

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

1. For  $t \geq 1$ , consider the following update rules:

- SGD with momentum:

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

where  $\epsilon > 0$  and  $\alpha \in (0, 1)$ .

- SGD with running average of  $\mathbf{g}_t$ :

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

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  $(\alpha, \epsilon)$  as a function of  $(\beta, \delta)$ .

Here, when we have SGD with momentum,

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

When we have SGD with running average,

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

Establishing the equivalence, we get

$$(\alpha, \epsilon) = (\beta, \delta(1 - \beta)) \text{ or } (\beta, \delta) = (\alpha, \frac{\epsilon}{1 - \alpha})$$

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

$$\mathbf{v}_t = \beta \mathbf{v}_{t-1} + (1 - \beta) \mathbf{g}_t$$

Expanding,  $\mathbf{v}_{t-1}$ , we get,

$$\mathbf{v}_t = \beta(\beta \mathbf{v}_{t-2} + (1 - \beta) \mathbf{g}_{t-1}) + (1 - \beta) \mathbf{g}_t$$

If we keep expanding the inner term  $\mathbf{v}_t$  recursively, eventually we get the following relation,

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

3. Assume  $\mathbf{g}_t$  has a stationary distribution independent of  $t$ . Show that the running average is biased, i.e.  $\mathbb{E}[\mathbf{v}_t] \neq \mathbb{E}[\mathbf{g}_t]$ . Propose a way to eliminate such a bias by rescaling  $\mathbf{v}_t$ .

Taking the expectation over the expression in previous question, we get,

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

, Since expectations are linear,

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

, Due to stationarity,  $\mathbb{E}[\mathbf{g}_i] = \mathbb{E}[\mathbf{g}_t] \quad \forall i \in (1, 2, \dots, t)$ , Hence, we can bring the expectation term out of summation,

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

From above, we can clearly say that the estimator is biased, the unbiased estimator would be the one compensating for the  $(1 - \beta^t)$  factor. So one can use  $\frac{\mathbf{v}_t}{1 - \beta^t}$  as an unbiased estimator.

**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  $\mathbf{X} \in \mathbb{R}^{n \times d}$ , weights  $\mathbf{w} \in \mathbb{R}^{d \times 1}$  and targets  $\mathbf{y} \in \mathbb{R}^{n \times 1}$ . Suppose that dropout is applied to the input (with probability  $1 - p$  of dropping the unit i.e. setting it to 0). Let  $\mathbf{R} \in \mathbb{R}^{n \times d}$  be the dropout mask such that  $\mathbf{R}_{ij} \sim \text{Bern}(p)$  is sampled i.i.d. from the Bernoulli distribution.

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

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

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

Here, let's say  $\mathbf{R}_i$  is the row  $i$  of the dropout mask  $\mathbf{R}$ . Similarly,  $\mathbf{X}_i$  represents datapoint  $i$  in the matrix  $\mathbf{X}$ .  $y_i$  represents the corresponding groundtruth for the datapoint  $\mathbf{X}_i$ . Here, since  $\mathbf{R}_{ij} \sim \text{Bern}(p)$ ,  $\text{Var}(\mathbf{R}_{ij}) = p(1 - p)$

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

Here,

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

Similarly,

$$\begin{aligned}\sum_{i=1}^n \text{Var}_{\mathbf{R}}[(y_i - \mathbf{w}^T(\mathbf{X}_i \odot \mathbf{R}_i))] &= \sum_{i=1}^n \mathbb{E}_{\mathbf{R}}[(y_i - \mathbf{w}^T(\mathbf{X}_i \odot \mathbf{R}_i) - \mathbb{E}_{\mathbf{R}}[(y_i - \mathbf{w}^T(\mathbf{X}_i \odot \mathbf{R}_i))])^2] \\ &= \sum_{i=1}^n \mathbb{E}_{\mathbf{R}}[(y_i - \mathbf{w}^T(\mathbf{X}_i \odot \mathbf{R}_i) - y_i + p\mathbf{X}_i\mathbf{w})^2] \\ &= \sum_{i=1}^n \mathbb{E}_{\mathbf{R}}[((p\mathbf{X}_i - \mathbf{R}_i \odot \mathbf{X}_i)\mathbf{w})^2] \\ &= \sum_{i=1}^n \mathbb{E}_{\mathbf{R}}[\mathbf{w}^T(p\mathbf{X}_i^T - \mathbf{R}_i^T \odot \mathbf{X}_i^T)(p\mathbf{X}_i - \mathbf{R}_i \odot \mathbf{X}_i)\mathbf{w}] \\ &= \sum_{i=1}^n (\mathbf{w}^T(\cancel{p^2\mathbf{X}_i^T\mathbf{X}_i} - p\mathbb{E}_{\mathbf{R}}[\mathbf{R}_i]\mathbf{X}_i^T\mathbf{X}_i \\ &\quad - p\mathbb{E}_{\mathbf{R}}[\mathbf{R}_i^T]\mathbf{X}_i^T\mathbf{X}_i + \mathbb{E}_{\mathbf{R}}[\mathbf{R}_i^T\mathbf{R}_i]\mathbf{X}_i^T\mathbf{X}_i)\mathbf{w}) \\ &= \sum_{i=1}^n (\mathbf{w}^T(-p^2\mathbf{X}_i^T\mathbf{X}_i + p\mathbf{X}_i^T\mathbf{X}_i)\mathbf{w}) \\ &\quad (\because \text{Mean of the squared Bernoulli random variable is same as mean of the Bernoulli random variable}) \\ &= \sum_{i=1}^n (\mathbf{w}^T(p(1-p)\mathbf{X}_i^T\mathbf{X}_i)\mathbf{w}) \\ &= \mathbf{w}^T(p(1-p)(\sum_{i=1}^n \mathbf{X}_i^T\mathbf{X}_i)\mathbf{w}) \\ &= p(1-p)(\mathbf{w}^T(\sum_{i=1}^n (\mathbf{X}^T\mathbf{X})_{ii})\mathbf{w}) = p(1-p)\|\Gamma\mathbf{w}\|^2\end{aligned}$$

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

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

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

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

Let's take derivative of expectation of loss derived in 2.1,

$$\begin{aligned} \nabla_{\mathbf{w}} \mathbb{E}[L(\mathbf{w})] &= \nabla_{\mathbf{w}} \|y - p\mathbf{X}\mathbf{w}\|^2 + p(1-p)\|\Gamma\mathbf{w}\|^2 = 0 \\ &= -p\mathbf{X}^T(y - p\mathbf{X}\mathbf{w}) + p(1-p)\Gamma^2\mathbf{w} = 0 \\ \mathbf{X}^T y &= p\mathbf{X}^T \mathbf{X}\mathbf{w} + (1-p)\Gamma^2\mathbf{w} \\ \mathbf{X}^T y &= p\mathbf{X}^T \mathbf{X}\mathbf{w} + \frac{1-p}{p}\Gamma^2 p\mathbf{w} \\ \mathbf{X}^T y &= (\mathbf{X}^T \mathbf{X} + (\frac{1-p}{p})\Gamma^2)p\mathbf{w} \end{aligned}$$

Taking inverse, we get,

$$\boxed{p\mathbf{w} = (\mathbf{X}^T \mathbf{X} + (\frac{1-p}{p})\Gamma^2)^{-1} \mathbf{X}^T y}$$

Above equation satisfies  $p\mathbf{w}^{\text{dropout}} = (\mathbf{X}^\top \mathbf{X} + \lambda^{\text{dropout}} \Gamma^2)^{-1} \mathbf{X}^\top \mathbf{y}$ , where  $\lambda^{\text{dropout}} = \frac{1-p}{p}$ . The range of  $p$  is  $[0, 1]$ . As  $p$  heads towards 0, we see that the regularization coefficient moves towards  $\infty$ .  $p = 0$  will force all the units to be dropped out, making it highly regularized system. Similarly,  $p = 1$  will effectively retain all the units, which effectively means no regularization. This is reflected through  $\lambda_{\text{dropout}} = 0$ . So  $p$  and  $\lambda_{\text{dropout}}$  have biased inverse relation as given by  $\lambda_{\text{dropout}} = \frac{1}{p} - 1$ .

3. Express the loss function for a linear regression problem without dropout and with  $L^2$  regularization, with regularization coefficient  $\lambda^{L^2}$ . Derive its closed form solution  $\mathbf{w}^{L^2}$ .  
Loss with L2 regularization,

$$\begin{aligned} L(\mathbf{w}) &= \|y - \mathbf{X}\mathbf{w}\|^2 + \lambda^{L^2} \|\mathbf{w}\|^2 \\ \frac{dL(\mathbf{w})}{d\mathbf{w}} &= (-\mathbf{X}^T)(y - \mathbf{X}\mathbf{w}^{L^2}) + \lambda^{L^2} \mathbf{w}^{L^2} = 0 \\ \frac{dL(\mathbf{w})}{d\mathbf{w}} &= -\mathbf{X}^T y + \mathbf{X}^T \mathbf{X}\mathbf{w}^{L^2} + \lambda^{L^2} \mathbf{w}^{L^2} = 0 \\ \mathbf{X}^T \mathbf{X}\mathbf{w}^{L^2} + \lambda^{L^2} \mathbf{w}^{L^2} &= \mathbf{X}^T y \\ \boxed{\mathbf{w}^{L^2} = (\mathbf{X}^T \mathbf{X} + \lambda^{L^2} I)^{-1} \mathbf{X}^T y} \end{aligned}$$

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).

In 2.3, the regularization coefficient merely reduces the overall value of weight. However, this weight reduction is uniform in nature. While in 2.2, the weight reduction is not uniform, as the regularization coefficient is scaled by  $\Gamma^2$ , which is a data dependent term. So we can view dropout as data-dependent scaled L2 regularization.

**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:

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

where  $\mathbf{a}^{(t)}$  are the pre-activations and  $\mathbf{h}^{(t)}$  are the activations for layer  $t$ ,  $g$  is an activation function,  $\mathbf{W}^{(t)}$  is a  $d^{(t)} \times d^{(t-1)}$  matrix, and  $\mathbf{b}^{(t)}$  is a  $d^{(t)} \times 1$  bias vector. The bias is initialized as a constant vector  $\mathbf{b}^{(t)} = [c, \dots, c]^\top$  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  $\text{Var}(a_i^{(t)}) = 1$ ,  $1 \leq i \leq d^{(t)}$ ) 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 possibilities).

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  $1 \leq i \leq d^{(t-1)}$ . Also, assume entries of  $\mathbf{h}^{(t-1)}$  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  $X \perp Y$

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

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

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

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

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

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

- (b) Write  $\mathbb{E}[a_i^{(t)}]$  and  $\text{Var}(a_i^{(t)})$  in terms of  $c, \mu, \sigma^2, \text{Var}(h_i^{(t-1)}), \mathbb{E}[h_i^{(t-1)}]$ . Calculating  $\mathbb{E}[a_i^{(t)}]$ , Here, let  $\mathbf{W}_i^{(t)}$  be  $i^{\text{th}}$  row of the matrix  $\mathbf{W}^{(t)}$

$$\mathbb{E}[a_i^{(t)}] = \mathbb{E}[\mathbf{W}_i^{(t)}\mathbf{h}^{(t-1)} + \mathbf{b}_i^{(t)}]$$

$$\mathbb{E}[a_i^{(t)}] = \mathbb{E}[\mathbf{W}_i^{(t)}]\mathbb{E}[\mathbf{h}^{(t-1)}] + \mathbb{E}[\mathbf{b}_i^{(t)}]$$

$$(\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \text{ if } X \perp Y \text{ and } \mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y])$$

Here, we know that  $W_{ij}^{(t)} \sim \mathcal{N}(\mu, \sigma^2)$ , hence  $\mathbb{E}[\mathbf{W}_i^{(t)}] = \mu \mathbf{1}_{d_{t-1}}^T$ , where  $\mathbf{1}_{d_{t-1}}^T$  is vector of 1's with length  $d_{t-1}$ . Similarly,  $\mathbb{E}[\mathbf{b}_i^t] = c$  as  $\mathbf{b}_i^t = c$ . Substituting that, we get

$$\mathbb{E}[a_i^{(t)}] = \mu \mathbf{1}_{d_{t-1}}^T \mathbb{E}[\mathbf{h}^{(t-1)}] + c = \sum_{i=1}^{d_{t-1}} \mu \mathbb{E}[h_i^{(t-1)}] + c$$

Calculating  $\text{Var}(a_i^{(t)})$ ,

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

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

$$\text{because } \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

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

$$\text{Var}(a_i^{(t)}) = \sum_{j=1}^{d_{t-1}} \text{Var}(W_{ij}^{(t)}) \text{Var}(h_j^{(t-1)}) + \text{Var}(W_{ij}^{(t)}) \mathbb{E}[h_j^{(t-1)}]^2 + \mathbb{E}[W_{ij}^{(t)}]^2 \text{Var}(h_j^{(t-1)}) + \text{Var}(\mathbf{b}_i^{(t)})$$

Here, we know that  $\text{Var}(W_{ij}^{(t)}) = \sigma^2$ ,  $\mathbb{E}[W_{ij}^{(t)}] = \mu$ ,  $\text{Var}(\mathbf{b}_i^{(t)}) = 0$ . Substituting above,

$$\text{Var}(a_i^{(t)}) = \sum_{j=1}^{d_{t-1}} \sigma^2 \text{Var}(h_j^{(t-1)}) + \sigma^2 \mathbb{E}[h_j^{(t-1)}]^2 + \mu^2 \text{Var}(h_j^{(t-1)})$$

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

We know that  $\mathbb{E}[h_i^{(t-1)}] = 0$  and  $\mathbb{E}[a_i^{(t)}] = 0$ , substituting that in answer of (b), we get following,

$$0 = \sum_{i=1}^{d_{t-1}} \mu 0 + c \implies \boxed{c = 0}$$

Plugging the same values for  $\text{Var}(a_i^{(t)})$  in answer of (b),

$$1 = \sum_{i=1}^{d_{t-1}} \mu^2 + \sigma^2 \implies \sigma^2 + \mu^2 = \frac{1}{d_{t-1}}$$

One combination that would work is when  $\mu = 0$ , then  $\sigma = 1/\sqrt{d_{t-1}}$ . Hence, we get the following values

$$\boxed{c = 0, \mu = 0, \sigma^2 = \frac{1}{d_{t-1}}}$$

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

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

Using definition of expectation computation,

$$\begin{aligned}
 \mathbb{E}[(h_i^{(t-1)})^2] &= \int_{-\infty}^{\infty} \max\{0, a_i^{(t-1)}\}^2 p(a_i^{(t-1)}) da_i^{(t-1)} \\
 &= \int_{-\infty}^{\infty} (a_i^{(t-1)})^2 p(a_i^{(t-1)}) da_i^{(t-1)} = \int_0^{\infty} (a_i^{(t-1)})^2 p(a_i^{(t-1)}) da_i^{(t-1)} \\
 &= \frac{1}{2} \int_{-\infty}^{\infty} (a_i^{(t-1)})^2 p(a_i^{(t-1)}) da_i^{(t-1)} \\
 &= \frac{1}{2} \mathbb{E}[(a_i^{(t-1)})^2] = \frac{1}{2} (\mathbb{E}[(a_i^{(t-1)})^2] - \mathbb{E}[(a_i^{(t-1)})]^2) = \frac{1}{2} \text{Var}(a_i^{(t-1)}) \\
 \boxed{\mathbb{E}[(h_i^{(t-1)})^2] = \frac{1}{2}} & \quad (\because \text{Var}(a_i^{(t-1)}) = 1)
 \end{aligned}$$

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

$$\begin{aligned}
 \text{Var}(a_i^{(t)}) &= \sum_{j=1}^{d_{t-1}} \text{Var}(W_{ij}^{(t)}) \text{Var}(h_j^{(t-1)}) + \text{Var}(W_{ij}^{(t)}) \mathbb{E}[h_j^{(t-1)}]^2 + \mathbb{E}[W_{ij}^{(t)}]^2 \text{Var}(h_j^{(t-1)}) + \text{Var}(\mathbf{b}_i^{(t)}) \\
 &= \sum_{j=1}^{d_{t-1}} \sigma^2 (\text{Var}(h_j^{(t-1)}) + \mathbb{E}[h_j^{(t-1)}]^2) = \sum_{j=1}^{d_{t-1}} \sigma^2 (\mathbb{E}[(h_j^{(t-1)})^2]) \\
 &= \frac{1}{2} \sum_{j=1}^{d_{t-1}} \sigma^2 = \frac{\sigma^2 d_{t-1}}{2} = 1 \\
 \Rightarrow \boxed{\sigma = \sqrt{\frac{2}{d_{t-1}}}} & \quad (2)
 \end{aligned}$$

Hence, we get  $c = 0$ ,  $\mu = 0$  and  $\sigma = \sqrt{\frac{2}{d_{t-1}}}$

- (c) What popular initialization scheme has this form? This is He normal initialization.
- (d) Why do you think this initialization would work well in practice? Answer in 1-2 sentences.  
If we initialize weights using this, the pre-activations will have a unit variance with 0 mean. That means that half of the neurons on average would be fired up. So this is a good way to ensure learning properly, and this would work well in practice.
3. For both assumptions (1,2) give values  $\alpha, \beta$  for  $W_{ij}^{(t)} \sim \text{Uniform}(\alpha, \beta)$  such that  $\mathbb{E}[a_i^{(t)}] = 0$  and  $\text{Var}(a_i^{(t)}) = 1$ .  
If we go for assumptions of (1,2), we have  $\mathbb{E}[\mathbf{W}_{ij}^{(t)}] = \frac{\alpha+\beta}{2}$ , to have  $\mathbb{E}[\mathbf{W}_{ij}^{(t)}] = 0$ , we have  $\alpha = -\beta$ ,

with  $c = 0$ , substituting in answer of (b), we get  $\mathbb{E}[a_i^{(t)}] = 0$ . Now, for variance calculation,

$$\begin{aligned}
 \text{Var}(a_i^{(t)}) &= \sum_{j=1}^{d_{t-1}} \text{Var}(W_{ij}^{(t)})\text{Var}(h_j^{(t-1)}) + \cancel{\text{Var}(W_{ij}^{(t)})\mathbb{E}[h_j^{(t-1)}]^2} + \cancel{\mathbb{E}[W_{ij}^{(t)}]^2\text{Var}(h_j^{(t-1)})} + \cancel{\text{Var}(b_i^{(t)})} \\
 &= \sum_{j=1}^{d_{t-1}} \frac{(\beta - \alpha)^2}{12} = \sum_{j=1}^{d_{t-1}} \frac{\alpha^2}{3} \\
 &= \frac{\alpha^2 d_{t-1}}{3} = 1 \\
 &\Rightarrow \alpha = -\sqrt{\frac{3}{d_{t-1}}} \\
 &\Rightarrow \boxed{\alpha = -\sqrt{\frac{3}{d_{t-1}}}, \beta = \sqrt{\frac{3}{d_{t-1}}}}
 \end{aligned}$$

This is Glorot uniform initialization.

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

1. Batch normalization, layer normalization and instance normalization all involve calculating the mean  $\mu$  and variance  $\sigma^2$  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:  $\mu_{batch}$ ,  $\mu_{layer}$ ,  $\mu_{instance}$ ,  $\sigma_{batch}^2$ ,  $\sigma_{layer}^2$ , and  $\sigma_{instance}^2$ .

$$\begin{bmatrix} \begin{bmatrix} 1, 3, 2 \\ 1, 2, 3 \end{bmatrix}, \begin{bmatrix} 3, 3, 2 \\ 2, 4, 4 \end{bmatrix}, \begin{bmatrix} 4, 2, 2 \\ 1, 2, 4 \end{bmatrix}, \begin{bmatrix} 3, 3, 2 \\ 3, 3, 2 \end{bmatrix} \end{bmatrix}$$

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.

In batch normalization, we take means for channels. So  $\mu_{batch}$  is a 2x1 vector because there are 2 channels in the input data. Mean and covariance are as below,

$$\boxed{\mu_{batch} = [2.5000, 2.5833], \sigma_{batch}^2 = [0.6364, 1.1742]}$$

In layer normalization, we take means for every sample. Here, we have 4 samples, as batchsize is 4. So  $\mu_{layer}$  is a 4x1 vector because there are 4 examples in the input data. Mean and covariance are as below,

$$\boxed{\mu_{layer} = [2.0000, 3.0000, 2.5000, 2.6667], \sigma_{layer}^2 = [0.8000, 0.8000, 1.5000, 0.2667]}$$

In instance normalization, we take means for every channel of every sample. Here, we have 4 samples, and 2 channels for each sample. So  $\mu_{instance}$  is a 4x2 matrix. Mean and covariance are as below,

$$\boxed{\mu_{instance} = \begin{bmatrix} 2.0000, 2.0000 \\ 2.6667, 3.3333 \\ 2.6667, 2.3333 \\ 2.6667, 2.6667 \end{bmatrix}, \sigma_{instance}^2 = \begin{bmatrix} 1.0000, 1.0000 \\ 0.3333, 1.3333 \\ 1.3333, 2.3333 \\ 0.3333, 0.3333 \end{bmatrix}}$$



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

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

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

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  $y = \mathbf{u}^\top \mathbf{x}$ . Assume the data  $\mathbf{x}$  (a random vector) is whitened ( $\text{Var}(\mathbf{x}) = \mathbf{I}$ ) and centered at 0 ( $\mathbb{E}[\mathbf{x}] = \mathbf{0}$ ). Show that  $\hat{y} = \mathbf{w}^\top \mathbf{x} + \beta$ .

$$\mu_y = \mathbb{E}[\mathbf{u}^\top \mathbf{x}] = \mathbf{u}^\top \mathbb{E}[\mathbf{x}] = 0$$

$$\sigma_y^2 = \text{Var}(\mathbf{u}^\top \mathbf{x}) = \mathbf{u}^\top \text{Var}(\mathbf{x}) \mathbf{u} = \|\mathbf{u}\|^2 \implies \sigma_y = \|\mathbf{u}\|$$

$$\hat{y} = \gamma \cdot \frac{y - \mu_y}{\sigma_y} + \beta = \gamma \cdot \frac{\mathbf{u}^\top \mathbf{x} - 0}{\|\mathbf{u}\|} = \left( \gamma \frac{\mathbf{u}^\top}{\|\mathbf{u}\|} \right) \mathbf{x} + \beta \implies \boxed{\hat{y} = \mathbf{w}^\top \mathbf{x} + \beta}$$

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

$$\nabla_{\mathbf{u}} L = \nabla_{\mathbf{u}} \mathbf{w} \nabla_{\mathbf{w}} L$$

$$\begin{aligned} \nabla_{\mathbf{u}} \mathbf{w} &= \nabla_{\mathbf{u}} \left( \gamma \frac{\mathbf{u}}{\|\mathbf{u}\|} \right) = \gamma \left( \nabla_{\mathbf{u}} \frac{\mathbf{u}}{\|\mathbf{u}\|} \right) \\ &= \gamma \left( \frac{\|\mathbf{u}\| \mathbf{I} - \mathbf{u} \frac{\mathbf{u}^\top}{\|\mathbf{u}\|}}{\|\mathbf{u}\|^2} \right) \quad (\because \text{gradient of the norm of the vector a unit vector}) \\ &= \frac{\gamma}{\|\mathbf{u}\|} \left( \mathbf{I} - \frac{\mathbf{u} \mathbf{u}^\top}{\|\mathbf{u}\|^2} \right) = \frac{\gamma}{\|\mathbf{u}\|} \mathbf{W}^\perp \implies \boxed{\nabla_{\mathbf{u}} \mathbf{w} = \frac{\gamma}{\|\mathbf{u}\|} \mathbf{W}^\perp} \\ \nabla_{\mathbf{u}} L &= \nabla_{\mathbf{u}} \mathbf{w} \nabla_{\mathbf{w}} L = \frac{\gamma}{\|\mathbf{u}\|} \mathbf{W}^\perp \nabla_{\mathbf{w}} L = s \mathbf{W}^\perp \nabla_{\mathbf{w}} L \text{ where } s = \frac{\gamma}{\|\mathbf{u}\|} \end{aligned} \quad (3)$$

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

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

1. Show that applying the activation function in this way is equivalent to the conventional way of applying the activation function:  $\mathbf{g}_t = \sigma(\mathbf{W} \mathbf{g}_{t-1} + \mathbf{U} \mathbf{x}_t + \mathbf{b})$  (i.e. express  $\mathbf{g}_t$  in terms of  $\mathbf{h}_t$ ). 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  $t - 1$ .

1. As a side note:  $\mathbf{W}^\perp$  is an orthogonal complement that projects the gradient away from the direction of  $\mathbf{w}$ , 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.

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

This holds if  $\mathbf{g}_{t-1} = \sigma(\mathbf{h}_{t-1})$ . Substituting this relation in above equation will automatically give us inductive relation  $\boxed{\mathbf{g}_t = \sigma(\mathbf{h}_t)}$ .

- \*2. Let  $\|\mathbf{A}\|$  denote the  $L_2$  operator norm<sup>2</sup> of matrix  $\mathbf{A}$  ( $\|\mathbf{A}\| := \max_{\mathbf{x}: \|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|$ ). 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.

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

Use the following properties of the  $L_2$  operator norm

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\| \|\mathbf{B}\| \quad \text{and} \quad \|\mathbf{A}\| = \sqrt{\lambda_1(\mathbf{A}^\top \mathbf{A})}$$

$$\begin{aligned} \left\| \frac{\partial \mathbf{h}_T}{\partial \mathbf{h}_0} \right\| &= \left\| \frac{\partial \mathbf{h}_T}{\partial \mathbf{h}_{T-1}} \right\| \cdot \left\| \frac{\partial \mathbf{h}_{T-1}}{\partial \mathbf{h}_{T-2}} \right\| \cdots \left\| \frac{\partial \mathbf{h}_1}{\partial \mathbf{h}_0} \right\| \\ \left\| \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}} \right\| &= \left\| \mathbf{W} \frac{\partial \sigma(\mathbf{h}_{t-1})}{\partial \mathbf{h}_{t-1}} \right\| \\ &\leq \|\mathbf{W}\| \cdot \left\| \frac{\partial \sigma(\mathbf{h}_{t-1})}{\partial \mathbf{h}_{t-1}} \right\| (\because \|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|) \\ &\leq \|\mathbf{W}\| \cdot \|\sigma'\| \\ &\leq \sqrt{\lambda_1(\mathbf{W}^\top \mathbf{W})} \cdot \gamma \\ &\leq \sqrt{\frac{\delta^2}{\gamma^2}} \cdot \gamma = \|\delta\| = \delta \\ \left\| \frac{\partial \mathbf{h}_T}{\partial \mathbf{h}_0} \right\| &\leq \delta \cdot \delta \cdots \delta = \delta^T \implies \boxed{\left\| \frac{\partial \mathbf{h}_T}{\partial \mathbf{h}_0} \right\| \rightarrow 0 \text{ as } T \rightarrow \infty} \quad (\because 0 \leq \delta \leq 1) \end{aligned}$$

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).

If we have  $\lambda_1(\mathbf{W}^\top \mathbf{W}) \geq \frac{\delta^2}{\gamma^2}$  and  $\left\| \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}} \right\| \leq \sqrt{\lambda_1(\mathbf{W}^\top \mathbf{W})} \cdot \gamma$ , we have no conclusive relation between  $\left\| \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}} \right\|$  and  $\frac{\delta^2}{\gamma^2}$ . Hence this condition is neither *necessary* nor *sufficient* for the gradient to explode

---

2. The  $L_2$  operator norm of a matrix  $\mathbf{A}$  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:

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

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

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

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.

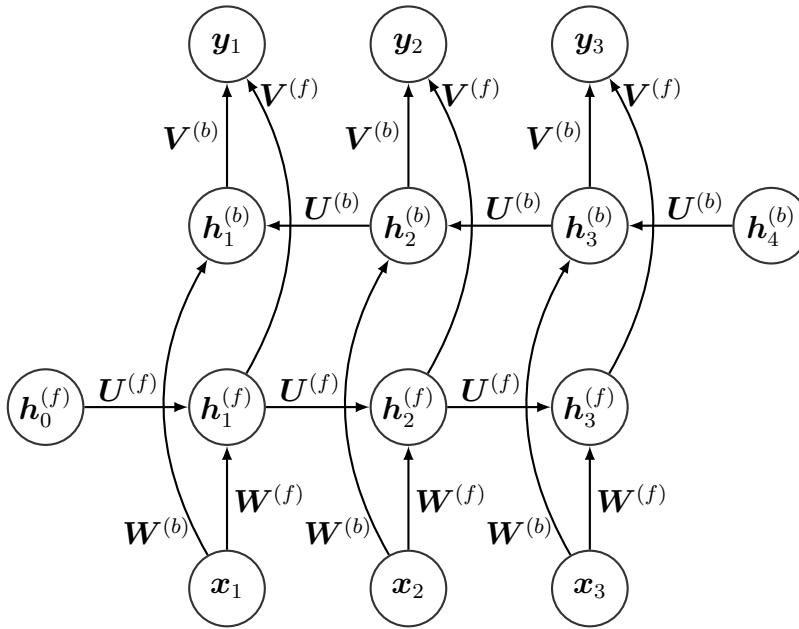


FIGURE 1 – Computational graph of the bidirectional RNN unrolled for three timesteps.

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

$$\begin{aligned} \nabla_{\mathbf{h}_t^{(f)}} L &= \nabla_{\mathbf{h}_t^{(f)}} L_t + \left( \frac{\partial \mathbf{h}_{t+1}^{(f)}}{\partial \mathbf{h}_t^{(f)}} \right)^\top \nabla_{\mathbf{h}_{t+1}^{(f)}} L \\ \nabla_{\mathbf{h}_t^{(b)}} L &= \nabla_{\mathbf{h}_t^{(b)}} L_t + \left( \frac{\partial \mathbf{h}_{t-1}^{(b)}}{\partial \mathbf{h}_t^{(b)}} \right)^\top \nabla_{\mathbf{h}_{t-1}^{(b)}} L \end{aligned}$$

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

Deriving  $\nabla_{\mathbf{h}_t^{(f)}} L_t$ ,

$$\begin{aligned}
 \nabla_{\mathbf{h}_t^{(f)}} L_t &= \nabla_{\mathbf{h}_t^{(f)}} \|\mathbf{z}_t - \mathbf{y}_t\|_2^2 \\
 &= 2\|\mathbf{z}_t - \mathbf{y}_t\|_2 \nabla_{\mathbf{h}_t^{(f)}} \|\mathbf{z}_t - \mathbf{y}_t\|_2 \\
 &= 2\|\mathbf{z}_t - \mathbf{y}_t\|_2 \nabla_{\mathbf{h}_t^{(f)}} \|\mathbf{z}_t - (\mathbf{V}^{(f)} \mathbf{h}_t^{(f)} + \mathbf{V}^{(b)} \mathbf{h}_t^{(b)})\|_2 \\
 &\implies \boxed{\nabla_{\mathbf{h}_t^{(f)}} L_t = -2\mathbf{V}^{(f)} \|\mathbf{z}_t - \mathbf{y}_t\|_2}
 \end{aligned}$$

Deriving  $\nabla_{\mathbf{h}_t^{(b)}} L_t$ ,

$$\begin{aligned}
 \nabla_{\mathbf{h}_t^{(b)}} L_t &= \nabla_{\mathbf{h}_t^{(b)}} \|\mathbf{z}_t - \mathbf{y}_t\|_2^2 \\
 &= 2\|\mathbf{z}_t - \mathbf{y}_t\|_2 \nabla_{\mathbf{h}_t^{(b)}} \|\mathbf{z}_t - \mathbf{y}_t\|_2 \\
 &= 2\|\mathbf{z}_t - \mathbf{y}_t\|_2 \nabla_{\mathbf{h}_t^{(b)}} \|\mathbf{z}_t - (\mathbf{V}^{(f)} \mathbf{h}_t^{(f)} + \mathbf{V}^{(b)} \mathbf{h}_t^{(b)})\|_2 \\
 &\implies \boxed{\nabla_{\mathbf{h}_t^{(b)}} L_t = -2\mathbf{V}^{(b)} \|\mathbf{z}_t - \mathbf{y}_t\|_2}
 \end{aligned}$$

Deriving  $\frac{\partial \mathbf{h}_{t+1}^{(f)}}{\partial \mathbf{h}_t^{(f)}}$ ,

$$\begin{aligned}
 \frac{\partial \mathbf{h}_{t+1}^{(f)}}{\partial \mathbf{h}_t^{(f)}} &= \frac{\partial \sigma(\mathbf{W}^{(f)} \mathbf{x}_{t+1} + \mathbf{U}^{(f)} \mathbf{h}_t^{(f)})}{\partial \mathbf{h}_t^{(f)}} \\
 &= \sigma(\mathbf{W}^{(f)} \mathbf{x}_{t+1} + \mathbf{U}^{(f)} \mathbf{h}_t^{(f)}) (1 - \sigma(\mathbf{W}^{(f)} \mathbf{x}_{t+1} + \mathbf{U}^{(f)} \mathbf{h}_t^{(f)})) \mathbf{U}^{(f)} \\
 &\implies \boxed{\frac{\partial \mathbf{h}_{t+1}^{(f)}}{\partial \mathbf{h}_t^{(f)}} = \mathbf{h}_{t+1}^{(f)} (1 - \mathbf{h}_{t+1}^{(f)}) \mathbf{U}^{(f)}}
 \end{aligned}$$

Deriving  $\frac{\partial \mathbf{h}_{t-1}^{(b)}}{\partial \mathbf{h}_t^{(b)}}$ ,

$$\begin{aligned}
 \frac{\partial \mathbf{h}_{t-1}^{(b)}}{\partial \mathbf{h}_t^{(b)}} &= \frac{\partial \sigma(\mathbf{W}^{(b)} \mathbf{x}_{t-1} + \mathbf{U}^{(b)} \mathbf{h}_t^{(b)})}{\partial \mathbf{h}_t^{(b)}} \\
 &= \sigma(\mathbf{W}^{(b)} \mathbf{x}_{t-1} + \mathbf{U}^{(b)} \mathbf{h}_t^{(b)}) (1 - \sigma(\mathbf{W}^{(b)} \mathbf{x}_{t-1} + \mathbf{U}^{(b)} \mathbf{h}_t^{(b)})) \mathbf{U}^{(b)} \\
 &\implies \boxed{\frac{\partial \mathbf{h}_{t-1}^{(b)}}{\partial \mathbf{h}_t^{(b)}} = \mathbf{h}_{t-1}^{(b)} (1 - \mathbf{h}_{t-1}^{(b)}) \mathbf{U}^{(b)}}
 \end{aligned}$$

3. Now derive  $\nabla_{\mathbf{W}^{(f)}} L$  and  $\nabla_{\mathbf{U}^{(b)}} L$  as functions of  $\nabla_{\mathbf{h}_t^{(f)}} L$  and  $\nabla_{\mathbf{h}_t^{(b)}} L$ , 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.*

$$\begin{aligned}\nabla_{\mathbf{W}^{(f)}} L &= \nabla_{\mathbf{W}^{(f)}} \mathbf{h}_t^{(f)} \nabla_{\mathbf{h}_t^{(f)}} L \\&= \nabla_{\mathbf{W}^{(f)}} \sigma(\mathbf{W}^{(f)} \mathbf{x}_t + \mathbf{U}^{(f)} \mathbf{h}_{t-1}^{(f)}) \nabla_{\mathbf{h}_t^{(f)}} L \\&= \sigma(\mathbf{W}^{(f)} \mathbf{x}_t + \mathbf{U}^{(f)} \mathbf{h}_{t-1}^{(f)}) (1 - \sigma(\mathbf{W}^{(f)} \mathbf{x}_t + \mathbf{U}^{(f)} \mathbf{h}_{t-1}^{(f)})) \mathbf{x}_t \nabla_{\mathbf{h}_t^{(f)}} L \\&= \mathbf{h}_t^{(f)} (1 - \mathbf{h}_t^{(f)}) \mathbf{x}_t \nabla_{\mathbf{h}_t^{(f)}} L \\&\implies \boxed{\nabla_{\mathbf{W}^{(f)}} L = \mathbf{h}_t^{(f)} (1 - \mathbf{h}_t^{(f)}) \mathbf{x}_t \nabla_{\mathbf{h}_t^{(f)}} L}\end{aligned}$$

$$\begin{aligned}\nabla_{\mathbf{U}^{(b)}} L &= \nabla_{\mathbf{U}^{(b)}} \mathbf{h}_t^{(b)} \nabla_{\mathbf{h}_t^{(b)}} L \\&= \nabla_{\mathbf{W}^{(b)}} \sigma(\mathbf{W}^{(b)} \mathbf{x}_t + \mathbf{U}^{(b)} \mathbf{h}_{t+1}^{(b)}) \nabla_{\mathbf{h}_t^{(b)}} L \\&= \sigma(\mathbf{W}^{(b)} \mathbf{x}_t + \mathbf{U}^{(b)} \mathbf{h}_{t+1}^{(b)}) (1 - \sigma(\mathbf{W}^{(b)} \mathbf{x}_t + \mathbf{U}^{(b)} \mathbf{h}_{t+1}^{(b)})) \mathbf{h}_{t+1}^{(b)} \nabla_{\mathbf{h}_t^{(b)}} L \\&= \mathbf{h}_t^{(b)} (1 - \mathbf{h}_t^{(b)}) \mathbf{h}_{t+1}^{(b)} \nabla_{\mathbf{h}_t^{(b)}} L \\&\implies \boxed{\nabla_{\mathbf{W}^{(b)}} L = \mathbf{h}_t^{(b)} (1 - \mathbf{h}_t^{(b)}) \mathbf{h}_{t+1}^{(b)} \nabla_{\mathbf{h}_t^{(b)}} L}\end{aligned}$$