Appendix Of Dynamic Personalized Federated Learning Based On Distribution Distance Measurement

Anonymous

Anonymous

1 Proof Of Theorem 1

Theorem 1. *Let the space of the meta-network parameter and clients' distribution representation are bounded in a minimum enclosing ball of radius R_e, and the dimension of them are N and E, respectively. Let the Lip_l, Lip_φ and Lip_r are the Lipschitz constant of the functions $L_c^p(x,y)$, $f(*;\varphi)$ and $f(r;*)$. Under the above stated assumptions, For all p, ε, δ with $p \in [1...P]$ and $0<\varepsilon$, $\delta<1$, if the period size S(all the C clients use the same period size in the Fed-3DA) satisfies: $T \triangleq \{S \geq \mathbb{O}(\frac{CE+N}{C\varepsilon^2}(\log R_e Lip_l(Lip_\varphi + Lip_r) - \log(\varepsilon\delta)))\}$, we have with probability at least 1-δ, any φ, r will satisfy $|\widehat{ER}(\varphi, r) - ER(\varphi, r)| \leq \varepsilon$.*

First, we analyze the meta-network from the perspective of p period. From the assumptions of the equation (4) in this paper, 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 is(C represents total client number):

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

We follow the assumptions in the section 4.2 of the previous study [3], 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:

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

From the definition 4 and theorem 4 in the sections 2.3 and 2.6 of [1], for multi-task learning, in order to minimize the average generalization error, the sample size s of a single task needs to satisfy:

$$s \geq \mathbb{O}\left(\frac{1}{n\varepsilon^2}\log\frac{\mathcal{C}\left(\varepsilon, \mathbb{H}_{\mathcal{L}}^{C}\right)}{\delta}\right)$$

where C represents task number, which equivalent to Fed-3DA client number, 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 our 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 [1] defined as:

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

From the above Lipschitz inequalities, we have:

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

Combined with the proof A in [3], the above results imply that if we want an ε-covering in the hypothesis distance d(*, *), 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, $\log\left(\mathcal{C}\left(\varepsilon, \mathbb{H}_{\mathcal{L}}^{C}\right)\right) = \mathbb{O}\left(\frac{CE+N}{\varepsilon}(\log R_e Lip_l^p(Lip_\varphi^p + Lip_r^p) - \log(\varepsilon\delta))\right)$.

So far, for any period p, after the condition T 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 empirical and expected loss: $\widehat{ER}(\varphi, r) = \frac{1}{P}\sum_{p=1}^{P} \widehat{ER}^{P}(\varphi^p, r^p)$, $ER(\varphi, r) = \frac{1}{P}\sum_{p=1}^{P} ER^p(\varphi^p, r^p)$, we have:

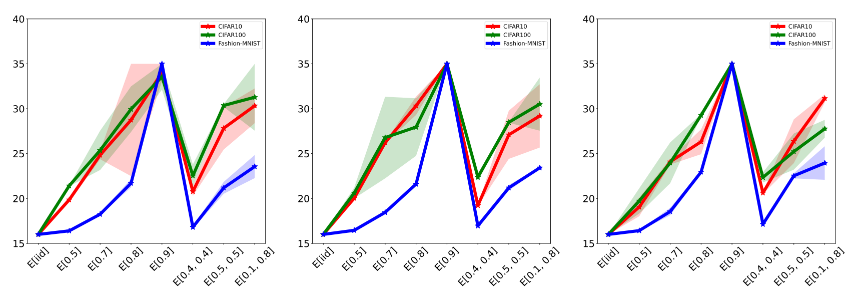
$$
\begin{aligned}
|\widehat{ER}(\varphi, r) - ER(\varphi, r)| &= |\frac{1}{P}\sum_{p=1}^{P} \widehat{ER}^{p}(\varphi^p, r^p) - \frac{1}{P}\sum_{p=1}^{P} ER^p(\varphi^p, r^p)| \\
&= \frac{1}{P}\sum_{p=1}^{P} |\widehat{ER^p}(\varphi^p, r^p) - ER^p(\varphi^p, r^p)| \\
&\le \frac{\varepsilon^1 + ... + \varepsilon^P}{P}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

This completes the proof.

2 Additional Experiments

2.1 Distribution Distance Between The Dataset Ruler And Dataset

The distribution distance between the *Dataset Ruler* and the datasets is shown in **figure 1**. In the single main class dataset, the *distance*$(DR, \mathbb{E}_{[iid]})$ and *distance*$(DR, \mathbb{E}_{[0.9]})$ obtain the minimum and maximum, the *distance* is positively correlated with the proportion of the main class. So does the two main class dataset. The dataset size has a slight effect on the trend of the distribution distance. Based on the above experimental results, we believe that the *Dataset Ruler* can be used as a uniform measure of the distribution distance under the different dataset type conditions.

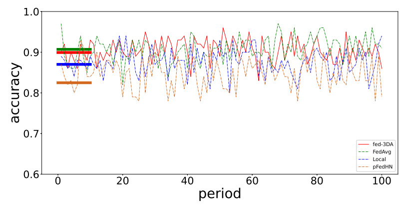


(a) dataset size=500 (b) dataset size=1000 (c) dataset size=2000

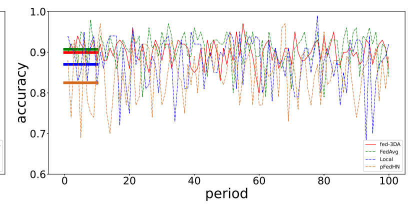
Fig. 1. *Distribution distance between the Dataset Ruler and dataset.* The *DR*s of *CIFAR10*, *CIFAR100* and *Fashion-MNIST* are all sampled from the *ImageNet* [2], while the samples are cropped to keep the size consistent and grayed when compared with the *Fashion-MNIST*.

2.2 Model Adaptability Under The Dynamic Distribution Data

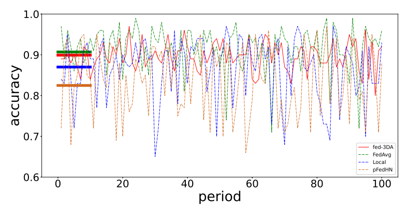
This section supplements the scenarios of the experiments in the main text, and the conclusions are consistent with the main text.



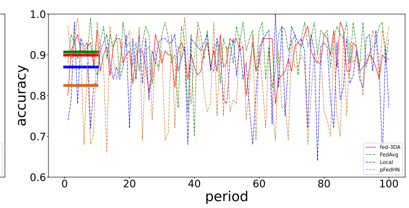
(a) $\mathbb{E}_{[iid]}$



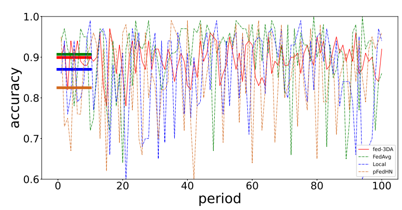
(b) $\mathbb{E}_{[0.5]}$



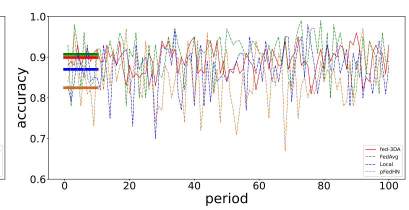
(c) $\mathbb{E}_{[0.7]}$



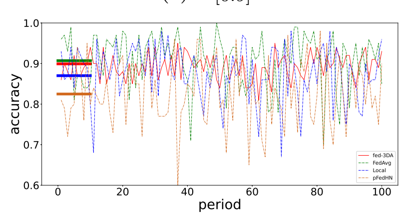
(d) $\mathbb{E}_{[0.8]}$



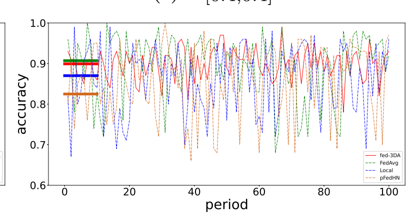
(e) $\mathbb{E}_{[0.9]}$



(f) $\mathbb{E}_{[0.4,0.4]}$



(g) $\mathbb{E}_{[0.5,0.5]}$



(h) $\mathbb{E}_{[0.1,0.8]}$

Fig. 2. *Federated client test accuracy on the Fashion-MNIST under the dynamic distribution*. Every approach test for 100 periods with 10 clients, the validation accuracy(*FedAvg*: 90.7%, *Fed-3DA*: 89.9%, *Local*: 87.1%, *pFedHN*: 82.5%) is marked with a short line.

References

1. Baxter, J.: A model of inductive bias learning. Journal of artificial intelligence research **12**, 149–198 (2000)

2. Deng, J.: A large-scale hierarchical image database. Proc. of IEEE Computer Vision and Pattern Recognition, 2009 (2009)
3. Shamsian, A., Navon, A., Fetaya, E., Chechik, G.: Personalized federated learning using hypernetworks. arXiv preprint arXiv:2103.04628 (2021)