A PROOF

A.1 PROOF OF THEOREM 1

Theorem 1: *For a m-layer fully connected neural network with d-dimensional input and Q-dimensional output, assume the Sigmoid activation functions and square loss are used, and the training data distribution $\boldsymbol{D} = [c_1/n, ..., c_Q/n]$. In the parameter space, if at some point such that all the model prediction outputs $\boldsymbol{y} = \boldsymbol{D}$, that point is a critical point by first order optimality.*

Proof: To proof the theorem, the key is showing that at such point the gradients are all zero. The square loss function L is defined as:

$$L = \sum_{i=1}^{n} L_i = \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{t}_i)^2, \tag{3}$$

where $\boldsymbol{y}_i$ and $\boldsymbol{t}_i$ are the prediction output and target with respect to ith input sample, respectively. Given that the neural network is m fully connected hidden layers, we denote $\boldsymbol{z}_j \in \mathbb{R}^{h_j}$ as the output vector from jth hidden layer with h_j neurons, and denote $\boldsymbol{W}_j$ as the corresponding weight matrix and $\boldsymbol{b}_j$ as the bias vector, where $j = 1, \cdots, m+1$. Note that $\boldsymbol{W}_{m+1}$ and $\boldsymbol{b}_{m+1}$ are the weight and bias of the output layer. Since the activation function is Sigmoid, take the derivative of L with respect to model parameters $\boldsymbol{W}$ and $\boldsymbol{b}$, we have,

$$\frac{dL}{d\boldsymbol{b}_{m+1}} = \sum_{i=1}^{n} \frac{dL_i}{d\boldsymbol{b}_{m+1}} = \sum_{i=1}^{n} [2(\boldsymbol{y}_i - \boldsymbol{t}_i)^T \circ [\boldsymbol{y}_i \circ (\boldsymbol{1} - \boldsymbol{y}_i)]]^T \tag{4}$$

$$\frac{dL}{d\boldsymbol{W}_{m+1}} = \sum_{i=1}^{n} \frac{dL_i}{d\boldsymbol{W}_{m+1}} = \sum_{i=1}^{n} [2(\boldsymbol{y}_i - \boldsymbol{t}_i)^T \circ [\boldsymbol{y}_i \circ (\boldsymbol{1} - \boldsymbol{y}_i)]]^T \boldsymbol{z}_m^T \tag{5}$$

$$\frac{dL}{d\boldsymbol{b}_j} = \sum_{i=1}^{n} \frac{dL_i}{d\boldsymbol{b}_j} = \sum_{i=1}^{n} [2(\boldsymbol{y}_i - \boldsymbol{t}_i)^T \circ [\boldsymbol{y}_i \circ (\boldsymbol{1} - \boldsymbol{y}_i)] \boldsymbol{W}_{m+1} \circ [\boldsymbol{z}_m \circ (\boldsymbol{1} - \boldsymbol{z}_m)] \cdots \boldsymbol{W}_{j+1} \circ [\boldsymbol{z}_j \circ (\boldsymbol{1} - \boldsymbol{z}_j)]]^T \tag{6}$$

$$\frac{dL}{d\boldsymbol{W}_j} = \sum_{i=1}^{n} \frac{dL_i}{d\boldsymbol{W}_j} = \sum_{i=1}^{n} [2(\boldsymbol{y}_i - \boldsymbol{t}_i)^T \circ [\boldsymbol{y}_i \circ (\boldsymbol{1} - \boldsymbol{y}_i)] \boldsymbol{W}_{m+1} \circ [\boldsymbol{z}_m \circ (\boldsymbol{1} - \boldsymbol{z}_m)] \cdots \boldsymbol{W}_{j+1} \circ [\boldsymbol{z}_j \circ (\boldsymbol{1} - \boldsymbol{z}_j)]]^T \boldsymbol{z}_{j-1}^T \tag{7}$$

$$\frac{dL}{d\boldsymbol{b}_1} = \sum_{i=1}^{n} \frac{dL_i}{d\boldsymbol{b}_1} = \sum_{i=1}^{n} [2(\boldsymbol{y}_i - \boldsymbol{t}_i)^T \circ [\boldsymbol{y}_i \circ (\boldsymbol{1} - \boldsymbol{y}_i)] \boldsymbol{W}_{m+1} \circ [\boldsymbol{z}_m \circ (\boldsymbol{1} - \boldsymbol{z}_m)] \cdots \boldsymbol{W}_2 \circ [\boldsymbol{z}_1 \circ (\boldsymbol{1} - \boldsymbol{z}_1)]]^T \tag{8}$$

$$\frac{dL}{d\boldsymbol{W}_1} = \sum_{i=1}^{n} \frac{dL_i}{d\boldsymbol{W}_1} = \sum_{i=1}^{n} [2(\boldsymbol{y}_i - \boldsymbol{t}_i)^T \circ [\boldsymbol{y}_i \circ (\boldsymbol{1} - \boldsymbol{y}_i)] \boldsymbol{W}_{m+1} \circ [\boldsymbol{z}_m \circ (\boldsymbol{1} - \boldsymbol{z}_m)] \cdots \boldsymbol{W}_2 \circ [\boldsymbol{z}_1 \circ (\boldsymbol{1} - \boldsymbol{z}_1)]]^T \boldsymbol{x}^T, \tag{9}$$

where $j = 1, 2, \cdots, m$, and $\circ$ denotes element-wise multiplication. Assume that the inputs of Sigmoid function are in range $[-1, 1]$, then it can be approximated as a linear function with constant slope $s \sim 0.25$. Thus,

$$\frac{dL}{d\boldsymbol{b}_{m+1}} = \sum_{i=1}^{n} \frac{dL_i}{d\boldsymbol{b}_{m+1}} = \sum_{i=1}^{n} [2(\boldsymbol{y}_i - \boldsymbol{t}_i)^T \circ [\boldsymbol{y}_i \circ (\boldsymbol{1} - \boldsymbol{y}_i)]]^T \tag{10}$$

$$\frac{dL}{d\boldsymbol{W}_{m+1}} = \sum_{i=1}^{n} \frac{dL_i}{d\boldsymbol{W}_{m+1}} = \sum_{i=1}^{n} [2(\boldsymbol{y}_i - \boldsymbol{t}_i)^T \circ [\boldsymbol{y}_i \circ (\boldsymbol{1} - \boldsymbol{y}_i)]]^T \boldsymbol{z}_m^T \tag{11}$$

$$\frac{dL}{d\boldsymbol{b}_j} = \sum_{i=1}^{n} \frac{dL_i}{d\boldsymbol{b}_j} = \sum_{i=1}^{n} [2(\boldsymbol{y}_i - \boldsymbol{t}_i)^T \circ [\boldsymbol{y}_i \circ (\boldsymbol{1} - \boldsymbol{y}_i)] \boldsymbol{W}_{m+1} s \cdots \boldsymbol{W}_{j+1} s]^T \tag{12}$$

$$\frac{dL}{d\boldsymbol{W}_j} = \sum_{i=1}^{n} \frac{dL_i}{d\boldsymbol{W}_j} = \sum_{i=1}^{n} [2(\boldsymbol{y}_i - \boldsymbol{t}_i)^T \circ [\boldsymbol{y}_i \circ (\boldsymbol{1} - \boldsymbol{y}_i)] \boldsymbol{W}_{m+1} s \cdots \boldsymbol{W}_{j+1} s]^T \boldsymbol{z}_{j-1}^T \tag{13}$$

$$\frac{dL}{d\boldsymbol{b}_1} = \sum_{i=1}^{n} \frac{dL_i}{d\boldsymbol{b}_1} = \sum_{i=1}^{n} [2(\boldsymbol{y}_i - \boldsymbol{t}_i)^T \circ [\boldsymbol{y}_i \circ (\boldsymbol{1} - \boldsymbol{y}_i)] \boldsymbol{W}_{m+1} s \cdots \boldsymbol{W}_2 s]^T \tag{14}$$

$$\frac{dL}{d\boldsymbol{W}_1} = \sum_{i=1}^{n} \frac{dL_i}{d\boldsymbol{W}_1} = \sum_{i=1}^{n} [2(\boldsymbol{y}_i - \boldsymbol{t}_i)^T \circ [\boldsymbol{y}_i \circ (\boldsymbol{1} - \boldsymbol{y}_i)] \boldsymbol{W}_{m+1} s \cdots \boldsymbol{W}_2 s]^T \boldsymbol{x}^T. \tag{15}$$

Now assume at some point in parameter space such that $\boldsymbol{y} = \boldsymbol{y}_{eq} = [c_1/n, ..., c_Q/n]$, plugging in $\boldsymbol{y}_{eq}$, Eq.(12) becomes

$$\begin{aligned} \sum_{i=1}^{n} \frac{dL_i}{d\boldsymbol{b}_j} =& \sum_{i=1}^{c_1} \frac{dL_i}{d\boldsymbol{b}_j} + \cdots + \sum_{i=1}^{c_Q} \frac{dL_i}{d\boldsymbol{b}_j} \\ =& \sum_{i=1}^{c_1} [2(\boldsymbol{y}_{eq} - \boldsymbol{t}_i)^T \circ [\boldsymbol{y}_{eq} \circ (\boldsymbol{1} - \boldsymbol{y}_{eq})] \boldsymbol{W}_{m+1} s \cdots \boldsymbol{W}_{j+1} s]^T + \cdots \\ &+ \sum_{i=1}^{c_Q} [2(\boldsymbol{y}_{eq} - \boldsymbol{t}_i)^T \circ [\boldsymbol{y}_{eq} \circ (\boldsymbol{1} - \boldsymbol{y}_{eq})] \boldsymbol{W}_{m+1} s \cdots \boldsymbol{W}_{j+1} s]^T. \end{aligned} \tag{16}$$

The matrix multiplication terms are the same, so we can just focus on the terms with $\boldsymbol{y}_{eq}$. Consider the first entry of Eq.(16) without constant terms, we have

$$\begin{aligned} &\sum_{i=1}^{c_1} [2(y_{eq1} - 1)[y_{eq1}(1 - y_{eq1})] + \cdots + \sum_{i=1}^{c_Q} [2(y_{eq1} - 0)[y_{eq1}(1 - y_{eq1})] \\ &= 2 \left((n - c_1) \frac{c_1^2 (n - c_1)}{n^3} + c_2 \frac{c_1^2 (n - c_1)}{n^3} + \cdots + c_Q \frac{c_1^2 (n - c_1)}{n^3} \right) = 0. \end{aligned} \tag{17}$$

This applies for any jth entry of Eq.(16). Therefore, Eq.(12) is equal to zero. With the same argument, we have Eq.(10) and Eq.(14) equal to zero. Now plug $\boldsymbol{y}_eq$ into Eq.(13), we have

$$\begin{aligned} \frac{dL}{d\boldsymbol{W}_j} =& \sum_{i=1}^{c_1} \frac{dL_i}{d\boldsymbol{W}_j} + \cdots + \sum_{i=1}^{c_Q} \frac{dL_i}{d\boldsymbol{W}_j} \\ =& \sum_{i=1}^{c_1} [2(\boldsymbol{y}_{eq} - \boldsymbol{t}_i)^T \circ [\boldsymbol{y}_{eq} \circ (\boldsymbol{1} - \boldsymbol{y}_{eq})] \boldsymbol{W}_{m+1} s \cdots \boldsymbol{W}_{j+1} s]^T \boldsymbol{z}_{j-1}^T + \cdots \\ &+ \sum_{i=1}^{c_Q} [2(\boldsymbol{y}_{eq} - \boldsymbol{t}_i)^T \circ [\boldsymbol{y}_{eq} \circ (\boldsymbol{1} - \boldsymbol{y}_{eq})] \boldsymbol{W}_{m+1} s \cdots \boldsymbol{W}_{j+1} s]^T \boldsymbol{z}_{j-1}^T. \end{aligned} \tag{18}$$

Since $\boldsymbol{z}_{j-1} = \boldsymbol{W}_{j-1}(\boldsymbol{W}_{j-2} \cdots (\boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_{j-2}) + \boldsymbol{b}_{j-1}$, terms without $\boldsymbol{x}$ are zero by previous analysis. Thus, we can focus on terms with $\boldsymbol{x}$ in Eq.(18). Consider the first entry of terms with $\boldsymbol{x}$, we have

$$\begin{aligned} &\frac{2c_1(n - c_1)}{n^2} \boldsymbol{W}_{m+1} s \cdots \boldsymbol{W}_1 s \left(-\frac{n - c_1}{n} \sum_{c_1} \boldsymbol{x}_i + \frac{c_1}{n} \sum_{c_2} \boldsymbol{x}_i + \cdots + \frac{c_1}{n} \sum_{c_Q} \boldsymbol{x}_i \right) \\ &= \frac{2c_1^2(n - c_1)}{n^3} \boldsymbol{W}_{m+1} s \cdots \boldsymbol{W}_1 s \left(-\frac{n - c_1}{n} \sum_{c_1} \boldsymbol{x}_i + \sum_{c_2} \boldsymbol{x}_i + \cdots + \sum_{c_Q} \boldsymbol{x}_i \right) \\ &= \frac{2c_1^2(n - c_1)}{n^3} \boldsymbol{W}_{m+1} s \cdots \boldsymbol{W}_1 s \left(\sum_{c_2} \boldsymbol{x}_i - \frac{c_2}{c_1} \sum_{c_1} \boldsymbol{x}_i + \cdots + \sum_{c_Q} \boldsymbol{x}_i - \frac{c_Q}{c_1} \sum_{c_1} \boldsymbol{x}_i \right). \end{aligned} \tag{19}$$

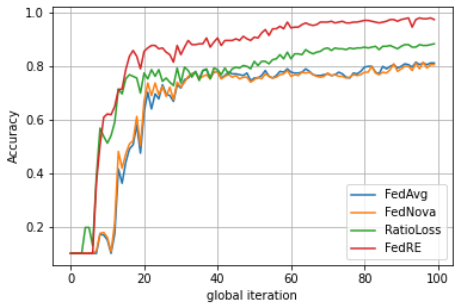
Eq.(19) equals to zero when $\frac{1}{c_j} \sum_{c_j} \boldsymbol{x}_i - \frac{1}{c_1} \sum_{c_1} \boldsymbol{x}_i$ for all $j = 1, 2, ..., Q$. That is, when The average of input samples from all classes are the same, the gradient Eq.(13) is zero. With the same argument, we have Eq.(11) and Eq.(15) are zero. Therefore, all the loss gradient with respect to model parameters from Eq.(4) to Eq.(9) are zero at such point, and this completes the proof of Theorem 1.

A.2 COMMENTS ON THEOREM 1

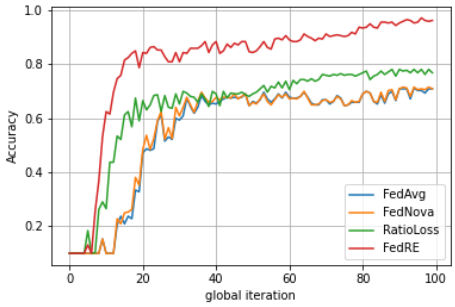
Theorem 1 shows that under the assumptions, if a point in parameter space makes all the prediction outputs equal to the class distribution, then such point is a critical point by first order optimality. There are two important assumptions in the proof. The first one is the linear approximation of Sigmoid. Since the output from Sigmoid function is in the range $[-1, 1]$, as long as the magnitudes of model parameters are not too large, the approximation would still hold. The second assumption is that sample averages from different classes are all the same, which seems to be a much stronger assumption. Nevertheless, in practical classification problem, the pixel values of input images are normalized, which makes the average difference between each class bounded. In addition, as we can see in Eq.(19), s is a constant that is smaller than 1, and following the previous argument that the magnitudes of model parameters are not too large. Thus, from the theoretical point of view, the error caused by nonzero average difference can be mitigated by increasing the number of hidden layers. This conclusion is validated in the experiment as shown in Fig.4. In summary, the assumptions for the proof of Theorem 1 are reasonable, and the theoretical analysis matches the observations in practical experiments.

B CONVERGENCE CURVE

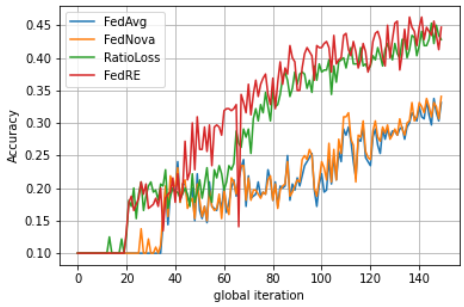
Four the convergence curves from experiments shown in Table 2 and 3 are provided in Fig.6 as the supplementary material. We can observe that the proposed method FedRE has the best performance among all 4 methods in all different settings.



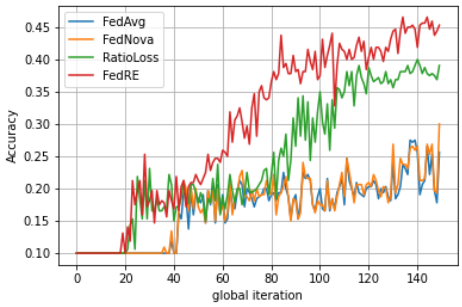
(a) MNIST, $\rho = 10$, $C = 0.3$, 1 minority class



(b) MNIST, $\rho = 10$, $C = 0.3$, 2 minority class



(c) CIFAR10, $\rho = 10$, $C = 0.3$, 1 minority class



(d) CIFAR10, $\rho = 10$, $C = 0.3$, 2 minority class

Figure 6: Four convergence curves of 4 algorithms on MNIST and CIFAR10 with different settings.