# FED-3DA: A DYNAMIC AND PERSONALIZED FEDERATED LEARNING FRAMEWORK APPENDIX

*Name of author*

Address - Line 1
Address - Line 2
Address - Line 3

## 1. PROOF OF THEOREM 1

**Theorem 1.** *Let the space of the meta-network parameter and clients' distribution representation be bounded in a minimum enclosing ball of radius $R_e$, and the dimension of them are $N$ and $E$. Let $Lip_l$, $Lip_\varphi$ and $Lip_r$ be the Lipschitz constant of functions $\mathcal{L}_c^p(x,y)$, $f(*;\varphi)$ and $f(r;*)$. For all p, ε, δ with $p \in [P]$ and 0 <ε, δ< 1, if the period size $S$ satisfies $S \geq \mathbb{O}(\frac{CE+N}{C\varepsilon^2}(\log R_e Lip_l(Lip_\varphi + Lip_r) - \log(\varepsilon\delta)))$, we have with probability at least 1-δ, any $\varphi$, $r$ will satisfy $|\widehat{ER}(\varphi,r) - ER(\varphi,r)| \leq \varepsilon$.*

*Proof.* **First, we analyze the meta-network from the perspective of the $p^{th}$ period**. From the assumptions of the equation (3), we have: $\theta^p = f(r^p;\varphi^p)$, $\theta^p = \{\theta_c^p\}_{c=1}^C$, $r^p = \{r_c^p\}_{c=1}^C$. The loss of the client $c$ in the $p^{th}$ period defined as:

$$\mathcal{L}_c^p(x_c^p, y_c^p;\ \theta_c^p) = \mathcal{L}_c^p(x_c^p, y_c^p;\ f(r_c^p;\ \varphi^p)).$$

The average loss for all clients in the $p^{th}$ period defined as:

$$\mathcal{L}^p(x^p, y^p;\ \theta^p) = \frac{1}{C}\sum_{c=1}^{C} \mathcal{L}_c^p(x_c^p, y_c^p;\ f(r_c^p;\ \varphi^p)),$$

where $C$ represents the total client number. We follow the assumptions in section 4.2 of the previous study [1], the space of the *meta-network* parameter and clients' distribution representation are bounded in a minimum enclosing ball of radius $R_e$, so the following *Lipschitz* conditions hold in the $p^{th}$ period:

$$\begin{aligned}
|f(r^p;\ \varphi^p) - f(r^p;\ \tilde{\varphi^p})| &\leq Lip_\varphi^p||\varphi^p - \tilde{\varphi^p}|| \\
|f(r^p;\ \varphi^p) - f(\tilde{r^p};\ \varphi^p)| &\leq Lip_r^p||r^p - \tilde{r^p}|| \\
|\mathcal{L}_c^p(x_c^p, y_c^p;\ \theta_c^p) - \mathcal{L}_c^p(x_c^p, y_c^p;\ \tilde{\theta_c^p})| &\leq Lip_l^p||\theta_c^p - \tilde{\theta_c^p}||
\end{aligned}$$

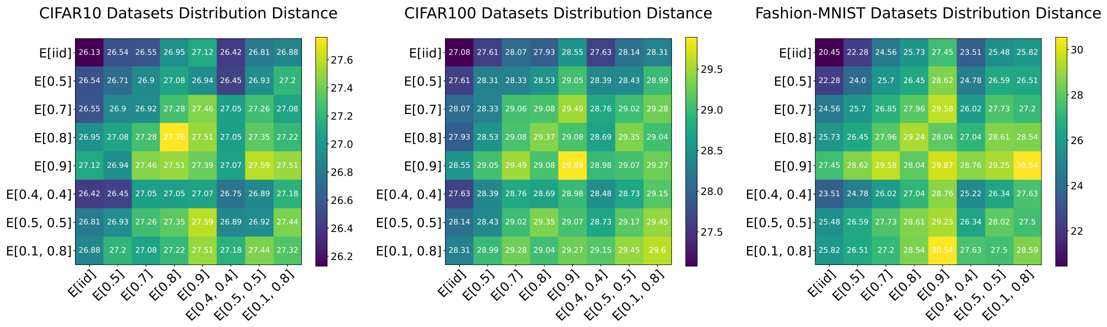
From the definition 4 and theorem 4 in the sections 2.3 and 2.6 of [2], for multi-task learning, in order to minimize the average generalization error, the sample size $s$ of a single task needs to satisfy: $s \geq \mathbb{O}\left(\frac{1}{n\varepsilon^2}\log\frac{\mathcal{C}(\varepsilon, \mathbb{H}_{\mathcal{L}}^C)}{\delta}\right)$, where $C$ represents task number, which equivalent to the client number in our Fed-3DA, so we reuse the letter $C$. $\mathcal{C}\left(\varepsilon, \mathbb{H}_{\mathcal{L}}^C\right)$ is the covering number for the permissible hypothesis space family $\mathbb{H}_{\mathcal{L}}^C$. In Fed-3DA, every hypothesis of $\mathbb{H}_{\mathcal{L}}^C$ is parameterized by $[r_1^p, ..., r_C^p;\ \varphi^p]$, and the hypothesis distance from [2] defined as:

$$\begin{aligned}
& d\left((r_1^p, ..., r_C^p;\ \varphi^p),\ (\tilde{r_1^p}, ..., \tilde{r_C^p};\ \tilde{\varphi^p})\right) \\
&= \mathop{\mathbb{E}}_{x_c^p, y_c^p \sim \mathcal{P}_c^p}\left[\frac{1}{C}\left|\sum_{c=1}^{C}\mathcal{L}_c^p\left(x_c^p, y_c^p;\ f\left(r_c^p;\ \varphi^p\right)\right) - \right.\right. \\
&\qquad \left.\left.\sum_{c=1}^{C}\mathcal{L}_c^p\left(x_c^p, y_c^p;\ f\left(\tilde{r_c^p};\ \tilde{\varphi^p}\right)\right)\right|\right] \\
&= \frac{1}{C}\sum_{c=1}^{C}\mathop{\mathbb{E}}_{x_c^p, y_c^p \sim \mathcal{P}_c^p}\left[\left|\mathcal{L}_c^p\left(x_c^p, y_c^p;\ f\left(r_c^p;\ \varphi^p\right)\right) - \right.\right. \\
&\qquad \left.\left.\mathcal{L}_c^p\left(x_c^p, y_c^p;\ f\left(\tilde{r_c^p};\ \tilde{\varphi^p}\right)\right)\right|\right] \\
&= \frac{1}{C}\sum_{c=1}^{C}\mathop{\mathbb{E}}_{x_c^p, y_c^p \sim \mathcal{P}_c^p}\left[\left|\mathcal{L}_c^p\left(x_c^p, y_c^p;\ \theta^p\right) - \mathcal{L}_c^p\left(x_c^p, y_c^p;\ \tilde{\theta^p}\right)\right|\right]
\end{aligned}$$

From the above *Lipschitz* inequalities, we have:

$$\begin{aligned}
& d\left((r_1^p, ..., r_C^p;\ \varphi^p),\ (\tilde{r_1^p}, ..., \tilde{r_C^p};\ \tilde{\varphi^p})\right) \\
&\leq Lip_l^p||\theta^p - \tilde{\theta^p}|| \\
&\leq Lip_l^p||f(r^p;\ \varphi^p) - f(\tilde{r^p};\ \tilde{\varphi^p})|| \\
&\leq Lip_l^p||f(r^p;\ \varphi^p) - f(r^p;\ \tilde{\varphi^p})|| + \\
&\quad Lip_l^p||f(r^p;\ \tilde{\varphi^p}) - f(\tilde{r^p};\ \tilde{\varphi^p})|| \\
&\leq Lip_l^p \cdot Lip_\varphi^p||\varphi^p - \tilde{\varphi^p}|| + Lip_l^p \cdot Lip_r^p||r^p - \tilde{r^p}||
\end{aligned}$$

Combined with the proof A in [1], the above results imply that if we want an $\varepsilon$-covering in the hypothesis distance $d(\cdot,\cdot)$, we need to select a parameter space in which the distance of pairs $(\varphi^p, \tilde{\varphi^p})$ and $(r^p, \tilde{r^p})$ are $\frac{\varepsilon}{2Lip_l^p(Lip_\varphi^p + Lip_r^p)}$, meanwhile,



CIFAR10 Datasets Distribution Distance

CIFAR100 Datasets Distribution Distance

Fashion-MNIST Datasets Distribution Distance

**Fig. 1**. *Distribution distance between the different types of datasets*. *CIFAR10*, *CIFAR100* and *Fashion-MNIST* datasets distribution distance based on the optimal transport strategy (**only with the feature, without the label**, *dataset size*=1000).

$\log\left(\mathcal{C}\left(\varepsilon, \mathbb{H}_{\mathcal{L}}^{C}\right)\right) = \mathbb{O}\left(\frac{CE+N}{\varepsilon}(\log R_e Lip_l^p(Lip_\varphi^p + Lip_r^p) - \log(\varepsilon\delta))\right)$.
So far, for any period $p$, after the condition in **theorem 1** is satisfied, we have:

$$\{|\widehat{ER}^p(\varphi^p, r^p) - ER^p(\varphi^p, r^p)| \leq \varepsilon^p\}_{p=1}^P,\ 0 < \{\varepsilon^p\}_{p=1}^P < 1$$

where $\widehat{ER}^p(\varphi^p, r^p)$ and $ER^p(\varphi^p, r^p)$ represent the empirical and expected loss in the $p^{th}$ period, respectively. In the sequence $\{\varepsilon^p\}_{p=1}^P$, we assume that: $\{\varepsilon^p \leq \varepsilon^* \mid p, * \in [1...P]\}$. **In all** $P$ **periods**, the difference between the empirical loss $\widehat{ER}(\varphi, r)$ and expected loss $ER(\varphi, r)$:

$$\begin{aligned}
&|\widehat{ER}(\varphi, r) - ER(\varphi, r)| \\
=&|\frac{1}{P}\sum_{p=1}^{P}\widehat{ER}^p(\varphi^p, r^p) - \frac{1}{P}\sum_{p=1}^{P}ER^p(\varphi^p, r^p)| \\
=&\frac{1}{P}\sum_{p=1}^{P}|\widehat{ER}^p(\varphi^p, r^p) - ER^p(\varphi^p, r^p)| \\
\leq&\frac{\varepsilon^1 + ... + \varepsilon^P}{P} \leq \frac{P\varepsilon^*}{P} = \varepsilon^* \in (0,\ 1)
\end{aligned}$$

This completes the proof. □

## 2. DISTRIBUTION DISTANCE BETWEEN DIFFERENT DATASETS

The distribution distance between different datasets is shown in **figure 1**. Taking CIFAR10 as an example, among the same distribution type of the datasets, the *distance*($\mathcal{P}_{[iid]}$, $\mathcal{P}_{[iid]}$)=26.13 and *distance*($\mathcal{P}_{[0.9]}$, $\mathcal{P}_{[0.9]}$)=27.51 obtain the minimum and maximum, and the distance increases with the increase of the main class proportion. So does the distance among the one main class datasets of the different types, the *distance*($\mathcal{P}_{[0.8]}$, $\mathcal{P}_{[0.9]}$)=27.76 obtain the maximum. Between the same. type of the two main classes datasets, the distribution distance increases with the increase of the sum of the main class proportion, the *distance*($\mathcal{P}_{[0.1,0.8]}$, $\mathcal{P}_{[0.1,0.8]}$)=27.32 obtain the maximum. The distribution distance behaves similarly on the CIFAR100 and Fashion-MNIST.

## 3. REFERENCES

[1] Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik, "Personalized federated learning using hypernetworks," *arXiv preprint arXiv:2103.04628*, 2021.

[2] Jonathan Baxter, "A model of inductive bias learning," *Journal of artificial intelligence research*, vol. 12, pp. 149–198, 2000.