Student : Ziqiang Wang

Question 1

1. We can express g_t in terms of h_t by :

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

To prove the induction step, we will assume that the expression holds for time step $t-1$:

$$\boldsymbol{g}_{t-1} = \sigma(\boldsymbol{h}_{t-1})$$

that is $\sigma^{-1}(\boldsymbol{g}_{t-1}) = \boldsymbol{h}_{t-1}$, where σ^{-1} is inverse activation function.
Then we need to show that the relationship also holds for time step t :

$$\boldsymbol{g}_t = \sigma^{-1}(\boldsymbol{h}_t)$$

We can start with the recurrence of h_t :

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

Substitute our induction assumption, $\sigma^{-1}(\boldsymbol{g}_{t-1}) = \boldsymbol{h}_{t-1}$:

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

Now apply the activation function σ to both sides :

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

According to the recurrence of $\boldsymbol{g}_t$, $\boldsymbol{g}_t = \sigma(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})$,the induction step is completed :

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

2. We can use the chain rule to express the gradient with respect to the initial hidden state as a product of gradients with respect to each intermediate hidden state :

$$\frac{\partial \boldsymbol{g}_T}{\partial \boldsymbol{g}_0} = \prod_{t=1}^{T} \frac{\partial \boldsymbol{g}_t}{\partial \boldsymbol{g}_{t-1}}$$

Using the recurrence relation for the hidden state, we have :

$$\frac{\partial \boldsymbol{g}_t}{\partial \boldsymbol{g}_{t-1}} = \frac{\partial \sigma'(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})}{\partial \boldsymbol{g}_{t-1}} = \sigma'(\boldsymbol{W}\boldsymbol{g}_{t-1} + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b})\boldsymbol{W}$$

Using the first property of the L2 operator norm in the question, we have :

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

Substitute the assumption, $|\sigma'(x)| \leq \gamma$:

$$\left|\left|\frac{\partial \boldsymbol{g}_t}{\partial \boldsymbol{g}_{t-1}}\right|\right| \leq \gamma ||\boldsymbol{W}||$$

Recursively apply this bound and the two properties, we get :

$$\begin{aligned}\left|\left|\frac{\partial \boldsymbol{g}_T}{\partial \boldsymbol{g}_0}\right|\right| &\leq \prod_{t=1}^{T}\left|\left|\frac{\partial \boldsymbol{g}_t}{\partial \boldsymbol{g}_{t-1}}\right|\right| \\ &\leq \gamma^T ||\boldsymbol{W}||^T \\ &= \gamma^T (\sqrt{\lambda_1(\boldsymbol{W}^\top \boldsymbol{W})})^T\end{aligned}$$

Substitute $\lambda_1(\boldsymbol{W}^\top \boldsymbol{W}) \leq \frac{\delta^2}{\gamma^2}$ where $\gamma > 0, 0 \leq \delta \leq 1$:

$$\left|\left|\frac{\partial \boldsymbol{g}_T}{\partial \boldsymbol{g}_0}\right|\right| \leq \gamma^T \left(\sqrt{\frac{\delta^2}{\gamma^2}}\right)^T = \gamma^T \sqrt{\frac{\delta^2}{\gamma^2}}^T = \delta^T$$

Thus, $\delta^T \to 0$ as $T \to \infty \implies \left|\left|\frac{\partial \boldsymbol{g}_T}{\partial \boldsymbol{g}_0}\right|\right| \to 0$ as $T \to \infty$

3. If the largest eigenvalue of the weights is larger than $\frac{\delta^2}{\gamma^2}$, then the gradients of the hidden state are likely to explode. however, this condition is necessary but not sufficient for the gradient to explode.

$$\left|\left|\frac{\partial \boldsymbol{g}_T}{\partial \boldsymbol{g}_0}\right|\right| \leq \gamma^T (\sqrt{\lambda_1(\boldsymbol{W}^\top \boldsymbol{W})})^T > \delta^T$$

Question 2

1. For the SGD with momentum, we have :

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

Since $\Delta\boldsymbol{\theta}_{t-1} = -\boldsymbol{v}_{t-1}$, we can write $\boldsymbol{v}_{t-1} = -\Delta\boldsymbol{\theta}_{t-1}$. Substituting this into the equation above, we have :

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

For the SGD with running average of $\boldsymbol{g}_t$, we have :

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

Since $\Delta\boldsymbol{\theta}_{t-1} = -\delta \boldsymbol{v}_{t-1}$, we can write $\boldsymbol{v}_{t-1} = -\frac{1}{\delta}\Delta\boldsymbol{\theta}_{t-1}$. Substituting this into the equation above, we have :

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

Now, to show that the two update rules are equivalent, we need to find a relationship between (α, ϵ) and (β, δ) by comparing the two expressions for $\Delta\boldsymbol{\theta}_t$:

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

To make these two expressions equal, we need :

$$\alpha = \beta \qquad \text{and} \qquad \epsilon = (1-\beta)\delta$$

2.

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

Continue this process for all t time steps, we have :

$$\boldsymbol{v}_t = \beta^t\boldsymbol{v}_0 + \sum_{i=1}^{t}(1-\beta)\beta^{t-i}\boldsymbol{g}_i$$

Since $\boldsymbol{v}_0$ is initialized as a vector of zeros, we can simplify the expression to :

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

3.

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

Taking the expectation of both sides :

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

Since $\boldsymbol{g}_t$ has a stationary distribution independent of t, we can have $\mathbb{E}[\boldsymbol{g}_i] = \mu_{\boldsymbol{g}}$, that is a constant value. Thus, we can rewrite the equation as :

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

Isolating $\mu_{\boldsymbol{g}}$:

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

Thus, we can estimate $\mathbb{E}[\boldsymbol{g}_i]$ using $\mathbb{E}[\boldsymbol{v}_t]$:

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

Question 3

1. We can express the one-step gradient descent update as follows :

$$x_1 = x_0 - \epsilon g$$

This question is to find the value of $\hat{f}_{x_0}(x_1)$ after the above update :

$$\hat{f}_{x_0}(x_1) = f(x_0) + (x_1 - x_0)^T g + \frac{1}{2}(x_1 - x_0)^T H (x_1 - x_0)$$

Substituting $x_1 = x_0 - \epsilon g$:

$$\hat{f}_{x_0}(x_1) = f(x_0) + (-\epsilon g)^T g + \frac{1}{2}(-\epsilon g)^T H (-\epsilon g)$$

Finally, simplifying the equation :

$$\hat{f}_{x_0}(x_1) = f(x_0) - \epsilon g^T g + \frac{1}{2}\epsilon^2 g^T H g$$

2. To determine whether gradient descent would work, we need to look at the sign of $\hat{f}_{x_0}(x_1) - f(x_0)$, which gives the change in the objective function after one step of gradient descent. We have :

$$\begin{aligned}\hat{f}_{x_0}(x_1) - f(x_0) &= f(x_0) - \epsilon g^T g + \frac{1}{2}\epsilon^2 g^T H g - f(x_0) \\ &= -\epsilon g^T g + \frac{1}{2}\epsilon^2 g^T H g\end{aligned}$$

Thus, gradient descent would work if and only if ϵ is small enough such that $-\epsilon g^T g + \frac{1}{2}\epsilon^2 g^T H g < 0$, or equivalently, $\epsilon < \frac{2g^T g}{g^T H g}$.

3. To derive a new optimization algorithm based on setting the gradient of $\hat{f}_{x_0}(\cdot)$ to zero, we can differentiate $\hat{f}_{x_0}(\cdot)$ with respect to x and set the resulting expression to zero :

$$\nabla_x \hat{f}_{x_0}(x) = g + \frac{1}{2} * 2H((x - x_0) = g + H(x - x_0) = 0$$

Isolating x :

$$x = x_0 - H^{-1}g$$

which is the Newton's Method.

Question 4

1. For BN, given that x is whitened to be independently distributed with zero mean and unit variance,i.e., $E[x] = 0$ and $Var[x] = 1$, we can get :

$$E[w^T x + b] = w^T E[x] + b = u^T(0) + b = b$$

$$Var[w^T x + b] = Var[w^T x] = E[(w^T x)^2] - (E[w^T x])^2 = E[w^T x w] - 0^2 = w^T E[xx^T]w = w^T w = \|w\|^2$$

Therefore, the output after BN is :

$$y_{BN} = \frac{w^T x + b - E[w^T x + b]}{\sqrt{Var[w^T x + b]}} = \frac{w^T x}{\|w\|}$$

For WN :

$$y_{WN} = (\frac{g}{\|u\|}u)^T x + b = g\frac{u^T x}{\|u\|} + b$$

Thus, in the condition of ignoring the learned scale and shift terms for both BN and WN, we can say in this case y_{WN} is equivalent to y_{BN}.

2. By the chain rule, we get :

$$\nabla_u L = \nabla_w L \cdot \nabla_u w$$

First we compute the derivative of w with respect to u :

$$\nabla_u w = \frac{g}{\|u\|}\left(I - \frac{uu^T}{\|u\|^2}\right)$$

where I is the identity matrix. The term $\frac{uu^T}{\|u\|^2}$ can be regarded as vv^T where v is a unit vector with the same direction as u. Considering a vector a, we have :

$$(vv^T)a = v(v^T a)$$

Here, $v^T a$ is a scalar that represents the component of a in the direction of v. Thus $\frac{uu^T}{\|u\|^2}$ (that is vv^T) represents the projection matrix onto the direction of u. Consequently, $I - \frac{uu^T}{\|u\|^2}$ (that is $I - vv^T$) is the orthogonal complement projection matrix, because :

$$(1 - vv^T)a = a - v(v^T a)$$

We denote this projection matrix as W^* :

$$W^* = I - \frac{uu^T}{\|u\|^2}$$

Then we get the $\nabla_u L$:

$$\nabla_u L = \nabla_w L \cdot \frac{g}{\|u\|} \cdot W^*$$

Since $\frac{g}{\|u\|}$ is a scalar, we can express $\nabla_u L$ as :

$$\nabla_u L = sW^* \cdot \nabla_w L$$

where $s = \frac{g}{\|u\|}$.

3. Assume the gradient update step for u with step size α is :

$$u_{t+1} = u_t - \alpha \nabla_u L_t$$

Substitute the $\nabla_u L$ from the last question :

$$u_{t+1} = u_t - \alpha s W^* \cdot \nabla_w L_t$$
$$\implies \|u_{t+1}\| = \|u_t - \alpha s W^* \cdot \nabla_w L_t\|$$

Because W^* is the orthogonal complement projection matrix, which projects any vector onto the subspace orthogonal to u_t, so u_t and $-\alpha s W^* \cdot \nabla_w L_t$ are orthogonal vectors. Then according to the Pythagorean theorem, we can get :

$$\|u_{t+1}\|^2 = \|u_t\|^2 + \alpha^2 s^2 \|W^* \cdot \nabla_w L_t\|^2$$

Because $0 \leq \alpha^2 s^2 \|W^* \cdot \nabla_w L_t\|^2$, thus :

$$\|u_t\|^2 \leq \|u_{t+1}\|^2$$
$$\implies \|u_t\| \leq \|u_{t+1}\|$$

This shows that $\|u\|$ becomes equal or larger after one gradient update step.