Report of Homework 1

Hou Lintao 2018310076

1. Part One: Code Implementation

1.1 Multilayer Perceptron (MLP)

The algorithms were completed in *mlp.py*. The hyperparameters of hidden dim, initial learning rate, batch size and epochs were set as 50, 0.1, 16 and 60. The learning rate *lr* will be 0.1 times the before value after 20 epochs each time. The loss curve was shown as Figure 1.1, and the test accuracy was 95.18% and exceeds 90%.

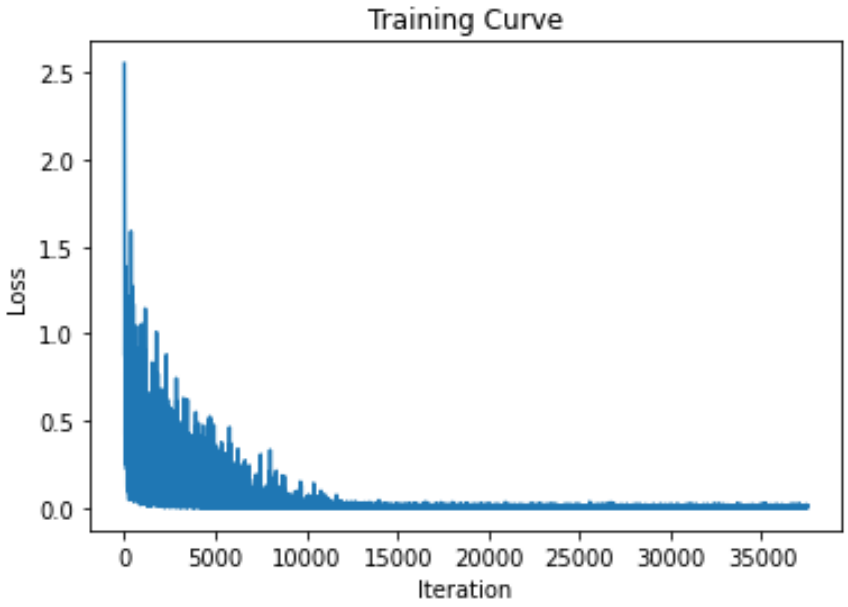
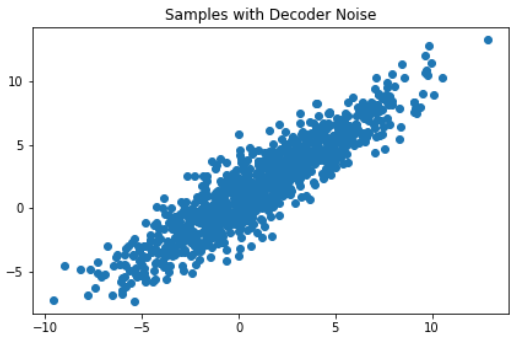


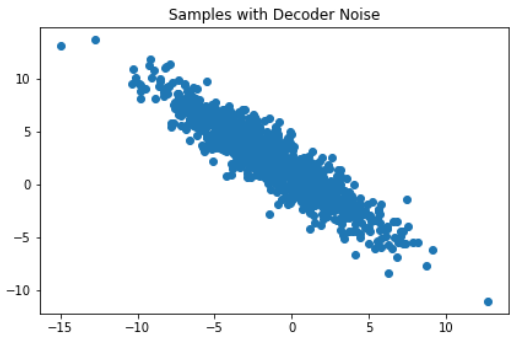
Figure 1.1 Loss curve and the test accuracy.

1.2 Variational Autoencoder (VAE)

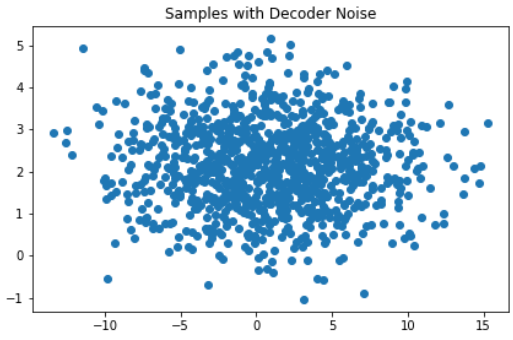
The algorithms were completed in *vae.py*. and the training results were shown in file folder visualization. The sample results with Decoder Noise were shown in Figure 1.2, and the data distribution is almost the same as the given data distribution.



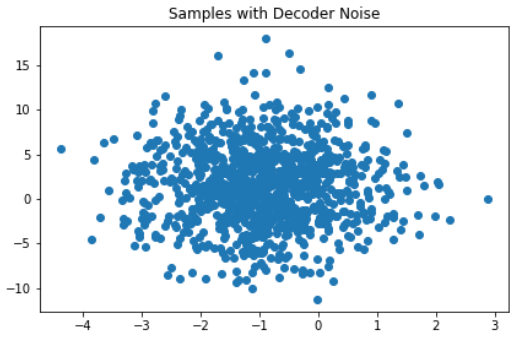
(1) Part a-1



(2) Part a-2



(3) Part b-1



(4) Part b-2

Figure 1.2 Samples with Decoder Noise

2. Part Two: Back-propagation

2.1 Block One: gradients of some basic layer

(i) The gradients of the output $y_i = \mathbf{BN}_{\gamma,\beta}(x_i)$ with respect to the parameters of γ, β:

$$\frac{\partial y_i}{\partial \gamma} = \frac{\partial(\gamma \hat{x}_i + \beta)}{\partial \gamma} = \hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}}$$

$$\frac{\partial y_i}{\partial \beta} = \frac{\partial(\gamma \hat{x}_i + \beta)}{\partial \beta} = 1$$

(ii) The forward computations of the dropout layer is:

$$y_{DP,j} = x_j \cdot \mathbf{M}_j$$

$$\mathbf{M}_j = \begin{cases} 0, & r_j < p \\ 1/(1-p), & r_j \geq p \end{cases}$$

$$= \frac{1}{1-p}\operatorname{sgn}(r_j - p)$$

The function $\operatorname{sgn}(x) = (x+|x|)/2$.

Therefore, the gradients of the output of a dropout layer with respect to its input are:

$$\frac{\partial y_j}{\partial x_j} = M_j$$

(iii) The forward computations of the softmax function is:

$$y_{soft,i} = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} = 1 - \frac{\sum_{j=1, j\neq i}^{n} e^{x_j}}{\sum_{j=1}^{n} e^{x_j}}$$

Therefore, the gradients of the output of a softmax function with respect to its input are calculated as following.

When $i = k$:

$$\frac{\partial y_{soft,i}}{\partial x_i} = \left(1 - \frac{\sum_{j=1, j\neq i}^{n} e^{x_j}}{\sum_{j=1}^{n} e^{x_j}}\right) \Big/ \partial x_i = \frac{\sum_{j=1, j\neq i}^{n} e^{x_j}}{\left(\sum_{j=1}^{n} e^{x_j}\right)^2} \frac{\partial \sum_{j=1}^{n} e^{x_j}}{\partial x_i}$$

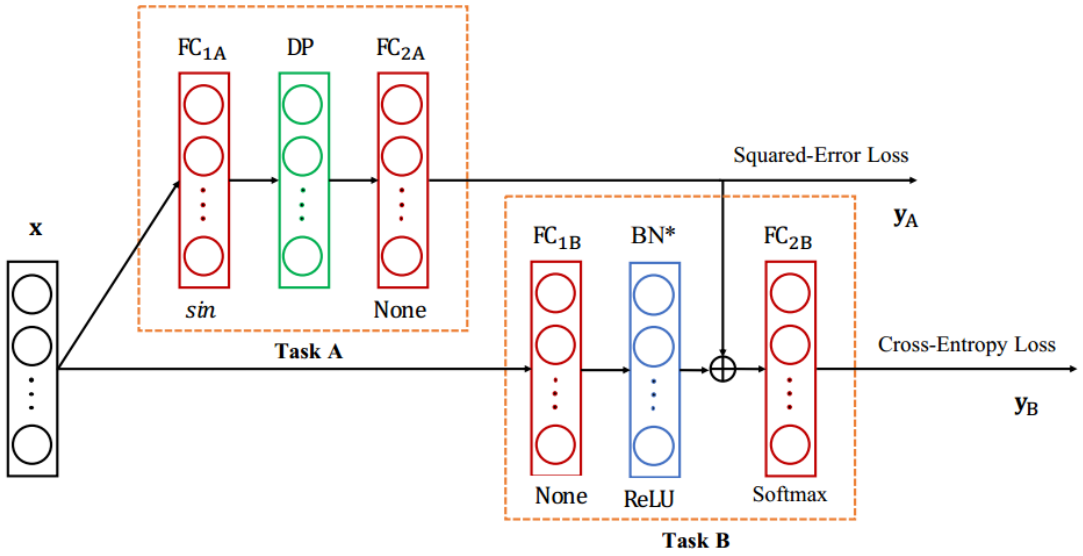
$$= \frac{\sum_{j=1, j\neq i}^{n} e^{x_j}}{\sum_{j=1}^{n} e^{x_j}} \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} = \left(1 - y_{soft,i}\right) y_{soft,i}$$

When $i \neq k$:

$$\frac{\partial y_{soft,i}}{\partial x_k} = \left(\partial \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}\right) \Big/ \partial x_k = -\frac{e^{x_i}}{\left(\sum_{j=1}^{n} e^{x_j}\right)^2} \frac{\partial \sum_{j=1}^{n} e^{x_j}}{\partial x_k}$$

$$= -\frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \frac{e^{x_k}}{\sum_{j=1}^{n} e^{x_j}} = -y_{soft,i} y_{soft,k}$$

2.2 Block Two: feed-forward and backpropagation of the multi-task network



(i) Task A:

The result of layer FC_{1A} is:

$$\mathbf{z}_{1A} = \theta_{1A}\mathbf{x} + \mathbf{b}_{1A}$$
$$\mathbf{x}_{1A} = \sin(\mathbf{z}_{1A})$$

The result of layer FC_{DP} is:

$$\mathbf{r} = random(\mathbf{x}_{1A}) \in U(0,1)$$
$$\mathbf{M} = \frac{1}{1-p}\operatorname{sgn}(\mathbf{r} - p \cdot \mathbf{1})$$
$$\mathbf{x}_{DP} = \mathbf{x}_{1A} \circ \mathbf{M}$$

The vector $\mathbf{r}$ is a random vector with uniform distribution between 0 and 1 for each neuron of $\mathbf{x}_{1A}$, and $\mathbf{1}$ is the vector that all entries are one. The definition of mask vector $\mathbf{M}$ can be seen in 2.1(ii).

The result of layer FC_{2A} is:

$$\hat{\mathbf{y}}_A = \mathbf{x}_{2A} = \theta_{2A}\mathbf{x}_{DP} + \mathbf{b}_{2A}$$

Task B:

The result of layer FC_{1B} is:

$$\mathbf{x}_{1B} = \theta_{1B}\mathbf{x}$$

The result of layer BN* is:

$$\mu = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_{1B}^{i}$$
$$\mathbf{x}_{BN} = \mathbf{x}_{1B} - \mu + \mathbf{b}_{1B}$$
$$\mathbf{x}_{BN^*} = \mathrm{ReLU}(\mathbf{x}_{BN}) = \max(\mathbf{x}_{BN}, 0)$$

The result of layer $\oplus$ is:

$$\mathbf{x}_{\oplus} = \mathbf{x}_{2\mathrm{A}} + \mathbf{x}_{\mathrm{BN}^*}$$

The result of layer $\mathrm{FC}_{2\mathrm{B}}$ is:

$$\mathbf{z}_{2\mathrm{B}} = \theta_{2\mathrm{B}}\mathbf{x}_{\oplus} + \mathbf{b}$$
$$\hat{\mathbf{y}}_{\mathrm{B}} = \mathbf{x}_{2\mathrm{B}} = \mathrm{Softmax}\left(\mathbf{z}_{2\mathrm{B}}\right)$$

(ii) The loss function of Task A and Task B is:

$$L = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{2}\left\|\hat{\mathbf{y}}_{\mathrm{A}}^{i} - \mathbf{y}_{\mathrm{A}}^{i}\right\|_2^2 - \sum_{k=1}^{b} y_{\mathrm{B},k}^{i}\log\hat{y}_{\mathrm{B},k}^{i}\right]$$

We split the loss function into two parts:

$$L = L_A + L_B$$
$$L_A = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left\|\hat{\mathbf{y}}_{\mathrm{A}}^{i} - \mathbf{y}_{\mathrm{A}}^{i}\right\|_2^2$$
$$L_B = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{b} y_{\mathrm{B},k}^{i}\log\hat{y}_{\mathrm{B},k}^{i}$$

For Task A and Task B, we can first calculate the residual of each layer.

Task B:

The residual of layer $\mathrm{FC}_{2\mathrm{B}}$:

$$\frac{\partial L}{\partial \hat{y}_{\mathrm{B,j}}^{i}} = \frac{\partial L_B}{\partial \hat{y}_{\mathrm{B,j}}^{i}} = -\frac{1}{m}\sum_{k=1}^{b}\frac{\partial\left(y_{\mathrm{B},k}^{i}\log\hat{y}_{\mathrm{B},k}^{i}\right)}{\partial \hat{y}_{B,j}^{i}} = -\frac{1}{m}\frac{y_{\mathrm{B},j}^{i}}{\hat{y}_{B,j}^{i}}$$

$$\frac{\partial L}{\partial z_{2B,j}^{i}} = \sum_{k=1}^{b}\frac{\partial L}{\partial \hat{y}_{\mathrm{B,k}}^{i}}\frac{\partial \hat{y}_{\mathrm{B,k}}^{i}}{\partial z_{2B,j}^{i}} = \sum_{k=1,k\neq j}^{b}\frac{\partial L}{\partial \hat{y}_{\mathrm{B,k}}^{i}}\frac{\partial \hat{y}_{\mathrm{B,k}}^{i}}{\partial z_{2B,j}^{i}} + \frac{\partial L}{\partial \hat{y}_{\mathrm{B,j}}^{i}}\frac{\partial \hat{y}_{\mathrm{B,j}}^{i}}{\partial z_{2B,j}^{i}}$$

$$= \frac{1}{m}\left[\sum_{k=1,k\neq j}^{b}\left(\frac{y_{\mathrm{B},k}^{i}}{\hat{y}_{B,k}^{i}}\hat{y}_{B,k}^{i}\hat{y}_{B,j}^{i}\right) + \frac{y_{\mathrm{B},j}^{i}}{\hat{y}_{B,j}^{i}}\left(\hat{y}_{B,j}^{i} - 1\right)\hat{y}_{B,j}^{i}\right]$$

$$= \frac{1}{m}\left[\sum_{k=1,k\neq j}^{b} y_{\mathrm{B},k}^{i}\hat{y}_{B,j}^{i} + y_{\mathrm{B},j}^{i}\left(\hat{y}_{B,j}^{i} - 1\right)\right]$$

$$= \frac{1}{m}\left(\hat{y}_{B,j}^{i}\sum_{k=1}^{b} y_{\mathrm{B},k}^{i} - y_{\mathrm{B},j}^{i}\right) = \frac{1}{m}\left(\hat{y}_{B,j}^{i} - y_{\mathrm{B},j}^{i}\right)$$

$$\left(\sum_{k=1}^{b} y_{\mathrm{B},k}^{i} = 1 \text{ for one-hot question}\right)$$

Therefore the residual of $\mathbf{z}_{2\mathrm{B}}^{i}$ is:

$$\frac{\partial L}{\partial \mathbf{z}_{2\mathrm{B}}^{i}} = \frac{1}{m}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i} - \mathbf{y}_{\mathrm{B}}^{i}\right)$$

The residual of layer $\oplus$:

$$\frac{\partial L}{\partial \mathbf{x}_{\oplus}^{i}} = \frac{\partial L}{\partial \mathbf{z}_{2\mathrm{B}}^{i}}\frac{\partial \mathbf{z}_{2\mathrm{B}}^{i}}{\partial \mathbf{x}_{\oplus}^{i}} = \frac{1}{m}\theta_{2\mathrm{B}}^{T}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i} - \mathbf{y}_{\mathrm{B}}^{i}\right)$$

The residual of layer BN*:

$$\frac{\partial L}{\partial \mathbf{x}_{\mathrm{BN}^*}^{i}} = \frac{\partial L}{\partial \mathbf{x}_{\oplus}^{i}}\frac{\partial \mathbf{x}_{\oplus}^{i}}{\partial \mathbf{x}_{\mathrm{BN}^*}^{i}} = \frac{1}{m}\theta_{2\mathrm{B}}^{T}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i} - \mathbf{y}_{\mathrm{B}}^{i}\right)$$

$$\frac{\partial L}{\partial \mathbf{x}_{\mathrm{BN}}^{i}}=\frac{\partial L}{\partial \mathbf{x}_{\mathrm{BN}^{*}}^{i}} \frac{\partial \mathbf{x}_{\mathrm{BN}^{*}}^{i}}{\partial \mathbf{x}_{\mathrm{BN}}^{i}}=\frac{1}{m} \theta_{2 \mathrm{B}}^{T}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i}-\mathbf{y}_{\mathrm{B}}^{i}\right) \circ \operatorname{sgn}\left(\mathbf{x}_{\mathrm{BN}}^{i}\right)$$

The residual of layer $\mathrm{FC}_{1\mathrm{B}}$:

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{x}_{1 \mathrm{B}}^{i}} &=\sum_{j=1}^{m} \frac{\partial L}{\partial \mathbf{x}_{\mathrm{BN}}^{j}} \frac{\partial \mathbf{x}_{\mathrm{BN}}^{j}}{\partial \mathbf{x}_{1 \mathrm{B}}^{i}}=\sum_{j=1, j \neq i}^{m} \frac{\partial L}{\partial \mathbf{x}_{\mathrm{BN}}^{j}} \frac{\partial\left(\mathbf{x}_{1 \mathrm{B}}^{j}-\mu+\mathbf{b}_{1 \mathrm{B}}\right)}{\partial \mathbf{x}_{1 \mathrm{B}}^{i}} \\
&=\sum_{j=1}^{m} \frac{\partial L}{\partial \mathbf{x}_{\mathrm{BN}}^{j}}\left(\frac{\partial \mathbf{x}_{1 \mathrm{B}}^{j}}{\partial \mathbf{x}_{1 \mathrm{B}}^{i}}-\frac{\partial \mu}{\partial \mathbf{x}_{1 \mathrm{B}}^{i}}+0\right) \\
&=\frac{\partial L}{\partial \mathbf{x}_{\mathrm{BN}}^{i}}-\frac{1}{m} \sum_{j=1}^{m} \frac{\partial L}{\partial \mathbf{x}_{\mathrm{BN}}^{j}} \\
&=\frac{1}{m} \theta_{2 \mathrm{B}}^{T}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i}-\mathbf{y}_{\mathrm{B}}^{i}\right) \circ \operatorname{sgn}\left(\mathbf{x}_{\mathrm{BN}}^{i}\right)-\frac{1}{m^{2}} \sum_{j=1}^{m}\left[\theta_{2 \mathrm{B}}^{T}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{j}-\mathbf{y}_{\mathrm{B}}^{j}\right) \circ \operatorname{sgn}\left(\mathbf{x}_{\mathrm{BN}}^{j}\right)\right]
\end{aligned}$$

Task A:

The residual of layer $\mathrm{FC}_{2\mathrm{A}}$:

$$\begin{aligned}
\frac{\partial L}{\partial \hat{\mathbf{y}}_{\mathrm{A}}^{i}} &=\frac{\partial L_{A}}{\partial \hat{\mathbf{y}}_{\mathrm{A}}^{i}}+\frac{\partial L_{B}}{\partial \hat{\mathbf{y}}_{\mathrm{A}}^{i}} \\
&=\frac{1}{m}\left(\hat{\mathbf{y}}_{\mathrm{A}}^{i}-\mathbf{y}_{\mathrm{A}}^{i}\right)+\frac{\partial L_{B}}{\partial \mathbf{x}_{\oplus}^{i}} \frac{\partial \mathbf{x}_{\oplus}^{i}}{\partial \hat{\mathbf{y}}_{\mathrm{A}}^{i}} \\
&=\frac{1}{m}\left[\left(\hat{\mathbf{y}}_{\mathrm{A}}^{i}-\mathbf{y}_{\mathrm{A}}^{i}\right)+\theta_{2 \mathrm{B}}^{T}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i}-\mathbf{y}_{\mathrm{B}}^{i}\right)\right]
\end{aligned}$$

The residual of layer DP:

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{x}_{\mathrm{DP}}^{i}} &=\frac{\partial L}{\partial \hat{\mathbf{y}}_{\mathrm{A}}^{i}} \frac{\partial \hat{\mathbf{y}}_{\mathrm{A}}^{i}}{\partial \mathbf{x}_{\mathrm{DP}}^{i}} \\
&=\frac{1}{m} \theta_{2 A}^{T}\left[\left(\hat{\mathbf{y}}_{\mathrm{A}}^{i}-\mathbf{y}_{\mathrm{A}}^{i}\right)+\theta_{2 \mathrm{B}}^{T}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i}-\mathbf{y}_{\mathrm{B}}^{i}\right)\right]
\end{aligned}$$

The residual of layer $\mathrm{FC}_{1\mathrm{A}}$:

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{x}_{1 \mathrm{A}}^{i}} &=\frac{\partial L}{\partial \mathbf{x}_{\mathrm{DP}}^{i}} \frac{\partial \mathbf{x}_{\mathrm{DP}}^{i}}{\partial \mathbf{x}_{1 \mathrm{A}}^{i}} \\
&=\frac{1}{m} \theta_{2 A}^{T}\left[\left(\hat{\mathbf{y}}_{\mathrm{A}}^{i}-\mathbf{y}_{\mathrm{A}}^{i}\right)+\theta_{2 \mathrm{B}}^{T}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i}-\mathbf{y}_{\mathrm{B}}^{i}\right)\right] \circ \mathbf{M}^{i} \\
&\left(\mathbf{M}^{i}=\frac{1}{1-p} \operatorname{sgn}(\mathbf{r}^{i}\left(\mathbf{x}_{1 \mathrm{A}}^{i}\right)-p \cdot \mathbf{1})\right) \\
\frac{\partial L}{\partial \mathbf{z}_{1 \mathrm{A}}^{i}} &=\frac{\partial L}{\partial \mathbf{x}_{1 \mathrm{A}}^{i}} \frac{\partial \mathbf{x}_{1 \mathrm{A}}^{i}}{\partial \mathbf{z}_{1 \mathrm{A}}^{i}}=\frac{\partial L}{\partial \mathbf{x}_{1 \mathrm{A}}^{i}} \circ \frac{\partial \sin\left(\mathbf{z}_{1 \mathrm{A}}^{i}\right)}{\partial \mathbf{z}_{1 \mathrm{A}}^{i}} \\
&=\frac{1}{m} \theta_{2 A}^{T}\left[\left(\hat{\mathbf{y}}_{\mathrm{A}}^{i}-\mathbf{y}_{\mathrm{A}}^{i}\right)+\theta_{2 \mathrm{B}}^{T}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i}-\mathbf{y}_{\mathrm{B}}^{i}\right)\right] \circ \mathbf{M}^{i} \circ \cos\left(\mathbf{z}_{1 \mathrm{A}}^{i}\right)
\end{aligned}$$

Backward Propagation:

Therefore, we can get the gradients of the overall loss in a mini-batch with respect to the parameters at each layer.

Task A:

The gradients of layer FC_{2A}:

$$\frac{\partial L}{\partial \theta_{2\mathrm{A}}}=\sum_{i=1}^{m}\frac{\partial L}{\partial \hat{\mathbf{y}}_{\mathrm{A}}^{i}}\frac{\partial \hat{\mathbf{y}}_{\mathrm{A}}^{i}}{\partial \theta_{2\mathrm{A}}}=\frac{1}{m}\sum_{i=1}^{m}\left[\left(\hat{\mathbf{y}}_{\mathrm{A}}^{i}-\mathbf{y}_{\mathrm{A}}^{i}\right)+\theta_{2\mathrm{B}}^{T}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i}-\mathbf{y}_{\mathrm{B}}^{i}\right)\right]\left(\mathbf{x}_{\mathrm{DP}}^{i}\right)^{T}$$

$$\frac{\partial L}{\partial \mathbf{b}_{2\mathrm{A}}}=\sum_{i=1}^{m}\frac{\partial L}{\partial \hat{\mathbf{y}}_{\mathrm{A}}^{i}}\frac{\partial \hat{\mathbf{y}}_{\mathrm{A}}^{i}}{\partial \mathbf{b}_{2\mathrm{A}}}=\frac{1}{m}\sum_{i=1}^{m}\left[\left(\hat{\mathbf{y}}_{\mathrm{A}}^{i}-\mathbf{y}_{\mathrm{A}}^{i}\right)+\theta_{2\mathrm{B}}^{T}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i}-\mathbf{y}_{\mathrm{B}}^{i}\right)\right]$$

The gradients of layer FC_{1A}:

$$\frac{\partial L}{\partial \theta_{1\mathrm{A}}}=\sum_{i=1}^{m}\frac{\partial L}{\partial \mathbf{z}_{1\mathrm{A}}^{i}}\frac{\partial \mathbf{z}_{1\mathrm{A}}^{i}}{\partial \theta_{1\mathrm{A}}}=\frac{1}{m}\sum_{i=1}^{m}\left\{\theta_{2A}^{T}\left[\left(\hat{\mathbf{y}}_{\mathrm{A}}^{i}-\mathbf{y}_{\mathrm{A}}^{i}\right)+\theta_{2\mathrm{B}}^{T}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i}-\mathbf{y}_{\mathrm{B}}^{i}\right)\right]\circ\mathbf{M}^{i}\circ\cos\left(\mathbf{z}_{1\mathrm{A}}^{i}\right)\left(\mathbf{x}^{i}\right)^{T}\right\}$$

$$\frac{\partial L}{\partial \mathbf{b}_{1\mathrm{A}}}=\sum_{i=1}^{m}\frac{\partial L}{\partial \mathbf{z}_{1\mathrm{A}}^{i}}\frac{\partial \mathbf{z}_{1\mathrm{A}}^{i}}{\partial \mathbf{b}_{1\mathrm{A}}}=\frac{1}{m}\sum_{i=1}^{m}\left\{\theta_{2A}^{T}\left[\left(\hat{\mathbf{y}}_{\mathrm{A}}^{i}-\mathbf{y}_{\mathrm{A}}^{i}\right)+\theta_{2\mathrm{B}}^{T}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i}-\mathbf{y}_{\mathrm{B}}^{i}\right)\right]\circ\mathbf{M}^{i}\circ\cos\left(\mathbf{z}_{1\mathrm{A}}^{i}\right)\right\}$$

Task B:

The gradients of layer FC_{2B}:

$$\frac{\partial L}{\partial \theta_{2\mathrm{B}}}=\sum_{i=1}^{m}\frac{\partial L}{\partial \mathbf{z}_{2\mathrm{B}}^{i}}\frac{\partial \mathbf{z}_{2\mathrm{B}}^{i}}{\partial \theta_{2\mathrm{B}}}=\frac{1}{m}\sum_{i=1}^{m}\left[\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i}-\mathbf{y}_{\mathrm{B}}^{i}\right)\left(\mathbf{x}_{\oplus}^{i}\right)^{T}\right]$$

$$\frac{\partial L}{\partial \mathbf{b}_{2\mathrm{B}}}=\sum_{i=1}^{m}\frac{\partial L}{\partial \mathbf{z}_{2\mathrm{B}}^{i}}\frac{\partial \mathbf{z}_{2\mathrm{B}}^{i}}{\partial \mathbf{b}_{2\mathrm{B}}}=\frac{1}{m}\sum_{i=1}^{m}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i}-\mathbf{y}_{\mathrm{B}}^{i}\right)$$

The gradients of layer FC_{1B} and BN*:

$$\begin{aligned}\frac{\partial L}{\partial \theta_{1\mathrm{B}}}&=\sum_{i=1}^{m}\frac{\partial L}{\partial \mathbf{x}_{1\mathrm{B}}^{i}}\frac{\partial \mathbf{x}_{1\mathrm{B}}^{i}}{\partial \theta_{1\mathrm{B}}}\\&=\frac{1}{m}\sum_{i=1}^{m}\left\{\left\{\theta_{2\mathrm{B}}^{T}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i}-\mathbf{y}_{\mathrm{B}}^{i}\right)\circ\operatorname{sgn}\left(\mathbf{x}_{\mathrm{BN}}^{i}\right)-\frac{1}{m}\sum_{j=1}^{m}\left[\theta_{2\mathrm{B}}^{T}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{j}-\mathbf{y}_{\mathrm{B}}^{j}\right)\circ\operatorname{sgn}\left(\mathbf{x}_{\mathrm{BN}}^{j}\right)\right]\right\}\left(\mathbf{x}^{i}\right)^{T}\right\}\end{aligned}$$

$$\frac{\partial L}{\partial \mathbf{b}_{1\mathrm{B}}}=\sum_{i=1}^{m}\frac{\partial L}{\partial \mathbf{x}_{\mathrm{BN}}^{i}}\frac{\partial \mathbf{x}_{\mathrm{BN}}^{i}}{\partial \mathbf{b}_{1\mathrm{B}}}=\frac{1}{m}\sum_{i=1}^{m}\left[\theta_{2\mathrm{B}}^{T}\left(\hat{\mathbf{y}}_{\mathrm{B}}^{i}-\mathbf{y}_{\mathrm{B}}^{i}\right)\circ\operatorname{sgn}\left(\mathbf{x}_{\mathrm{BN}}^{i}\right)\right]$$