Question 1 (6-10)**.** The goal of this question is for you to understand the reasoning behind different parameter initializations for deep networks, particularly to think about the ways that the initialization affects the activations (and therefore the gradients) of the network. Consider the following equation for the t-th layer of a deep network:

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

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

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

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

1. Assume $\mathbb{E}[\boldsymbol{h}_i^{(t-1)}] = 0$ and $\mathrm{Var}(\boldsymbol{h}_i^{(t-1)}) = 1$ for $1 \le i \le d^{(t-1)}$. Assume entries of $\boldsymbol{h}^{(t-1)}$ are uncorrelated (the answer should not depend on g).
   (a) Gaussian: give the values for c, μ, and σ^2 as a function of $d^{(t-1)}$.
   (b) Uniform: give the values for c, α, and β as a function of $d^{(t-1)}$.
2. Assume that the preactivations of the previous layer satisfy $\mathbb{E}[\boldsymbol{a}_i^{(t-1)}] = 0$, $\mathrm{Var}(\boldsymbol{a}_i^{(t-1)}) = 1$ and $\boldsymbol{a}_i^{(t-1)}$ has a symmetric distribution for $1 \le i \le d^{(t-1)}$. Assume entries of $\boldsymbol{a}^{(t-1)}$ are uncorrelated. Consider the case of ReLU activation: $g(x) = \max\{0, x\}$.
   (a) Gaussian: give the values for c, μ, and σ^2 as a function of $d^{(t-1)}$.
   (b) Uniform: give the values for c, α, and β as a function of $d^{(t-1)}$.
   (c) What popular initialization scheme has this form ?
   (d) Why do you think this initialization would work well in practice ? Answer in 1-2 sentences.

Answer 1. Assume $\boldsymbol{W}_i^{(t)}$ is the ith row of $\boldsymbol{W}^{(t)}$ for $1 \le i \le d^{(t)}$.

$$\begin{aligned}
\mathbb{E}[\boldsymbol{a}_i^{(t)}] &= \mathbb{E}[\boldsymbol{W}_i^{(t)}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_i^{(t)}] \\
&= \mathbb{E}[\boldsymbol{W}_i^{(t)}\boldsymbol{h}^{(t-1)}] \; + \; \mathbb{E}[\boldsymbol{b}_i^{(t)}] \qquad (\because \mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]) \\
&= \mathbb{E}[\boldsymbol{W}_i^{(t)}]\; \mathbb{E}[\boldsymbol{h}^{(t-1)}] \; + \; \mathbb{E}[\boldsymbol{b}_i^{(t)}] \qquad (\because \mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \text{ if } X \perp Y)
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{a}_i^{(t)}) &= \mathrm{Var}(\boldsymbol{W}_i^{(t)}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_i^{(t)}) \\
&= \mathrm{Var}(\boldsymbol{W}_i^{(t)}\boldsymbol{h}^{(t-1)}) \; + \; \mathrm{Var}(\boldsymbol{b}_i^{(t)}) \qquad (\because \mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)) \\
&= \mathrm{Var}(\boldsymbol{W}_i^{(t)})\; \mathrm{Var}(\boldsymbol{h}^{(t-1)}) \; + \mathrm{Var}(\boldsymbol{W}_i^{(t)})\; \mathbb{E}[\boldsymbol{h}^{(t-1)}]^2 + \mathbb{E}[\boldsymbol{W}_i^{(t)}]^2 \mathrm{Var}(\boldsymbol{h}^{(t-1)}) + \mathrm{Var}(\boldsymbol{b}_i^{(t)}) \\
&\qquad (\because \mathrm{Var}(XY) = \mathrm{Var}(X)\mathrm{Var}(Y) + \mathrm{Var}(X)\mathbb{E}[Y]^2 + \mathbb{E}[X]^2\mathrm{Var}(Y) \text{ if } X \perp Y)
\end{aligned}$$

1.

$$\mathbb{E}[\boldsymbol{h}_i^{(t-1)}] = 0, \mathrm{Var}(\boldsymbol{h}_i^{(t-1)}) = 1 \text{ for } 1 \le i \le d^{(t-1)}$$

$$\implies \mathbb{E}[\boldsymbol{a}_i^{(t)}] = \mathbb{E}[\boldsymbol{W}_i^{(t)}]\, \cancelto{\mathbf{0}_{d^{(t-1)}}}{\mathbb{E}[\boldsymbol{h}^{(t-1)}]} + \cancelto{c}{\mathbb{E}[\boldsymbol{b}_i^{(t)}]} = c,$$

$$(\boldsymbol{b}_i^{(t)} = c \implies \mathbb{E}[\boldsymbol{b}_i^{(t)}] = c)$$

$$\mathrm{Var}(\boldsymbol{a}_i^{(t)}) = \mathrm{Var}(\boldsymbol{W}_i^{(t)})\, \cancelto{\mathbf{1}_{d^{(t-1)}}}{\mathrm{Var}(\boldsymbol{h}^{(t-1)})} + \mathrm{Var}(\boldsymbol{W}_i^{(t)})\, \cancelto{\mathbf{0}_{d^{(t-1)}}}{\mathbb{E}[\boldsymbol{h}^{(t-1)}]^2} + \mathbb{E}[\boldsymbol{W}_i^{(t)}]^2 \cancelto{\mathbf{1}_{d^{(t-1)}}}{\mathrm{Var}(\boldsymbol{h}^{(t-1)})}$$

$$+ \cancelto{0}{\mathrm{Var}(\boldsymbol{b}_i^{(t)})} \quad (\boldsymbol{b}_i^{(t)} = c \implies \mathrm{Var}(\boldsymbol{b}_i^{(t)}) = 0)$$

$$= \mathrm{Var}(\boldsymbol{W}_i^{(t)})\ \mathbf{1}_{d^{(t-1)}} + \mathbb{E}[\boldsymbol{W}_i^{(t)}]^2\ \mathbf{1}_{d^{(t-1)}}$$

(a)

$$\boldsymbol{W}_{ij}^{(t)} \sim \mathcal{N}(\mu, \sigma^2) \implies \mathbb{E}[\boldsymbol{W}_i^{(t)}] = \mu \mathbf{1}_{d^{(t-1)}}^\top,\ \mathrm{Var}(\boldsymbol{W}_i^{(t)}) = \sigma^2 \mathbf{1}_{d^{(t-1)}}^\top$$

$$\implies \mathbb{E}[\boldsymbol{a}_i^{(t)}] = c,$$

$$\mathrm{Var}(\boldsymbol{a}_i^{(t)}) = \mathrm{Var}(\boldsymbol{W}_i^{(t)})\mathbf{1}_{d^{(t-1)}} + \mathbb{E}[\boldsymbol{W}_i^{(t)}]^2 \mathbf{1}_{d^{(t-1)}} = \sigma^2\ \mathbf{1}_{d^{(t-1)}}^\top \mathbf{1}_{d^{(t-1)}} + \mu^2\ \mathbf{1}_{d^{(t-1)}}^\top \mathbf{1}_{d^{(t-1)}}$$

$$= (\sigma^2 + \mu^2) d^{(t-1)} \qquad (\because \mathbf{1}_{d^{(t-1)}}^\top \mathbf{1}_{d^{(t-1)}} = d^{(t-1)})$$

$$\mathbb{E}[\boldsymbol{a}_i^{(t)}] = 0, \mathrm{Var}(\boldsymbol{a}_i^{(t)}) = 1 \implies c = 0,\ \sigma^2 + \mu^2 = 1/d^{(t-1)}$$

One possible combination of μ and σ that satisfies this is $c = 0,\ \mu = 0,\ \sigma = 1/\sqrt{d^{(t-1)}}$. This is Glorot Normal initialization, considering only fan_in.

(b)

$$\boldsymbol{W}_{ij}^{(t)} \sim U(\alpha, \beta) \implies \mathbb{E}[\boldsymbol{W}_i^{(t)}] = \frac{\alpha+\beta}{2}\mathbf{1}^\top,\ \mathrm{Var}(\boldsymbol{W}_i^{(t)}) = \frac{(\beta-\alpha)^2}{12}\mathbf{1}^\top$$

$$\implies \mathbb{E}[\boldsymbol{a}_i^{(t)}] = c,$$

$$\mathrm{Var}(\boldsymbol{a}_i^{(t)}) = \mathrm{Var}(\boldsymbol{W}_i^{(t)})\mathbf{1} + \mathbb{E}[\boldsymbol{W}_i^{(t)}]^2\mathbf{1} = \tfrac{(\beta-\alpha)^2}{12}\mathbf{1}^\top\mathbf{1} + (\tfrac{\alpha+\beta}{2})^2\mathbf{1}^\top\mathbf{1}$$

$$= (\tfrac{\alpha^2+\beta^2-2\alpha\beta}{12})d^{(t-1)} + (\tfrac{\alpha^2+\beta^2+2\alpha\beta}{4})d^{(t-1)} = (\tfrac{\alpha^2+\beta^2-2\alpha\beta+3\alpha^2+3\beta^2+6\alpha\beta}{12})d^{(t-1)}$$

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

$$\mathbb{E}[\boldsymbol{a}_i^{(t)}] = 0, \mathrm{Var}(\boldsymbol{a}_i^{(t)}) = 1 \implies c = 0, (\alpha^2+\beta^2+\alpha\beta)d^{(t-1)}/3 = 1$$

One possible combination of α and β that satisfies this is $=0,\ \alpha = -\sqrt{\frac{3}{d^{(t-1)}}},\ \beta = \sqrt{\frac{3}{d^{(t-1)}}}$. This is Glorot Uniform initialization, considering only fan_in.

2. $\mathbb{E}[\boldsymbol{a}_i^{(t-1)}] = 0$, $\mathrm{Var}(\boldsymbol{a}_i^{(t-1)}) = 1$ 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. ReLU activation: $g(x) = \max\{0, x\}$.

We know that $\mathbb{E}[\boldsymbol{b}_i^{(t)}] = c$ and $\mathrm{Var}(\boldsymbol{b}_i^{(t)}) = 0\ \because\ \boldsymbol{b}_i = c\ \forall\ 1 < i < d^{(t)}$.

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

Here,

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{h}_i^{(t)})^2] &= \int_{-\infty}^{+\infty} \max\{0, \boldsymbol{a}_i^{(t)}\}^2 \; p(\boldsymbol{a}_i^{(t)}) \; d\boldsymbol{a}_i^{(t)} = \int_{0}^{+\infty} (\boldsymbol{a}_i^{(t)})^2 \; p(\boldsymbol{a}_i^{(t)}) \; d\boldsymbol{a}_i^{(t)} \\
&= \frac{1}{2} \int_{-\infty}^{+\infty} (\boldsymbol{a}_i^{(t)})^2 \; p(\boldsymbol{a}_i^{(t)}) \; d\boldsymbol{a}_i^{(t)} \qquad [\because \boldsymbol{a}_i^{(t)} \text{is symmetric}] \\
&= \frac{1}{2} \; \mathbb{E}[(\boldsymbol{a}_i^{(t)})^2] = \frac{1}{2} \Big(\mathbb{E}[(\boldsymbol{a}_i^{(t)})^2] - \mathbb{E}[\boldsymbol{a}_i^{(t)}]^2 \Big) \;\; [\because \mathbb{E}[\boldsymbol{a}_i^{(t)}] = 0 \text{ acc. to the question}] \\
&= \frac{1}{2} \text{Var}(\boldsymbol{a}_i^{(t)}) = \frac{1}{2} \qquad [\because \text{Var}(\boldsymbol{a}_i^{(t)}) = 1 \text{ acc. to the question}]
\end{aligned}$$

$$\therefore \text{Var}(\boldsymbol{h}_i^{(t)}) = \mathbb{E}[(\boldsymbol{h}_i^{(t)})^2] - \mathbb{E}[\boldsymbol{h}_i^{(t)}]^2 = \tfrac{1}{2} - \mathbb{E}[\boldsymbol{h}_i^{(t)}]^2 \implies \text{Var}(\boldsymbol{h}^{(t)}) = \tfrac{1}{2} \mathbf{1}_{d^{(t)}} - \mathbb{E}[\boldsymbol{h}^{(t)}]^2$$

From the equations derived at the beginning,

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{a}_i^{(t)}] = \mathbb{E}[\boldsymbol{W}_i^{(t)}] \; \mathbb{E}[\boldsymbol{h}^{(t-1)}] \; + \; \mathbb{E}[\boldsymbol{b}_i^{(t)}] = \mathbb{E}[\boldsymbol{W}_i^{(t)}] \; \mathbb{E}[\boldsymbol{h}^{(t-1)}] + c \qquad [\because \mathbb{E}[\boldsymbol{b}_i^{(t)}] = c] \\
&\implies 0 = \mathbb{E}[\boldsymbol{W}_i^{(t)}] \; \mathbb{E}[\boldsymbol{h}^{(t-1)}] + c \\
&\implies \mathbb{E}[\boldsymbol{W}_i^{(t)}] \; \mathbb{E}[\boldsymbol{h}^{(t-1)}] = -c \\
&\implies (\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] \mathbf{1}_{d^{(t-1)}}^\top) \; (\mathbb{E}[\boldsymbol{h}_i^{(t-1)}] \mathbf{1}_{d^{(t-1)}}) = -c \\
&\implies \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] \; \mathbb{E}[\boldsymbol{h}_i^{(t-1)}] \; \mathbf{1}_{d^{(t-1)}}^\top \mathbf{1}_{d^{(t-1)}} = -c \\
&\implies \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] \; \mathbb{E}[\boldsymbol{h}_i^{(t-1)}] \; d^{(t-1)} = -c \\
&\implies \mathbb{E}[\boldsymbol{h}_i^{(t-1)}] = -\frac{c}{\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}] \; d^{(t-1)}}
\end{aligned}$$

$$\begin{aligned}
&\text{Var}(\boldsymbol{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 + \mathbb{E}[\boldsymbol{W}_i^{(t)}]^2 \text{Var}(\boldsymbol{h}^{(t-1)}) + \text{Var}(\boldsymbol{b}_i^{(t)}) \\
&\implies 1 = \text{Var}(\boldsymbol{W}_i^{(t)}) \Big(\frac{1}{2} \mathbf{1}_{d^{(t-1)}} - \mathbb{E}[\boldsymbol{h}^{(t-1)}]^2 \Big) + \text{Var}(\boldsymbol{W}_i^{(t)}) \; \mathbb{E}[\boldsymbol{h}^{(t-1)}]^2 \\
&\qquad + \; \mathbb{E}[\boldsymbol{W}_i^{(t)}]^2 \Big(\frac{1}{2} \mathbf{1}_{d^{(t-1)}} - \mathbb{E}[\boldsymbol{h}^{(t-1)}]^2 \Big) + 0 \\
&\implies 1 \; = \; \frac{1}{2} \text{Var}(\boldsymbol{W}_i^{(t)}) \mathbf{1}_{d^{(t-1)}} \cancel{- \text{Var}(\boldsymbol{W}_i^{(t)}) \; \mathbb{E}[\boldsymbol{h}^{(t-1)}]^2} + \cancel{\text{Var}(\boldsymbol{W}_i^{(t)}) \; \mathbb{E}[\boldsymbol{h}^{(t-1)}]^2} \\
&\qquad + \frac{1}{2} \mathbb{E}[\boldsymbol{W}_i^{(t)}]^2 \mathbf{1}_{d^{(t-1)}} - \mathbb{E}[\boldsymbol{W}_i^{(t)}]^2 \; \mathbb{E}[\boldsymbol{h}^{(t-1)}]^2 \\
&\implies 1 \; = \; \frac{1}{2} \text{Var}(\boldsymbol{W}_i^{(t)}) \mathbf{1}_{d^{(t-1)}} + \frac{1}{2} \mathbb{E}[\boldsymbol{W}_i^{(t)}]^2 \mathbf{1}_{d^{(t-1)}} - \mathbb{E}[\boldsymbol{W}_i^{(t)}]^2 \; \mathbb{E}[\boldsymbol{h}^{(t-1)}]^2 \\
&\implies 1 \; = \; \frac{1}{2} \text{Var}(\boldsymbol{W}_{ij}^{(t)}) \mathbf{1}_{d^{(t-1)}}^\top \mathbf{1}_{d^{(t-1)}} \; + \; \frac{1}{2} \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]^2 \mathbf{1}_{d^{(t-1)}}^\top \mathbf{1}_{d^{(t-1)}} \\
&\qquad - \; (\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]^2 \mathbf{1}_{d^{(t-1)}}^\top) \; (\mathbb{E}[\boldsymbol{h}_i^{(t-1)}]^2 \mathbf{1}_{d^{(t-1)}}) \\
&\implies 1 \; = \; \frac{1}{2} \; \text{Var}(\boldsymbol{W}_{ij}^{(t)}) \; d^{(t-1)} \; + \; \frac{1}{2} \; \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]^2 \; d^{(t-1)} \; - \; \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]^2 \; \mathbb{E}[\boldsymbol{h}_i^{(t-1)}]^2 \; d^{(t-1)} \\
&\implies 1 \; = \; \frac{1}{2} \; \text{Var}(\boldsymbol{W}_{ij}^{(t)}) \; d^{(t-1)} \; + \; \frac{1}{2} \; \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]^2 \; d^{(t-1)} \; - \; \cancel{\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]^2} \Big(- \frac{c}{\cancel{\mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]} \; d^{(t-1)}} \Big)^2 \; d^{(t-1)} \\
&\implies 1 \; = \; \frac{1}{2} \; \text{Var}(\boldsymbol{W}_{ij}^{(t)}) \; d^{(t-1)} \; + \; \frac{1}{2} \; \mathbb{E}[\boldsymbol{W}_{ij}^{(t)}]^2 \; d^{(t-1)} \; - \; \frac{c^2}{d^{(t-1)}}
\end{aligned}$$

(a)

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

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

One possible combination of c, μ and σ that satisfies this is $c = 0,\ \mu = 0,\ \sigma = \sqrt{\frac{2}{d^{(t-1)}}}$. This is He Normal initialization.

(b)

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

$$\implies 1 = \frac{1}{2}\,\frac{(\beta-\alpha)^2}{12}\,d^{(t-1)} + \frac{1}{2}\left(\frac{\alpha+\beta}{2}\right)^2 d^{(t-1)} - \frac{c^2}{d^{(t-1)}}$$
$$\implies c^2 = \frac{1}{2}\,\frac{\alpha^2+\beta^2-2\alpha\beta}{12}\,(d^{(t-1)})^2 + \frac{1}{2}\,\frac{\alpha^2+\beta^2+2\alpha\beta}{4}\,(d^{(t-1)})^2 - d^{(t-1)}$$
$$\implies c^2 = \frac{1}{2}\,\frac{\alpha^2+\beta^2-2\alpha\beta+3\alpha^2+3\beta^2+6\alpha\beta}{12}\,(d^{(t-1)})^2 - d^{(t-1)}$$
$$\implies c^2 = \frac{1}{2}\,\frac{4\alpha^2+4\beta^2+4\alpha\beta}{12}\,(d^{(t-1)})^2 - d^{(t-1)}$$
$$\implies c^2 = \frac{\alpha^2+\beta^2+\alpha\beta}{6}\,(d^{(t-1)})^2 - d^{(t-1)}$$

One possible combination of c, μ and σ that satisfies this is $c = 0,\ \alpha = -\sqrt{\frac{6}{d^{(t-1)}}},\ \beta = \sqrt{\frac{6}{d^{(t-1)}}}$. This is He Uniform initialization.

(c) The values proposed above are He initialization.

(d) This would work well in practice since ReLU has proven to be a very popular non-linearity because of its convenient gradient, and it is desirable to have unit variance of pre-activations of non-linear activations. This was missing in the previous (Glorot) initialization scheme.

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 $\Gamma_{ii} = (\boldsymbol{X}^\top \boldsymbol{X})^{1/2}_{ii}$. Show that the *expectation (over $\boldsymbol{R}$)* of the loss function can be rewritten as $L(\boldsymbol{w}) = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma \boldsymbol{w}||^2$.
3. Show that the solution $\boldsymbol{w}^{\text{dropout}}$ that minimizes the expected loss from question 2.2 satisfies

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

   where λ^{dropout} is a regularization coefficient depending on p. How does the value of p affect the regularization coefficient, λ^{dropout} ?
4. Express the solution $\boldsymbol{w}^{L_2}$ for a linear regression problem without dropout and with L^2 regularization, with regularization coefficient λ^{L_2} in closed form.
5. Compare the results of 2.3 and 2.4: identify specific differences in the equations you arrive at, and discuss qualitatively what the equations tell you about the similarities and differences in the effects of weight decay and dropout (1-3 sentences).

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

2. $\mathbb{E}_{\boldsymbol{R}}[L(\boldsymbol{w})] = \mathbb{E}_{\boldsymbol{R}}\big[||\boldsymbol{y} - (\boldsymbol{R} * \boldsymbol{X}).\boldsymbol{w}||^2\big] = \sum_{i=1}^n \mathbb{E}_{\boldsymbol{R}}\big[\big(\boldsymbol{y}_i - (\boldsymbol{R}_i * \boldsymbol{X}_i).\boldsymbol{w}\big)^2\big]$
   We know (from Assignment 0) that:

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

$$\begin{aligned}\therefore \mathbb{E}_{\boldsymbol{R}}[L(\boldsymbol{w})] &= \sum_{i=1}^n (\mathbb{E}_{\boldsymbol{R}}\big[\boldsymbol{y}_i - (\boldsymbol{R}_i * \boldsymbol{X}_i).\boldsymbol{w}\big])^2 + \text{Var}_{\boldsymbol{R}}\big(\boldsymbol{y}_i - (\boldsymbol{R}_i * \boldsymbol{X}_i).\boldsymbol{w}\big) \\ &= \sum_{i=1}^n (\boldsymbol{y}_i - (\mathbb{E}_{\boldsymbol{R}}[\boldsymbol{R}_i] * \boldsymbol{X}_i).\boldsymbol{w})^2 + \text{Var}_{\boldsymbol{R}}\big(\boldsymbol{y}_i - (\boldsymbol{R}_i * \boldsymbol{X}_i).\boldsymbol{w}\big)\end{aligned}$$

   $\sum_{i=1}^n (\boldsymbol{y}_i - (\mathbb{E}_{\boldsymbol{R}}(\boldsymbol{R}_i) * \boldsymbol{X}_i).\boldsymbol{w})^2 = \sum_{i=1}^n (\boldsymbol{y}_i - p\boldsymbol{X}_i.\boldsymbol{w})^2 = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2$

   $\sum_{i=1}^n \text{Var}_{\boldsymbol{R}}\big(\boldsymbol{y}_i - (\boldsymbol{R}_i * \boldsymbol{X}_i).\boldsymbol{w}\big) = \sum_{i=1}^n \mathbb{E}_{\boldsymbol{R}}\big[\big(\boldsymbol{y}_i - (\boldsymbol{R}_i * \boldsymbol{X}_i).\boldsymbol{w} - \mathbb{E}_{\boldsymbol{R}}[\boldsymbol{y}_i - (\boldsymbol{R}_i * \boldsymbol{X}_i).\boldsymbol{w}]\big)^2\big]$
   $= \sum_{i=1}^n \mathbb{E}_{\boldsymbol{R}}\big[\big(\boldsymbol{y}_i - (\boldsymbol{R}_i * \boldsymbol{X}_i).\boldsymbol{w} - \boldsymbol{y}_i + p\boldsymbol{X}_i.\boldsymbol{w}\big)^2\big] = \sum_{i=1}^n \mathbb{E}_{\boldsymbol{R}}\big[\big((p\boldsymbol{X}_i - \boldsymbol{R}_i * \boldsymbol{X}_i).\boldsymbol{w}\big)^2\big]$
   $= \sum_{i=1}^n \mathbb{E}_{\boldsymbol{R}}\big[\boldsymbol{w}^\top.(p\boldsymbol{X}_i^\top - \boldsymbol{R}_i^\top * \boldsymbol{X}_i^\top).(p\boldsymbol{X}_i - \boldsymbol{R}_i * \boldsymbol{X}_i).\boldsymbol{w}\big]$
   $= \sum_{i=1}^n (\boldsymbol{w}^\top.(p^2\boldsymbol{X}^\top\boldsymbol{X}_i - p\mathbb{E}_{\boldsymbol{R}}[\boldsymbol{R}_i]\boldsymbol{X}_i^\top\boldsymbol{X}_i - p\mathbb{E}_{\boldsymbol{R}}[\boldsymbol{R}_i^\top]\boldsymbol{X}_i^\top\boldsymbol{X}_i + \mathbb{E}[\boldsymbol{R}^\top\boldsymbol{R}](\boldsymbol{X}_i^\top\boldsymbol{X}_i)).\boldsymbol{w})$
   $= \sum_{i=1}^n (\boldsymbol{w}^\top.(\cancel{p^2\boldsymbol{X}_i^\top\boldsymbol{X}_i} - \cancel{p^2\boldsymbol{X}_i^\top\boldsymbol{X}_i} - p^2\boldsymbol{X}_i^\top\boldsymbol{X}_i + p(\boldsymbol{X}_i^\top\boldsymbol{X}_i)).\boldsymbol{w})$
   [$\because$ expectation of square of Bernoulli random variable of mean p is also p]
   $= \sum_{i=1}^n (\boldsymbol{w}^\top.p(1-p)(\boldsymbol{X}_i^\top\boldsymbol{X}_i).\boldsymbol{w})$
   $= p(1-p)(\boldsymbol{w}^\top.\,\text{diag}(\boldsymbol{X}^\top\boldsymbol{X}).\boldsymbol{w})$
   $= p(1-p)||\Gamma\boldsymbol{w}||^2$

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

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

$$\frac{d\mathbb{E}_{\boldsymbol{R}}[L(\boldsymbol{w})]}{d\boldsymbol{w}} = 0 \implies 2(-p\boldsymbol{X}^\top)(\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}^{\text{dropout}}) + p(1-p)2\Gamma^\top(\Gamma\boldsymbol{w}^{\text{dropout}}) = 0$$
$$\implies -2p\boldsymbol{X}^\top\boldsymbol{y} + 2p^2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}^{\text{dropout}} + 2p(1-p)(\Gamma^\top\Gamma)\boldsymbol{w}^{\text{dropout}} = 0$$

$\Gamma^\top\Gamma = \big(\operatorname{diag}(\boldsymbol{X}^\top\boldsymbol{X})^{1/2}\big)^\top\big(\operatorname{diag}(\boldsymbol{X}^\top\boldsymbol{X})^{1/2}\big) = \operatorname{diag}(\boldsymbol{X}^\top\boldsymbol{X})^{1/2}\operatorname{diag}(\boldsymbol{X}^\top\boldsymbol{X})^{1/2}$
$= \big(\operatorname{diag}(\boldsymbol{X}^\top\boldsymbol{X})^{1/2}\big)^2 = \Gamma^2$

$$\implies -\cancel{2}p\boldsymbol{X}^\top\boldsymbol{y} + \cancel{2}p^2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}^{\text{dropout}} + \cancel{2}p(1-p)\Gamma^2\boldsymbol{w}^{\text{dropout}} = 0$$
$$\implies p^2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}^{\text{dropout}} + p(1-p)\Gamma^2\boldsymbol{w}^{\text{dropout}} = p\boldsymbol{X}^\top\boldsymbol{y}$$
$$\implies \big(p\boldsymbol{X}^\top\boldsymbol{X} + (1-p)\Gamma^2\big)p\boldsymbol{w}^{\text{dropout}} = p\boldsymbol{X}^\top\boldsymbol{y}$$
$$\implies \Big(\boldsymbol{X}^\top\boldsymbol{X} + \Big(\frac{1}{p} - 1\Big)\Gamma^2\Big)p\boldsymbol{w}^{\text{dropout}} = \boldsymbol{X}^\top\boldsymbol{y}$$
$$\implies p\boldsymbol{w}^{\text{dropout}} = \Big(\boldsymbol{X}^\top\boldsymbol{X} + \Big(\frac{1}{p} - 1\Big)\Gamma^2\Big)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

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

When $p = 1$, $\lambda^{\text{dropout}} = 0 \implies$ no contribution from the regularization term in $L(\boldsymbol{w})$ involving Γ.

As $p \to 0$, $\lambda^{\text{dropout}} \to \infty$, so the contribution from the regularization increases.

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

$$\frac{dL(\boldsymbol{w})}{d\boldsymbol{w}} = 0 \implies 2(-\boldsymbol{X}^\top)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}^{L_2}) + 2\lambda^{L_2}\boldsymbol{w}^{L_2} = 0$$
$$\implies -\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}^{L_2} + \lambda^{L_2}\boldsymbol{w}^{L_2} = 0$$
$$\implies \boldsymbol{w}^{L_2} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{L_2})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

5. From the equation in 2.4, it is clear that the regularization offered by $L2$ merely reduces the value of the optimal closed-form weight — by increasing the denominator value by λ^{L_2}, the overall values of the weights are decreased.

   However, from the equation in 2.3, it is clear that the regularization offered by dropout is data-dependent. Γ qualitatively represents the standard deviation of the data in each of its dimensions. Hence, in the dimensions where the data varies more, the values of the weights are reduced more.

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 $\boldsymbol{v}_t$ to eliminate such a bias by rescaling.

Answer 3. 1. Expressing both update rules in recursive form:

— SGD with momentum: $\boldsymbol{v}_t = \alpha\boldsymbol{v}_{t-1} + \epsilon\boldsymbol{g}_t,\ \Delta\boldsymbol{\theta}_t = -\boldsymbol{v}_t$, where $\epsilon > 0$ and $\alpha \in (0, 1)$.
$\Delta\boldsymbol{\theta}_t = -\boldsymbol{v}_t = -\alpha\boldsymbol{v}_{t-1} - \epsilon\boldsymbol{g}_t = \alpha\Delta\boldsymbol{\theta}_{t-1} - \epsilon\boldsymbol{g}_t \qquad [\because \Delta\boldsymbol{\theta}_{t-1} = -\boldsymbol{v}_{t-1}]$

— SGD with running average of $\boldsymbol{g}_t$: $\boldsymbol{v}_t = \beta\boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t,\ \Delta\boldsymbol{\theta}_t = -\delta\boldsymbol{v}_t$, where $\beta \in (0, 1)$ and $\delta > 0$.
$\Delta\boldsymbol{\theta}_t = -\delta\boldsymbol{v}_t = -\delta\beta\boldsymbol{v}_{t-1} - \delta(1-\beta)\boldsymbol{g}_t = \beta\Delta\boldsymbol{\theta}_{t-1} - \delta(1-\beta)\boldsymbol{g}_t \qquad [\because \Delta\boldsymbol{\theta}_{t-1} = -\delta\boldsymbol{v}_{t-1}]$

Thus, we can see that both update rules are equivalent, with $(\alpha,\ \epsilon) = (\beta,\ \delta(1-\beta))$.

2. Running average update rule has:

$$\begin{aligned}
\boldsymbol{v}_t &= \beta\boldsymbol{v}_{t-1} + (1-\beta)\boldsymbol{g}_t = \beta(\beta\boldsymbol{v}_{t-2} + (1-\beta)\boldsymbol{g}_{t-1}) + (1-\beta)\boldsymbol{g}_t \\
&= \beta^2\boldsymbol{v}_{t-2} + \beta(1-\beta)\boldsymbol{g}_{t-1} + (1-\beta)\boldsymbol{g}_t = \beta^2(\beta\boldsymbol{v}_{t-3} + (1-\beta)\boldsymbol{g}_{t-2}) + \beta(1-\beta)\boldsymbol{g}_{t-1} + (1-\beta)\boldsymbol{g}_t \\
&= \beta^3\boldsymbol{v}_{t-3} + \beta^2(1-\beta)\boldsymbol{g}_{t-2} + \beta(1-\beta)\boldsymbol{g}_{t-1} + (1-\beta)\boldsymbol{g}_t \\
&\ \ \ldots \\
&= \beta^t\boldsymbol{v}_0 + \sum_{i=1}^{t}\beta^{t-i}(1-\beta)\boldsymbol{g}_i \\
&= (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\boldsymbol{g}_i \qquad [\because \boldsymbol{v}_0 = 0]
\end{aligned}$$

3. If each sample of the gradient $\boldsymbol{g}_t$ is independent of time t, we know from 3.2 that:

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

$$\begin{aligned}
\mathbb{E}[\boldsymbol{v}_t] &= \mathbb{E}\Big[(1-\beta)\sum_{i=1}^{t}\beta^{t-i}\boldsymbol{g}_i\Big] \\
&= (1-\beta)\ \sum_{i=1}^{t}\beta^{t-i}\ \mathbb{E}[\boldsymbol{g}_t] \qquad [\because \mathbb{E} \text{ of sum is sum of } \mathbb{E}, \text{ and } \beta \text{ is a constant}] \\
&= (1-\beta)\ \mathbb{E}[\boldsymbol{g}_t]\sum_{i=1}^{t}\beta^{t-i} = (1-\beta)\ \mathbb{E}[\boldsymbol{g}_t]\ \sum_{i=0}^{t-1}\beta^{i} = (1-\beta)\ \mathbb{E}[\boldsymbol{g}_t]\ \frac{(1-\beta^t)}{(1-\beta)} \\
&= \mathbb{E}[\boldsymbol{g}_t]\ (1-\beta^t)
\end{aligned}$$

Thus, we can see that the running average is biased.
We can see that this bias can be removed if we rescale $\boldsymbol{v}_t$ by this bias (as is done in Adam for each of the moments calculated):

$$\tilde{\boldsymbol{v}}_t = \boldsymbol{v}_t/(1-\beta^t) \qquad \Delta\boldsymbol{\theta}_t = -\delta\tilde{\boldsymbol{v}}_t$$

Question 4 (5-5-5)**.** This question is about weight normalization. We consider the following parameterization of a weight vector $\boldsymbol{w}$:

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

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

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

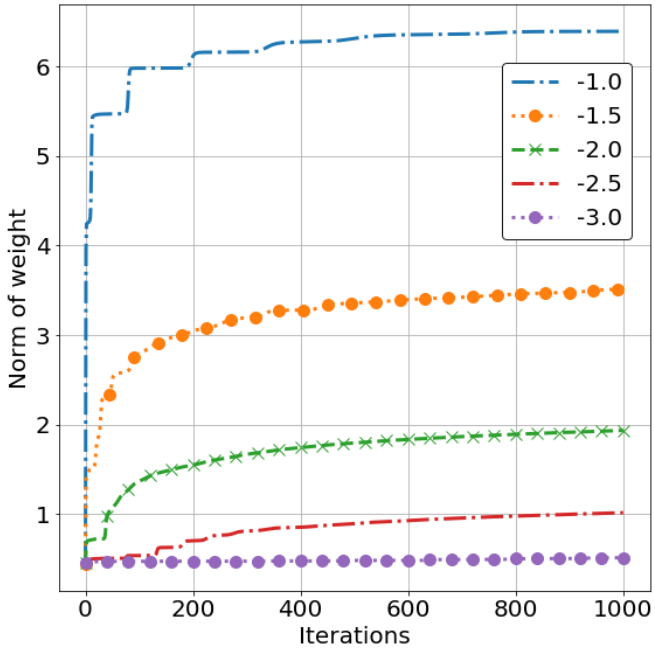


FIGURE 1 – Norm of parameters with different learning rate.

Answer 4. 1. Given: $\boldsymbol{w} := \gamma \frac{\boldsymbol{u}}{||\boldsymbol{u}||}$, $\hat{y} = \gamma \cdot \frac{y-\mu_y}{\sigma_y} + \beta$, $y = \boldsymbol{u}^\top \boldsymbol{x}$, $\mathbb{E}[\boldsymbol{x}] = 0$, $\text{Var}(\boldsymbol{x}) = \boldsymbol{I}$.

$\mu_y = \mathbb{E}[y] = \mathbb{E}[\boldsymbol{u}^\top \boldsymbol{x}] = \boldsymbol{u}^\top \mathbb{E}[\boldsymbol{x}] = 0$
$\text{Var}(y) = \text{Var}(\boldsymbol{u}^\top \boldsymbol{x}) = \boldsymbol{u}^\top \text{Var}(\boldsymbol{x}) \boldsymbol{u} = \boldsymbol{u}^\top \boldsymbol{I} \boldsymbol{u} = \boldsymbol{u}^\top \boldsymbol{u} = ||\boldsymbol{u}||^2 \implies \sigma_y = ||\boldsymbol{u}||$

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

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. To show that $\nabla_{\boldsymbol{u}} L = s\boldsymbol{W}^{\perp}\nabla_{\boldsymbol{w}} L$, given $\boldsymbol{W}^{\perp} = \Big(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^{\top}}{||\boldsymbol{u}||^2}\Big)$.

$$\nabla_{\boldsymbol{u}} L = \nabla_{\boldsymbol{u}} \boldsymbol{w} \cdot \nabla_{\boldsymbol{w}} L = \nabla_{\boldsymbol{u}} \Big(\gamma \frac{\boldsymbol{u}}{||\boldsymbol{u}||}\Big) \cdot \nabla_{\boldsymbol{w}} L = \gamma \Big(\nabla_{\boldsymbol{u}} \frac{\boldsymbol{u}}{||\boldsymbol{u}||}\Big) \cdot \nabla_{\boldsymbol{w}} L$$

$$\begin{aligned}
\nabla_{\boldsymbol{u}} \frac{\boldsymbol{u}}{||\boldsymbol{u}||} &= \frac{||\boldsymbol{u}|| \frac{d\boldsymbol{u}}{d\boldsymbol{u}} - \boldsymbol{u} \frac{d||\boldsymbol{u}||}{d\boldsymbol{u}}}{||\boldsymbol{u}||^2} = \frac{1}{||\boldsymbol{u}||^2} \Big(||\boldsymbol{u}|| \boldsymbol{I} - \boldsymbol{u} \, \frac{d}{d\boldsymbol{u}} \big((\boldsymbol{u}^{\top}\boldsymbol{u})^{\frac{1}{2}} \big) \Big) \\
&= \frac{1}{||\boldsymbol{u}||^2} \Big(||\boldsymbol{u}|| \boldsymbol{I} - \boldsymbol{u} \cdot \frac{1}{2} (\boldsymbol{u}^{\top}\boldsymbol{u})^{-\frac{1}{2}} \cdot \frac{d(\boldsymbol{u}^{\top}\boldsymbol{u})}{d\boldsymbol{u}} \Big) = \frac{1}{||\boldsymbol{u}||^2} \Big(||\boldsymbol{u}|| \boldsymbol{I} - \boldsymbol{u} \cdot \frac{1}{2||\boldsymbol{u}||} \cdot 2\boldsymbol{u}^{\top} \Big) \\
&= \frac{1}{||\boldsymbol{u}||^2} \Big(||\boldsymbol{u}|| \boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^{\top}}{||\boldsymbol{u}||} \Big) = \frac{1}{||\boldsymbol{u}||} \Big(\boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{u}^{\top}}{||\boldsymbol{u}||^2} \Big) = \frac{1}{||\boldsymbol{u}||} \boldsymbol{W}^{\perp}
\end{aligned}$$

$\therefore \nabla_{\boldsymbol{u}} L = \gamma \Big(\frac{1}{||\boldsymbol{u}||} \boldsymbol{W}^{\perp} \Big) \cdot \nabla_{\boldsymbol{w}} L = \frac{\gamma}{||\boldsymbol{u}||} \boldsymbol{W}^{\perp} \nabla_{\boldsymbol{w}} L = s\boldsymbol{W}^{\perp} \nabla_{\boldsymbol{w}} L$, where $s = \frac{\gamma}{||\boldsymbol{u}||}$.

3. Weight update using gradient descent:

$$\boldsymbol{u} \;\leftarrow\; \boldsymbol{u} - \eta \nabla_{\boldsymbol{u}} L \;=\; \boldsymbol{u} - \eta s \boldsymbol{W}^{\perp} \nabla_{\boldsymbol{w}} L \;=\; \boldsymbol{u} - \frac{\eta\gamma}{||\boldsymbol{u}||} \boldsymbol{W}^{\perp} \nabla_{\boldsymbol{w}} L$$

We know that $\boldsymbol{u}$ and $\boldsymbol{W}^{\perp}$ are orthogonal to each other. Hence, the updated $\boldsymbol{u}$ is a vector addition of two vectors $\boldsymbol{u}$ and $-\frac{\eta\gamma}{||\boldsymbol{u}||} \boldsymbol{W}^{\perp} \nabla_{\boldsymbol{w}} L$ that are orthogonal to each other. We know by Pythagoras theorem that the length of the diagonal, is always greater than both the sides of the right-angled triangle. In this case, there is a right angle formed by the vectors $\boldsymbol{u}$ and $-\frac{\eta\gamma}{||\boldsymbol{u}||} \boldsymbol{W}^{\perp} \nabla_{\boldsymbol{w}} L$, hence their vector sum, the diagonal of said right-angled triangle, is always greater than the previous weight, one of the sides of the triangle. Hence, the norm of the weights must increase with more gradient updates.

However, we also observe that the second vector has $||\boldsymbol{u}||$ in its denominator. This means that as the norm of the weights increases, this denominator also increases, and so the second vector's magnitude decreases. Hence, the magnitude of the vector sum would be changing lesser and lesser from the magnitude of the first vector, $\boldsymbol{u}$. Hence, we also observe that the rate of increase of the weight norm decreases with updates.

Finally, we observe that the learning rate is in the numerator of the second vector. Hence, at every update, the magnitude of the second vector is greater for a greater rate. By Pythagoras' theorem, this implies that the magnitude of the sum of this vector with a vector orthogonal to it would also be greater. Hence, the rate of increase of the norm of the weights is greater for greater learning rate.

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 $\boldsymbol{h}_t$).
*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

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

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. $\boldsymbol{g}_t = \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1}+\boldsymbol{U}\boldsymbol{x}_t+\boldsymbol{b}) = \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1}+\boldsymbol{h}_t-\boldsymbol{W}\sigma(\boldsymbol{h}_{t-1})) = \sigma(\boldsymbol{h}_t+\boldsymbol{W}(\boldsymbol{g}_{t-1}-\sigma(\boldsymbol{h}_{t-1})))$
If we assume that $\boldsymbol{g}_{t-1} = \sigma(\boldsymbol{h}_{t-1})$, then $\boldsymbol{g}_t = \sigma(\boldsymbol{h}_t + \boldsymbol{W}(\boldsymbol{0})) = \sigma(\boldsymbol{h}_t)$
$\therefore \boldsymbol{g}_t = \sigma(\boldsymbol{h}_t)$ (assuming/necessitating $\boldsymbol{g}_0 = \sigma(\boldsymbol{h}_0)$)

2.

$$\left|\left|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_0}\right|\right| = \left|\left|\frac{\partial \boldsymbol{h}_T}{\partial \boldsymbol{h}_{T-1}}\right|\right| . \left|\left|\frac{\partial \boldsymbol{h}_{T-1}}{\partial \boldsymbol{h}_{T-2}}\right|\right| ... \left|\left|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}}\right|\right| ... \left|\left|\frac{\partial \boldsymbol{h}_1}{\partial \boldsymbol{h}_0}\right|\right|$$

$$\left|\left|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}}\right|\right| = \left|\left|\frac{\partial\big(\boldsymbol{W}\sigma(\boldsymbol{h}_{t-1}) + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}\big)}{\partial \boldsymbol{h}_{t-1}}\right|\right| = \left|\left|\boldsymbol{W}\ \frac{\partial \sigma(\boldsymbol{h}_t)}{\partial \boldsymbol{h}_{t-1}}\right|\right| = ||\boldsymbol{W}\ \sigma'||$$

$$\leq ||\boldsymbol{W}||\ ||\sigma'|| \qquad [\ \because ||\boldsymbol{A}\boldsymbol{B}|| \leq ||\boldsymbol{A}||\,||\boldsymbol{B}||\ ]$$

$$\leq ||\boldsymbol{W}|| \cdot \gamma \qquad [\ \because |\sigma'| \leq \gamma\ ]$$

$$= \sqrt{\lambda_1(\boldsymbol{W}^\top \boldsymbol{W})} \cdot \gamma \qquad [\ \because ||\boldsymbol{A}|| = \sqrt{\lambda_1(\boldsymbol{A}^\top \boldsymbol{A})}\ ]$$

$$\leq \sqrt{\frac{\delta^2}{\gamma^2}} \cdot \gamma \qquad [\ \because \lambda_1(\boldsymbol{W}^\top \boldsymbol{W}) \leq \frac{\delta^2}{\gamma^2}\ ]$$

$$= \left|\frac{\delta}{\gamma}\right| \cdot \gamma = \frac{\delta}{\gamma} \cdot \gamma \qquad [\ \because \delta > 0, \gamma > 0\ ]$$

$$= \delta$$

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

3. $\left|\left|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}}\right|\right| \leq \sqrt{\lambda_1(\boldsymbol{W}^\top \boldsymbol{W})} \cdot \gamma$; if $\lambda_1(\boldsymbol{W}^\top \boldsymbol{W}) > \frac{\delta^2}{\gamma^2}$, this does not say anything about the relation between $\left|\left|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}}\right|\right|$ and $\frac{\delta^2}{\gamma^2}$. Hence, this offers neither a necessary nor a sufficient condition for gradients to vanish or explode.

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

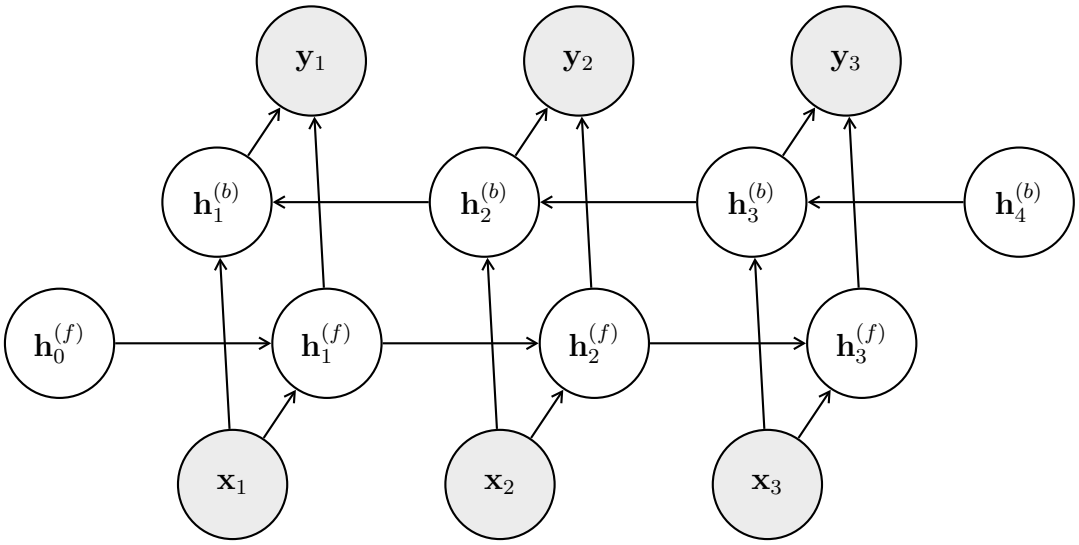
$$\begin{aligned}\boldsymbol{h}_t^{(f)} &= \sigma(\boldsymbol{W}^{(f)}\boldsymbol{x}_t + \boldsymbol{U}^{(f)}\boldsymbol{h}_{t-1}^{(f)}) \\ \boldsymbol{h}_t^{(b)} &= \sigma(\boldsymbol{W}^{(b)}\boldsymbol{x}_t + \boldsymbol{U}^{(b)}\boldsymbol{h}_{t+1}^{(b)}) \\ \boldsymbol{y}_t &= \boldsymbol{V}^{(f)}\boldsymbol{h}_t^{(f)}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

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

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

*2. Let $\boldsymbol{z}_t$ be the true target of the prediction $\boldsymbol{y}_t$ and consider the sum of squared loss $L = \sum_t L_t$ where $L_t = ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2$. Express the gradients $\nabla_{\boldsymbol{h}_t^{(f)}} L$ and $\nabla_{\boldsymbol{h}_t^{(b)}} L$ recursively (in terms of $\nabla_{\boldsymbol{h}_{t+1}^{(f)}} L$ and $\nabla_{\boldsymbol{h}_{t-1}^{(b)}} L$ respectively). Then derive $\nabla_{\boldsymbol{W}^{(f)}} L$ and $\nabla_{\boldsymbol{U}^{(b)}} L$.

Answer 6. 1. The unfolded bidirectional recurrent neural network for 3 time steps is:



2.

$$\nabla_{\boldsymbol{h}_t^{(f)}} L = \nabla_{\boldsymbol{h}_t^{(f)}} \sum_{i=t}^{T} L_i \qquad [\ \because \text{ there is no path from } \boldsymbol{h}_t^{(f)} \text{ to any } L_{i<t}\ ]$$

$$= \nabla_{\boldsymbol{h}_t^{(f)}} L_t + \nabla_{\boldsymbol{h}_t^{(f)}} \sum_{i=t+1}^{T} L_i$$

$$= \nabla_{\boldsymbol{h}_t^{(f)}} L_t + \nabla_{\boldsymbol{h}_t^{(f)}} \boldsymbol{h}_{t+1}^{(f)} \cdot \nabla_{\boldsymbol{h}_{t+1}^{(f)}} \sum_{i=t+1}^{T} L_i$$

$$= \nabla_{\boldsymbol{h}_t^{(f)}} L_t + \nabla_{\boldsymbol{h}_t^{(f)}} \boldsymbol{h}_{t+1}^{(f)} \cdot \nabla_{\boldsymbol{h}_{t+1}^{(f)}} L \qquad [\ \because \nabla_{\boldsymbol{h}_{t+1}^{(f)}} \sum_{i=t+1}^{T} L_i = \nabla_{\boldsymbol{h}_{t+1}^{(f)}} L\ ]$$

$$\nabla_{\boldsymbol{h}_t^{(f)}} L_t = \nabla_{\boldsymbol{h}_t^{(f)}} ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2 = 2||\boldsymbol{z}_t - \boldsymbol{y}_t||_2 \ \nabla_{\boldsymbol{h}_t^{(f)}} \big(\boldsymbol{z}_t - (\boldsymbol{V}^{(f)}\boldsymbol{h}_t^{(f)} + \boldsymbol{V}^{(b)}\boldsymbol{h}_t^{(b)})\big)$$
$$= 2||\boldsymbol{z}_t - \boldsymbol{y}_t||_2(-\boldsymbol{V}^{(f)}) = -2\boldsymbol{V}^{(f)}||\boldsymbol{z}_t - \boldsymbol{y}_t||_2$$

$$\nabla_{\boldsymbol{h}_t^{(f)}} \boldsymbol{h}_{t+1}^{(f)} = \nabla_{\boldsymbol{h}_t^{(f)}} \big(\sigma(\boldsymbol{W}^{(f)}\boldsymbol{x}_{t+1} + \boldsymbol{U}^{(f)}\boldsymbol{h}_t^{(f)})\big)$$
$$= \sigma(\boldsymbol{W}^{(f)}\boldsymbol{x}_{t+1} + \boldsymbol{U}^{(f)}\boldsymbol{h}_t^{(f)}) \cdot \big(1 - \sigma(\boldsymbol{W}^{(f)}\boldsymbol{x}_{t+1} + \boldsymbol{U}^{(f)}\boldsymbol{h}_t^{(f)})\big) \cdot \boldsymbol{U}^{(f)}$$
$$[\ \because \nabla_a \sigma(a) = \sigma(a)(1 - \sigma(a)),\ \text{and then applying chain rule}\ ]$$
$$= \boldsymbol{h}_{t+1}^{(f)}(1 - \boldsymbol{h}_{t+1}^{(f)})\boldsymbol{U}^{(f)}$$

$$\therefore \nabla_{\boldsymbol{h}_t^{(f)}} L = -2\boldsymbol{V}^{(f)}||\boldsymbol{z}_t - \boldsymbol{y}_t||_2 \ + \ \boldsymbol{h}_{t+1}^{(f)}(1 - \boldsymbol{h}_{t+1}^{(f)})\boldsymbol{U}^{(f)} \cdot \nabla_{\boldsymbol{h}_{t+1}^{(f)}} L$$

Similarly,

$$\nabla_{\boldsymbol{h}_t^{(b)}} L = \nabla_{\boldsymbol{h}_t^{(b)}} \sum_{i=0}^{t} L_i \qquad [\ \because \text{ there is no path from } \boldsymbol{h}_t^{(b)} \text{ to any } L_{i>t}\ ]$$

$$= \nabla_{\boldsymbol{h}_t^{(b)}} L_t + \nabla_{\boldsymbol{h}_t^{(b)}} \sum_{i=0}^{t-1} L_i$$

$$= \nabla_{\boldsymbol{h}_t^{(b)}} L_t + \nabla_{\boldsymbol{h}_t^{(b)}} \boldsymbol{h}_{t-1}^{(b)} \cdot \nabla_{\boldsymbol{h}_{t-1}^{(b)}} \sum_{i=0}^{t-1} L_i$$

$$= \nabla_{\boldsymbol{h}_t^{(b)}} L_t + \nabla_{\boldsymbol{h}_t^{(b)}} \boldsymbol{h}_{t-1}^{(b)} \cdot \nabla_{\boldsymbol{h}_{t-1}^{(b)}} L \qquad [\ \because \nabla_{\boldsymbol{h}_{t-1}^{(b)}} \sum_{i=0}^{t-1} L_i = \nabla_{\boldsymbol{h}_{t-1}^{(b)}} L\ ]$$

$$\nabla_{\boldsymbol{h}_t^{(b)}} L_t = \nabla_{\boldsymbol{h}_t^{(b)}} ||\boldsymbol{z}_t - \boldsymbol{y}_t||_2^2 = 2||\boldsymbol{z}_t - \boldsymbol{y}_t||_2 \ \nabla_{\boldsymbol{h}_t^{(b)}} \big(\boldsymbol{z}_t - (\boldsymbol{V}^{(f)}\boldsymbol{h}_t^{(f)} + \boldsymbol{V}^{(b)}\boldsymbol{h}_t^{(b)})\big)$$
$$= 2||\boldsymbol{z}_t - \boldsymbol{y}_t||_2(-\boldsymbol{V}^{(b)}) = -2\boldsymbol{V}^{(b)}||\boldsymbol{z}_t - \boldsymbol{y}_t||_2$$

$$\nabla_{\boldsymbol{h}_t^{(b)}} \boldsymbol{h}_{t-1}^{(b)} = \nabla_{\boldsymbol{h}_t^{(b)}} \big(\sigma(\boldsymbol{W}^{(b)}\boldsymbol{x}_{t-1} + \boldsymbol{U}^{(b)}\boldsymbol{h}_t^{(b)})\big)$$
$$= \sigma(\boldsymbol{W}^{(b)}\boldsymbol{x}_{t-1} + \boldsymbol{U}^{(b)}\boldsymbol{h}_t^{(b)}) \cdot \big(1 - \sigma(\boldsymbol{W}^{(b)}\boldsymbol{x}_{t-1} + \boldsymbol{U}^{(b)}\boldsymbol{h}_t^{(b)})\big) \cdot \boldsymbol{U}^{(b)}$$
$$[\ \because \nabla_a \sigma(a) = \sigma(a)(1 - \sigma(a)),\ \text{and then applying chain rule}\ ]$$
$$= \boldsymbol{h}_{t-1}^{(b)}(1 - \boldsymbol{h}_{t-1}^{(b)})\boldsymbol{U}^{(b)}$$

$$\therefore \nabla_{\boldsymbol{h}_t^{(b)}} L = -2\boldsymbol{V}^{(b)}||\boldsymbol{z}_t - \boldsymbol{y}_t||_2 \ + \ \boldsymbol{h}_{t-1}^{(b)}(1 - \boldsymbol{h}_{t-1}^{(b)})\boldsymbol{U}^{(b)} \cdot \nabla_{\boldsymbol{h}_{t-1}^{(b)}} L$$

$$\nabla_{\boldsymbol{W}^{(f)}} L = \nabla_{\boldsymbol{W}^{(f)}} \boldsymbol{h}_t^{(f)} \cdot \nabla_{\boldsymbol{h}_t^{(f)}} L \qquad [\because \text{ the path from L to } \boldsymbol{W}^{(f)} \text{ only involves } \boldsymbol{h}_t^{(f)} (\text{and not } \boldsymbol{h}_t^{(b)}) \,]$$

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

$$\implies \nabla_{\boldsymbol{W}^{(f)}} L = \boldsymbol{h}_t^{(f)} (1 - \boldsymbol{h}_t^{(f)}) \; \boldsymbol{x}_t \; \big(-2 \boldsymbol{V}^{(f)} || \boldsymbol{z}_t - \boldsymbol{y}_t ||_2 \; + \; \boldsymbol{h}_{t+1}^{(f)} (1 - \boldsymbol{h}_{t+1}^{(f)}) \boldsymbol{U}^{(f)} \cdot \nabla_{\boldsymbol{h}_{t+1}^{(f)}} L \big)$$

Similarly,

$$\nabla_{\boldsymbol{U}^{(b)}} L = \nabla_{\boldsymbol{U}^{(b)}} \boldsymbol{h}_t^{(b)} \cdot \nabla_{\boldsymbol{h}_t^{(b)}} L \qquad [\because \text{ the path from L to } \boldsymbol{U}^{(b)} \text{ only involves } \boldsymbol{h}_t^{(b)} (\text{and not } \boldsymbol{h}_t^{(f)}) \,]$$

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

$$\implies \nabla_{\boldsymbol{U}^{(b)}} L = \boldsymbol{h}_t^{(b)} (1 - \boldsymbol{h}_t^{(b)}) \; \boldsymbol{h}_{t+1}^{(b)} \; \big(-2 \boldsymbol{V}^{(b)} || \boldsymbol{z}_t - \boldsymbol{y}_t ||_2 \; + \; \boldsymbol{h}_{t-1}^{(b)} (1 - \boldsymbol{h}_{t-1}^{(b)}) \boldsymbol{U}^{(b)} \cdot \nabla_{\boldsymbol{h}_{t-1}^{(b)}} L \big)$$