A MATHEMATICAL DEMONSTRATION

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 Equation 9.

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

To bound the divergence between the weights obtained by the FedAVG algorithm ω^f_{mT} and the optimal weights ω^*_{mT} on the test dataset, an intermediate variable ω^c_{mT} is introduced in Equation 10 to assist the proof. The ω^c_{mT} 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.

$$\begin{aligned} ||\omega^f_{mT} - \omega^*_{mT}|| &\leq ||\omega^f_{mT} - \omega^c_{mT} + \omega^c_{mT} - \omega^*_{mT}|| \\ &\leq ||\omega^f_{mT} - \omega^c_{mT}|| + ||\omega^c_{mT} - \omega^*_{mT}||. \end{aligned} \tag{10}$$

An optimization step in local SGD is shown in Equation 11, where p^k_l is the local data distribution of client k and η is the learning rate.

$$\begin{aligned} \omega^k_t &= \omega^k_{t-1} - \eta\nabla_\omega l(\omega) \\ &= \omega^k_{t-1} - \eta \sum_{i=1}^{C} p^k_l(y=i)\nabla_\omega \mathbb{E}_{x|y=i}[-\log f_i(x,\omega^k_{t-1})]. \end{aligned} \tag{11}$$

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

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

We will next derive the boundaries of $||\omega^f_{mT} - \omega^c_{mT}||$ and $||\omega^c_{mT} - \omega^*_{mT}||$ in Section and Section separately.

A.1 Boundary of $||\omega^f_{mT} - \omega^c_{mT}||$

$$\begin{aligned} &||\omega^f_{mT} - \omega^c_{mT}|| = ||\frac{1}{K}\sum_{k=1}^{K}\omega^k_{mT} - \omega^c_{mT}|| \\ =&||\frac{1}{K}\sum_{k=1}^{K}(\omega^k_{mT-1} - \eta\sum_{i=1}^{C}p^k_l(y=i)\nabla_\omega\mathbb{E}_{x|y=i}[-\log f_i(x,\omega^k_{mT-1})]) \\ &-(\omega^c_{mT-1} - \eta\sum_{i=1}^{C}p_o(y=i)\nabla_\omega\mathbb{E}_{x|y=i}[-\log f_i(x,\omega^c_{mT-1})])|| \\ \leq&||\frac{1}{K}\sum_{k=1}^{K}\omega^k_{mT-1} - \omega^c_{mT-1}|| + \eta||\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{C}p^k_l(y=i) \\ &(\nabla_\omega\mathbb{E}_{x|y=i}[-\log f_i(x,\omega^k_{mT-1})] - \nabla_\omega\mathbb{E}_{x|y=i}[-\log f_i(x,\omega^c_{mT-1})])|| \\ \overset{(1)}{\leq}&\frac{1}{K}\sum_{k=1}^{K}||\omega^k_{mT-1}-\omega^c_{mT-1}|| + \frac{\eta\lambda}{K}\sum_{k=1}^{K}\sum_{i=1}^{C}p^k_l(y=i)||\omega^k_{mT-1}-\omega^c_{mT-1}|| \\ =&\frac{1}{K}\sum_{k=1}^{K}(1+\eta\lambda)||\omega^k_{mT-1} - \omega^c_{mT-1}|| \end{aligned} \tag{13}$$

The inequality (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

$$\begin{aligned} &||\omega^k_{mT-1} - \omega^c_{mT-1}|| \\ =&||\omega^k_{mT-2} - \eta\sum_{i=1}^{C}p^k_l(y=i)\nabla_\omega\mathbb{E}_{x|y=i}[-\log f_i(x,\omega^k_{mT-2})] \\ &-\omega^c_{mT-2} + \eta\sum_{i=1}^{C}p_o(y=i)\nabla_\omega\mathbb{E}_{x|y=i}[-\log f_i(x,\omega^c_{mT-2})]|| \\ \leq&||\omega^k_{mT-2} - \omega^c_{mT-2}|| + \eta||\sum_{i=1}^{C}p^k_l(y=i)(\nabla_\omega\mathbb{E}_{x|y=i}[-\log f_i(x,\omega^k_{mT-2})] \\ &-\nabla_\omega\mathbb{E}_{x|y=i}[-\log f_i(x,\omega^c_{mT-2})])|| \\ &+\eta||\sum_{i=1}^{C}(p^k_l(y=i) - p_o(y=i))\nabla_\omega\mathbb{E}_{x|y=i}[-\log f_i(x,\omega^c_{mT-2})]|| \\ \leq&||\omega^k_{mT-2} - \omega^c_{mT-2}|| + \eta\sum_{i=1}^{C}p^k_l(y=i)\lambda||\omega^k_{mT-2} - \omega^c_{mT-2}|| \\ &+\eta\mathrm{g}(\omega^c_{mT-2})||p^k_l - p_o||_1 \\ =&(1+\eta\lambda)||\omega^k_{mT-2} - \omega^c_{mT-2}|| + \eta\mathrm{g}(\omega^c_{mT-2})||p^k_l - p_o||_1 \end{aligned} \tag{14}$$

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

Then, by induction, we have

$$\begin{aligned} &||\omega^k_{mT-1} - \omega^c_{mT-1}|| \\ \leq&(1+\eta\lambda)||\omega^k_{mT-2} - \omega^c_{mT-2}|| + \eta\mathrm{g}(\omega^c_{mT-2})||p^k_l - p_o||_1 \\ \leq&(1+\eta\lambda)^2||\omega^k_{mT-3} - \omega^c_{mT-3}|| + (1+\eta\lambda)\eta\mathrm{g}(\omega^c_{mT-3})||p^k_l - p_o||_1 \\ &+\eta\mathrm{g}(\omega^c_{mT-2})||p^k_l - p_o||_1 \\ \leq&(1+\eta\lambda)^{T-1}||\omega^k_{(m-1)T} - \omega^c_{(m-1)T}|| \\ &+\eta\sum_{j=2}^{T}\mathrm{g}(\omega^c_{mT-j})(1+\eta\lambda)^{j-2}||p^k_l - p_o||_{11} \end{aligned} \tag{15}$$

Therefore, we have

$$\begin{aligned} ||\omega^f_{mT} - \omega^c_{mT}|| &\leq \frac{1}{K}\sum_{i=1}^{K}(1+\eta\lambda)||\omega^k_{mT-1} - \omega^c_{mT-1}|| \\ &\leq\frac{1}{K}\sum_{i=1}^{K}[(1+\eta\lambda)^T||\omega^k_{(m-1)T} - \omega^c_{(m-1)T}|| \end{aligned}$$

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

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

The boundary of $||\omega_{mT}^c - \omega_{mT}^*||$ is derived in Equation 17, with the idea in Equation 13.

$$\begin{aligned}
&\|\omega_{mT}^c - \omega_{mT}^*\| \\
=&\|\omega_{mT-1}^c - \eta\sum_{i=1}^{C}p_o(y=i)\nabla_\omega\mathbb{E}_{x|y=i}[-\log f_i(x,\omega_{mT-1}^c)] \\
&-\omega_{mT-1}^* + \eta\sum_{i=1}^{C}p_u(y=i)\nabla_\omega\mathbb{E}_{x|y=i}[-\log f_i(x,\omega_{mT-1}^*)]\| \\
\leq&\|\omega_{mT-1}^c - \omega_{mT-1}^*\| + \eta\|\sum_{i=1}^{C}p_o(y=i)\nabla_\omega\mathbb{E}_{x|y=i}[-\log f_i(x,\omega_{mT-1}^c)] \\
&-\sum_{i=1}^{C}p_u(y=i)\nabla_\omega\mathbb{E}_{x|y=i}[-\log f_i(x,\omega_{mT-1}^*)]\| \\
\leq&\|\omega_{mT-1}^c - \omega_{mT-1}^*\| + \eta\|\sum_{i=1}^{C}p_o(y=i)(\nabla_\omega\mathbb{E}_{x|y=i}[-\log f_i(x,\omega_{mT-1}^c)] \\
&-\nabla_\omega\mathbb{E}_{x|y=i}[-\log f_i(x,\omega_{mT-1}^*)])\| \\
&+\eta\|\sum_{i=1}^{C}(p(y=i)-p_u(y=i))\nabla_\omega\mathbb{E}_{x|y=i}[-\log f_i(x,\omega_{mT-1}^*)]\| \\
\leq&\|\omega_{mT-1}^c - \omega_{mT-1}^*\| + \eta\sum_{i=1}^{C}p_o(y=i)\lambda\|\omega_{mT-1}^c - \omega_{mT-1}^*\| \\
&+\eta\mathrm{g}(\omega_{mT-1}^*)||p_o - p_u||_1 \\
=&(1+\eta\lambda)\|\omega_{mT-1}^c - \omega_{mT-1}^*\| + \eta\mathrm{g}(\omega_{mT-1}^*)||p_o - p_u||_1 \\
\leq&(1+\eta\lambda)^T\|\omega_{(m-1)T}^c - \omega_{(m-1)T}^*\| \\
&+\eta||p_o - p_u||_1(\sum_{j=1}^{T}(1+\eta\lambda)^{j-1}\mathrm{g}(\omega_{mT-j}^*))
\end{aligned} \tag{17}$$