(\boldsymbol{h}^{(t-1)})\mathbb{E}[\boldsymbol{W}_i^{(t)}]^2 \\ &= \text{Var}(W_{ij}^{(t)})\mathbf{1}_{d^{(t-1)}}^T \text{Var}(h_i^{(t-1)})\mathbf{1}_{d^{(t-1)}} + \mathbb{E}[W_{ij}^{(t)}]^2 \mathbf{1}_{d^{(t-1)}}^T \text{Var}(h_i^{(t-1)})\mathbf{1}_{d^{(t-1)}}\end{aligned}$$

Assuming $\text{Var}(a_i^{(t)}) = 1$ we find the condition

$$1 = \text{Var}(W_{ij}^{(t)})d^{(t-1)} + \mathbb{E}[W_{ij}^{(t)}]^2 d^{(t-1)} \tag{25}$$

For solving this equation we must calculate $\mathbb{E}[W_{ij}^{(t)}]$ to find $\text{Var}(W_{ij})$ using

$$\text{Var}(W_{ij}) = \mathbb{E}[(W_{ij}^{(t)})^2] - \mathbb{E}[W_{ij}^{(t)}]^2 \tag{26}$$

with $W_{ij} \sim U(\alpha, \beta)$ and

$$U(x) = \begin{cases} \frac{1}{\beta - \alpha}, \text{ if } \alpha \leq x \leq \beta \\ 0, \text{ otherwise} \end{cases}$$

In this order of ideas we can calculate the integrals

$$\mathbb{E}[W_{ij}^{(t)}] = \frac{1}{\beta - \alpha} \int_\alpha^\beta x \, dx = \frac{\beta + \alpha}{2} \tag{27}$$

$$\mathbb{E}[(W_{ij}^{(t)})^2] = \frac{1}{\beta - \alpha} \int_\alpha^\beta x^2 \, dx = \frac{\beta^2 + \alpha\beta + \alpha^2}{3} \tag{28}$$

Replacing Eq.27 and Eq.28 on Eq.26 we obtain

$$\text{Var}(W_{ij}) = \frac{(\beta - \alpha)^2}{12}$$

and finally substituting on Eq.25

$$\frac{1}{d^{(t-1)}} = \frac{(\beta-\alpha)^2}{12} + \frac{(\beta+\alpha)^2}{4}$$
$$\frac{3}{d^{(t-1)}} = \beta^2 + \alpha\beta + \alpha^2$$

A solution could be $\alpha = -\sqrt{\frac{3}{d^{(t-1)}}}$, $\beta = \sqrt{\frac{3}{d^{(t-1)}}}$ which is a glorot distribution.

Assumption 2
From the Eq.22 we know that $\mathbb{E}[a_i^{(t)}] = 0$ if $c = 0$ and from Eq.24

$$1 = \frac{1}{2} d^{(t-1)} \left[\mathrm{Var}(W_{ij}^{(t)}) + \mathbb{E}[W_{ij}^{(t)}]^2 \right] - \frac{c^2}{d^{(t-1)}} \Rightarrow$$
$$1 = \frac{1}{2} d^{(t-1)} \left[\mathrm{Var}(W_{ij}^{(t)}) + \mathbb{E}[W_{ij}^{(t)}]^2 \right] \Rightarrow$$
$$1 = \frac{1}{2} d^{(t-1)} \left[\frac{(\beta-\alpha)^2}{12} + \frac{(\beta+\alpha)^2}{4} \right] \Rightarrow$$
$$\frac{6}{d^{(t-1)}} = \beta^2 + \alpha\beta + \alpha^2$$

A solution could be $c = 0$, $\alpha = -\sqrt{\frac{6}{d^{(t-1)}}}$, $\beta = \sqrt{\frac{6}{d^{(t-1)}}}$ it is known as He uniform distribution.

Question 4 (4-6-6)**.** This question is about normalization techniques.

1. Batch normalization, layer normalization and instance normalization all involve calculating the mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma^2}$ with respect to different subsets of the tensor dimensions. Given the following 3D tensor, calculate the corresponding mean and variance tensors for each normalization technique: $\boldsymbol{\mu}_{batch}$, $\boldsymbol{\mu}_{layer}$, $\boldsymbol{\mu}_{instance}$, $\boldsymbol{\sigma}^2_{batch}$, $\boldsymbol{\sigma}^2_{layer}$, and $\boldsymbol{\sigma}^2_{instance}$.

$$\left[\begin{bmatrix} 1,3,2 \\ 1,2,3 \end{bmatrix}, \begin{bmatrix} 3,3,2 \\ 2,4,4 \end{bmatrix}, \begin{bmatrix} 4,2,2 \\ 1,2,4 \end{bmatrix}, \begin{bmatrix} 3,3,2 \\ 3,3,2 \end{bmatrix} \right]$$

The size of this tensor is 4 x 2 x 3 which corresponds to the batch size, number of channels, and number of features respectively.

Solution

- **Batch normalization**: Considering the tensor of the exercise as $\boldsymbol{X} \in \mathbb{R}^{T \times C \times W}$ where T is the number of batches (4), C the number of channels (2), and W the number of features (3) the mean batch normalization $\boldsymbol{\mu}$ is defined as

$$\mu_i = \frac{1}{TW} \sum_{t=1}^{T} \sum_{l=1}^{W} X_{til} = \frac{1}{12} \sum_{t=1}^{4} \sum_{l=1}^{3} X_{til}$$

doing the calculations we obtain $\boldsymbol{\mu}_{batch} = [2.5, 2.583]$. On the other hand, the batch variance $\boldsymbol{\sigma}^{\boldsymbol{2}}_{bacth}$ is defined as

$$\sigma_i^2 = \frac{1}{TW} \sum_{t=1}^{T} \sum_{l=1}^{W} (X_{til} - \mu_i)^2 = \frac{1}{12} \sum_{t=1}^{4} \sum_{l=1}^{3} (X_{til} - \mu_i)^2$$

obtaining $\boldsymbol{\sigma}^{\boldsymbol{2}}_{batch} = [0.583, 1.076]$

- **Layer normalization**: Using the same notation, mean layer normalization $\boldsymbol{\mu}_{layer}$ is defined as

$$\mu_t = \frac{1}{CW}\sum_{i=1}^{C}\sum_{l=1}^{W} X_{til} = \frac{1}{6}\sum_{i=1}^{2}\sum_{l=1}^{3} X_{til}$$

arriving to $\boldsymbol{\mu}_{layer} = [2.0, 3.0, 2.5, 2.67]$. The layer variance is defined as

$$\sigma_t^2 = \frac{1}{CW}\sum_{i=1}^{C}\sum_{l=1}^{W}(X_{til} - \mu_t)^2 = \frac{1}{6}\sum_{i=1}^{2}\sum_{l=1}^{3}(X_{til} - \mu_t)^2$$

obtaining $\boldsymbol{\sigma}^{\mathbf{2}}_{layer} = [0.67, 0.67, 1.25, 0.22]$

- **Instance normalization:** Finally the mean instance normalization is defined as

$$\mu_{ti} = \frac{1}{W}\sum_{l=1}^{W} X_{til} = \frac{1}{3}\sum_{l=1}^{3} X_{til}$$

so the matrix $\boldsymbol{\mu}_{instance}$ is

$$\boldsymbol{\mu}_{instance} = \begin{bmatrix} 2.0, 2.0 \\ 2.67, 3.33 \\ 2.67, 3.33 \\ 2.67, 2.67 \end{bmatrix}$$

and the variance is defined as

$$\sigma_{ti} = \frac{1}{W}\sum_{l=1}^{W}(X_{til} - \mu_{ti})^2 = \frac{1}{3}\sum_{l=1}^{3}(X_{til} - \mu_{ti})^2$$

achieving the result

$$\boldsymbol{\sigma}^{\mathbf{2}}_{instance} = \begin{bmatrix} 0.67, 0.67 \\ 0.22, 0.88 \\ 0.88, 1.55 \\ 0.22, 0.22 \end{bmatrix}$$

2. For the next two subquestions, we consider the following parametrization of a weight vector $\boldsymbol{w}$:

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

where γ is scalar parameter controlling the magnitude and $\boldsymbol{u}$ is a vector controlling the direction of $\boldsymbol{w}$.

Consider one layer of a neural network, and omit the bias parameter. To carry out batch normalization, one normally standardizes the preactivation and performs elementwise scale and shift $\hat{y} = \gamma \cdot \frac{y-\mu_y}{\sigma_y} + \beta$ where $y = \boldsymbol{u}^\top \boldsymbol{x}$. Assume the data $\boldsymbol{x}$ (a random vector) is whitened ($\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{I}$) and centered at 0 ($\mathbb{E}[\boldsymbol{x}] = \mathbf{0}$). Show that $\hat{y} = \boldsymbol{w}^\top \boldsymbol{x} + \beta$.

Solution
We know that $y = \boldsymbol{u}^T\boldsymbol{x}$ so $\mathrm{Var}(y) = \mathrm{Var}(\boldsymbol{u}^T\boldsymbol{x})$, using the definition of variance

$$\mathrm{Var}(\boldsymbol{u}^T\boldsymbol{x}) = \mathbb{E}[(\boldsymbol{u}^T\boldsymbol{x})^2] - \mathbb{E}[\boldsymbol{u}^T\boldsymbol{x}]$$

but $\mathbb{E}[\boldsymbol{x}] = 0$ therefore

$$\text{Var}(\boldsymbol{u}^T\boldsymbol{x}) = \mathbb{E}[(\boldsymbol{u}^T\boldsymbol{x})^2] = \mathbb{E}[(\boldsymbol{u}^T\boldsymbol{x})(\boldsymbol{x}^T\boldsymbol{u})] = ||\boldsymbol{u}||$$

On the other hand, we know that $\text{Var}(\boldsymbol{u}^T\boldsymbol{x}) = \sigma_y$ so $\sigma_y = ||\boldsymbol{u}||$. By definition in the exercise $\hat{y} = \gamma \cdot \frac{y-\mu_y}{\sigma_y} + \beta$ so

$$\hat{y} = \gamma \cdot \frac{y - \mu_y}{\sigma_y} + \beta = \gamma \left(\frac{\boldsymbol{u}^T\boldsymbol{x} - \mu_y}{||\boldsymbol{u}||} \right) + \beta$$

but $\mu_y = \mathbb{E}[y] = 0$ and we conclude that

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

3. Show that the gradient of a loss function $L(\boldsymbol{u}, \gamma, \beta)$ with respect to $\boldsymbol{u}$ can be written in the form $\nabla_{\boldsymbol{u}} L = s\boldsymbol{W}^{\perp}\nabla_{\boldsymbol{w}} L$ for some s, where $\boldsymbol{W}^{\perp} = \left(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^{\top}}{||\boldsymbol{u}||^2}\right)$. Note that [1] $\boldsymbol{W}^{\perp}\boldsymbol{u} = \boldsymbol{0}$.

Solution
Applying chain rule $\nabla_{\boldsymbol{u}} L = \nabla_{\boldsymbol{w}} L \, \nabla_{\boldsymbol{u}} \boldsymbol{W}$. Furthermore, $\boldsymbol{W} = \gamma \frac{\boldsymbol{u}}{||\boldsymbol{u}||} = \gamma \frac{\boldsymbol{u}}{(\boldsymbol{u}^T\boldsymbol{u})^{1/2}}$ so

$$\begin{aligned}
\nabla_{\boldsymbol{u}} \boldsymbol{W} &= \frac{\gamma}{(\boldsymbol{u}^T\boldsymbol{u})} \left[(\boldsymbol{u}^T\boldsymbol{u})^{1/2} - \frac{1}{2} \frac{\boldsymbol{u}}{(\boldsymbol{u}^T\boldsymbol{u})^{1/2}} d(\boldsymbol{u}^T\boldsymbol{u}) \right] \\
&= \frac{\gamma}{||\boldsymbol{u}||^2} \left[||\boldsymbol{u}|| - \frac{1}{2} \frac{\boldsymbol{u}}{||\boldsymbol{u}||} 2\boldsymbol{u}^T \right] \\
&= \frac{\gamma}{||\boldsymbol{u}||} \left[\mathbf{I} - \frac{\boldsymbol{u}\boldsymbol{u}^T}{||\boldsymbol{u}||^2} \right]
\end{aligned}$$

and we arrive to $\nabla_{\boldsymbol{u}} L = s\boldsymbol{W}^{\perp}\nabla_{\boldsymbol{W}} L$ with $s \equiv \frac{\gamma}{||\boldsymbol{u}||}$ and $\boldsymbol{W}^{\perp} = \left(\mathbf{I} - \frac{\boldsymbol{u}\boldsymbol{u}^T}{||\boldsymbol{u}||^2}\right)$

Question 5 (4-6-4)**.** This question is about activation functions and vanishing/exploding gradients in recurrent neural networks (RNNs). Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function. When the argument is a vector, we apply σ element-wise. Consider the following recurrent unit:

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

1. Show that applying the activation function in this way is equivalent to the conventional way of applying the activation function: $\boldsymbol{g}_t = \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$ (i.e. express $\boldsymbol{g}_t$ in terms of $\boldsymbol{h}_t$). More formally, you need to prove it using mathematical induction. You only need to prove the induction step in this question, assuming your expression holds for time step $t - 1$.

Solution We have that

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

if we assume $\boldsymbol{g}_{t-1} = \sigma(\boldsymbol{h}_{t-1})$ we obtain $\boldsymbol{g}_t = \sigma(\boldsymbol{h}_t)$

1. As a side note: $\boldsymbol{W}^{\perp}$ is an orthogonal complement that projects the gradient away from the direction of $\boldsymbol{w}$, which is usually (empirically) close to a dominant eigenvector of the covariance of the gradient. This helps to condition the landscape of the objective that we want to optimize.

*2. 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.

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

Use the following properties of the L_2 operator norm

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

Solution We begin with the condition $\boldsymbol{h}_t = \boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}$, so we can say that

$$\begin{aligned} \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_0} &= \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} \frac{\partial \boldsymbol{h}_{t-1}}{\partial \boldsymbol{h}_0} \\ &= \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} \frac{\partial \boldsymbol{h}_{t-1}}{\partial \boldsymbol{h}_{t-2}} \frac{\partial \boldsymbol{h}_{t-2}}{\partial \boldsymbol{h}_0} \\ &= \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} \frac{\partial \boldsymbol{h}_{t-1}}{\partial \boldsymbol{h}_{t-2}} \cdots \frac{\partial \boldsymbol{h}_1}{\partial \boldsymbol{h}_0} \end{aligned} \tag{29}$$

but every derivative in the equation above satisfies

$$\left|\left|\frac{\partial \boldsymbol{h}_{t-i}}{\partial \boldsymbol{h}_{t-i-1}}\right|\right| = \left|\left|\boldsymbol{W}\frac{\partial \sigma}{\partial \boldsymbol{h}_{t-i-1}}\right|\right| \leq ||\boldsymbol{W}|| \left|\left|\frac{\partial \sigma}{\partial \boldsymbol{h}_{t-i-1}}\right|\right|$$

but $||\boldsymbol{W}|| = \sqrt{\lambda_1 \boldsymbol{W}^T \boldsymbol{W}}$, and $\left|\left|\frac{\partial \sigma}{\partial \boldsymbol{h}_{t-i-1}}\right|\right| \leq \delta$, therefore using Eq.29 we obtain the result

$$\begin{aligned} \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_0} &= \delta \times \delta \times \cdots \delta \ (T \text{ times}) \\ &= \delta^T \end{aligned}$$

As $0 \leq \delta < 1$ we conclude that

$$\left|\left|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_0}\right|\right| \to 0$$

when $T \to \infty$

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).

Solution
From the last exercise we have the condition

$$\left|\left|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_0}\right|\right| \leq \sqrt{\lambda_1 \boldsymbol{W}^T \boldsymbol{W}}\gamma$$

But if $\lambda_1 > \frac{\delta^2}{\gamma^2}$ we do not have a conclusive bound condition for the norm of $\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_0}$, so this condition must be a necessary condition for gradient exploding but it is not sufficient.

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.

Question 6 (4-8-8)**.** Consider the following Bidirectional RNN:

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

where the superscripts f and b correspond to the forward and backward RNNs respectively and σ denotes the logistic sigmoid function. Let $\boldsymbol{z}_t$ be the true target of the prediction $\boldsymbol{y}_t$ and consider the sum of squared loss $L = \sum_t L_t$ where $L_t = ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2$.

In this question our goal is to obtain an expression for the gradients $\nabla_{\boldsymbol{W}^{(f)}} L$ and $\nabla_{\boldsymbol{U}^{(b)}} L$.

1. First, complete the following computational graph for this RNN, unrolled for 3 time steps (from $t = 1$ to $t = 3$). Label each node with the corresponding hidden unit and each edge with the corresponding weight. Note that it includes the initial hidden states for both the forward and backward RNNs.

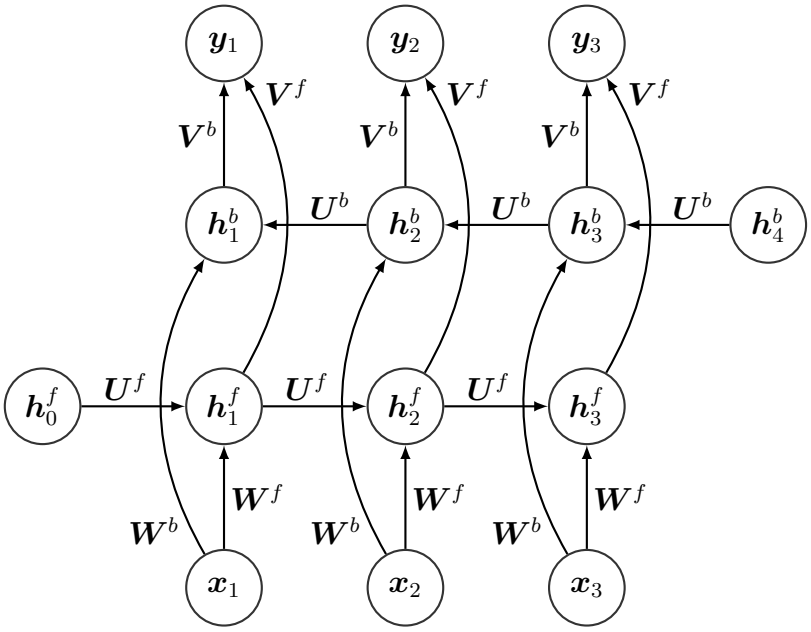

FIGURE 1 – Computational graph of the bidirectional RNN unrolled for three timesteps.

2. Using total derivatives we can express the gradients $\nabla_{\boldsymbol{h}_t^{(f)}} L$ and $\nabla_{\boldsymbol{h}_t^{(b)}} L$ recursively in terms of $\nabla_{\boldsymbol{h}_{t+1}^{(f)}} L$ and $\nabla_{\boldsymbol{h}_{t-1}^{(b)}} L$ as follows:

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

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

Derive an expression for $\nabla_{\boldsymbol{h}_t^{(f)}} L_t$, $\nabla_{\boldsymbol{h}_t^{(b)}} L_t$, $\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}}$ and $\frac{\partial \boldsymbol{h}_{t-1}^{(b)}}{\partial \boldsymbol{h}_t^{(b)}}$.

Solution

For the derivatives we will use the set of equations:

$$\boldsymbol{h}_t^{(f)} = \sigma(\boldsymbol{W}^{(f)}\boldsymbol{x}_t + \boldsymbol{U}^{(f)}\boldsymbol{h}_{t-1}^{(f)}) \tag{30}$$

$$\boldsymbol{h}_t^{(b)} = \sigma(\boldsymbol{W}^{(b)}\boldsymbol{x}_t + \boldsymbol{U}^{(b)}\boldsymbol{h}_{t+1}^{(b)}) \tag{31}$$

$$\boldsymbol{y}_t = \boldsymbol{V}^{(f)}\boldsymbol{h}_t^{(f)} + \boldsymbol{V}^{(b)}\boldsymbol{h}_t^{(b)} \tag{32}$$

$$L_t = ||z_t - y_t||_2 \tag{33}$$

therefore

$$\begin{aligned}\nabla_{\boldsymbol{h}_t^{(f)}} L = \frac{\partial L_t}{\partial h_t^{(f)}} &= \frac{\partial L_t}{\partial y_t}\left(\frac{\partial y_t}{\partial h_t^{(f)}}\right)^T \\ &= -2(\boldsymbol{V}^{(f)})^T(z_t - y_t)\end{aligned}$$

$$\begin{aligned}\nabla_{\boldsymbol{h}_t^{(b)}} L = \frac{\partial L_t}{\partial h_t^{(b)}} &= \frac{\partial L_t}{\partial y_t}\left(\frac{\partial y_t}{\partial h_t^{(b)}}\right)^T \\ &= -2(\boldsymbol{V}^{(b)})^T(z_t - y_t)\end{aligned}$$

For making the derivative $\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}}$ we must consider that $\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \boldsymbol{h}_t^{(f)}} = \frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \sigma}\frac{\partial \sigma}{\partial \boldsymbol{h}_t^{(f)}}$ but

$$\frac{\partial \boldsymbol{h}_{t+1}^{(f)}}{\partial \sigma} = \sigma(arg)(1 - \sigma(arg)) = \boldsymbol{h}_{t+1}^{(f)}(1 - \boldsymbol{h}_{t+1}^{(f)})$$

with $arg \equiv \boldsymbol{W}^{(f)}\boldsymbol{x}_t + \boldsymbol{U}^{(f)}\boldsymbol{h}_{t-1}^{(f)}$ and $\frac{\partial \sigma}{\partial \boldsymbol{h}_t^{(f)}} = \boldsymbol{U}^{(f)}$ summarizing the results we have

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

where *diag* indicates the diagonal matrix containing the elements $\boldsymbol{h}_{t+1}^{(f)}(1 - \boldsymbol{h}_{t+1}^{(f)})$. This is the Jacobian of the sigmoid function associated with the hidden unit i at time $t + 1$. In a similar way we obtain

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

3. Now derive $\nabla_{\boldsymbol{W}^{(f)}} L$ and $\nabla_{\boldsymbol{U}^{(b)}} L$ as functions of $\nabla_{\boldsymbol{h}_t^{(f)}} L$ and $\nabla_{\boldsymbol{h}_t^{(b)}} L$, respectively.
Hint: It might be useful to consider the contribution of the weight matrices when computing the recurrent hidden unit at a particular time t and how those contributions might be aggregated.

Solution

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

for achieving the last result we must remember that

$$\frac{\partial \boldsymbol{h}_t^{(f)}}{\partial \boldsymbol{W}^{(f)}} = diag(\boldsymbol{h}_t^{(f)}(1 - \boldsymbol{h}_t^{(f)}))\boldsymbol{x}^T$$

Finally

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

for achieving the last result we must remember that

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