AI6104 – Mathematics for AI

Neural Network Assignment

Zhang Huan

G1903429B

(1) Network definition

a. The structure of neural network is shown in the figure 1. It is a 4 - layer back propagation neural network. Some instructions are made here.

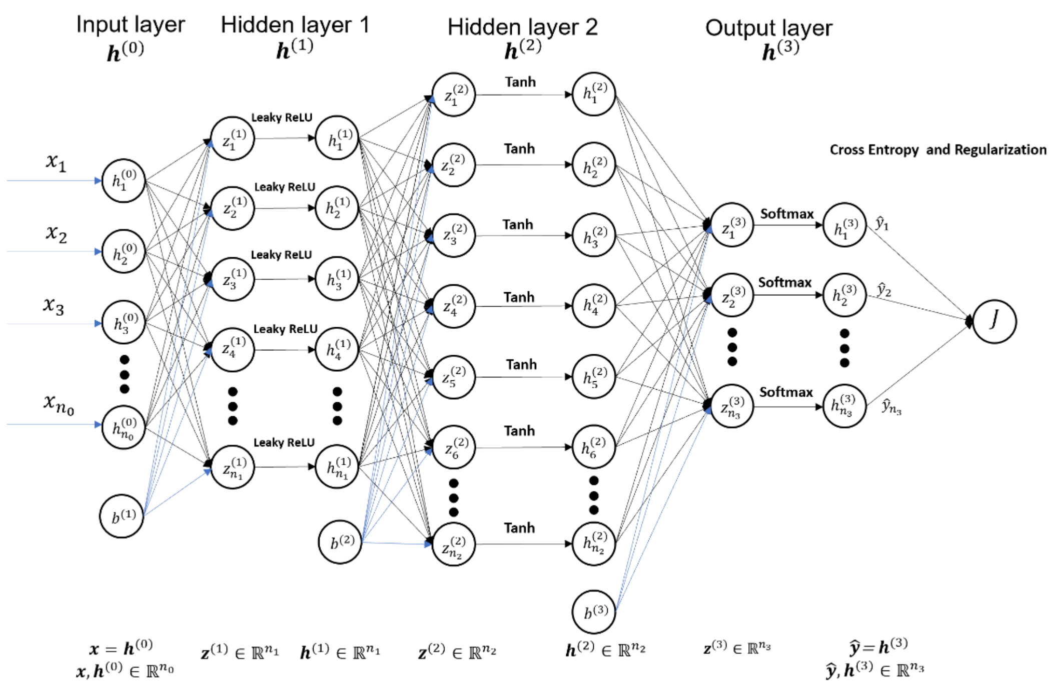

Figure 1: 4-layer BP network

The input layer is defined as $\boldsymbol{h}^{(0)}$ ($\boldsymbol{x}$) and the output layer is $\boldsymbol{h}^{(\mathrm{L})}$ ($\hat{\boldsymbol{y}}$). In my case, $\mathrm{L} = 3$. Therefore, there are $\mathrm{L}\text{-}1 = 2$ hidden layers. Some definitions are shown below:

$$\begin{cases} L+1 & \rightarrow \text{ number of layers (including input and ouput layer)} \\ [n_0, n_1, n_2, \cdots, n_{L-1}, n_L] & \rightarrow \text{ number of dimension in each layer} \\ [\varphi^{(1)}, \varphi^{(2)}, \cdots, \varphi^{(L-1)}, \varphi^{(L)}] & \rightarrow \text{ activate activation function} \end{cases} \tag{1.1}$$

Note that $n_0 = m$ and $n_L = s$. There are some instructions for variables.

$$\begin{cases} \boldsymbol{h}^{(l)} = \varphi^{(l)}(\boldsymbol{z}^{(l)}) \\ \boldsymbol{z}^{(l)} = \sum_{i=1}^{n_{l-1}} \boldsymbol{w}_i^{(l)} \boldsymbol{h}_i^{(l-1)} + \boldsymbol{b}^{(l)} \\ l = 1,2,\cdots,L \rightarrow l \text{ is current layer} \\ \boldsymbol{h}^{(0)} = \boldsymbol{x} \\ \boldsymbol{h}^{(L)} = \boldsymbol{y} \end{cases} \tag{1.2}$$

Parameters for learning are shown as below:

$$\begin{cases} \theta = (\theta_1, \theta_2, \cdots, \theta_L) \\ \theta_l = \left(\boldsymbol{w}^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}, \boldsymbol{b}^{(l)} \in \mathbb{R}^{n_l}\right) \\ l = 1,2,\cdots,L \end{cases} \tag{1.3}$$

$\boldsymbol{w}^{(l)}$ and $\boldsymbol{b}^{(l)}$ are weights and bias in layer l respectively.

b. There are 3 activation functions in this network.
   $\varphi^{(1)}$ is a **Leaky ReLU** function.

$$\varphi^{(1)}(x) = \begin{cases} x, & if\ x \geq 0 \\ \gamma x, & if\ x < 0 \end{cases} \quad x \in R, \gamma > 0 \text{ (in wiki, } \gamma = 0.01)$$

   $\varphi^{(2)}$ is a **Tanh** function.

$$\varphi^{(2)}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, x \in R$$

   $\varphi^{(3)}$ is a **Softmax** function.

$$\varphi^{(3)}(\boldsymbol{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \text{ for } i = 1, \cdots, K \text{ } and \text{ } \boldsymbol{z} = (z_1, \cdots, z_K) \in \mathbb{R}^K$$

c. Before specifying an objective function, I will introduce the train dataset.
   $\boldsymbol{x}^{(n)}$ and $\boldsymbol{y}^{(n)}$ are column vectors.

$$\begin{cases} \{\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\}_{n=1}^{N} \\ \boldsymbol{x}^{(n)} \in \mathbb{R}^{M}, \text{note M} = \text{n}_0 \\ \boldsymbol{y}^{(n)} \in \mathbb{R}^{S}, \text{note S} = \text{n}_L \end{cases} \tag{1.4}$$

Therefore, there are N train samples. My objective function is defined as below:

$$\min_{\theta} J(\theta) = L(\theta) + R(\theta) \tag{1.5}$$

Where $L(\theta)$ is cross entropy cost term and $R(\theta)$ is ℓ_2 norm regularized term.

$$L(\theta) = -\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{y}^{(n)} log \widehat{\boldsymbol{y}}^{(n)} \tag{1.6}$$

$$R(\theta) = \frac{\lambda}{2N}\|\theta\|_2^2 = \frac{\lambda}{2N}\|\boldsymbol{w}\|_2^2 \tag{1.7}$$

Where $\widehat{\boldsymbol{y}}^{(i)} = f(\boldsymbol{x}^{(i)}, \theta)$, $\boldsymbol{w}$ is the weight for $\{\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\}_{n=1}^{N}$, $\|\cdot\|_2$ is ℓ_2 norm, λ is a penalty for regularization. The norm regularized term is divided by 2 because it is convenient after differentiating.

(2) Gradient calculation

Rewrite (1.6) and (1.7):

$$L(\theta) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{s=1}^{S} \boldsymbol{y}_s^{(n)} log \widehat{\boldsymbol{y}}_s^{(n)} \tag{2.1}$$

$$R(\theta) = \frac{\lambda}{2N}\sum_{l=1}^{L}\left\|\boldsymbol{w}^{(l)}\right\|_2^2 = \frac{\lambda}{2N}\sum_{l=1}^{L}\sum_{i=1}^{n_l}\sum_{j=1}^{n_{l-1}}\left(w_{i,j}^{(l)}\right)^2 \tag{2.2}$$

Where S is the dimension of $\boldsymbol{y}^{(n)}$ and $\widehat{\boldsymbol{y}}^{(n)}$, $\boldsymbol{w}^{(l)}$ is the weight in layer l. n_l and n_{l-1} are the number of layer in layer l and $l-1$. Let us first consider one train sample ($\{\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\}$), so $N = 1$. For convenience, suppose $\boldsymbol{y}^{(n)} = \boldsymbol{y}$ and $\widehat{\boldsymbol{y}}^{(n)} = \widehat{\boldsymbol{y}}$, we have:

$$L(\theta) = -\sum_{s=1}^{S} y_s log \hat{y}_s \tag{2.3}$$

Part 1 ($\frac{\partial L(\theta)}{\partial \theta^{(l)}}$):

The target is to calculate $\frac{\partial J(\theta)}{\partial \theta^{(l)}}$, we will focus on $\frac{\partial L(\theta)}{\partial \theta^{(l)}}$ first. For convenience, according to definition (1.3), the problem is changed to calculate $\frac{\partial L(\theta)}{\partial w_{i,j}^{(l)}}$ and $\frac{\partial L(\theta)}{\partial b_i^{(l)}}$. Use chain rule and definition (1.2)

$$\frac{\partial L(\theta)}{\partial w_{i,j}^{(l)}} = \frac{\partial L(\theta)}{\partial z_i^{(l)}} \cdot \frac{\partial z_i^{(l)}}{\partial w_{i,j}^{(l)}}, \qquad \frac{\partial L(\theta)}{\partial b_i^{(l)}} = \frac{\partial L(\theta)}{\partial z_i^{(l)}} \cdot \frac{\partial z_i^{(l)}}{\partial b_i^{(l)}} \tag{2.4}$$

Where $z_i^{(l)}$ is the i_{th} element of $\boldsymbol{z}^{(l)}$, from definition (1.2) :

$$z_i^{(l)} = w_{i,1}^{(l)} h_1^{(l-1)} + w_{i,2}^{(l)} h_2^{(l-1)} + \cdots + w_{i,j}^{(l)} h_j^{(l-1)} + \cdots + w_{i,n_{l-1}}^{(l)} h_{n_{l-1}}^{(l-1)} + b_i^{(l)} \tag{2.5}$$

Therefore,

$$\frac{\partial z_i^{(l)}}{\partial w_{i,j}^{(l)}} = h_j^{(l-1)}, \frac{\partial z_i^{(l)}}{\partial b_i^{(l)}} = 1 \tag{2.6}$$

Suppose:

$$\delta_j^{(l)} \equiv \frac{\partial L(\theta)}{\partial z_j^{(l)}} \tag{2.7}$$

Then, from (2.4):

$$\frac{\partial L(\theta)}{\partial w_{i,j}^{(l)}} = \delta_i^{(l)} h_j^{(l-1)}, \quad \frac{\partial L(\theta)}{\partial b_i^{(l)}} = \delta_i^{(l)} \tag{2.8}$$

Write (2.4) in matrix form:

$$\frac{\partial L(\theta)}{\partial \boldsymbol{w}^{(l)}} = \begin{bmatrix} \frac{\partial L(\theta)}{\partial w_{1,1}^{(l)}} & \frac{\partial L(\theta)}{\partial w_{1,2}^{(l)}} & \cdots & \frac{\partial L(\theta)}{\partial w_{1,n_{l-1}}^{(l)}} \\ \frac{\partial L(\theta)}{\partial w_{2,1}^{(l)}} & \frac{\partial L(\theta)}{\partial w_{2,2}^{(l)}} & \cdots & \frac{\partial L(\theta)}{\partial w_{2,n_{l-1}}^{(l)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial L(\theta)}{\partial w_{n_l,1}^{(l)}} & \frac{\partial L(\theta)}{\partial w_{n_l,2}^{(l)}} & \cdots & \frac{\partial L(\theta)}{\partial w_{n_l,n_{l-1}}^{(l)}} \end{bmatrix} = \begin{bmatrix} \delta_1^{(l)} h_1^{(l-1)} & \delta_1^{(l)} h_2^{(l-1)} & \cdots & \delta_1^{(l)} h_{n_{l-1}}^{(l-1)} \\ \delta_2^{(l)} h_1^{(l-1)} & \delta_2^{(l)} h_2^{(l-1)} & \cdots & \delta_2^{(l)} h_{n_{l-1}}^{(l-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n_l}^{(l)} h_1^{(l-1)} & \delta_{n_l}^{(l)} h_2^{(l-1)} & \cdots & \delta_{n_l}^{(l)} h_{n_{l-1}}^{(l-1)} \end{bmatrix} = \begin{bmatrix} \delta_1^{(l)} \\ \delta_2^{(l)} \\ \vdots \\ \delta_{n_l}^{(l)} \end{bmatrix} \begin{bmatrix} h_1^{(l-1)} & h_2^{(l-1)} & \cdots & h_{n_{l-1}}^{(l-1)} \end{bmatrix} = \boldsymbol{\delta}^{(l)} \left(\boldsymbol{h}^{(l-1)}\right)^T \tag{2.9}$$

Similarly,

$$\frac{\partial L(\theta)}{\partial \boldsymbol{b}^{(l)}} = \boldsymbol{\delta}^{(l)} \tag{2.10}$$

Let us consider how to calculate $\boldsymbol{\delta}^{(l)}$, if $l = L = 3$, then

$$\delta_j^{(L)} \equiv \frac{\partial L(\theta)}{\partial z_j^{(L)}} = \frac{\partial(\sum_{i=1}^{S} y_i log\hat{y}_i)}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial z_j^{(L)}} = \sum_{i=1}^{S} (-\frac{y_i}{\hat{y}_i}) \cdot \frac{\partial \hat{y}_i}{\partial z_j^{(L)}}$$

$$\frac{\partial \hat{y}_i}{\partial z_j^{(L)}} = \frac{\partial \varphi^{(3)}(z_j^{(L)})}{\partial z_j^{(L)}} = \begin{cases} \hat{y}_i(1-\hat{y}_i), i = j \\ -\hat{y}_i\hat{y}_j, \quad\quad i \neq j \end{cases}, \varphi^{(3)} \text{ is softmax function}$$

$$\delta_j^{(L)} = \left(-\frac{y_i}{\hat{y}_i}\right)\hat{y}_i(1-\hat{y}_i)_{i=j} + \sum_{i=1, i\neq j}^{S} \left(-\frac{y_i}{\hat{y}_i}\right) \cdot -\hat{y}_i\hat{y}_j = -y_j + y_j\hat{y}_j + \sum_{i=1, i\neq j}^{S} y_i\hat{y}_j = -y_j + \hat{y}_j \sum_{i=1}^{S} y_i = \hat{y}_j - y_j \quad (2.11)$$

Note that y_i is one-hot vector, so $\sum_{i=1}^{S} y_i = 1$. Rewrite (2.11) in vector form

$$\boldsymbol{\delta}^{(L)} = \hat{\boldsymbol{y}} - \boldsymbol{y} \quad (2.12)$$

Now I will show how to calculate $\boldsymbol{\delta}^{(l)}$ if $1 \leq l < L$. From (2.5), $z_i^{(l+1)}$ can be represented a function of $z_j^{(l)}$. Therefore, suppose:

$$z_1^{(l+1)} = F_1(z_j^{(l)})$$

$$z_2^{(l+1)} = F_2(z_j^{(l)})$$

$$\vdots$$

$$z_{n_l}^{(l+1)} = F_{n_l}(z_j^{(l)})$$

Where $F_n(\cdot)$ means a function which has only one independent variable $z_j^{(l)}$, then according chain rule:

$$\delta_j^{(l)} \equiv \frac{\partial L(\theta)}{\partial z_j^{(l)}} = \frac{\partial L(\theta)}{\partial z_1^{(l+1)}} \frac{\partial z_1^{(l+1)}}{\partial z_j^{(l)}} + \frac{\partial L(\theta)}{\partial z_2^{(l+1)}} \frac{\partial z_2^{(l+1)}}{\partial z_j^{(l)}} + \cdots + \frac{\partial L(\theta)}{\partial z_{n_l}^{(l+1)}} \frac{\partial z_{n_l}^{(l+1)}}{\partial z_j^{(l)}}$$

$$\delta_j^{(l)} = \sum_{i=1}^{n_l} \frac{\partial L(\theta)}{\partial z_i^{(l+1)}} \frac{\partial z_i^{(l+1)}}{\partial z_j^{(l)}} = \sum_{i=1}^{n_l} \frac{\partial L(\theta)}{\partial z_i^{(l+1)}} \frac{\partial z_i^{(l+1)}}{\partial h_j^{(l)}} \frac{\partial h_j^{(l)}}{\partial z_j^{(l)}}$$

$$\frac{\partial L(\theta)}{\partial z_i^{(l+1)}} = \delta_i^{(l+1)}$$

$$\text{According to (2.5),}\qquad \frac{\partial z_i^{(l+1)}}{\partial h_j^{(l)}} = w_{i,j}^{(l+1)}$$

$$\frac{\partial h_j^{(l)}}{\partial z_j^{(l)}} = \varphi^{(l)\prime}(z_j^{(l)})$$

Therefore

$$\delta_j^{(l)} = \frac{\partial L(\theta)}{\partial z_j^{(l)}}\left(\sum_{i=1}^{n_l} \delta_i^{(l+1)} w_{i,j}^{(l+1)}\right)\varphi^{(l)\prime}\left(z_j^{(l)}\right) \tag{2.13}$$

Write in vector form

$$\boldsymbol{\delta}^{(l)} = \frac{\partial L(\theta)}{\partial \boldsymbol{z}^{(l)}} = \left(\boldsymbol{w}^{(l+1)}\right)^T \boldsymbol{\delta}^{(l+1)} \odot \varphi^{(l)\prime}\left(\boldsymbol{z}^{(l)}\right) \tag{2.14}$$

Where $\odot$ is Hadamard product.

Use (2.12) , (2.14) and definition of $\varphi^{(2)}, \varphi^{(1)}$, the result of each layer can be shown as below:

$$\boldsymbol{\delta}^{(3)} = \widehat{\boldsymbol{y}} - \boldsymbol{y}$$

$$\boldsymbol{\delta}^{(2)} = \left(\boldsymbol{w}^{(3)}\right)^T \boldsymbol{\delta}^{(3)} \odot \varphi^{(2)\prime}\left(\boldsymbol{z}^{(2)}\right) = \left(\boldsymbol{w}^{(3)}\right)^T \boldsymbol{\delta}^{(3)} \odot tanh^2(\boldsymbol{z}^{(2)}) \tag{2.15}$$

$$\boldsymbol{\delta}^{(1)} = \left(\boldsymbol{w}^{(2)}\right)^T \boldsymbol{\delta}^{(2)} \odot \varphi^{(1)\prime}\left(\boldsymbol{z}^{(1)}\right) = \left(\boldsymbol{w}^{(2)}\right)^T \boldsymbol{\delta}^{(2)} \odot \boldsymbol{p}$$

Where $\boldsymbol{p} = \varphi^{(1)\prime}\left(\boldsymbol{z}^{(1)}\right) \in \mathbb{R}^{n_1}$, $p_i(i = 1, \cdots, n_1) = \begin{cases} 1, z_i^{(1)} \geq 0 \\ \gamma, z_i^{(1)} < 0 \end{cases}$.

Combine (2.9) (2.10) and (2.15), we have:

$$\frac{\partial L(\theta)}{\partial \boldsymbol{w}^{(3)}} = \boldsymbol{\delta}^{(3)}\left(\boldsymbol{h}^{(2)}\right)^T = (\widehat{\boldsymbol{y}} - \boldsymbol{y}) \cdot \left(\boldsymbol{h}^{(2)}\right)^T$$

$$\frac{\partial L(\theta)}{\partial \boldsymbol{w}^{(2)}} = \boldsymbol{\delta}^{(2)}\left(\boldsymbol{h}^{(1)}\right)^T = \left(\boldsymbol{w}^{(3)}\right)^T \boldsymbol{\delta}^{(3)} \odot tanh^2(\boldsymbol{z}^{(2)}) \cdot \left(\boldsymbol{h}^{(1)}\right)^T$$

$$\frac{\partial L(\theta)}{\partial \boldsymbol{w}^{(1)}} = \boldsymbol{\delta}^{(1)}\left(\boldsymbol{h}^{(0)}\right)^T = \left(\boldsymbol{w}^{(2)}\right)^T \boldsymbol{\delta}^{(2)} \odot \boldsymbol{p} \cdot (\boldsymbol{x})^T \tag{2.16}$$

$$\frac{\partial L(\theta)}{\partial \boldsymbol{b}^{(3)}} = \boldsymbol{\delta}^{(3)} = \widehat{\boldsymbol{y}} - \boldsymbol{y}$$

$$\frac{\partial L(\theta)}{\partial \boldsymbol{b}^{(2)}} = \boldsymbol{\delta}^{(2)} = \left(\boldsymbol{w}^{(3)}\right)^T \boldsymbol{\delta}^{(3)} \odot tanh^2(\boldsymbol{z}^{(2)})$$

$$\frac{\partial L(\theta)}{\partial \boldsymbol{b}^{(1)}} = \boldsymbol{\delta}^{(1)} = \left(\boldsymbol{w}^{(2)}\right)^T \boldsymbol{\delta}^{(2)} \odot \boldsymbol{p}$$

Part 2 ($\frac{\partial R(\theta)}{\partial \theta^{(l)}}$):

This part is simple compare to part 1. Just do it directly:

$$\frac{\partial R(\theta)}{\partial w_{i,j}^{(l)}} = \frac{\frac{\lambda}{2} \partial\left(\sum_{l=1}^{L} \sum_{i=1}^{n_l} \sum_{j=1}^{n_{l-1}} \left(w_{i,j}^{(l)}\right)^2\right)}{\partial w_{i,j}^{(l)}} = \lambda w_{i,j}^{(l)}$$

Therefore,

$$\frac{\partial R(\theta)}{\partial \boldsymbol{w}^{(l)}} = \lambda \boldsymbol{w}^{(l)} \tag{2.17}$$

$$\frac{\partial R(\theta)}{\partial \boldsymbol{b}^{(l)}} = \boldsymbol{0} \tag{2.18}$$

Part 3 ($\frac{\partial J(\theta)}{\partial \theta^{(l)}}$):

For one train sample $\{\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\}$, from $(2.9), (2.10), (2.16)$ and (2.17), we have (for showing equation clearly, I use $\boldsymbol{\delta}^{(l)}$ to show results in (2.16)):

$$\frac{\partial J(\theta)}{\partial \boldsymbol{w}^{(l)}} = \boldsymbol{\delta}^{(l)} \left(\boldsymbol{h}^{(l-1)}\right)^T + \lambda \boldsymbol{w}^{(l)} \tag{2.19}$$

$$\frac{\partial J(\theta)}{\partial \boldsymbol{b}^{(l)}} = \boldsymbol{\delta}^{(l)} \tag{2.20}$$

Now expand to N train samples, for $\{\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\}_{n=1}^{N}$, we have:

$$\frac{\partial J(\theta)}{\partial \boldsymbol{w}^{(l)}} = \frac{1}{N}\left(\sum_{n=1}^{N} \boldsymbol{\delta}^{(l,n)} \left(\boldsymbol{h}^{(l-1,n)}\right)^T + \lambda \boldsymbol{w}^{(l)}\right) \tag{2.21}$$

$$\frac{\partial J(\theta)}{\partial \boldsymbol{b}^{(l)}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{\delta}^{(l,n)} \tag{2.22}$$

Note that the chain rule is not applicable for all cases in matrix derivatives. Therefore, I do not use the chain rule in this question for matrix derivatives although the result is the same as that using chain rule in matrix derivatives.

(3) Training equation

Parameters are updated based on gradient descent. Here I use Root mean square prop (RMSprop). Therefore:

$$\begin{cases} \theta_{(t+1)}^{(l)} = \theta_{(t)}^{(l)} - \dfrac{\alpha}{\sqrt{S_t + \epsilon}} \cdot \nabla\theta|_{\theta=\theta_{(t)}^{(l)}} \\ S_t = \beta S_{t-1} + (1-\beta)\left(\nabla\theta|_{\theta=\theta_{(t)}^{(l)}}\right)^2 \\ \nabla\theta|_{\theta=\theta_{(t)}^{(l)}} = \dfrac{\partial J(\theta)}{\partial\theta_{(t)}^{(l)}} = \dfrac{\partial L(\theta)}{\partial\theta_{(t)}^{(l)}} + \dfrac{\partial R(\theta)}{\partial\theta_{(t)}^{(l)}} \end{cases} \tag{3.1.1}$$

Where $\theta_{(t)}^{(l)}$ means the t_{th} iteration in training, S_t is mean-square term, $S_0 = 0$, $\beta = 0.9$, $\epsilon = 10^{-6}$ is to prevent denominator to be 0 and $\alpha = 0.001$ is the learning rate. For convenience, I will ignore subscript (t) in $\boldsymbol{w}$ and $\boldsymbol{b}$ in next equations.

Training equations for all parameters in my network will be shown below. N samples $\{\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\}_{n=1}^{N}$ are trained. := is used in update equation which means the definition. I will show in the most detail form from layer 3 to layer 1.

Suppose current iteration is t. To write RMSprop in a matrix form, a matrix should be defined:

$$\boldsymbol{s}_t^{(l)} = \begin{bmatrix} \dfrac{1}{\sqrt{\left(S_t^{(l)}\right)_{1,1} + \epsilon}} & \dfrac{1}{\sqrt{\left(S_t^{(l)}\right)_{1,2} + \epsilon}} & \cdots & \dfrac{1}{\sqrt{\left(S_t^{(l)}\right)_{1,n_{l-1}} + \epsilon}} \\ \dfrac{1}{\sqrt{\left(S_t^{(l)}\right)_{2,1} + \epsilon}} & \dfrac{1}{\sqrt{\left(S_t^{(l)}\right)_{2,2} + \epsilon}} & \cdots & \dfrac{1}{\sqrt{\left(S_t^{(l)}\right)_{2,n_{l-1}} + \epsilon}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{1}{\sqrt{\left(S_t^{(l)}\right)_{n_l,1} + \epsilon}} & \dfrac{1}{\sqrt{\left(S_t^{(l)}\right)_{n_l,2} + \epsilon}} & \cdots & \dfrac{1}{\sqrt{\left(S_t^{(l)}\right)_{n_l,n_{l-1}} + \epsilon}} \end{bmatrix}, \quad \boldsymbol{s}_t^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}} \tag{3.1.2}$$

Where $(S_t^{(l)})_{i,j}$ is the updating; parameter for $(w_{(t)}^{(l)})_{i,j}$ in t_{th} iteration. $(S_t^{(l)})_{i,j} = \beta(S_{t-1}^{(l)})_{i,j} + (1-\beta)\left(\nabla w|_{w=(w_{(t)}^{(l)})_{i,j}}\right)^2,\ (S_0^{(l)})_{i,j} = 0$.

Similarly,

$$\boldsymbol{T}_t^{(l)} = \begin{bmatrix} \frac{1}{\sqrt{\left(T_t^{(l)}\right)_1 + \epsilon}} \\ \frac{1}{\sqrt{\left(T_t^{(l)}\right)_2 + \epsilon}} \\ \vdots \\ \frac{1}{\sqrt{\left(T_t^{(l)}\right)_{n_l} + \epsilon}} \end{bmatrix}, \boldsymbol{T}_t^{(l)} \in \mathbb{R}^{n_l} \tag{3.1.3}$$

Where $(T_t^{(l)})_{i,j}$ is the update parameter for $(b_{(t)}^{(l)})_i$ in t_{th} iteration. $(T_t^{(l)})_i = \beta(T_{t-1}^{(l)})_i + (1-\beta)\left(\nabla b|_{b=(b_{(t)}^{(l)})_i}\right)^2,\ (T_0^{(l)})_i = 0$.

Training equation for each layer

Use results in question (2) and definition (3.1.1) (3.1.2) (3.1.3). Training equations can be shown as below.

Layer 3 (output layer):

$\boldsymbol{w}^{(3)}$**:**

$$\boldsymbol{w}^{(3)} := \boldsymbol{w}^{(3)} - \alpha \cdot \boldsymbol{s}_t^{(3)} \odot \nabla \boldsymbol{w}|_{\boldsymbol{w}=\boldsymbol{w}^{(3)}}, \qquad \boldsymbol{w}^{(3)} \in \mathbb{R}^{n_3 \times n_2}$$

$$\nabla \boldsymbol{w}|_{\boldsymbol{w}=\boldsymbol{w}^{(3)}} = \frac{\partial J(\theta)}{\partial \boldsymbol{w}^{(3)}} = \frac{1}{N}\left(\sum_{n=1}^{N} (\widehat{\boldsymbol{y}}^{(n)} - \boldsymbol{y}^{(\boldsymbol{n})})\left(\boldsymbol{h}^{(2,n)}\right)^T + \lambda \boldsymbol{w}^{(3)}\right)$$

$$\boldsymbol{w}^{(3)} := \boldsymbol{w}^{(3)} - \frac{\alpha}{N} \cdot \boldsymbol{s}_t^{(3)} \odot \left(\sum_{n=1}^{N} (\widehat{\boldsymbol{y}}^{(n)} - \boldsymbol{y}^{(\boldsymbol{n})})\left(\boldsymbol{h}^{(2,n)}\right)^T + \lambda \boldsymbol{w}^{(3)}\right) \tag{3.2.1}$$

Write in a simpler equation:

$$\boldsymbol{w}^{(3)} := \boldsymbol{w}^{(3)} - \frac{\alpha}{N} \cdot \boldsymbol{s}_t^{(3)} \odot \left(\sum_{n=1}^{N} \boldsymbol{\delta}^{(3,n)}\left(\boldsymbol{h}^{(2,n)}\right)^T + \lambda \boldsymbol{w}^{(3)}\right) \tag{3.2.2}$$

$\boldsymbol{b}^{(3)}$**:**

$$\boldsymbol{b}^{(3)} := \boldsymbol{b}^{(3)} - \alpha \cdot \boldsymbol{T}_t^{(3)} \odot \nabla \boldsymbol{b}|_{\boldsymbol{b}=\boldsymbol{b}^{(3)}}, \qquad \boldsymbol{b}^{(3)} \in \mathbb{R}^{n_3}$$

$$\nabla \boldsymbol{b}|_{\boldsymbol{b}=\boldsymbol{b}^{(3)}} = \frac{\partial J(\theta)}{\partial \boldsymbol{b}^{(3)}} = \frac{1}{N}\left(\sum_{n=1}^{N} (\widehat{\boldsymbol{y}}^{(n)} - \boldsymbol{y}^{(\boldsymbol{n})})\right)$$

$$\boldsymbol{b}^{(3)} \coloneqq \boldsymbol{b}^{(3)} - \frac{\alpha}{N} \cdot \boldsymbol{T}_t^{(3)} \odot \left(\sum_{n=1}^{N} \left(\widehat{\boldsymbol{y}}^{(n)} - \boldsymbol{y}^{(\boldsymbol{n})} \right) \right) \tag{3.3.1}$$

Write in a simpler equation:

$$\boldsymbol{b}^{(3)} \coloneqq \boldsymbol{b}^{(3)} - \frac{\alpha}{N} \cdot \boldsymbol{T}_t^{(3)} \odot \left(\sum_{n=1}^{N} \boldsymbol{\delta}^{(3,n)} \right) \tag{3.3.2}$$

Layer 2 (hidden layer 2):

$\boldsymbol{w}^{(2)}$**:**

$$\boldsymbol{w}^{(2)} := \boldsymbol{w}^{(2)} - \alpha \cdot \boldsymbol{s}_t^{(2)} \odot \nabla \boldsymbol{w}\big|_{\boldsymbol{w}=\boldsymbol{w}^{(2)}}, \qquad \boldsymbol{w}^{(2)} \in \mathbb{R}^{n_2 \times n_1}$$

$$\nabla \boldsymbol{w}\big|_{\boldsymbol{w}=\boldsymbol{w}^{(2)}} = \frac{\partial J(\theta)}{\partial \boldsymbol{w}^{(2)}} = \frac{1}{N} \left(\sum_{n=1}^{N} \left(\boldsymbol{w}^{(3)} \right)^T \cdot (\widehat{\boldsymbol{y}}^{(n)} - \boldsymbol{y}^{(n)}) \odot tanh^2 (\boldsymbol{z}^{(2,n)}) \cdot \left(\boldsymbol{h}^{(1,n)} \right)^T + \lambda \boldsymbol{w}^{(2)} \right)$$

$$\boldsymbol{w}^{(2)} \coloneqq \boldsymbol{w}^{(2)} - \frac{\alpha}{N} \cdot \boldsymbol{s}_t^{(2)} \odot \left(\sum_{n=1}^{N} \left(\boldsymbol{w}^{(3)} \right)^T \cdot \left(\widehat{\boldsymbol{y}}^{(n)} - \boldsymbol{y}^{(\boldsymbol{n})} \right) \odot tanh^2 \left(\boldsymbol{z}^{(2,n)} \right) \cdot \left(\boldsymbol{h}^{(1,n)} \right)^T + \lambda \boldsymbol{w}^{(2)} \right) \tag{3.4.1}$$

Write in a simpler equation:

$$\boldsymbol{w}^{(2)} \coloneqq \boldsymbol{w}^{(2)} - \frac{\alpha}{N} \cdot \boldsymbol{s}_t^{(2)} \odot \left(\sum_{n=1}^{N} \boldsymbol{\delta}^{(2,n)} \left(\boldsymbol{h}^{(1,n)} \right)^T + \lambda \boldsymbol{w}^{(2)} \right) \tag{3.4.2}$$

$\boldsymbol{b}^{(2)}$**:**

$$\boldsymbol{b}^{(2)} := \boldsymbol{b}^{(2)} - \alpha \cdot \boldsymbol{T}_t^{(2)} \odot \nabla \boldsymbol{b}\big|_{\boldsymbol{b}=\boldsymbol{b}^{(2)}}, \qquad \boldsymbol{b}^{(2)} \in \mathbb{R}^{n_2}$$

$$\nabla \boldsymbol{b}\big|_{\boldsymbol{b}=\boldsymbol{b}^{(2)}} = \frac{\partial J(\theta)}{\partial \boldsymbol{b}^{(2)}} = \frac{1}{N} \left(\sum_{n=1}^{N} \left(\boldsymbol{w}^{(3)} \right)^T \cdot \left(\widehat{\boldsymbol{y}}^{(n)} - \boldsymbol{y}^{(n)} \right) \odot tanh^2 \left(\boldsymbol{z}^{(2,n)} \right) \right)$$

$$\boldsymbol{b}^{(2)} \coloneqq \boldsymbol{b}^{(2)} - \frac{\alpha}{N} \cdot \boldsymbol{T}_t^{(2)} \odot \left(\sum_{n=1}^{N} \left(\boldsymbol{w}^{(3)} \right)^T \cdot \left(\widehat{\boldsymbol{y}}^{(n)} - \boldsymbol{y}^{(n)} \right) \odot tanh^2 \left(\boldsymbol{z}^{(2,n)} \right) \right) \tag{3.5.1}$$

Write in a simpler equation:

$$\boldsymbol{b}^{(2)} \coloneqq \boldsymbol{b}^{(2)} - \frac{\alpha}{N} \cdot \boldsymbol{T}_t^{(2)} \odot \left(\sum_{n=1}^{N} \boldsymbol{\delta}^{(2,n)} \right) \tag{3.5.2}$$

Layer 1 (hidden layer 1):

$\boldsymbol{w}^{(1)}$**:**

$$\boldsymbol{w}^{(1)} := \boldsymbol{w}^{(1)} - \alpha \cdot \boldsymbol{S}_t^{(1)} \odot \nabla \boldsymbol{w}|_{\boldsymbol{w}=\boldsymbol{w}^{(1)}}, \qquad \boldsymbol{w}^{(1)} \in \mathbb{R}^{n_1 \times n_0}$$

$$\nabla \boldsymbol{w}|_{\boldsymbol{w}=\boldsymbol{w}^{(1)}} = \frac{\partial J(\theta)}{\partial \boldsymbol{w}^{(1)}} = \frac{1}{N}\left(\sum_{n=1}^{N} \left(\boldsymbol{w}^{(2)}\right)^T \left(\left(\boldsymbol{w}^{(3)}\right)^T \cdot (\widehat{\boldsymbol{y}}^{(n)} - \boldsymbol{y}^{(n)}) \odot tanh^2(\boldsymbol{z}^{(2,n)})\right) \odot \boldsymbol{p} \cdot \left(\boldsymbol{x}^{(n)}\right)^T + \lambda \boldsymbol{w}^{(1)}\right)$$

$$\boldsymbol{w}^{(1)} \coloneqq \boldsymbol{w}^{(1)} - \frac{\alpha}{N} \cdot \boldsymbol{S}_t^{(1)} \odot \left(\sum_{n=1}^{N} \left(\boldsymbol{w}^{(2)}\right)^T \left(\left(\boldsymbol{w}^{(3)}\right)^T \cdot \left(\widehat{\boldsymbol{y}}^{(n)} - \boldsymbol{y}^{(n)}\right) \odot tanh^2\left(\boldsymbol{z}^{(2,n)}\right)\right) \odot \boldsymbol{p} \cdot \left(\boldsymbol{x}^{(n)}\right)^T + \lambda \boldsymbol{w}^{(1)}\right) \tag{3.6.1}$$

Write in a simpler equation:

$$\boldsymbol{w}^{(1)} \coloneqq \boldsymbol{w}^{(1)} - \frac{\alpha}{N} \cdot \boldsymbol{S}_t^{(1)} \odot \left(\sum_{n=1}^{N} \boldsymbol{\delta}^{(1,n)} \left(\boldsymbol{x}^{(n)}\right)^T + \lambda \boldsymbol{w}^{(1)}\right) \tag{3.6.2}$$

$\boldsymbol{b}^{(1)}$**:**

$$\boldsymbol{b}^{(1)} := \boldsymbol{b}^{(1)} - \alpha \cdot \boldsymbol{T}_t^{(1)} \odot \nabla \boldsymbol{b}|_{\boldsymbol{b}=\boldsymbol{b}^{(1)}}, \qquad \boldsymbol{b}^{(1)} \in \mathbb{R}^{n_1}$$

$$\nabla \boldsymbol{b}|_{\boldsymbol{b}=\boldsymbol{b}^{(1)}} = \frac{\partial J(\theta)}{\partial \boldsymbol{b}^{(1)}} = \frac{1}{N}\left(\sum_{n=1}^{N} \left(\boldsymbol{w}^{(2)}\right)^T \left(\left(\boldsymbol{w}^{(3)}\right)^T \cdot (\widehat{\boldsymbol{y}}^{(n)} - \boldsymbol{y}^{(n)}) \odot tanh^2(\boldsymbol{z}^{(2,n)})\right) \odot \boldsymbol{p}\right)$$

$$\boldsymbol{b}^{(1)} \coloneqq \boldsymbol{b}^{(1)} - \frac{\alpha}{N} \cdot \boldsymbol{T}_t^{(1)} \odot \left(\sum_{n=1}^{N} \left(\boldsymbol{w}^{(2)}\right)^T \left(\left(\boldsymbol{w}^{(3)}\right)^T \cdot \left(\widehat{\boldsymbol{y}}^{(n)} - \boldsymbol{y}^{(n)}\right) \odot tanh^2\left(\boldsymbol{z}^{(2,n)}\right)\right) \odot \boldsymbol{p}\right) \tag{3.7.1}$$

Write in a simpler equation:

$$\boldsymbol{b}^{(1)} \coloneqq \boldsymbol{b}^{(1)} - \frac{\alpha}{N} \cdot \boldsymbol{T}_t^{(1)} \odot \left(\sum_{n=1}^{N} \boldsymbol{\delta}^{(1,n)}\right) \tag{3.7.2}$$

Where $\boldsymbol{z}^{(i,n)}, \boldsymbol{h}^{(i,n)}$ and $\boldsymbol{\delta}^{(i,n)}$ mean $\boldsymbol{z}^{(i)}, \boldsymbol{h}^{(i)}$ and $\boldsymbol{\delta}^{(i)}$ in train sample $\{\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\}$. The definition of $\boldsymbol{p}$ is shown in equation (2.15).