A   DERIVATION OF WEIGHT DIVERGENCE

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

$$l(\omega) = \mathbb{E}_{\mathbf{x}, y \sim p} \Big[\sum_{i=1}^{C} \mathbb{I}_{y=i} (-\log f_i(\mathbf{x}, \omega)) \Big]$$
$$= \sum_{i=1}^{C} p(y=i) \mathbb{E}_{\mathbf{x}|y=i} [-\log f_i(\mathbf{x}, \omega)]. \tag{9}$$

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.

$$\|\omega_{mT}^f - \omega_{mT}^*\| \leq \|\omega_{mT}^f - \omega_{mT}^c + \omega_{mT}^c - \omega_{mT}^*\|$$
$$\leq \|\omega_{mT}^f - \omega_{mT}^c\| + \|\omega_{mT}^c - \omega_{mT}^*\|. \tag{10}$$

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.

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

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.

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

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 $\|\omega_{mT}^f - \omega_{mT}^c\|$

$$\|\omega_{mT}^f - \omega_{mT}^c\| = \|\frac{1}{K} \sum_{k\sim1}^{K} \omega_{mT}^k - \omega_{mT}^c\|$$

$$= \|\frac{1}{K} \sum_{k=1}^{K} (\omega_{mT-1}^k - \eta \sum_{i=1}^{C} p_l^k(y=i) \nabla_\omega \mathbb{E}_{\mathbf{x}|y=i} [-\log f_i(x, \omega_{mT-1}^k)])$$

$$- (\omega_{mT-1}^c - \eta \sum_{i=1}^{C} p_o(y=i) \nabla_\omega \mathbb{E}_{\mathbf{x}|y=i} [-\log f_i(x, \omega_{mT-1}^c)])\|$$

$$\leq \|\frac{1}{K} \sum_{k=1}^{K} \omega_{mT-1}^k - \omega_{mT-1}^c\| + \eta \|\frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{C} p_l^k(y=i)$$

$$(\nabla_\omega \mathbb{E}_{\mathbf{x}|y=i} [-\log f_i(x, \omega_{mT-1}^k)] - \nabla_\omega \mathbb{E}_{\mathbf{x}|y=i} [-\log f_i(x, \omega_{mT-1}^c)])\|$$

$$\overset{\textcircled{1}}{\leq} \frac{1}{K} \sum_{k=1}^{K} \|\omega_{mT-1}^k - \omega_{mT-1}^c\| + \frac{\eta\lambda}{K} \sum_{k=1}^{K} \sum_{i=1}^{C} p_l^k(y=i) \|\omega_{mT-1}^k - \omega_{mT-1}^c\|$$

$$= \frac{1}{K} \sum_{k=1}^{K} (1+\eta\lambda) \|\omega_{mT-1}^k - \omega_{mT-1}^c\|. \tag{13}$$

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

$$\|\omega_{mT-1}^k - \omega_{mT-1}^c\|$$

$$= \|\omega_{mT-2}^k - \eta \sum_{i=1}^{C} p_l^k(y=i) \nabla_\omega \mathbb{E}_{x|y=i} [-\log f_i(x, \omega_{mT-2}^k)]$$

$$- \omega_{mT-2}^c + \eta \sum_{i=1}^{C} p_o(y=i) \nabla_\omega \mathbb{E}_{x|y=i} [-\log f_i(x, \omega_{mT-2}^c)]\|$$

$$\leq \|\omega_{mT-2}^k - \omega_{mT-2}^c\| + \eta \| \sum_{i=1}^{C} p_l^k(y=i) (\nabla_\omega \mathbb{E}_{x|y=i} [-\log f_i(x, \omega_{mT-2}^k)]$$

$$- \nabla_\omega \mathbb{E}_{x|y=i} [-\log f_i(x, \omega_{mT-2}^c)])\|$$

$$+ \eta \| \sum_{i=1}^{C} (p_l^k(y=i) - p_o(y=i)) \nabla_\omega \mathbb{E}_{x|y=i} [-\log f_i(x, \omega_{mT-2}^c)]\|$$

$$\leq \|\omega_{mT-2}^k - \omega_{mT-2}^c\| + \eta \sum_{i=1}^{C} p_l^k(y=i) \lambda \|\omega_{mT-2}^k - \omega_{mT-2}^c\|$$

$$+ \eta g(\omega_{mT-2}^c) \|p_l^k - p_o\|_1$$

$$= (1+\eta\lambda) \|\omega_{mT-2}^k - \omega_{mT-2}^c\| + \eta g(\omega_{mT-2}^c) \|p_l^k - p_o\|_1 \tag{14}$$

Note that $g(\omega) = \max_{i=1}^{C} \|\nabla_\omega \mathbb{E}_{x|y=i} [-\log f_i(x, \omega)]\|$. 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

$$\|\omega_{mT-1}^k - \omega_{mT-1}^c\|$$

$$\leq (1+\eta\lambda) \|\omega_{mT-2}^k - \omega_{mT-2}^c\| + \eta g(\omega_{mT-2}^c) \|p_l^k - p_o\|_1$$

$$\leq (1+\eta\lambda)^2 \|\omega_{mT-3}^k - \omega_{mT-3}^c\| + (1+\eta\lambda) \eta g(\omega_{mT-3}^c) \|p_l^k - p_o\|_1$$

$$+ \eta g(\omega_{mT-2}^c) \|p_l^k - p_o\|_1 \tag{15}$$

$$\leq (1+\eta\lambda)^{T-1} \|\omega_{(m-1)T}^k - \omega_{(m-1)T}^c\|$$

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

Therefore, we have

$$\|\omega_{mT}^f - \omega_{mT}^c\| \leq \frac{1}{K} \sum_{i=1}^{K} (1+\eta\lambda) \|\omega_{mT-1}^k - \omega_{mT-1}^c\|$$

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

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

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).

$$||\omega_{mT}^c - \omega_{mT}^*||$$

$$= ||\omega_{mT-1}^c - \eta \sum_{i=1}^{C} p_o(y=i) \nabla_\omega \mathbb{E}_{x|y=i}[-\log f_i(x, \omega_{mT-1}^c)]$$

$$- \omega_{mT-1}^* + \eta \sum_{i=1}^{C} p_u(y=i) \nabla_\omega \mathbb{E}_{x|y=i}[-\log f_i(x, \omega_{mT-1}^*)]||$$

$$\leq ||\omega_{mT-1}^c - \omega_{mT-1}^*|| + \eta || \sum_{i=1}^{C} p_o(y=i) \nabla_\omega \mathbb{E}_{x|y=i}[-\log f_i(x, \omega_{mT-1}^c)]$$

$$- \sum_{i=1}^{C} p_u(y=i) \nabla_\omega \mathbb{E}_{x|y=i}[-\log f_i(x, \omega_{mT-1}^*)]||$$

$$\leq ||\omega_{mT-1}^c - \omega_{mT-1}^*|| + \eta || \sum_{i=1}^{C} p_o(y=i) (\nabla_\omega \mathbb{E}_{x|y=i}[-\log f_i(x, \omega_{mT-1}^c)]$$

$$- \nabla_\omega \mathbb{E}_{x|y=i}[-\log f_i(x, \omega_{mT-1}^*)])||$$

$$+ \eta || \sum_{i=1}^{C} (p(y=i) - p_u(y=i)) \nabla_\omega \mathbb{E}_{x|y=i}[-\log f_i(x, \omega_{mT-1}^*)]||$$

$$\leq ||\omega_{mT-1}^c - \omega_{mT-1}^*|| + \eta \sum_{i=1}^{C} p_o(y=i)\lambda ||\omega_{mT-1}^c - \omega_{mT-1}^*||$$

$$+ \eta \mathbf{g}(\omega_{mT-1}^*)||p_o - p_u||_1$$

$$= (1 + \eta\lambda)||\omega_{mT-1}^c - \omega_{mT-1}^*|| + \eta \mathbf{g}(\omega_{mT-1}^*)||p_o - p_u||_1$$

$$\leq (1 + \eta\lambda)^T ||\omega_{(m-1)T}^c - \omega_{(m-1)T}^*||$$

$$+ \eta ||p_o - p_u||_1 (\sum_{j=1}^{T} (1 + \eta\lambda)^{j-1} \mathbf{g}(\omega_{mT-j}^*)). \tag{17}$$

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

$$||\omega_{mT}^f - \omega_{mT}^*||$$

$$\leq \frac{1}{K} \sum_{k=1}^{K} [(1 + \eta\lambda)^T ||\omega_{(m-1)T}^k - \omega_{(m-1)T}^c||$$

$$+ \eta ||p_l^k - p_o||_1 (\sum_{j=2}^{T} \mathbf{g}(\omega_{mT-j}^c)(1 + \eta\lambda)^{j-1})] \tag{18}$$

$$+ (1 + \eta\lambda)^T ||\omega_{(m-1)T}^c - \omega_{(m-1)T}^*||$$

$$+ \eta ||p_o - p_u||_1 (\sum_{j=1}^{T} \mathbf{g}(\omega_{mT-j}^*)(1 + \eta\lambda)^{j-1}).$$