FEDRF-ADAPT: ROBUST AND COMMUNICATION-EFFICIENT FEDERATED DOMAIN ADAPTATION VIA RANDOM FEATURES

Yuanjie Wang[1], *Zhanbo Feng*[2], *Zhenyu Liao*[1]

[1] Huazhong University of Science and Technology, Wuhan, China
[2] Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Federated domain adaptation (FDA) allows one to train large-scale machine learning models over networked systems that adapt to novel target domains. Existing FDA methods suffer from an excessive communication overhead in aligning feature distributions between source and target domains. In this paper, we propose FedRF-Adapt, a communication-efficient FDA protocol that enjoys a *sample-size-independent* communicational complexity and is robust to limited network reliability. Extensive numerical experiments are provided to support the advantageous performance of FedRF-Adapt.

***Index Terms*—** Random features, maximum mean discrepancy, federated domain adaptation

1. INTRODUCTION

Federated learning, by performing local computation and between-user information exchange, allows to train large-scale machine learning (ML) models in a collaborative manner, without sharing the private data of end-users. Despite its rapid growth within the ML community, the practical use of federated learning is limited by its poor generalization performance in the presence of *domain shift* [1], when, e.g., a novel end-user is present in the network. In the respect, domain adaptation (DA) appears as a compelling technique to learn to "align" features (by minimizing their Maximum Mean Discrepancy, MMD, distance [2,3] say) from source and target domains in a common space. Downstream tasks such as classification and regression can then be further performed on these aligned features.

In this paper, we focus on the federate domain adaptation (FDA) approach that exploits DA to resolve the issue of domain shift in federated ML model training. Existing FDA protocols either rely on feature alignment [4] or adversarial learning [1], that both necessitate extensive information (e.g., features of source and target data and/or model parameters). This generally leads to a huge communication overhead and quickly becomes burdensome as the number of users and/or training samples increase. To address the computational and communicational challenges inherent in FDA, the FedKA method was introduced in [5]. This method leverages a less computationally intensive feature extractor to compute the MMD distance between source and target features. Building upon this idea, the authors in [6] further refined the approach by approximating the MMD distance instead of computing it exactly. Despite these advancements, a notable communication overhead persists, particularly due to the necessity of exchanging features or gradient information. This communicational complexity in general grows rapidly with the size of training samples.

In this paper, we propose Federated Random Features-based Adaptation (FedRF-Adapt), a novel FDA scheme with significantly less communication overhead (that is almost *independent* of the sample size), strong robustness against (the possibly limited) network reliability, and added privacy protection. FedRF-Adapt achieves these by leveraging the efficient RF-MMD method (to be discussed in details in Section 2.1 below) that compresses exchanged messages via low-rank approximation and randomization technique. In comparison to existing MMD-based FDA protocols, FedRF-Adapt offers significant reductions in both communication and computation complexity, while maintaining commendable performance.

Notations. We denote scalars by lowercase letters, vectors by bold lowercase, and matrices by bold uppercase. We denote the transpose operator by $(\cdot)^\mathsf{T}$, and use $\|\cdot\|_2$ to denote the Euclidean norm for vectors and spectral/operator norm for matrices. For a random variable z, $\mathbb{E}[z]$ denotes the expectation of z. We use $\mathbf{1}_p$ and $\mathbf{I}_p$ for the vector of all ones of dimension p and the identity matrix of dimension $p \times p$, respectively. We use Θ, O and Ω notations as in classical computer science literature [7, 8].

2. SETUP AND OUR APPROACH

The Maximum Mean Discrepancy (MMD) was first proposed as a test statistic in [2, 3] to assess whether data points are drawn in a i.i.d. fashion from the same distribution, by evaluating their features in a Reproducing Kernel Hilbert Space (RKHS). It has then gained wide popularity as the preferred optimization metric to align distinct feature distributions across diverse domains, of direct use in DA [9, 10]. Given source $\mathbf{X}_S$ and target dataset $\mathbf{X}_T$, their MMD distance can be empirically estimated as

$$\begin{aligned}\mathsf{MMD}(\mathbf{X}_S, \mathbf{X}_T) &= \left\| \frac{1}{n_S} \sum_{i=1}^{n_S} \phi(\mathbf{x}_i) - \frac{1}{n_T} \sum_{j=1}^{n_T} \phi(\mathbf{x}_j) \right\|_{\mathcal{H}}^2 \\ &= \boldsymbol{\ell}^\mathsf{T} \mathbf{K} \boldsymbol{\ell}, \end{aligned} \tag{1}$$

where label vector $\boldsymbol{\ell} \in \mathbb{R}^n$ with its ith entry given by

$$\ell_i = \frac{1}{n_S} 1_{\mathbf{x}_i \in \mathbf{X}_S} - \frac{1}{n_T} 1_{\mathbf{x}_i \in \mathbf{X}_T}, \tag{2}$$

by lifting the source $\mathbf{x}_i \in \mathbf{X}_S$ and target $\mathbf{x}_j \in \mathbf{X}_T$ data to some *predefined* RKHS $\mathcal{H}$ via the kernel trick $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}} = K(\mathbf{x}_i, \mathbf{x}_j)$ [11], to form the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$, with $n = n_S + n_T$.

This work is supported by the National Natural Science Foundation of China (via NSFC-62206101 and NSFC-12141107), the Fundamental Research Funds for the Central Universities of China (2021XXJS110), the Interdisciplinary Research Program of HUST (2023JCYJ012), and the Key Research and Development Program of Guangxi (GuiKe-AB21196034). Corresponding author: Zhenyu Liao (email: zhenyu_liao@hust.edu.cn).

979-8-3503-7451-3/24/$31.00 ©2024 IEEE

Note, however, from its definition in (1) that the computation of MMD necessitates access to data from both source and target domains. This, in the case of multi-source FDA, results in substantial communication costs between clients.

In the following, we propose a computationally efficient approach to MMD distance based on random features technique.

2.1. Our Approach: Random Features-based MMD

In the following, we introduce RF-MMD, a random features-based approach to computationally efficient MMD. Here, we focus on random Fourier features (RFFs) and Gaussian kernel. The same idea applies to other shift-invariant kernels (such as the Laplacian and Cauchy kernels), see [12]. We refer the readers to [13] for a review.

Definition 1 (Random Fourier features, [12])**.** *For data matrix* $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ *of size n, the random Fourier feature (RFF) matrix* $\Sigma \in \mathbb{R}^{2N \times n}$ *of* $\mathbf{X}$ *is given by*

$$\Sigma = \frac{1}{\sqrt{N}} \begin{bmatrix} \cos(\Omega \mathbf{X}) \\ \sin(\Omega \mathbf{X}) \end{bmatrix} \in \mathbb{R}^{2N \times n}, \tag{3}$$

with N the number of random features, $\Omega \in \mathbb{R}^{N \times p}$ *a random matrix having i.i.d. standard Gaussian entries with mean zero and variance* $1/\sigma^2$*, i.e.,* $[\Omega]_{ij} \sim \mathcal{N}(0, 1/\sigma^2)$*, and cosine and sine functions* $\cos(\cdot), \sin(\cdot)$ *applied entry-wise on* $\Omega\mathbf{X}$*.*

Denote $\dim(\mathbf{K}) \equiv \operatorname{tr} \mathbf{K} / \|\mathbf{K}\|_2$ the intrinsic dimension of the Gaussian kernel matrix $\mathbf{K} = \mathbf{K}_{\text{Gauss}}$, it is known that an order of $N = \Theta(\dim(\mathbf{K}) \log(n))$ RFFs suffices to well approximate the kernel matrix $\mathbf{K}$ in a spectral norm sense, as given in the following result.

Theorem 1 (RFFs approximation of Gaussian kernels, [14, Section 6.5])**.** *For random Fourier features* $\Sigma \in \mathbb{R}^{2N \times n}$ *of data* $\mathbf{X} \in \mathbb{R}^{p \times n}$ *as defined in Definition 1, one has*

$$\mathbb{E}\|\Sigma^{\mathsf{T}}\Sigma - \mathbf{K}\|_2 \leq C \left(\sqrt{\frac{n \log(n)}{N}} \|\mathbf{K}\|_2 + \frac{n \log(n)}{N} \right),$$

holds for some universal constant $C > 0$ *independent of N and n, with* $\mathbf{K} = \mathbf{K}_{\text{Gauss}} = \{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2\sigma^2))\}_{i,j=1}^n$ *the Gaussian kernel matrix of* $\mathbf{X}$*.*

As a direct consequence of Theorem 1, it can be shown that an order of $\Theta(\log(n))$ random features suffice to well approximate the MMD distance in (1), as in the following result.

Corollary 1 (RFFs approximation of MMD distance)**.** *Let*

$$\mathsf{RF\text{-}MMD}(\mathbf{X}_S, \mathbf{X}_T) = \ell^{\mathsf{T}}\Sigma^{\mathsf{T}}\Sigma\ell = \|\Sigma\ell\|_2^2 \tag{4}$$

denote the approximated MMD distance between source $\mathbf{X}_S$ *and target data* $\mathbf{X}_T$ *using RFFs with* $\mathbf{X} = [\mathbf{X}_S, \mathbf{X}_T]$ *as in Definition 1 and* $\ell \in \mathbb{R}^n$ *defined in (2) with* $n = n_S + n_T$*. Then, we have, for* MMD *distance defined in (1) with* $n_s, n_T = \Theta(n)$ *that*

$$\mathbb{E}[|\mathsf{RF\text{-}MMD}(\mathbf{X}_S, \mathbf{X}_T) - \mathsf{MMD}(\mathbf{X}_S, \mathbf{X}_T)|] \leq \varepsilon \tag{5}$$

holds for $N \geq C \log(n)/(\dim(\mathbf{K})\varepsilon^2)$ *with some constant* $C > 0$ *independent of n.*

Proof of Corollary 1. Given a desired error $\varepsilon \in (0,1)$, taking $N \geq C' \dim(\mathbf{K}) \log(n)/\varepsilon^2$ in Theorem 1 for some universal constant $C' > 0$ independent of n, with $\dim(\mathbf{K}) \equiv \operatorname{tr}\mathbf{K}/\|\mathbf{K}\|_2 = n/\|\mathbf{K}\|_2$ the *intrinsic dimension* of Gaussian kernel matrix $\mathbf{K} = \mathbf{K}_{\text{Gauss}}$, we have that the *expected* error satisfies $\mathbb{E}\|\Sigma^{\mathsf{T}}\Sigma - \mathbf{K}\|_2 \leq \varepsilon\|\mathbf{K}\|_2$. Further note that $\mathbb{E}[|\mathsf{RF\text{-}MMD} - \mathsf{MMD}|] \leq \|\ell\|^2 \cdot \mathbb{E}\|\Sigma^{\mathsf{T}}\Sigma - \mathbf{K}\|_2 \leq \frac{n\varepsilon}{n_S n_T}\|\mathbf{K}\|_2$, a change of variable in ε allows us to conclude the proof of Corollary 1. □

Corollary 1 tells us that a number of $N = \Theta(\log(n))$ random features are sufficient to well approximate the MMD distance in (1). More importantly, the computation of RF-MMD requires *only* the matrix-vector product $\Sigma\ell \in \mathbb{R}^{2N}$ instead of quadratic form in (1) of the original kernel matrix $\mathbf{K}$ of size n by n. This, as we shall see in Section 3, leads to a significant reduction in the communication complexity of FDA.

3. MAIN RESULTS

In this section, we extend the RF-MMD approach in Corollary 1 to a multi-source FDA scenario, and introduce the FedRF-Adapt protocol that offers a significant reduction in the FDA communication overhead, and strong robustness to network condition.

3.1. FedRF-Adapt: multi-source FDA via RF-MMD

Consider the following multi-source FDA classification problem: for K source domains $\mathcal{D}_S^{(i)}$ with corresponding data and labels $(\mathbf{X}_S^{(i)}, \mathbf{Y}_S^{(i)}), i \in \{1, \ldots, K\}$. we aim to leverage source domain information to classify the target data $\mathbf{X}_T$ on a solitary target domain $\mathcal{D}_T$. The proposed FedRF-Adapt scheme proposes to perform the FDA classification according to the following two steps:

(i) **local domain alignment** with RF-MMD, that learns to align source and target features by minimizing their MMD distance, via the exchange of $\Sigma\ell \in \mathbb{R}^{2N}$ as in Corollary 1; and

(ii) **global parameter aggregation** via FedAvg [15], that aggregates the source classifiers for final decision on the target data.

We discuss these two step in details as follows.

3.1.1. Local domain alignment

In each training round t, some randomly selected source clients are chosen and communicate with the target client, to exchange messages $\{\Sigma_S^{(i)}\ell_S^{(i)}\}_{i \in \mathcal{S}_t}$ and $\Sigma_T\ell_T$, where $\mathcal{S}_t \subset \{1, \ldots, K\}$ is a randomly drawn index set (that can even be a null set). These messages are then used to minimize the following objective functions:

$$L_S^{(i)} = L_C^{(i)} + \lambda L_{\text{MMD}}^{(i)}(\mathcal{D}_S^{(i)}, \mathcal{D}_T), \text{ for } i \in \mathcal{S}_t, \tag{6}$$

$$L_S^{(j)} = L_C^{(j)}, \text{ otherwise}, \tag{7}$$

for selected and non-selected source clients, respectively, with $\lambda > 0$ some hyper-parameter, $L_C^{(i)}$ the classification loss computed at source client i, and

$$L_{\text{MMD}}^{(i)}(\mathcal{D}_S^{(i)}, \mathcal{D}_T) = \mathsf{RF\text{-}MMD}(\mathbf{F}_S^{(i)}, \mathbf{F}_T), \tag{8}$$

the approximated RF-MMD distance between source $\mathbf{F}_S^{(i)}$ and target features $\mathbf{F}_T$ (obtained from local feature exactors $\mathbf{G}_S^{(i)}$ and $\mathbf{G}_T$, respectively, see an illustration in Figure 1) as in (4). The target client is trained by minimizing the MMD loss at round t as

$$L_T = \sum_{i \in \mathcal{S}_t} L_{\text{MMD}}^{(i)}(\mathcal{D}_S^{(i)}, \mathcal{D}_T). \tag{9}$$

All clients update their models locally. To compute the RFFs as in Definition 1, the Gaussian random matrix Ω is locally accessible to *all* clients through a shared random seed $\mathfrak{S}$.

The carefully designed random communication mechanism (with random index set $\mathcal{S}_t$) significantly enhances the robustness of FedRF-Adapt against poor network condition. Precisely, under ideal communication conditions, no message drop occurs and this

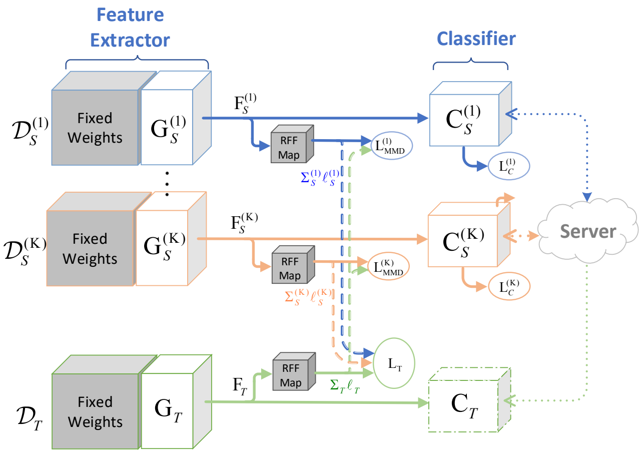

Fig. 1: Illustration of the proposed FedRF-Adapt protocol includes **feature extractors** (G_S and G_T for source and target, respectively) and **classifiers** (C_S and C_T). Only messages of type $\Sigma\ell \in \mathbb{R}^{2N}$ are exchanged between clients, and the classifiers are aggregated through a trustworthy server.

is modeled with $\mathcal{S}_t = \{1, \ldots, K\}$ as setting (I) in Table 2 and 3 of the numerical experiments in Section 4. In most real-world FDA scenarios, however, there may occur client and/or message drops during training (as a consequence of the unreliable network) and this is modeled by taking $\mathcal{S}_t$ as a random subset of $\{1, \ldots, K\}$. This corresponds to setting (II) and setting (III) in Table 2 and 3 of Section 4. Notably, we observe comparably good performance of the proposed FedRF-Adapt approach in such challenging environments under poor network condition.

3.1.2. *Global classifier aggregation*

To obtain a classifier at target client, we adopt the FedAvg protocol [15] to aggregate the parameters of selected source classifiers via a trustworthy server, and then sync the aggregated classifier C on both the source and target clients. Since the classifier aggregation can also be communicational burdensome, in the FedRF-Adapt scheme we propose to aggregate the classifiers $\{C_S\}_{i\in\mathcal{S}_t}$ *only* every $T_C \gg 1$ time intervals, to further reduce the communication overhead. This *infrequent* aggregation, as we shall see below in setting (III) of in Table 2 and 3, as well as in Figure 2, of the numerical experiments in Section 4 while significantly reduce the communication complexity, does *not* degenerate the performance of FedRF-Adapt.

The FedRF-Adapt training procedure is summarized in Algorithm 1 and illustrated in Figure 1.

3.2. Advantages of FedRF-Adapt

Here, we discuss the advantages of the proposed FedRF-Adapt in terms of its communication efficiency and robustness, as well as added privacy protection, when compared to popular federate DA methods such as FADA [1], FedKA [5], and FDA [6]. These results are summarized in Table 1.

By first applying random features technique to approximate the MMD distance and then "compressing" the RFFs into a single vector of the form $\Sigma\ell \in \mathbb{R}^{2N}$ with the RF-MMD approach in (4), the communication overhead of FedRF-Adapt in each round is *independent* of the sample size n, as opposed to existing federated DA methods

Algorithm 1 FedRF-Adapt training protocol

```
Input: Feature extractors and classifiers of $K$ sources $\{G_S^{(i)}, C_S^{(i)}\}_{i=1}^K$ and of target $G_T, C_T$.
Output: Target classifier $C_T$.
Initialization
Determine $T_C$ the time interval for aggregation.
Generate a random seed $\mathfrak{S}$ and send to all clients.
for each round $t = 1, 2, \ldots$ do
    Sample a random subset $\mathcal{S}_t$ from the index set $\{1, \ldots, K\}$.
    ▸ Local domain alignment:
    Each source client $i \in \mathcal{S}_t$:
        Sample a mini-batch from $(X_S^{(i)}, Y_S^{(i)})$.
        Compute source message $\Sigma_S^{(i)}\ell_S^{(i)}$ as in (4).
        Update $\{G_S^{(i)}, C_S^{(i)}\}$ by minimizing $\{L_S^{(i)}\}_{i\in\mathcal{S}_t}$ as in (6) or $\{L_S^{(i)}\}_{i\notin\mathcal{S}_t}$ as in (7).
    Target client:
        Sample a mini-batch from $X_T$.
        Compute target message $\Sigma_T\ell_T$ as in (4).
        Update $G_T$ by minimizing $L_T$ as in (9).
    ▸ Parameter aggregation and model update:
    if $t \% T_C = 0$ then
        For each client $i \in \mathcal{S}_t$, aggregate $C_S^{(i)}$ to get C.
        Assign C to $C_T$ and $\{C_S^{(i)}\}_{i\in\mathcal{S}_t}$.
    end if
end for
return Target classifier $C_T$.
```

listed in Table 1. Also, the computation of the RFFs Σ in Equation (3) involves periodic trigonometric functions and Gaussian random matrix Ω, making it impossible for a malicious third party to reveal users' privacy data X.

Table 1: Comparison between different federated DA methods, with K the number of clients, n the sample size, N the dimension of features in different methods, and $P \geq 1$ is the ciphertext size of Paillier encryption used in FDA [6].

Federated DA methods	FADA [1]	FedKA [5]	FDA [6]	**FedRF-Adapt (ours)**
Communication overhead	$O(KnN)$	$O(KnN)$	$O(KnNP)$	$O(KN)$
Robustness	✗	✗	✗	✓
Added privacy	✗	✗	✓	✓

We further demonstrate, with extensive numerical experiments in Table 2, Figure 2 and 3 of Section 4, that the proposed FedRF-Adapt shows excellent robustness in unreliable networks with random message and/or client dropouts.

4. NUMERICAL EXPERIMENTS

In this section, we provide comprehensive numerical results on the proposed FedRF-Adapt protocol on commonly used Office-Caltech [16] and Digit-Five [17] datasets, showing its advantageous performance as well as communicational efficiency and robustness. The code to reproduce the results in this section is available at https://github.com/yjwang346/FedRF-Adapt.

In particular, FedRF-Adapt improves over our previous FedRF-TCA protocol [18] by adopting a simpler FDA framework that further reduces communication overhead between clients, see [18] for further discussions and numerical experiments on FedRF-TCA.

Table 2: Classification accuracy (%) on Office-Caltech [16]. Baseline repeated from [1, 18]. Setting (I): all clients aggregate the classifier in each communication round; (II): only a random subset $\mathcal{S}_t$ of source clients are involved; (III): as for (II) with classifier aggregation interval $T_C = 100$. Best performance shown in **boldface**.

Methods	C,D,W→A	A,D,W→C	A,C,W→D	A,C,D→W	Avg
ResNet101 [19]	81.9	87.9	85.7	86.9	85.6
AdaBN [20]	82.2	88.2	85.9	87.4	85.7
AutoDIAL [21]	83.3	87.7	85.6	87.1	85.9
f-DAN[1] [22]	82.7	88.1	86.5	86.5	85.9
f-DANN[2] [23]	83.5	88.5	85.9	87.1	86.3
FADA [1]	84.2	88.7	87.1	88.1	87.1
FedRF-TCA [18] (III)	**94.5**	**98.6**	**98.8**	90.0	**95.5**
FedRF-Adapt (I)	92.6	85.3	97.6	**97.0**	93.1
FedRF-Adapt (II)	93.4	84.8	97.7	96.9	93.2
FedRF-Adapt (III)	92.7	82.8	96.5	96.2	92.1

Table 3: Classification accuracy (%) on Digit-Five [17]. Baseline repeated from [1]. "→mt" means "mm,sv,sy,up→mt." Settings (I), (II), and (III) as in Table 2.

Methods	→mt	→mm	→up	→sv	→sy	Avg
Source Only	75.4	49.6	75.5	22.7	44.3	53.5
f-DAN[1] [22]	86.4	57.5	90.8	45.3	58.4	67.7
f-DANN[2] [23]	86.1	59.5	89.7	44.3	53.4	66.6
FADA [1]	91.4	62.5	91.7	**50.5**	**71.8**	73.6
FedRF-TCA [18] (III)	97.4	64.3	89.5	41.9	44.4	67.5
FedRF-Adapt (I)	98.5	**76.3**	95.4	46.5	52.1	**73.8**
FedRF-Adapt (II)	98.5	74.3	95.1	45.1	52.9	73.2
FedRF-Adapt (III)	**98.5**	75.5	**95.7**	46.0	50.4	73.2

In Table 2 and 3, setting (I) represents the most ideal federate DA scenario where all source clients exchange information with the target client in *each* round of communication. Settings (II) and (III), on the other hand, consider more practical scenarios for which (random) message and/or client dropouts occur. It can be seen from Table 2 and 3 that the performance of the proposed FedRF-Adapt protocol under setting (II) and (III) is equally good as setting (I), demonstrating the excellent robustness of FedRF-Adapt against unreliable network conditions.

Further note that under setting (III), the source classifiers are aggregated *only* every $T_C \gg 1$ rounds, leading to additional reduction in communication overhead. We further show in Figure 2 that the robustness to network reliability can be consistently observed across various communication interval choices T_C in Algorithm 1, with a performance fluctuation less than 1% for T_C ranging from 50 to 800.

5. CONCLUSION

In this paper, we propose RF-MMD as a computational efficient "proxy" to the original MMD distance. We further extend RF-MMD to a FDA setting by introducing FedRF-Adapt, that is both communication-efficient and robust to unreliable network conditions. Numerical experiments show that the proposed FedRF-Adapt scheme yields performance comparable to state-of-the-art FDA methods with a significant reduction in communication overhead.

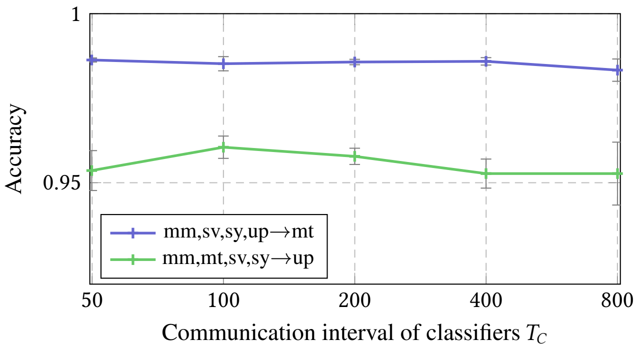

Fig. 2: Classification accuracy (mean ± standard deviation) of FedRF-Adapt with different communication intervals $T_C \in \{50, 100, 200, 400, 800\}$, with in total 1 650 rounds of communication, under Setting (III) of Table 3.

Note that RF-MMD avoids the stringent requirement of simultaneous access to all data for MMD distance computation, and it shows promise for wider applications in other MMD-based methods.

6. REFERENCES

[1] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko, "Federated adversarial domain adaptation," in *International Conference on Learning Representations*, 2019.

[2] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf, "Algorithmic Learning Theory," *Lecture Notes in Computer Science*, pp. 13–31, 2007.

[3] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola, "A kernel method for the two-sample-problem," *Advances in neural information processing systems*, vol. 19, 2006.

[4] Lei Song, Chunguang Ma, Guoyin Zhang, and Yun Zhang, "Privacy-preserving unsupervised domain adaptation in federated setting," *IEEE Access*, vol. 8, pp. 143233–143240, 2020.

[5] Yuwei Sun, Ng Chong, and Hideya Ochiai, "Feature distribution matching for federated domain generalization," in *Asian Conference on Machine Learning*. PMLR, 2023, pp. 942–957.

[6] Hua Kang, Zhiyang Li, and Qian Zhang, "Communicational and computational efficient federated domain adaptation," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 12, pp. 3678–3689, 2022.

[7] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein, *Introduction to algorithms*, MIT press, 2022.

[8] Nicolaas Govert De Bruijn, *Asymptotic methods in analysis*, vol. 4, Courier Corporation, 1981.

[9] Sinno Jialin Pan, James T Kwok, Qiang Yang, et al., "Transfer learning via dimensionality reduction.," in *AAAI*, 2008, vol. 8, pp. 677–682.

[10] Sinno Jialin Pan and Qiang Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[11] Bernhard Schölkopf and Alexander J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press, 2018.

[1] Here, f-DAN is a federated DA method based on DAN [22].

[2] Here, f-DANN is a federated DA method based on DANN [23].

[12] Ali Rahimi and Benjamin Recht, "Random Features for Large-Scale Kernel Machines," in *Advances in Neural Information Processing Systems*. 2008, vol. 20 of *NIPS'08*, pp. 1177–1184, Curran Associates, Inc.

[13] Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan A. K. Suykens, "Random features for kernel approximation: A survey on algorithms, theory, and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2021.

[14] Joel A. Tropp, "An Introduction to Matrix Concentration Inequalities," *Foundations and Trends® in Machine Learning*, vol. 8, no. 1-2, pp. 1–230, 2015.

[15] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[16] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2066–2073.

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[18] Zhanbo Feng, Yuanjie Wang, Jie Li, Fan Yang, Jiong Lou, Tiebin Mi, Robert Qiu, Zhenyu Liao, et al., "Robust and communication-efficient federated domain adaptation via random features," *arXiv preprint arXiv:2311.04686*, 2023.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, IEEE.

[20] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou, "Revisiting batch normalization for practical domain adaptation," *arXiv preprint arXiv:1603.04779*, 2016.

[21] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulo, "Autodial: Automatic domain alignment layers," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5067–5075.

[22] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.

[23] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.