Notations: $\odot$ 表示 Hadamard product，$\otimes$ 表示 kronecker product，$sgn(x)=\begin{cases}1 & x\geq 0\\ 0 & \text{else}\end{cases}$，$f(X)=\begin{bmatrix} f(x_{1,1}) & \cdots & f(x_{1,n})\\ \vdots & & \vdots \\ f(x_{m,1}) & \cdots & f(x_{m,n})\end{bmatrix}\in\mathbb{R}^{m\times n}$

1、(1) 由 $y_i=BN_{\gamma,\beta}(\hat{x}_i)=\gamma\hat{x}_i+\beta$ 可知

$\frac{\partial y_i}{\partial\gamma}=\hat{x}_i$，$\frac{\partial y_i}{\partial\beta}=1$，其中 $\hat{x}_i$ 的计算即按照BN的处理流程，即：$\begin{cases}\mu_B=\frac{1}{m}\sum_{i=1}^{m}x_i\\ \sigma_B^2=\frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_B)^2\\ \hat{x}_i=\frac{x_i-\mu_B}{\sqrt{\sigma_B^2+\varepsilon}}\end{cases}$

(2) 考虑 softmax 函数 $f:\mathbb{R}^N\to\mathbb{R}^C$（即C分类），对于 $x\in\mathbb{R}^N$，$y=f(x)\in\mathbb{R}^C$，其中

$y_k=P(\text{类别}k|x,W)=\frac{\exp(W_k^Tx)}{\sum_{n=1}^{C}\exp(W_n^Tx)}$，其中 $W_i\in\mathbb{R}^N$（$1\leq i\leq C$ 表示权重）

考虑输出y对输入x的偏导：$\frac{\partial y_k}{\partial x}=\begin{bmatrix}\frac{\partial y_k}{\partial x_1}\\ \vdots\\ \frac{\partial y_k}{\partial x_N}\end{bmatrix}$

$$\frac{\partial y_k}{\partial x_i}=\frac{1}{(\sum_{n=1}^{C}\exp(W_n^Tx))^2}\left((\sum_{n=1}^{C}\exp(W_n^Tx))\exp(W_k^Tx)W_{k,i}-\exp(W_k^Tx)(\sum_{n=1}^{C}W_{n,i}\exp(W_n^Tx))\right)$$

$$=W_{k,i}y_k-y_k\cdot\frac{\sum_{n=1}^{C}W_{n,i}\exp(W_n^Tx)}{\sum_{n=1}^{C}\exp(W_n^Tx)}=y_k\left(W_{k,i}-\frac{\sum_{n=1}^{C}W_{n,i}\exp(W_n^Tx)}{\sum_{n=1}^{C}\exp(W_n^Tx)}\right)$$

∴ 输出y相对输入x的Jacobi矩阵 $\left[\frac{\partial y}{\partial x^T}\right]_{k,i}=\frac{\partial y_k}{\partial x_i}$ 如上所示

(3) 对于前向传播过程，首先计算单个样本对应的预测 $\hat{y}_A^i\in\mathbb{R}^a$ 与 $\hat{y}_B^i\in\mathbb{R}^b$

令 FC_{1A}，DP，FC_{1B}，BN^* 的输出分别为 $x_{1A}^i\in\mathbb{R}^s$，$x_{DP}^i\in\mathbb{R}^s$，$x_{1B}^i\in\mathbb{R}^t$，$x_{BN}^i\in\mathbb{R}^t$，

则有 $\begin{cases}x_{1A}^i=\sin(\theta_{1A}x^i+b_{1A})\\ x_{DP}^i=M\odot x_{1A}^i\\ \hat{y}_A^i=\theta_{2A}x_{DP}^i+b_{2A}\end{cases}$ 对应Task A的前向传播结果 $\Rightarrow$ $\hat{y}_A^i=\theta_{2A}(M\odot\sin(\theta_{1A}x^i+b_{1A}))+b_{2A}$

对于Task B，则有 $x_{1B}^i=\theta_{1B}x^i$

在计算BN^*输出时，$\mu=\frac{1}{m}\sum_{i=1}^{m}x_{1B}^i$

故 $x_{BN}^i=ReLu(x_{1B}^i-\mu+b_{1B})$

$=ReLu(\theta_{1B}x^i-\frac{1}{m}\theta_{1B}\sum_{i=1}^{m}x^i+b_{1B})$

∴ $\hat{y}_B^i=softmax(\theta_{2B}(\hat{y}_A^i+x_{BN}^i)+b_{2B})$

更进一步，$\hat{y}_B^i=softmax(\theta_{2B}(\hat{y}_A^i+x_{BN}^i)+b_{2B})=\frac{\exp(\theta_{2B}(\hat{y}_A^i+x_{BN}^i)+b_{2B})}{1_b^T\exp(\theta_{2B}(\hat{y}_A^i+x_{BN}^i)+b_{2B})}$

其中 $\hat{y}_{B,k}^i=\frac{\exp(\theta_{2B,k}(\hat{y}_A^i+x_{BN}^i)+b_{2B})}{\sum_{p=1}^{b}\exp(\theta_{2B,p}(\hat{y}_A^i+x_{BN}^i)+b_{2B})}$ 其中 θ_{2B} 表示矩阵的第i行

(4) 将损失函数L写成向量的形式：$L=\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{2}\|\hat{y}_A^i-y_A^i\|_2^2-(y_B^i)^T\log\hat{y}_B^i\right]$

∴ 可得 $dL=\sum_{i=1}^{m}(\frac{\partial L}{\partial\hat{y}_A^i})^Td\hat{y}_A^i+\sum_{i=1}^{m}(\frac{\partial L}{\partial\hat{y}_B^i})^Td\hat{y}_B^i$

$$=\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_A^i-y_A^i)^Td\hat{y}_A^i+\frac{1}{m}\sum_{i=1}^{m}(y_B^i)^T(\frac{1}{\hat{y}_B^i}\odot d\hat{y}_B^i)$$

$$=\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_A^i-y_A^i)^Td\hat{y}_A^i+\frac{1}{m}\sum_{i=1}^{m}[(y_B^i)^T\odot\frac{1}{(\hat{y}_B^i)^T}]d\hat{y}_B^i]$$

下面考虑将 $d\hat{y}_A^i$ 与 $d\hat{y}_B^i$ 使用 $d\theta_{1A}, d\theta_{2A}, d\theta_{1B}, d\theta_{2B}$ 表示

$$d\hat{y}_B^i = \hat{y}_B^i \odot (d\theta_{2B}(\hat{y}_A^i + X_{BN}^i) + db_{2B}) - \hat{y}_B^i \cdot \hat{y}_B^{iT}(d\theta_{2B}(\hat{y}_A^i + X_{BN}^i) + db_{2B})$$

$$\Rightarrow dL = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_A^i - y_A^i)^T d\hat{y}_A^i + \frac{1}{m}\sum_{i=1}^{m}(y_B^{iT} \odot \frac{1}{\hat{y}_B^{iT}})\left(\hat{y}_B^i \odot (d\theta_{2B}(\hat{y}_A^i + X_{BN}^i) + db_{2B}) - \hat{y}_B^i \hat{y}_B^{iT}(d\theta_{2B}(\hat{y}_A^i + X_{BN}^i) + db_{2B})\right)$$

$$= \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_A^i - y_A^i)^T d\hat{y}_A^i + \frac{1}{m}\sum_{i=1}^{m}\left(y_B^{iT}(d\theta_{2B}(\hat{y}_A^i + X_{BN}^i) + db_{2B}) - \hat{y}_B^{iT}(d\theta_{2B}(\hat{y}_A^i + X_{BN}^i) + db_{2B})\right)$$

由于 $\hat{y}_A^i$ 与 θ_{2B} 无关，故 $dL = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_A^i - y_A^i)^T d\hat{y}_A + tr\left(\left(\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_A^i + X_{BN}^i)(y_B^i - \hat{y}_B^i)^T\right) d\theta_{2B}\right) + tr\left(\frac{1}{m}\sum_{i=1}^{m}(y_B^i - \hat{y}_B^i)^T db_{2B}\right)$

故 $\frac{\partial L}{\partial \theta_{2B}} = \frac{1}{m}\sum_{i=1}^{m}(y_B^i - \hat{y}_B^i)(\hat{y}_A^i + X_{BN}^i)^T$，$\frac{\partial L}{\partial b_{2B}} = \frac{1}{m}\sum_{i=1}^{m}(y_B^i - \hat{y}_B^i)$

之后，求解 $\frac{\partial L}{\partial \theta_{1B}}$，先进一步表示 dy_B^i，表示为 $d\hat{y}_A^i$ 与 dX_{BN}^i 的微分形式

$$d\hat{y}_B^i = d\left(\frac{\exp(\theta_{2B}(\hat{y}_A^i + X_{BN}^i) + b_{2B})}{1_b^T \exp(\theta_{2B}(\hat{y}_A^i + X_{BN}^i) + b_{2B})}\right) = \hat{y}_B^i \odot (\theta_{2B}(d\hat{y}_A^i + dX_{BN}^i)) - \hat{y}_B^i \cdot \hat{y}_B^{iT}\theta_{2B}(d\hat{y}_A^i + dX_{BN}^i)$$

$$= \hat{y}_B^i \odot (\theta_{2B} d\hat{y}_A^i) + \hat{y}_B^i \odot (\theta_{2B} dX_{BN}^i) - \hat{y}_B^i \hat{y}_B^{iT}\theta_{2B} d\hat{y}_A^i - \hat{y}_B^i \hat{y}_B^{iT}\theta_{2B} dX_{BN}^i$$

由于 $\hat{y}_A^i$ 不是 θ_{1B} 与 b_{1B} 的函数，故只用考虑 dX_{BN}^i

一步表示 $X_{BN}^i = ReLu(X_{1B}^i - \frac{1}{m}\sum_{s=1}^{m} X_{1B}^s + b_{1B})$

$$= ReLu(\theta_{1B}(X^i - \frac{1}{m}\sum_{s=1}^{m} X^s) + b_{1B})$$

故 $dX_{BN}^i = sgn(\theta_{1B}(X^i - \bar{X}) + b_{1B}) \odot (d\theta_{1B}(X^i - \bar{X}) + db_{1B})$ $(\bar{X} = \frac{1}{m}\sum_{s=1}^{m} X^s)$

代入可得：$dL = C + \frac{1}{m}\sum_{i=1}^{m}(y_B^{iT} \odot \frac{1}{\hat{y}_B^{iT}})(\hat{y}_B^i \odot (\theta_{2B} dX_{BN}^i) - \hat{y}_B^i \hat{y}_B^{iT}\theta_{2B} dX_{BN}^i)$

$$= C + \frac{1}{m}\sum_{i=1}^{m}(y_B^i - \hat{y}_B^i)^T \theta_{2B} dX_{BN}^i$$

$$= C + \frac{1}{m}\sum_{i=1}^{m}(y_B^i - \hat{y}_B^i)^T \theta_{2B}\left(sgn(\theta_{1B}(X^i - \bar{X}) + b_{1B}) \odot (d\theta_{1B}(X^i - \bar{X}) + db_{1B})\right)$$

$$= C + \frac{1}{m}\sum_{i=1}^{m} tr\left(\left[(y_B^i - \hat{y}_B^i)^T \theta_{2B} \odot sgn^T(\theta_{1B}(X^i - \bar{X}) + b_{1B})\right](d\theta_{1B}(X^i - \bar{X}) + db_{1B})\right)$$

$$= C + \frac{1}{m}\sum_{i=1}^{m} tr\left((X^i - \bar{X})\left[(y_B^i - \hat{y}_B^i)^T \theta_{2B} \odot sgn^T(\theta_{1B}(X^i - \bar{X}) + b_{1B})\right] d\theta_{1B}\right)$$

$$+ \frac{1}{m}\sum_{i=1}^{m} tr\left(\left[(y_B^i - \hat{y}_B^i)^T \theta_{2B} \odot sgn^T(\theta_{1B}(X^i - \bar{X}) + b_{1B})\right] db_{1B}\right)$$ 其中 C 为无关项

故 $\frac{\partial L}{\partial \theta_{1B}} = \frac{1}{m}\sum_{i=1}^{m}\left(\theta_{2B}^T(y_B^i - \hat{y}_B^i) \odot sgn(\theta_{1B}(X^i - \bar{X}) + b_{1B})\right)(X^i - \bar{X})^T$

$$\frac{\partial L}{\partial b_{1B}} = \frac{1}{m}\sum_{i=1}^{m}\left(\theta_{2B}^T(y_B^i - \hat{y}_B^i) \odot sgn(\theta_{1B}(X^i - \bar{X}) + b_{1B})\right)$$

之后求解 $\frac{\partial L}{\partial \theta_{2A}}, \frac{\partial L}{\partial b_{2A}}, \frac{\partial L}{\partial \theta_{1A}}, \frac{\partial L}{\partial b_{1A}}$

由于 $\hat{y}_A$ 与 $\hat{y}_B^i$ 与 $\theta_{2A}, \theta_{2B}, b_{2A}, b_{2B}$ 均有关，故需同时展开

$$d\hat{y}_A^i = d(\theta_{2A} X_{Dp} + b_{2A}) = d\theta_{2A} X_{Dp} + db_{2A}$$

$$d\hat{y}_B^i = \hat{y}_B^i \odot (\theta_{2B}\, d\hat{y}_A^i) + \hat{y}_B^i \odot (\theta_{2B}\, dX_{BN}^i) - \hat{y}_B^i \hat{y}_B^{iT} \theta_{2B}\, d\hat{y}_A^i - \hat{y}_B^i \hat{y}_B^{iT} \theta_{2B}\, dX_{BN}^i$$

首先求解 $\frac{\partial L}{\partial \hat{y}_A^i}$，

$$dL = \frac{1}{m}\sum_{i=1}^{m} (\hat{y}_A^i - y_A^i)^T d\hat{y}_A^i + \frac{1}{m}\sum_{i=1}^{m} [(y_B^i)^T \odot \frac{1}{\hat{y}_B^{iT}}\, d\hat{y}_B^i],$$

$$= \frac{1}{m}\sum_{i=1}^{m} (\hat{y}_A^i - y_A^i)^T d\hat{y}_A^i + \frac{1}{m}\sum_{i=1}^{m} [(y_B^i)^T \odot \frac{1}{\hat{y}_B^{iT}} \left(\hat{y}_B^i \odot (\theta_{2B}\, d\hat{y}_A^i) - \hat{y}_B^i \hat{y}_B^{iT} \theta_{2B}\, d\hat{y}_A^i\right)] + C$$ （C代表与 $d\hat{y}_A$ 无关的项）

$$= \frac{1}{m}\sum_{i=1}^{m} (\hat{y}_A^i - y_A^i)^T d\hat{y}_A^i + \frac{1}{m}\sum_{i=1}^{m} (y_B^i - \hat{y}_B^i)^T \theta_{2B}\, d\hat{y}_A^i + C$$

$$= \frac{1}{m}\sum_{i=1}^{m} \left((\hat{y}_A^i - y_A^i) + \theta_{2B}^T (y_B^i - \hat{y}_B^i)\right)^T d\hat{y}_A^i + C$$

又 $d\hat{y}_A^i = d(\theta_{2A} X_{Dp}^i + b_{2A}) = d\theta_{2A} X_{Dp}^i + db_{2A}$

$$\Rightarrow dL = \frac{1}{m}\sum_{i=1}^{m} \left((\hat{y}_A^i - y_A^i) + \theta_{2B}^T (y_B^i - \hat{y}_B^i)\right)^T (d\theta_{2A} X_{Dp}^i + db_{2A})$$

$$= tr\left(\frac{1}{m}\sum_{i=1}^{m} \left((\hat{y}_A^i - y_A^i) + \theta_{2B}^T (y_B^i - \hat{y}_B^i)\right) X_{Dp}^{iT})^T d\theta_{2A}\right)$$

$$+ tr\left(\frac{1}{m}\sum_{i=1}^{m} \left((\hat{y}_A^i - y_A^i) + \theta_{2B}^T (y_B^i - \hat{y}_B^i)\right)^T db_{2A}\right)$$

故 $\frac{\partial L}{\partial \theta_{2A}} = \frac{1}{m}\sum_{i=1}^{m} \left[(\hat{y}_A^i - y_A^i) + \theta_{2B}^T (y_B^i - \hat{y}_B^i)\right] X_{Dp}^{iT}$

$\frac{\partial L}{\partial b_{2A}} = \frac{1}{m}\sum_{i=1}^{m} \left((\hat{y}_A^i - y_A^i) + \theta_{2B}^T (y_B^i - \hat{y}_B^i)\right)$

同理，继续求解残差

$$y_A^i = \theta_{2A} X_{Dp}^i + b_{2A} = \theta_{2A}(M \odot X_{1A}^i) + b_{2A}$$

故 $dy_A^i = \theta_{2A}(M \odot dX_{1A}^i)$

又 $dX_{1A}^i = d[\sin(\theta_{1A} X^i + b_{1A})] = \cos(\theta_{1A} X^i + b_{1A}) \odot (d\theta_{1A} X^i + db_{1A})$

故 $dy_A^i = \theta_{2A}(M \odot (\cos(\theta_{1A} X^i + b_{1A}) \odot (d\theta_{1A} X^i + db_{1A})))$

$$= \theta_{2A}((M \odot \cos(\theta_{1A} X^i + b_{1A})) \odot (d\theta_{1A} X^i + db_{11}))$$

代入得 $dL = \frac{1}{m}\sum_{i=1}^{m} \left((\hat{y}_A^i - y_A^i) + \theta_{2B}^T (y_B^i - \hat{y}_B^i)\right)^T \theta_{2A}((M \odot \cos(\theta_{1A} X^i + b_{1A})) \odot (d\theta_{1A} X^i + db_{1A}))$

$$= tr\left(\frac{1}{m}\sum_{i=1}^{m} ((M \odot \cos(\theta_{1A} X^i + b_{1A}))^T \odot (((\hat{y}_A^i - y_A^i) + \theta_{2B}^T (y_B^i - \hat{y}_B^i))^T \theta_{2A})(d\theta_{1A} X^i + db_{1A})\right)$$

$$= tr\left(\frac{1}{m}\sum_{i=1}^{m} X_i ((M \odot \cos(\theta_{1A} X^i + b_{1A}))^T \odot (((\hat{y}_A^i - y_A^i) + \theta_{2B}^T (y_B^i - \hat{y}_B^i))^T \theta_{2A}) d\theta_{1A}\right)$$

$$+ tr\left(((M \odot \cos(\theta_{1A} X^i + b_{1A}))^T \odot (((\hat{y}_A^i - y_A^i) + \theta_{2B}^T (y_B^i - \hat{y}_B^i))^T \theta_{2A}) db_{1A}\right)$$

故 $$\frac{\partial L}{\partial \theta_{1A}} = \frac{1}{m}\sum_{i=1}^{m}\left(\left(M \odot \cos(\theta_{1A} x^{i} + b_{1A})\right) \odot \left(\theta_{2A}^{T} (\hat{y}_A^{i} - y_A^{i}) + \theta_{2B}^{T} (y_B^{i} - \hat{y}_B^{i}) \right) \right) x^{iT},$$

$$\frac{\partial L}{\partial b_{1A}} = \frac{1}{m}\sum_{i=1}^{m}\left(\left(M \odot \cos(\theta_{1A} x^{i} + b_{1A})\right) \odot \left(\theta_{2A}^{T} (\hat{y}_A^{i} - y_A^{i}) + \theta_{2B}^{T} (y_B^{i} - \hat{y}_B^{i}) \right) \right),$$