

A MATHEMATICAL DEMONSTRATION

We consider the classification problem with C classes in Federated Learning. The function $f: \mathcal{X} \rightarrow \mathcal{Z}$ maps data \mathbf{x} to the probability simplex \mathcal{Z} and $\mathcal{Z} = \{z \mid \sum_{i=1}^C z_i = 1; z_i \geq 0, \forall i \in [C]\}$, where z_i is the probability of class i . The population cross-entropy loss $l(\omega)$ is defined in (9).

$$\begin{aligned} l(\omega) &= \mathbb{E}_{\mathbf{x}, y \sim p} \left[\sum_{i=1}^C \mathbb{I}_{y=i} (-\log f_i(\mathbf{x}, \omega)) \right] \\ &= \sum_{i=1}^C p(y=i) \mathbb{E}_{\mathbf{x} \mid y=i} [-\log f_i(\mathbf{x}, \omega)]. \end{aligned} \quad (9)$$

To bound the divergence between the weights obtained by the FedAVG algorithm ω_{mT}^f and the optimal weights ω_{mT}^* on the test dataset, an intermediate variable ω_{mT}^c is introduced in (10) to assist the proof. The ω_{mT}^c physically represents the weights trained over the data from the selected clients in a centralized manner. The m is the round number and T is the number of optimization steps conducted in each round.

$$\begin{aligned} \|\omega_{mT}^f - \omega_{mT}^*\| &\leq \|\omega_{mT}^f - \omega_{mT}^c + \omega_{mT}^c - \omega_{mT}^*\| \\ &\leq \|\omega_{mT}^f - \omega_{mT}^c\| + \|\omega_{mT}^c - \omega_{mT}^*\|. \end{aligned} \quad (10)$$

The optimization step in local SGD is shown in (11), where p_l^k is the local data distribution of client k and η is the learning rate.

$$\begin{aligned} \omega_t^k &= \omega_{t-1}^k - \eta \nabla_{\omega} l(\omega) \\ &= \omega_{t-1}^k - \eta \sum_{i=1}^C p_l^k(y=i) \nabla_{\omega} \mathbb{E}_{\mathbf{x} \mid y=i} [-\log f_i(\mathbf{x}, \omega_{t-1}^k)]. \end{aligned} \quad (11)$$

The optimization step in centralized SGD is shown in (12) and $p_o(y=j) = \sum_{k \in \mathcal{S}} p_l^k(y=j) / |\mathcal{S}|$, which is the population data distribution.

$$\omega_t^c = \omega_{t-1}^c - \eta \sum_{i=1}^C p_o(y=i) \nabla_{\omega} \mathbb{E}_{\mathbf{x} \mid y=i} [-\log f_i(\mathbf{x}, \omega_{t-1}^c)]. \quad (12)$$

We will next derive the boundaries of $\|\omega_{mT}^f - \omega_{mT}^c\|$ and $\|\omega_{mT}^c - \omega_{mT}^*\|$ in Section A.1 and Section A.2 separately.

A.1 Boundary of $\|\omega_{mT}^f - \omega_{mT}^c\|$

$$\begin{aligned} \|\omega_{mT}^f - \omega_{mT}^c\| &= \left\| \frac{1}{K} \sum_{k=1}^K \omega_{mT}^k - \omega_{mT}^c \right\| \\ &= \left\| \frac{1}{K} \sum_{k=1}^K (\omega_{mT-1}^k - \eta \sum_{i=1}^C p_l^k(y=i) \nabla_{\omega} \mathbb{E}_{\mathbf{x} \mid y=i} [-\log f_i(\mathbf{x}, \omega_{mT-1}^k)]) \right. \\ &\quad \left. - (\omega_{mT-1}^c - \eta \sum_{i=1}^C p_o(y=i) \nabla_{\omega} \mathbb{E}_{\mathbf{x} \mid y=i} [-\log f_i(\mathbf{x}, \omega_{mT-1}^c)]) \right\| \\ &\leq \left\| \frac{1}{K} \sum_{k=1}^K \omega_{mT-1}^k - \omega_{mT-1}^c \right\| + \eta \left\| \frac{1}{K} \sum_{k=1}^K \sum_{i=1}^C p_l^k(y=i) \right. \\ &\quad \left. (\nabla_{\omega} \mathbb{E}_{\mathbf{x} \mid y=i} [-\log f_i(\mathbf{x}, \omega_{mT-1}^k)]) - \nabla_{\omega} \mathbb{E}_{\mathbf{x} \mid y=i} [-\log f_i(\mathbf{x}, \omega_{mT-1}^c)]) \right\| \end{aligned}$$

$$\begin{aligned} &\stackrel{\textcircled{1}}{\leq} \frac{1}{K} \sum_{k=1}^K \|\omega_{mT-1}^k - \omega_{mT-1}^c\| + \frac{\eta\lambda}{K} \sum_{k=1}^K \sum_{i=1}^C p_l^k(y=i) \|\omega_{mT-1}^k - \omega_{mT-1}^c\| \\ &= \frac{1}{K} \sum_{k=1}^K (1 + \eta\lambda) \|\omega_{mT-1}^k - \omega_{mT-1}^c\|. \end{aligned} \quad (13)$$

The inequality $\textcircled{1}$ holds because we assume $\nabla_{\omega} \mathbb{E}_{\mathbf{x} \mid y=i} [-\log f_i(\mathbf{x}, \omega)]$ is λ -Lipschitz for $\mathbf{x}, y \sim p$. In that case $\|\nabla_{\omega} \mathbb{E}_{\mathbf{x} \mid y=i} [-\log f_i(\mathbf{x}, \omega_1)] - \nabla_{\omega} \mathbb{E}_{\mathbf{x} \mid y=i} [-\log f_i(\mathbf{x}, \omega_2)]\| \leq \lambda \|\omega_1 - \omega_2\|$. Then we have

$$\begin{aligned} &\|\omega_{mT-1}^k - \omega_{mT-1}^c\| \\ &= \|\omega_{mT-2}^k - \eta \sum_{i=1}^C p_l^k(y=i) \nabla_{\omega} \mathbb{E}_{\mathbf{x} \mid y=i} [-\log f_i(\mathbf{x}, \omega_{mT-2}^k)] \\ &\quad - \omega_{mT-2}^c + \eta \sum_{i=1}^C p_o(y=i) \nabla_{\omega} \mathbb{E}_{\mathbf{x} \mid y=i} [-\log f_i(\mathbf{x}, \omega_{mT-2}^c)]\| \\ &\leq \|\omega_{mT-2}^k - \omega_{mT-2}^c\| + \eta \left\| \sum_{i=1}^C p_l^k(y=i) (\nabla_{\omega} \mathbb{E}_{\mathbf{x} \mid y=i} [-\log f_i(\mathbf{x}, \omega_{mT-2}^k)] \right. \\ &\quad \left. - \nabla_{\omega} \mathbb{E}_{\mathbf{x} \mid y=i} [-\log f_i(\mathbf{x}, \omega_{mT-2}^c)]) \right\| \\ &\quad + \eta \left\| \sum_{i=1}^C (p_l^k(y=i) - p_o(y=i)) \nabla_{\omega} \mathbb{E}_{\mathbf{x} \mid y=i} [-\log f_i(\mathbf{x}, \omega_{mT-2}^c)] \right\| \\ &\leq \|\omega_{mT-2}^k - \omega_{mT-2}^c\| + \eta \sum_{i=1}^C p_l^k(y=i) \lambda \|\omega_{mT-2}^k - \omega_{mT-2}^c\| \\ &\quad + \eta \mathbf{g}(\omega_{mT-2}^c) \|p_l^k - p_o\|_1 \\ &= (1 + \eta\lambda) \|\omega_{mT-2}^k - \omega_{mT-2}^c\| + \eta \mathbf{g}(\omega_{mT-2}^c) \|p_l^k - p_o\|_1 \end{aligned} \quad (14)$$

Note that $\mathbf{g}(\omega) = \max_{i=1}^C \|\nabla_{\omega} \mathbb{E}_{\mathbf{x} \mid y=i} [-\log f_i(\mathbf{x}, \omega)]\|$. Formula (14) implies that the weight divergence after each step of optimization of client k is restricted by the weight divergence from the last step plus a term which is related to the discrepancy between p_l^k and p_o .

Then, by induction, we have

$$\begin{aligned} &\|\omega_{mT-1}^k - \omega_{mT-1}^c\| \\ &\leq (1 + \eta\lambda) \|\omega_{mT-2}^k - \omega_{mT-2}^c\| + \eta \mathbf{g}(\omega_{mT-2}^c) \|p_l^k - p_o\|_1 \\ &\leq (1 + \eta\lambda)^2 \|\omega_{mT-3}^k - \omega_{mT-3}^c\| + (1 + \eta\lambda) \eta \mathbf{g}(\omega_{mT-3}^c) \|p_l^k - p_o\|_1 \\ &\quad + \eta \mathbf{g}(\omega_{mT-2}^c) \|p_l^k - p_o\|_1 \\ &\leq (1 + \eta\lambda)^{T-1} \|\omega_{(m-1)T}^k - \omega_{(m-1)T}^c\| \\ &\quad + \eta \sum_{j=2}^T \mathbf{g}(\omega_{mT-j}^c) (1 + \eta\lambda)^{j-2} \|p_l^k - p_o\|_1. \end{aligned} \quad (15)$$

Therefore, we have

$$\begin{aligned} \|\omega_{mT}^f - \omega_{mT}^c\| &\leq \frac{1}{K} \sum_{i=1}^K (1 + \eta\lambda) \|\omega_{mT-1}^k - \omega_{mT-1}^c\| \\ &\leq \frac{1}{K} \sum_{i=1}^K [(1 + \eta\lambda)^T \|\omega_{(m-1)T}^k - \omega_{(m-1)T}^c\| \end{aligned}$$

$$+ \eta \|p_l^k - p_o\|_1 (\eta \sum_{j=2}^T \mathbf{g}(\omega_{mT-j}^c)(1 + \lambda)^{j-1})). \quad (16)$$

A.2 Boundary of $\|\omega_{mT}^c - \omega_{mT}^*\|$

The boundary of $\|\omega_{mT}^c - \omega_{mT}^*\|$ is derived in (17). The derivation process in (17) is similar to the process in (13).

$$\begin{aligned} & \|\omega_{mT}^c - \omega_{mT}^*\| \\ &= \|\omega_{mT-1}^c - \eta \sum_{i=1}^C p_o(y=i) \nabla_{\omega} \mathbb{E}_{x|y=i} [-\log f_i(x, \omega_{mT-1}^c)] \\ & \quad - \omega_{mT-1}^* + \eta \sum_{i=1}^C p_u(y=i) \nabla_{\omega} \mathbb{E}_{x|y=i} [-\log f_i(x, \omega_{mT-1}^*)]\| \\ &\leq \|\omega_{mT-1}^c - \omega_{mT-1}^*\| + \eta \left\| \sum_{i=1}^C p_o(y=i) \nabla_{\omega} \mathbb{E}_{x|y=i} [-\log f_i(x, \omega_{mT-1}^c)] \right. \\ & \quad \left. - \sum_{i=1}^C p_u(y=i) \nabla_{\omega} \mathbb{E}_{x|y=i} [-\log f_i(x, \omega_{mT-1}^*)] \right\| \\ &\leq \|\omega_{mT-1}^c - \omega_{mT-1}^*\| + \eta \left\| \sum_{i=1}^C p_o(y=i) (\nabla_{\omega} \mathbb{E}_{x|y=i} [-\log f_i(x, \omega_{mT-1}^c)] \right. \\ & \quad \left. - \nabla_{\omega} \mathbb{E}_{x|y=i} [-\log f_i(x, \omega_{mT-1}^*)]) \right\| \\ & \quad + \eta \left\| \sum_{i=1}^C (p(y=i) - p_u(y=i)) \nabla_{\omega} \mathbb{E}_{x|y=i} [-\log f_i(x, \omega_{mT-1}^*)] \right\| \end{aligned}$$

$$\begin{aligned} &\leq \|\omega_{mT-1}^c - \omega_{mT-1}^*\| + \eta \sum_{i=1}^C p_o(y=i) \lambda \|\omega_{mT-1}^c - \omega_{mT-1}^*\| \\ & \quad + \eta \mathbf{g}(\omega_{mT-1}^*) \|p_o - p_u\|_1 \\ &= (1 + \eta \lambda) \|\omega_{mT-1}^c - \omega_{mT-1}^*\| + \eta \mathbf{g}(\omega_{mT-1}^*) \|p_o - p_u\|_1 \\ &\leq (1 + \eta \lambda)^T \|\omega_{(m-1)T}^c - \omega_{(m-1)T}^*\| \\ & \quad + \eta \|p_o - p_u\|_1 \left(\sum_{j=1}^T (1 + \eta \lambda)^{j-1} \mathbf{g}(\omega_{mT-j}^*) \right). \quad (17) \end{aligned}$$

According to (10), (16) and (17), we have the boundary of the weight divergence $\|\omega_{mT}^f - \omega_{mT}^*\|$ as shown in (18).

$$\begin{aligned} & \|\omega_{mT}^f - \omega_{mT}^*\| \\ &\leq \frac{1}{K} \sum_{k=1}^K [(1 + \eta \lambda)^T \|\omega_{(m-1)T}^k - \omega_{(m-1)T}^c\| \\ & \quad + \eta \|p_l^k - p_o\|_1 (\eta \sum_{j=2}^T \mathbf{g}(\omega_{mT-j}^c)(1 + \eta \lambda)^{j-1})] \\ & \quad + (1 + \eta \lambda)^T \|\omega_{(m-1)T}^c - \omega_{(m-1)T}^*\| \\ & \quad + \eta \|p_o - p_u\|_1 \left(\sum_{j=1}^T \mathbf{g}(\omega_{mT-j}^*) (1 + \eta \lambda)^{j-1} \right). \quad (18) \end{aligned}$$

Text in blue is newly added.