An Optimal Transport Approach to Personalized Federated Learning

Farzan Farnia*, Amirhossein Reisizadeh*, Ramtin Pedarsani, *Senior Member, IEEE*, Ali Jadbabaie, *Fellow, IEEE*

***Abstract*—Federated learning is a distributed machine learning paradigm, which aims to train a model using the local data of many distributed clients. A key challenge in federated learning is that the data samples across the clients may not be identically distributed. To address this challenge, personalized federated learning with the goal of tailoring the learned model to the data distribution of every individual client has been proposed. In this paper, we focus on this problem and propose a novel personalized Federated Learning scheme based on Optimal Transport (`FedOT`) as a learning algorithm that learns the optimal transport maps for transferring data points to a common distribution as well as the prediction model under the applied transport map. To formulate the `FedOT` problem, we extend the standard optimal transport task between two probability distributions to multi-marginal optimal transport problems with the goal of transporting samples from multiple distributions to a common probability domain. We then leverage the results on multi-marginal optimal transport problems to formulate `FedOT` as a min-max optimization problem and analyze its generalization and optimization properties. We discuss the results of several numerical experiments to evaluate the performance of `FedOT` under heterogeneous data distributions in federated learning problems.**

I. Introduction

The proliferation of mobile devices requires learning algorithms capable of training a prediction model using data distributed across local users in a network. Federated learning [1] is a recent learning paradigm where several users are connected to a central server and train a machine learning model through their communications with the server. While standard federated learning algorithms perform successfully under identically distributed training data at different users, this assumption does not usually hold in practical federated learning settings in which the training samples are collected by multiple agents with different backgrounds, e.g. speech and text data gathered from a multi-lingual community. To address the heterogeneity of users' data distributions, federated learning under heterogeneous data has received great attention in the machine learning community [2]–[6].

A recently studied approach for federated learning under non-identically distributed data is to adapt the globally trained model to the particular distribution of every local user. Based on this approach, instead of learning a common model shared by all the users, the learning algorithm tailors the trained model to the samples observed by every user in the network. As such personalized federated learning algorithms lead to different trained models at different users, an important baseline for their evaluation is a locally-performing learning algorithm in which every user fits a separate model to only her own data. Therefore, the conditions under which the users can improve upon such a non-federated purely local baseline play a key role in the design of a successful personalized federated learning method.

In a general federated learning setting with arbitrarily different users' distributions, the users do not necessarily benefit from cooperation through federated learning. For example, if the users aim for orthogonal classification objectives, their cooperation according to standard federated learning algorithms can even lead to worse performance than their locally trained models. To characterize conditions under which a mutually beneficial cooperation is feasible, a standard assumption in the literature is to bound the distance between the distributions of different users. However, such assumptions on the closeness of the distributions raise the question of whether federated learning will remain beneficial if the users' distributions do not stay in a small distance from each other.

In this work, we study the above question through the lens of optimal transport theory and demonstrate that a well-designed federated learning algorithm can still improve upon the users' locally-trained models as long as the transportation maps between the users' distributions can be properly learned from the training data. We show that this condition relaxes the bounded distance assumption used in the literature and further applies to any federated learning setting where the learners only have some rudimentary knowledge of the statistical nature of distribution shifts, e.g. under affine convolutional filters applied to change the color, brightness, and intensity of image data.

To learn the personalized models under the above condition, we introduce `FedOT` as a *Federated learning framework based on Optimal Transport*. According to `FedOT`, the users simultaneously learn the transportation maps for transferring their samples to a common probability domain and fit a global classifier to the transferred training data. To personalize the

Manuscript received January 1, 2022; revised April 20, 2022; accepted June 2, 2022. This work was supported by a CUHK Faculty of Engineering Direct Grant for Research, ONR grant N00014-20-1-2394, MIT-IBM Watson AI lab, NSF grants CNS 2003035, CCF 1909320, and UC Office of President Grant LFR-18-548175.

Farzan Farnia is with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Sha Tin, Hong Kong SAR (email: farnia@cse.cuhk.edu.hk).

Amirhossein Reisizadeh is with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (email: amirr@mit.edu).

Ramtin Pedarsani is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, Santa Barbara, CA 93106 USA (email: ramtin@ece.ucsb.edu).

Ali Jadbabaie is with the Institute for Data, Systems and Society and the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (email: jadbabai@mit.edu).

*: Contributed equally

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. The material includes the proofs of the paper's theoretical statements. Also, the paper's code is accessible at the GitHub repository https://github.com/farzanfarnia/FedOT.

This article has been accepted for publication in IEEE Journal on Selected Areas in Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2022.3182355

globally trained model to the specific distribution of every user, FedOT combines the global classifier with the learned transportation maps needed for transferring samples from the original distributions of local users to the common distribution.

In order to formulate and solve FedOT, we leverage optimal transport theory to reduce FedOT's learning task to a min-max optimization problem. To this end, we focus on an extension of standard optimal transport problems between two probability domains to a structured multi-marginal optimal transport task for mapping several different distributions to a common probability domain. In Section II, we review several key definitions and results from multi-marginal optimal transport theory for which we provide a unified set of notations and novel proofs. We generalize standard duality results in optimal transport theory to the multi-marginal setting, which results in a min-max formulation of FedOT. The main results in this section not only guide us toward formulating a minimax optimization problem for the FedOT framework (Theorem 1), but also provide intuition on how to design the function spaces in the FedOT minimax approach (Theorem 2). Specifically, we leverage the intuition offered by Theorem 2 to reduce the size of function spaces in the FedOT minimax problem and improve the generalization and optimization performance of the FedOT learners.

Next, we show that FedOT's min-max formulation is capable of being decomposed into a distributed form, and thus FedOT provides a scalable federated learning framework. We further analyze the generalization and optimization properties of the proposed FedOT approach. Under the condition that the sample complexity of learning the classifier dominates the complexity of finding the transportation maps, we prove that FedOT enjoys a better generalization performance in comparison to locally trained models. In addition, we show that the formulated min-max optimization problem can be solved to a stationary min-max solution by a standard distributed gradient descent ascent (GDA) algorithm. Therefore, the min-max formulation leads to a tractable distributed optimization problem, since the iterative GDA updates can be decomposed into a distributed form.

Finally, we discuss the results of our numerical experiments comparing the performance of FedOT with several standard federated learning schemes. Our experimental results demonstrate the success of FedOT under various types of distribution changes including affine distribution shifts and image color transformations. We can summarize the main contributions of this work as follows:

- Introducing FedOT as an optimal transport-based approach to the federated learning problem under heterogeneous data,
- Extending standard results of optimal transport theory to the multi-marginal optimal transport problem with the goal of transporting the input distributions to a common probability domain,
- Analyzing the generalization and optimization properties of FedOT and establishing conditions under which FedOT improves upon locally-learned models,
- Demonstrating the efficacy of FedOT through several numerical experiments on standard image recognition datasets and neural network architectures.

Related Work on Federated Learning and Min-Max Optimization. There has been a vast variety of tools and techniques used to address the prersonalization challenge in federated learning. As discussed before, utilizing only a shared global model for all the clients fails to capture the discrepancies in users' data distributions. On the other hand, local models would not benefit from the samples of other clients if a mere local training is implemented. Therefore, a combination of the two trained models, global and local ones, would naturally provide a degree of personalization [7]–[9] which is also known as model interpolation.

Meta-learning-based approaches to federated learning under heterogeneous data distributions have been proposed by the related works [10]–[12]. According to these approaches, a local and personalized model is adapted for each client by performing a few gradient steps on a common global model. This family of federated learning algorithms have been shown to be successful in handling unstructured distribution shifts where the learners have no prior knowledge of the structure of distribution shifts in the underlying network. On the other hand, the main focus of our proposed FedOT framework is on the learning scenarios where the learners have some prior knowledge of the type of distribution shifts.

In a data interpolation approach to personalized federated learning [7], [8], a local model is trained for each client by minimizing the loss over a mixture of local and global distributions. [13], [14] propose to learn a common representation for personalized federated learning. Similarly, [15] develop a personalized federated learning approach through a group of hypernetworks to update the neural net classifier. While our work pursues a similar goal of learning a common representation, it introduces a novel minimax learning algorithm by leveraging optimal transport theory.

Cluster-based federated learning methods based on clustering users with similar underlying distributions have also been explored in several related works [16]–[19] to overcome the challenge of heterogeneous data in federated learning. As another approach, [20] propose applying local batch normalization to train personalized neural network classifiers. In a slightly different approach to handle the data heterogeneity challenge in federated learning, [21]–[23] propose different min-max formulations to train robust models against non-i.i.d. samples. Aside its federated learning applications, nonconvex-concave min-max optimization and its complexity guarantees have been extensively studied in the literature [24]–[27].

Related Work on Optimal Transport Frameworks in Machine Learning. A large body of related works apply optimal transport theory to address various statistical learning problems. These applications include generative adversarial networks (GANs) [28]–[30], distributionally robust supervised learning [31]–[33], learning mixture models [34], [35], and combining neural network models [36]. Multi-marginal optimal transport costs [37] have also been studied in other machine learning contexts including GANs [38], domain adaptation [39], and Wasserstein barycenters [40]–[42].

II. Multi-input Optimal Transport Problems

A useful approach to learning under heterogeneous data distributions is to transport the different input distributions to a shared probability domain and then learn a supervised learning model for the shared probability domain. This task can be cast as a multi-input optimal transport problem, since the goal is to map the input distributions to a common distribution. In this section, we review the key definitions and tools from multi-input optimal transport theory to address the transportation task. The results in this section guide us toward formulating a minimax optimization problem for federated learning under heterogeneous distributions, and further help to reduce the statistical and computational complexities of the learning problem through leveraging prior knowledge of the structure of distribution shifts in the federated learning setting.

In the literature, the optimal transport problem is typically defined for transporting samples between two probability domains [43]. For a cost function $c(x,x')$ measuring the cost of transporting x to x', optimal transport cost $W_c(P,Q)$ is defined through finding the coupling that leads to the minimum expected cost of transporting samples between P, Q:

$$W_c(P,Q) := \min_{\pi\in\Pi(P,Q)} \mathbb{E}_{(X,X')\sim\pi}\bigl[c(X,X')\bigr].$$

Here $\Pi(P,Q)$ denotes the set of all joint distributions on (X,X') that are marginally distributed as P and Q. Note that the above optimal transport cost quantifies the optimal expected cost of mapping samples between the domains P and Q.

However, for several problems of interest in machine learning one needs to extend the above definition to multi-input cost functions where the goal is to transport samples across multiple distributions. To define the n-ary optimal transport cost, a standard extension [37] is to consider an n-ary cost function $c(x_1,\cdots,x_n)$ and define the n-ary optimal transport map as:

$$W_c(P_1,\cdots,P_n) := \min_{\pi\in\Pi(P_1,\cdots,P_n)} \mathbb{E}_{\pi}\bigl[c(X_1,\cdots,X_n)\bigr],$$

where $\Pi(P_1,\cdots,P_n)$ denotes the set of joint distributions on $(X_1,\dots,X_n)$ that are marginally distributed as $P_1,\dots,P_n$, respectively.

Inspired by the personalized federated learning problem where our goal is to map the different input distributions to a common probability domain, we focus on the following type of n-ary cost functions throughout this paper, which is also referred to as the infimal convolution cost [37]. The optimal transport costs resulting from the following type of n-ary costs preserve the key features of standard optimal transport costs with binary cost $\tilde{c}(x,x')$:

$$c(x_1,\cdots,x_n) = \min_{x'} \sum_{i=1}^{n} \tilde{c}(x',x_i). \tag{1}$$

Such an n-ary cost function lets us focus on n-ary transportation problems where the goal is to transport all the n inputs to a single point that minimizes the total cost of transportation. The following proposition by [44] connects the n-ary optimal transport costs to binary optimal transport costs.

Proposition 1 ([44], Prop. 3)**.** *Consider the n-ary cost in (1). Then,*

$$W_c(P_1,\cdots,P_n) = \min_{Q} \sum_{i=1}^{n} W_{\tilde{c}}(Q,P_i). \tag{2}$$

Proof: We defer the proof to the Appendix. ■

We note that if the binary cost function is chosen as a powered norm difference $\tilde{c}(\mathbf{x},\mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\|^q$, then the proposed multi-marginal optimal transport cost simplifies to the well-known family of Wasserstein barycenters. Next, we present a generalization of the Kantorovich duality theorem to n-ary optimal transport costs with the characterized cost function. This result has been already shown in the optimal transport theory literature [44], and we present our new proof of the result in the Appendix. In the following theorem, we use the standard definition of the c-transform of a real-valued function ϕ as $\phi^{\tilde{c}}(x) := \min_{x'}\ \tilde{c}(x,x') + \phi(x')$.

Theorem 1. *For the n-ary cost in (1), we have the following duality result where each variable $\phi_i : \mathbb{R}^d \rightarrow \mathbb{R}$ denotes a real-valued function:*

$$W_c(P_1,\cdots,P_n) = \max_{\substack{\phi_{1:n}:\\ \forall\mathbf{x}:\ \sum_i \phi_i(\mathbf{x})=0}} \sum_{i=1}^{n} \mathbb{E}_{P_i}\bigl[\,\phi_i^{\tilde{c}}(\mathbf{X})\,\bigr].$$

Proof: We defer the proof to the Appendix. ■

In above and henceforth, we use the short-hand notation $a_{1:n} := \{a_1,\cdots,a_n\}$, for n vectors $a_1,\cdots,a_n$. Next, we apply the above result to standard norm-based cost functions and simplify the dual maximization problem for these Wasserstein costs:

Example 1. *For the 1-Wasserstein cost $c_1(\mathbf{x}_1,\cdots,\mathbf{x}_n) = \min_{\mathbf{x}'}\sum_i \|\mathbf{x}_i-\mathbf{x}'\|$, we have*

$$W_{c_1}(P_1,\cdots,P_n) = \max_{\substack{\phi_{1:n}:\ \text{1-Lipschitz}\\ \forall\mathbf{x}:\ \sum_i \phi_i(\mathbf{x})\le 0}} \sum_{i=1}^{n} \mathbb{E}_{P_i}\bigl[\phi_i(\mathbf{X})\bigr].$$

Note that in the special case $n=2$, the triangle inequality implies that $c_1(\mathbf{x}_1,\mathbf{x}_2) = \|\mathbf{x}_1-\mathbf{x}_2\|$ which leads to standard 1-Wasserstein distance in the optimal transport theory literature [43].

Example 2. *For the 2-Wasserstein cost $c_2(\mathbf{x}_1,\cdots,\mathbf{x}_n) = \min_{\mathbf{x}'}\sum_i \|\mathbf{x}_i-\mathbf{x}'\|_2^2$, we have*

$$W_{c_2}(P_1,\cdots,P_n) = \max_{\substack{\phi_{1:n}:\ \text{convex},\ \forall\mathbf{x}:\\ \frac{1}{n}\sum_i \phi_i(\mathbf{x})\le\frac{1}{2}\|\mathbf{x}\|_2^2}} \sum_{i=1}^{n} \mathbb{E}_{P_i}\Bigl[\frac{1}{2}\|\mathbf{X}\|^2 - \phi_i^{\star}(\mathbf{X})\Bigr] \tag{3}$$

In the above, $\phi^{\star}$ denotes the Fenchel conjugate defined as $\phi^{\star}(\mathbf{x}) := \sup_{\mathbf{x}'} \mathbf{x}^{\top}\mathbf{x}' - \phi(\mathbf{x}')$. For the special case $n=2$, one can see $c_2(\mathbf{x}_1,\mathbf{x}_2) = \frac{1}{2}\|\mathbf{x}_1-\mathbf{x}_2\|_2^2$ which results in the standard 2-Wasserstein distance in the literature [43].

The next result shows that in the case of the 2-Wasserstein cost the optimal potential function $\phi^*_{1:n}$ will transport samples to a common probability domain matching the distribution Q^* in (2) with the optimal sum of Wasserstein costs to the input distributions. This result has been previously shown in [44], and we present a new proof in the Appendix.

Theorem 2. *Suppose that $\phi^*_1,\cdots,\phi^*_n$ denote the optimal solutions to (3) for 2-Wasserstein dual optimization problem.*

Then,

$$\forall\, 1 \le i, j \le n: \quad \nabla\phi_i^{\star\star}(\mathbf{X}_i) \overset{\text{dist}}{=} \nabla\phi_j^{\star\star}(\mathbf{X}_j).$$

In the above, each $\mathbf{X}_i$ denotes the ith random variable distributed according to P_i and $\overset{\text{dist}}{=}$ means the two random variables share an identical distribution.

Proof: We defer the proof to the Appendix. ■

As implied by the above theorem, the gradients of optimal potential functions lead to transportation maps for transporting samples from the different input distributions to a common probability domain. As we discuss later, transporting input samples to a common probability distribution can help to reduce the generalization error of a distributed learning task.

III. FedOT: Federated Learning based on Optimal Transport

A. Federated Learning Setting

We focus on a federated learning scenario with n local nodes connected to a single parameter server. We assume that every node $i \in [n]$ observes m training samples $\{(\mathbf{x}_{i,j}, y_{i,j})\}_{j=1}^m$ which are independently sampled from distribution P_i. Note that the input distributions are in general different, leading to a non-i.i.d. federated learning problem.

To model the heterogeneity of the distributions across the network, we suppose that for each node i, there exists an invertible transportation map $\psi_i : \mathbb{R}^d \rightarrow \mathbb{R}^d$ that maps a sample $(\mathbf{X}_i, Y_i)$ observed by node i to a common distribution, i.e.,

$$\forall\, 1 \le i, j \le n: \ \big(\psi_i(\mathbf{X}_i), Y_i\big) \overset{\text{dist}}{=} \big(\psi_j(\mathbf{X}_j), Y_j\big).$$

In the above, $\overset{\text{dist}}{=}$ denotes an identical probability distribution for the transported samples. Therefore, the mappings $\psi_{1:n}$ transfer the input distributions across the network to a common probability domain. Furthermore, we assume that there exists a space of functions $\Psi = \{\psi_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ parameterized by $\boldsymbol{\theta}$ containing the underlying transportation map ψ_i's in our described federated learning setting.

In the above federated learning setting, one can simplify the federated learning problem to finding a prediction rule $f_{\mathbf{w}} \in \mathcal{F}$ which predicts label Y from the transported data vector in the shared probability domain of $\psi_i(X_i)$'s. Here $\mathcal{F} = \{f_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}\}$ is the set of models for training the prediction rule parameterized by the vector $\mathbf{w}$. Since $\psi_i(X_i)$'s are identically distributed across the network, the collected transported samples from *all* the nodes can be used to train the prediction rule $f_{\mathbf{w}}$. Note that after finding the optimal classification rule $f_{\mathbf{w}^*}$, every node i can personalize the classification rule by combining the transportation function ψ_i and $f_{\mathbf{w}^*}$. Here, the personalized classifier for node i will be $f_{\mathbf{w}^*}(\psi_i(\cdot))$.

Remark 1. *According to the Brenier's theorem [43], [45], the existence of the invertible transportation maps $\psi_i : \mathbb{R}^d \rightarrow \mathbb{R}^d$ for $i = 1, \ldots, n$ mapping client distribution P_i's to a common domain is guaranteed under the regularity assumption that the input distributions are absolutely continuous with respect to one another. Furthermore, we note that our analysis requires this assumption only for the underlying client distributions and does not need the condition for the empirical distributions of training samples.*

Remark 2. *While the described setting requires the same marginal distribution P_Y for every client's label variable Y, the optimal transport-based framework can be further extended to cases with heterogeneous marginal distributions. To do this, we need to extend the assumption on the clients' feature distribution $P_{\mathbf{X}}$ to the clients' conditional feature distribution $P_{\mathbf{X}|Y=y}$ for every label outcome $y \in \mathcal{Y}$. In the extended setting, we further assume that for every $y \in \mathcal{Y}$, invertible transportation map $\psi_{y,i}$'s exist such that the conditional feature distribution $P_{\psi_{y,i}(\mathbf{X}_i)|Y_i=y}$ is identical for different clients. In this work, our main focus is on the setting with heterogeneous feature distributions, as the gain attained by the optimal transport approach is obtained through leveraging the structures on the features distribution shifts. Nevertheless, we still note that the optimal transport approach can be further extended to learning settings with different marginal distributions on the label variable Y.*

B. FedOT as a Min-Max Optimization Problem

In order to train a personalized classification rule $f_{\mathbf{w}}$ and transportation maps $\psi_{\boldsymbol{\theta}_{1:n}}$, we consider the following optimization problem:

$$\min_{\mathbf{w}, \boldsymbol{\theta}_{1:n}} \widehat{\mathcal{L}}(\mathbf{w}, \boldsymbol{\theta}_{1:n}), \text{ s.t. } W_c\big(P_{\psi_{\boldsymbol{\theta}_1}(\mathbf{X}_1)}, \cdots, P_{\psi_{\boldsymbol{\theta}_n}(\mathbf{X}_n)}\big) \le \varepsilon.$$

In the above problem, we denote the empirical risk under transport maps $\psi_{\boldsymbol{\theta}_{1:n}}$ as

$$\widehat{\mathcal{L}}(\mathbf{w}, \boldsymbol{\theta}_{1:n}) := \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} \ell\big(f_{\mathbf{w}}(\psi_{\boldsymbol{\theta}_i}(\mathbf{x}_{i,j})), y_{i,j}\big),$$

which quantifies the empirical risk associated with the mn transported data samples across the n nodes and $W_c(\cdot, \cdots, \cdot)$ denotes the n-ary optimal transport cost which measures the distance among the input distributions. Ideally, one wants the n-ary optimal transport cost to take a zero value that is necessary for having the same probability distribution for different $\psi_{\boldsymbol{\theta}_i}(\mathbf{X}_i)$'s. However, due to the generalization error in estimating the optimal transport cost from finite training data we allow an ϵ-bounded optimal transport cost in the above formulation.

In our analysis, we transfer the constraint bounding the optimal transport cost to the objective via a Lagrangian penalty and study the following optimization problem for a non-negative constant $\lambda \ge 0$:

$$\min_{\mathbf{w}, \boldsymbol{\theta}_{1:n}} \widehat{\mathcal{L}}(\mathbf{w}, \boldsymbol{\theta}_{1:n}) + \lambda W_c\big(P_{\psi_{\boldsymbol{\theta}_1}(\mathbf{X}_1)}, \cdot\cdot, P_{\psi_{\boldsymbol{\theta}_n}(\mathbf{X}_n)}\big).$$

In order to solve the above optimization problem, we apply the generalized Kantorovich duality in Theorem 1 and reduce the above optimization problem to a min-max optimization task:

$$\min_{\mathbf{w}, \boldsymbol{\theta}_{1:n}} \max_{\substack{\phi_{1:n}: \\ \forall \mathbf{x}: \sum_i \phi_i(\mathbf{x})=0}} \widehat{\mathcal{L}}(\mathbf{w}, \boldsymbol{\theta}_{1:n}, \phi_{1:n}) := \qquad (6)$$
$$\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} \ell\big(f_{\mathbf{w}}(\psi_{\boldsymbol{\theta}_i}(\mathbf{x}_{i,j})), y_{i,j}\big) + \lambda\phi_i^{\tilde{c}}(\psi_{\boldsymbol{\theta}_i}(\mathbf{x}_{i,j})).$$

We call the above min-max framework *Federated Learning based on Optimal Transport (*`FedOT`*)*. We note that `FedOT` represents a family of federated learning algorithms for different cost functions.

To solve the above min-max problem of `FedOT` for neural network function variables $\phi_{1:n}$, we enforce the zero sum condition in the above problem through constraining every neural net in $\phi_{1:n}$ to share the same weights for all the layers before the last layer and satisfy a zero summation of the weights of the last layers. Here, for activation function $\rho(\cdot)$ and weight matrices $\mathbf{U} := [U_1, \ldots, U_L]$, we let $\phi_{\mathbf{U}}$ represent the neural network's mapping to the last layer and $\mathbf{v}_{1:n}$ stand for the weights of the last layers with a zero sum, i.e., $\sum_i \mathbf{v}_i = \mathbf{0}$, and hence we use the following function variables:

$$\phi_i(\mathbf{x}) := \mathbf{v}_i^\top \phi_{\mathbf{U}}(\mathbf{x}), \quad \phi_{\mathbf{U}}(\mathbf{x}) := \rho(U_L\rho(\cdots\rho(U_1\mathbf{x})\cdots)) \\ \text{s.t.} \quad \sum_{i=1}^{n} \mathbf{v}_i = \mathbf{0}.$$

In the following, we characterize the `FedOT` learning problems for 1-Wasserstein and 2-Wasserstein cost functions as earlier defined in Examples 1 and 2.

Example 3. *Consider the* `FedOT` *problem with the* 1-*Wasserstein cost in Example 1. This formulation with neural net ϕ_i's leads to the* 1-*FedOT min-max problem:*

$$\min_{\mathbf{w},\boldsymbol{\theta}_{1:n}} \max_{\substack{\mathbf{v}_{1:n},\mathbf{U}: \\ \mathbf{v}_i^\top \phi_{\mathbf{U}} \textit{ 1-Lipschitz}, \\ \sum_i \mathbf{v}_i = \mathbf{0}}} \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\Big[\ell\big(f_{\mathbf{w}}(\psi_{\boldsymbol{\theta}_i}(\mathbf{x}_{i,j})), y_{i,j}\big) \\ + \lambda \mathbf{v}_i^T \phi_{\mathbf{U}}(\psi_{\boldsymbol{\theta}_i}(\mathbf{x}_{i,j}))\Big] \tag{7}$$

Example 4. *Consider the* `FedOT` *problem with the* 2-*Wasserstien cost in Example 2. This formulation leads to the* 2-*FedOT min-max problem:*

$$\min_{\mathbf{w},\boldsymbol{\theta}_{1:n}} \max_{\substack{\mathbf{v}_{1:n},\mathbf{U}: \\ \mathbf{v}_i^\top \phi_{\mathbf{U}} \textit{ 1-convex}, \\ \sum_i \mathbf{v}_i = \mathbf{0}}} \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\Big[\ell\big(f_{\mathbf{w}}(\psi_{\boldsymbol{\theta}_i}(\mathbf{x}_{i,j})), y_{i,j}\big) \tag{8} \\ + \frac{\lambda}{2}\|\psi_{\boldsymbol{\theta}_i}(\mathbf{x}_{i,j})\|^2 - \lambda\big(\mathbf{v}_i^\top \phi_{\mathbf{U}}\big)^\star(\psi_{\boldsymbol{\theta}_i}(\mathbf{x}_{i,j}))\Big]$$

Here, a function $g(\mathbf{x})$ is called 1-*convex if $g(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}\|_2^2$ is a convex function. Also, $(\mathbf{v}_i^\top \phi_{\mathbf{U}})^\star$ denotes the Fenchel conjugate of $\mathbf{v}_i^\top \phi_{\mathbf{U}}$.*

Next, we reduce (8) to an L_2-regularized min-max optimization problem with no Fenchel conjugates.

Proposition 2. *Suppose that the maximization variables in (8) are constrained such that $\mathbf{v}_i^\top \phi_{\mathbf{U}}$ is γ-smooth, i.e., $\nabla_{\mathbf{x}} \mathbf{v}_i^\top \phi_{\mathbf{U}}(\mathbf{x})$ is γ-Lipschitz w.r.t. $\mathbf{x}$, and the operator norm of every layer of neural net $\phi_{\mathbf{U}}$ satisfies $\|U_i\|_2 \le 1$. Then, the min-max objective in (8) is lower-bounded by:*

$$\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\Big[\ell\big(f_{\mathbf{w}}(\psi_{\boldsymbol{\theta}_i}(\mathbf{x}_{i,j})), y_{i,j}\big) \tag{9} \\ + \lambda \mathbf{v}_i^\top \phi_{\mathbf{U}}\big(\psi_{\boldsymbol{\theta}_i}(\mathbf{x}_{i,j})\big) - \frac{\lambda}{1-\gamma}\big(\|\mathbf{v}_i\|_2^2 + \|\mathbf{U}\|_F^2\big)\Big],$$

where $\|\mathbf{U}\|_F$ denotes the Frobenius norm of $\mathbf{U} = [U_1, \ldots, U_L]$ defined as $\|\mathbf{U}\|_F^2 := \sum_{i=1}^{L} \|U_i\|_F^2$.

Proof: We defer the proof to the Appendix. ■

Note that if $\mathbf{v}_i^\top \phi_{\mathbf{U}}\big(\psi_{\boldsymbol{\theta}_i}(\mathbf{x})\big)$ is γ'-smooth as a function of $\mathbf{v}_i, \mathbf{U}$ where $\gamma' < \frac{1}{1-\gamma}$, then the min-max objective in (9) will be $\lambda\big(\frac{1}{1-\gamma} - \gamma'\big)$-strongly concave in terms of the maximization variables, resulting in a nonconvex strongly-concave min-max problem. We later show a federated gradient descent ascent (GDA) algorithm can solve such a min-max problem to find a first-order stationary min-max solution.

IV. Generalization and Optimization Properties of `FedOT`

A. Generalization Guarantees

As discussed in the previous section, `FedOT` formulates the federated learning problem through the min-max optimization problem in (III-B). In the heterogeneous case where every agent i observes samples drawn from a different distribution P_i, the min-max objective of (III-B) provides an empirical estimation of the following true min-max objective:

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\theta}_{1:n}, \phi_{1:n}) := \\ \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P_i}\big[\,\ell\big(f_{\mathbf{w}}(\psi_{\boldsymbol{\theta}_i}(\mathbf{X})), Y\big) + \lambda\phi_i^{\tilde{c}}(\psi_{\boldsymbol{\theta}_i}(\mathbf{X}))\,\big].$$

With no assumptions on the optimal transport functions, estimating the above objective for all ϕ_i's will require an exponentially growing number of training samples in the dimension of data variable $\mathbf{X}$ [46]. In order to mitigate such an exponential complexity, we assume that for any feasible underlying P_i, the optimal potential functions $\phi^*_{1:n}$ belong to a set of functions Φ with bounded complexity. Under the assumption that for all feasible θ_i's, $\phi_i^* \in \Phi$ is satisfied for optimal ϕ_i's one can equivalently solve the min-max problem (III-B) with the additional constraints $\forall i: \ \phi_i \in \Phi$, which as will be shown attains a bounded generalization error.

In our generalization analysis, we use the following standard definition of the covering number $\mathcal{N}(\mathcal{F}, \epsilon, \|\cdot\|_\infty)$ of a set of functions $\mathcal{F}$ with respect to the L_∞-norm:

$$\mathcal{N}(\mathcal{F}, \epsilon, \|\cdot\|_\infty) := \min\big\{N \in \mathbb{N} : \\ \text{an } \epsilon\text{-covering of } \mathcal{F} \text{ exists w.r.t. } \|\cdot\|_\infty \text{ with size } N\big\}.$$

In order to simplify our theoretical statements, we use the following notation in our theorems where $M := \sup_{f\in\mathcal{F},\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$ and $\mathcal{V}(\mathcal{F}) := \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}, M\epsilon, \|\cdot\|_\infty)}\, d\epsilon$.

Theorem 3. *Suppose that the loss function ℓ is L_ℓ-Lipschitz and the expected loss is bounded by M under all feasible distributions. Assume that for any $\mathbf{w} \in \mathcal{W}$, $\phi \in \Phi$, $\boldsymbol{\theta} \in \Theta$, $f_{\mathbf{w}}$, ϕ, $\psi_{\boldsymbol{\theta}}$ are $L_{\mathbf{w}}$, L_ϕ, L_θ-Lipschitz. Then, $\forall \delta > 0$ with probability*

Algorithm 1: `FedOT-GDA`

```
Initialize initial models $(\mathbf{w}_0, \mathbf{v}_0)$, stepsizes $\eta_1, \eta_2$, number of local updates $\tau$
for $t = 1, \cdots, T-1$ do
    if $t \nmid \tau$ then
        $\mathbf{w}^i_{t+1} = \mathbf{w}^i_t - \eta_1 \widetilde{\nabla}_{\mathbf{w}} \widehat{\mathcal{L}}_i(\mathbf{w}^i_t, \mathbf{v}^i_t)$ and $\mathbf{v}^i_{t+1} = \mathbf{v}^i_t + \eta_2 \widetilde{\nabla}_{\mathbf{v}} \widehat{\mathcal{L}}_i(\mathbf{w}^i_t, \mathbf{v}^i_t)$
    end
    else
        $\mathbf{w}^i_{t+1} = \frac{1}{n}\sum_{k=1}^{n} \left[\mathbf{w}^k_t - \eta_1 \widetilde{\nabla}_{\mathbf{w}} \widehat{\mathcal{L}}_k(\mathbf{w}^k_t, \mathbf{v}^k_t)\right]$ and $\mathbf{v}^i_{t+1} = \frac{1}{n}\sum_{k=1}^{n} \left[\mathbf{v}^k_t + \eta_2 \widetilde{\nabla}_{\mathbf{v}} \widehat{\mathcal{L}}_k(\mathbf{w}^k_t, \mathbf{v}^k_t)\right]$
    end
end
Output $\overline{\mathbf{w}}_T = \frac{1}{n}\sum_{i=1}^{n} \mathbf{w}^i_T$ and $\overline{\mathbf{v}}_T = \frac{1}{n}\sum_{i=1}^{n} \mathbf{v}^i_T$
```

at least $1-\delta$ *the following holds for all* $\mathbf{w} \in \mathcal{W}$ *in* (7)

$$\Bigg| \min_{\boldsymbol{\theta}_{1:n}} \max_{\substack{\phi_{1:n}\in\Phi:\\ \forall \mathbf{x}:\, \sum_i \phi_i(\mathbf{x})=0}} \mathcal{L}(\mathbf{w}, \boldsymbol{\theta}_{1:n}, \phi_{1:n}) - \min_{\boldsymbol{\theta}_{1:n}} \max_{\substack{\phi_{1:n}\in\Phi:\\ \forall \mathbf{x}:\, \sum_i \phi_i(\mathbf{x})=0}} \widehat{\mathcal{L}}(\mathbf{w}, \boldsymbol{\theta}_{1:n}, \phi_{1:n}) \Bigg|$$
$$\leq \mathcal{O}\Bigg(L_\ell L_{\mathbf{w}} M \sqrt{\frac{\big(\mathcal{V}(\mathcal{W}) + \mathcal{V}(\Theta)\big)^2 \log(1/\delta)}{mn}} + \lambda L_\phi L_\theta M \sqrt{\frac{\big(\mathcal{V}(\Phi) + \mathcal{V}(\Theta)\big)^2 \log(n/\delta)}{m}} + \frac{M L_w L_\ell}{\lambda} \Bigg).$$

Proof: We defer the proof to the Appendix. ■

The above theorem suggests that the sample complexity will scale linearly with mn, which is the total number of samples observed in the network, under the condition that $\mathcal{V}(\Phi) + \mathcal{V}(\Theta) < \frac{\mathcal{V}(\mathcal{W})}{n}$, i.e., if the complexity measure of the classifier function space $\mathcal{W}$ is lower-bounded by the product of the number of users and the total complexity measure of Φ and Θ.

B. Optimization Guarantees

To solve `FedOT` nonconvex-strongly-concave minimax problem (9), we propose a gradient descent-ascent (GDA) method in Algorithm 1, namely `FedOT-GDA`, and further analyze its optimization properties. For the purpose of readability, we present our method and results using the following notation for the minimax formulation:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} \widehat{\mathcal{L}}(\mathbf{w}, \mathbf{v}) := \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathcal{L}}_i(\mathbf{w}, \mathbf{v}), \tag{10}$$

where each $\widehat{\mathcal{L}}_i$ denotes the local loss function corresponding to node i's samples. Here, $\mathbf{w}$ and $\mathbf{v}$ respectively denote the minimization and maximization variables described in (9), i.e. $\mathbf{w} = \{\mathbf{w}, \boldsymbol{\theta}_{1:n}\}$ and $\mathbf{v} = \{\mathbf{v}_{1:n}, \mathbf{U}\}$. We propose the following iterative GDA routine summarized in Algorithm 1. Let us denote by $(\mathbf{w}^i_t, \mathbf{v}^i_t)$ the local variable corresponding to node i at iteration t. In every round, each node i updates its local models $(\mathbf{w}^i_t, \mathbf{v}^i_t)$ using the stepsizes η_1, η_2 for τ successive iterations. Then, all updated local variables are uploaded to the parameter server and the corresponding averages are sent back to local nodes as the initial point for the next round of updates. There, $\widetilde{\nabla}_{\mathbf{w}} \widehat{\mathcal{L}}_i$ and $\widetilde{\nabla}_{\mathbf{v}} \widehat{\mathcal{L}}_i$ denote stochastic gradients of local losses w.r.t. their first and second arguments. It is important to note that `FedOT-GDA` imposes small communication (with periodic synchronization) and computation burden (by one gradient computation per iteration) on the network which is essential in federated learning methods.

As mentioned in Section III, for smooth enough loss functions, the minimax objective in (9) is nonconvex-strongly-concave. That is, $\widehat{\mathcal{L}}(\mathbf{w}, \mathbf{v})$ in (10) is nonconvex in $\mathbf{w}$ and strongly-concave in $\mathbf{v}$. The following set of assumptions formally characterizes the setting.

Assumption 1. *(i)* $\mathcal{V}$ *is a convex and bounded set with a diameter* D. *(ii) Local functions* $\widehat{\mathcal{L}}_i(\mathbf{w}, \mathbf{v})$ *have* L*-Lipchits gradients and are* μ*-strongly concave in* $\mathbf{v}$*. That is, for both* $* \in \{\mathbf{w}, \mathbf{v}\}$

$$\|\nabla_* \widehat{\mathcal{L}}_i(\mathbf{w}, \mathbf{v}) - \nabla_* \widehat{\mathcal{L}}_i(\mathbf{w}', \mathbf{v}')\|^2 \leq L^2 \left(\|\mathbf{w} - \mathbf{w}'\|^2 + \|\mathbf{v} - \mathbf{v}'\|^2\right).$$

We denote the condition number by $\kappa := L/\mu$. *(iii) (Gradient Diversity) There are constants* $\rho_{\mathbf{w}}$ *and* $\rho_{\mathbf{v}}$ *such that for both* $* \in \{\mathbf{w}, \mathbf{v}\}$*, we have that* $\frac{1}{n}\sum_{i=1}^{n} \|\nabla_* \widehat{\mathcal{L}}_i(\mathbf{w}, \mathbf{v}) - \nabla_* \widehat{\mathcal{L}}(\mathbf{w}, \mathbf{v})\|^2 \leq \rho_*^2$.

Since the global loss function $\widehat{\mathcal{L}}(\mathbf{w}, \mathbf{v})$ is nonconvex w.r.t. the minimization variable $\mathbf{w}$, we aim to find ϵ-stationary solutions for the primal function $\Lambda(\mathbf{w}) := \max_{\mathbf{v}\in\mathcal{V}} \widehat{\mathcal{L}}(\mathbf{w}, \mathbf{v})$. Next theorem characterizes the convergence rate of the proposed `FedOT-GDA` in Algorithm 1 to find a stationary solution for $\min_{\mathbf{w}\in\mathcal{W}} \Lambda(\mathbf{w})$.

Theorem 4. *Consider the iterates* $\{\mathbf{w}^i_t, \mathbf{v}^i_t\}$ *in Algorithm 1 and let Assumption 1 hold. Moreover, assume that the local stochastic gradients are unbiased and variance bounded, i.e.,* $\mathbb{E}\|\widetilde{\nabla}_* \widehat{\mathcal{L}}_i(\mathbf{w}, \mathbf{v}) - \nabla_* \widehat{\mathcal{L}}_i(\mathbf{w}, \mathbf{v})\|^2 \leq \sigma_*^2$ *for* $* \in \{\mathbf{w}, \mathbf{v}\}$*. Then, there exists iteration* $t \in \{0, \cdots, T-1\}$ *for which*

$$\mathbb{E}\|\nabla \Lambda(\overline{\mathbf{w}}_t)\|^2 \leq \mathcal{O}\Bigg(\frac{\Delta_\Lambda}{\eta_1 T} + \frac{\kappa^3 L D^2}{\eta_2 T} + \eta_1 \frac{\sigma_{\mathbf{w}}^2}{n} + \eta_2 \kappa^2 L \frac{\sigma_{\mathbf{v}}^2}{n} + \eta_\sigma^2 \kappa^2 L^2 \tau + \eta_\rho^2 \kappa^2 L^2 \tau^2 \Bigg),$$

Dataset	MNIST			CIFAR-10			Colored-MNIST	
Method	$m=50$	$m=100$	$m=500$	$m=50$	$m=100$	$m=500$	$m=50$	$m=500$
FedOT, $\tau=1$	**87.0%**	**95.6%**	**97.0%**	**42.2%**	**51.6%**	61.8%	86.0%	96.6%
FedOT, $\tau=5$	85.4%	94.4%	96.4%	40.8%	51.2%	**63.0%**	**88.6%**	**97.4%**
FedAvg, $\tau=1$	72.0%	78.4%	86.8%	22.8%	26.4%	37.2%	75.8%	90.8%
FedAvg, $\tau=5$	64.8%	72.6%	82.2%	18.8%	25.0%	36.6%	73.2%	91.4%
L-FedAvg, $\tau=1$	66.4%	74.2%	88.0%	17.8%	23.0%	39.0%	71.0%	91.2%
L-FedAvg, $\tau=5$	61.2%	71.2%	85.0%	16.0%	22.6%	36.8%	69.8%	92.0%
FedMI, $\tau=1$	64.0%	75.8%	87.4%	21.0%	27.0%	40.2%	64.0%	91.8%
FedMI, $\tau=5$	61.8%	74.0%	85.4%	17.6%	25.4%	37.0%	62.2%	92.6%
Fed-FOMAML, $\tau=1$	52.2%	81.0%	89.0%	14.8%	31.4%	46.4%	66.8%	94.6%
Fed-FOMAML, $\tau=5$	44.0%	77.8%	88.2%	12.0%	28.6%	45.6%	58.4%	94.2%

TABLE I: AlexNet results: Average test accuracy under affine distribution shifts (MNIST & CIFAR-10) and color transformations (Colored-MNIST) and different training set sizes per user m computed for FedOT vs. the baseline methods including FedAvg, Local-FedAvg (L-FedAvg), Federated Model Interpolation (FedMI), and Federated First-Order Model Agnostic Meta Learning Fed-FOMAML.

where $\eta_\sigma^2 := \eta_1^2\sigma_{\mathbf{w}}^2 + \eta_2^2\sigma_{\mathbf{v}}^2$, $\eta_\rho^2 := \eta_1^2\rho_{\mathbf{w}}^2 + \eta_2^2\rho_{\mathbf{v}}^2$ *and* $\Delta_\Lambda := \Lambda(\mathbf{w}_0) - \min_{\mathbf{w}\in\mathcal{W}} \Lambda(\mathbf{w})$.

Proof: We defer the theorem's proof to the Appendix. ■

The result of Theorem 4 indicates that after T iterations of FedOT-GDA in Algorithm 1 and for proper choices of the stepsizes $\eta_1 = \mathcal{O}(1/\sqrt{T})$ and $\eta_2 = \mathcal{O}(1/\sqrt{T})$, an ϵ-stationary solution $\overline{\mathbf{w}}$ for the min-max problem (10) (and hence (9)) can be obtained for which $\mathbb{E}\|\nabla\Lambda(\overline{\mathbf{w}})\|^2 \le \mathcal{O}(1/\sqrt{T})$. However, we still note that this result requires the inner maximization objective to be strongly-concave. Extending this result to general nonconvex-nonconcave settings is an interesting future direction to this work.

V. Numerical Results

We evaluated the empirical performance of our proposed FedOT method on standard image recognition datasets including MNIST [47], CIFAR-10 [48], and Colored-MNIST [49]. We used the standard AlexNet [50] and InceptionNet [51] neural network architectures in our experiments which we implemented in the TensorFlow platform [52]. For the federated learning setting, we used a network of $n = 100$ users and ran every experiment with three user-based training size values: $m = 50, 100, 500$. We also tested two values of $\tau = 1, 5$ for the number of local steps before every synchronization. In our experiments, we simulated the following two types of distribution shifts:

1) **Affine distribution shifts**: Here, we drew $n = 100$ random isotropic Gaussian vectors $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \sigma I_d)$ with $\sigma = 1$ and n random uniformly-distributed vectors $\mathbf{s}_i \sim \mathrm{Unif}([0.5, 1.5]^d)$ and manipulated every training sample $\mathbf{x}_{i,j}$ at the ith node as follows
$$\forall i, j:\; \mathbf{x}'_{i,j} = \mathrm{diag}\{\mathbf{s}_i\}\mathbf{x}_{i,j} + \mathbf{z}_i.$$
2) **Color-based distribution shifts**: We experimented color-based shifts on MNIST samples. Here, we used a threshold of $\zeta = 10^{-4}$ to detect near-zero pixel values for every MNIST sample. Then, we drew n pairs of uniformly-distributed vectors $\mathbf{a}_i, \mathbf{b}_i \in \mathrm{Unif}([0, 1]^3)$ (corresponding to the three RGB channels) and manipulated every pixel (l_1, l_2) as follows:
$$\forall i, j, l_1, l_2:\; \mathbf{x}'_{i,j,l_1,l_2} = \begin{cases} \mathbf{a}_i & \text{if } x_{i,j,l_1,l_2} \le \zeta, \\ x_{i,j,l_1,l_2}\mathbf{b}_i & \text{if } x_{i,j,l_1,l_2} > \zeta. \end{cases}$$

We use the insight offered by Theorem 2 to design the class of potential functions in these numerical experiments. As shown in Theorem 2, the optimal potential function will also be the integral of the optimal transport maps which will be an affine transformation under affine distribution shifts and a piecewise affine transformation under color-based distribution shifts. Therefore, we used the following class of functions Φ and Θ in our experiments:

1) For affine distribution shifts, we applied affine transformations $\psi_{\boldsymbol{\theta}_{1:n}}$ and quadratic potential functions $\phi_{1:n}$, where $\forall i$,
$$\psi_{\boldsymbol{\theta}_i}(\mathbf{x}) = \Theta_{i,1}\mathbf{x} + \boldsymbol{\theta}_{i,0},\;\; \phi_{\mathbf{v}_i}(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top V_{i,0}\mathbf{x} + \mathbf{v}_{i,1}^\top\mathbf{x},$$
$$\text{s.t.}\quad \sum_{i=1}^n V_{i,0} = \mathbf{0} \quad\text{and}\quad \sum_{i=1}^n \mathbf{v}_{i,1} = \mathbf{0}.$$
2) For color-based distribution shifts, we considered one-hidden layer neural networks with ReLU activation ($\mathrm{ReLU}(z) = \max\{z, 0\}$) for both $\psi_{\boldsymbol{\theta}_{1:n}}$ and potential functions $\phi_{1:n}$, where
$$\begin{aligned}\forall i: \psi_{\boldsymbol{\theta}_i}(\mathbf{x}) &= \mathrm{ReLU}(\Theta_{i,2}\mathbf{x} + \boldsymbol{\theta}_{i,1}) + \boldsymbol{\theta}_{i,0}, \phi_{\mathbf{v}_i}(\mathbf{x}) \\ &= \mathbf{v}_{i,2}^\top\,\mathrm{ReLU}(V_1\mathbf{x} + \mathbf{v}_0),\;\text{s.t.}\; \sum_{i=1}^n \mathbf{v}_{i,2} = \mathbf{0}.\end{aligned}$$

We used the FedOT-GDA algorithm (Algorithm 1), that is a distributed mini-batch stochastic GDA, for solving the regularized FedOT's min-max problem as formulated in Proposition 2. We used a batch-size of 20 for every user and tuned the minimization and maximization stepsize parameters $\eta_1 = \eta_2 = 10^{-4}$ while applying 10 maximization steps per minimization step. For the L_2-regularization penalty, we tuned

Dataset	MNIST			CIFAR-10			Colored-MNIST	
Method	$m=50$	$m=100$	$m=500$	$m=50$	$m=100$	$m=500$	$m=50$	$m=500$
FedOT, $\tau=1$	**76.6%**	**83.2%**	**91.0%**	**50.4%**	**59.2%**	70.4%	**77.4%**	**97.0%**
FedOT, $\tau=5$	73.0%	82.6%	90.6%	48.4%	57.8%	**72.2%**	72.0%	96.6%
FedAvg, $\tau=1$	70.8%	78.8%	84.2%	29.2%	34.6%	44.0%	69.8%	89.8%
FedAvg, $\tau=5$	66.2%	75.0%	83.4%	25.0%	32.8%	45.2%	67.2%	90.6%
L-FedAvg, $\tau=1$	67.4%	78.0%	84.6%	23.4%	32.8%	43.8%	65.4%	92.2%
L-FedAvg, $\tau=5$	63.0%	76.8%	83.8%	19.6%	30.4%	46.6%	63.6%	92.4%
FedMI, $\tau=1$	58.2%	73.6%	82.6%	23.6%	33.6%	44.8%	61.4%	92.0%
FedMI, $\tau=5$	54.6%	74.6%	83.2%	19.2%	34.0%	45.2%	59.8%	92.6%
Fed-FOMAML, $\tau=1$	58.0%	80.2%	86.6%	16.8%	34.0%	49.4%	67.0%	94.2%
Fed-FOMAML, $\tau=5$	46.2%	73.8%	86.0%	16.2%	32.6%	48.8%	65.6%	94.2%

TABLE II: InceptionNet results: Average test accuracy under affine distribution shifts (MNIST & CIFAR-10) and color transformations (Colored-MNIST) and different training set sizes per user m computed for FedOT vs. the baseline methods including FedAvg, Local-FedAvg (L-FedAvg), Federated Model Interpolation (FedMI), and Federated First-Order Model Agnostic Meta Learning Fed-FOMAML.

a coefficient of $\lambda = 4$ for the CIFAR-10 experiments and $\lambda = 1$ for the MNIST experiments. For baseline methods, we used the the following three methods: (1) standard FedAvg [1], (2) localized FedAvg (L-FedAvg) where each client personalizes the final shared model of FedAvg by locally updating it via 500 additional local iterations, (3) federated model interpolation (FedMI) [8] where each client averages the global and its own local models, and (4) federated first-order model agnostic meta learning (Fed-FOMAML) [11] applying a first-order meta learning approach to update the local models. Note that our evaluation metric is the test accuracy averaged over the individual distributions of the $n = 100$ nodes.

Table I includes the test accuracy scores of our experiments with the AlexNet architecture. In these experiments, we applied affine distribution shifts for the MNIST and CIFAR-10 experiments and used color transformation shifts for the Colored-MNIST experiments. As shown by our numerical results, FedOT consistently outperformed the baseline methods in all the experiments and with a definitive margin which was above 15% in six of the eight experimental settings. Similarly, Table II shows that FedOT also achieves the best performance for the InceptionNet architecture. Overall, our numerical results indicate that FedOT can lead to a significant performance improvement when the learners can learn and reverse the underlying distribution shifts via the optimal transport-based framework.

VI. Conclusion

In this paper, we introduced the optimal transport-based FedOT framework to address the federated learning problem under heterogeneous data distributions. The FedOT framework leverages multi-input optimal transport costs to measure the discrepancy between the input distributions and also learn the transportation maps needed for transferring the input distributions to a common probability domain. We demonstrated that such a transportation to a common distribution offers an improved generalization and optimization performance in learning the personalized prediction models. In addition, the optimal transport-based analysis results in an upper-bound on the statistical complexity of the federated learning problem. The applied approach can be potentially useful for bounding the sample complexity of learning under heterogeneous data distributions which appear in other transfer and meta learning settings, and can complement information theoretic tools for deriving lower bounds on the statistical complexity. An interesting future direction is to analyze the tightness of the generalization error bound in Section IV through developing information theoretic lower-bounds on the sample complexity of learning under different input distributions.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[2] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[3] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, 2018.

[4] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[5] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.

[6] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for on-device federated learning," *arXiv preprint arXiv:1910.06378*, 2019.

[7] Y. Deng, M. M. Kamani, and M. Mahdavi, "Adaptive personalized federated learning," *arXiv preprint arXiv:2003.13461*, 2020.

[8] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh, "Three approaches for personalization with applications to federated learning," *arXiv preprint arXiv:2002.10619*, 2020.

[9] F. Hanzely and P. Richtárik, "Federated learning of a mixture of global and local models," *arXiv preprint arXiv:2002.05516*, 2020.

[10] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," *Advances in neural information processing systems*, vol. 30, 2017.

[11] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[12] Y. Jiang, J. Konečnỳ, K. Rush, and S. Kannan, "Improving federated learning personalization via model agnostic meta learning," *arXiv preprint arXiv:1909.12488*, 2019.

This article has been accepted for publication in IEEE Journal on Selected Areas in Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2022.3182355

[13] P. P. Liang, T. Liu, L. Ziyin, N. B. Allen, R. P. Auerbach, D. Brent, R. Salakhutdinov, and L.-P. Morency, "Think locally, act globally: Federated learning with local and global representations," *arXiv preprint arXiv:2001.01523*, 2020.

[14] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," *arXiv preprint arXiv:2102.07078*, 2021.

[15] A. Shamsian, A. Navon, E. Fetaya, and G. Chechik, "Personalized federated learning using hypernetworks," *arXiv preprint arXiv:2103.04628*, 2021.

[16] A. Ghosh, J. Hong, D. Yin, and K. Ramchandran, "Robust federated learning in a heterogeneous environment," *arXiv preprint arXiv:1906.06629*, 2019.

[17] M. Xie, G. Long, T. Shen, T. Zhou, X. Wang, and J. Jiang, "Multi-center federated learning," *arXiv preprint arXiv:2005.01026*, 2020.

[18] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," *arXiv preprint arXiv:2006.04088*, 2020.

[19] T. Diamandis, Y. Eldar, A. Fallah, F. Farnia, and A. Ozdaglar, "A wasserstein minimax framework for mixed linear regression," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2697–2706.

[20] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "Fedbn: Federated learning on non-iid features via local batch normalization," *arXiv preprint arXiv:2102.07623*, 2021.

[21] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4615–4625.

[22] A. Reisizadeh, F. Farnia, R. Pedarsani, and A. Jadbabaie, "Robust federated learning: The case of affine distribution shifts," *arXiv preprint arXiv:2006.08907*, 2020.

[23] Y. Deng, M. M. Kamani, and M. Mahdavi, "Distributionally robust federated averaging," *arXiv preprint arXiv:2102.12660*, 2021.

[24] T. Lin, C. Jin, and M. I. Jordan, "On gradient descent ascent for nonconvex-concave minimax problems," *arXiv preprint arXiv:1906.00331*, 2019.

[25] J. Yang, N. Kiyavash, and N. He, "Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems," *arXiv preprint arXiv:2002.09621*, 2020.

[26] M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn, "Solving a class of non-convex min-max games using iterative first order methods," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 905–14 916.

[27] Y. Deng and M. Mahdavi, "Local stochastic gradient descent ascent: Convergence analysis and communication efficiency," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1387–1395.

[28] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.

[29] M. Sanjabi, J. Ba, M. Razaviyayn, and J. D. Lee, "On the convergence and robustness of training gans with regularized optimal transport," *arXiv preprint arXiv:1802.08249*, 2018.

[30] S. Feizi, F. Farnia, T. Ginart, and D. Tse, "Understanding gans in the lqg setting: Formulation, generalization and stability," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 304–311, 2020.

[31] J. Lee and M. Raginsky, "Minimax statistical learning with wasserstein distances," *arXiv preprint arXiv:1705.07815*, 2017.

[32] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," in *Operations Research & Management Science in the Age of Analytics*. INFORMS, 2019, pp. 130–166.

[33] J. Blanchet, Y. Kang, and K. Murthy, "Robust wasserstein profile inference and applications to machine learning," *Journal of Applied Probability*, vol. 56, no. 3, pp. 830–857, 2019.

[34] S. Kolouri, G. K. Rohde, and H. Hoffmann, "Sliced wasserstein distance for learning gaussian mixture models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3427–3436.

[35] Y. Balaji, R. Chellappa, and S. Feizi, "Normalized wasserstein for mixture distributions with applications in adversarial learning and domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6500–6508.

[36] S. P. Singh and M. Jaggi, "Model fusion via optimal transport," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[37] B. Pass, "Multi-marginal optimal transport: theory and applications," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 49, no. 6, pp. 1771–1790, 2015.

[38] J. Cao, L. Mo, Y. Zhang, K. Jia, C. Shen, and M. Tan, "Multi-marginal wasserstein gan," *arXiv preprint arXiv:1911.00888*, 2019.

[39] L. Hui, X. Li, J. Chen, H. He, and J. Yang, "Unsupervised multi-domain image translation with domain-specific encoders/decoders," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2044–2049.

[40] M. Cuturi and A. Doucet, "Fast computation of wasserstein barycenters," in *International conference on machine learning*. PMLR, 2014, pp. 685–693.

[41] S. Claici, E. Chien, and J. Solomon, "Stochastic wasserstein barycenters," in *International Conference on Machine Learning*. PMLR, 2018, pp. 999–1008.

[42] A. Kroshnin, N. Tupitsa, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and C. Uribe, "On the complexity of approximating wasserstein barycenters," in *International conference on machine learning*. PMLR, 2019, pp. 3530–3540.

[43] C. Villani, *Optimal transport: old and new*. Springer, 2009, vol. 338.

[44] G. Carlier and I. Ekeland, "Matching for teams," *Economic theory*, vol. 42, no. 2, pp. 397–418, 2010.

[45] R. J. McCann and N. Guillen, "Five lectures on optimal transportation: geometry, regularity and applications," *Analysis and geometry of metric measure spaces: lecture notes of the séminaire de Mathématiques Supérieure (SMS) Montréal*, pp. 145–180, 2011.

[46] V. M. Panaretos and Y. Zemel, "Statistical aspects of wasserstein distances," *Annual review of statistics and its application*, vol. 6, pp. 405–431, 2019.

[47] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[48] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[49] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[52] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

Farzan Farnia (Member, IEEE) received the first and second bachelor's degrees in electrical engineering and mathematics from Sharif University of Technology in 2013, the master's degree in electrical engineering from Stanford University in 2015, and the Ph.D. degree in Electrical Engineering from Stanford University in 2019. He is currently an Assistant Professor of Computer Science and Engineering at The Chinese University of Hong Kong. Prior to joining CUHK, he was a postdoctoral research associate at the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, between 2019-2021. His research interests span statistical learning theory, information theory, and convex optimization. He has been the recipient of the Stanford Graduate Fellowship (Sequoia Capital Fellowship) between 2013-2016 and the Numerical Technology Founders Prize as the second top performer of Stanford's electrical engineering PhD qualifying exams in 2014.

Amirhossein Reisizadeh (Member, IEEE) is currently a Postdoctoral Associate at the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology. He received his Ph.D. and master's degree both in Electrical and Computer Engineering from UC Santa Barbara and UCLA, in 2021 and 2016, respectively. He also received his bachelor's degree from Sharif University of Technology, Tehran, Iran, in 2014. He was a finalist in the Qualcomm Innovation Fellowship program in 2019. His primary research interests include optimization for machine learning, distributed and federated learning, and data-driven decision making.

Ramtin Pedarsani (Senior Member, IEEE) is an Assistant Professor in the ECE Department at the University of California, Santa Barbara. He received the B.Sc. degree in electrical engineering from the University of Tehran, Tehran, Iran, in 2009, the M.Sc. degree in communication systems from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 2011, and his Ph.D. from the University of California, Berkeley, in 2015. His research interests include machine learning, information and coding theory, networks, and transportation systems. Ramtin is a recipient of the IEEE Communications Society and Information Theory Society joint paper award in 2020 and the IEEE international conference on communications (ICC) best paper award in 2014.

Ali Jadbabaie (Fellow, IEEE) received the B.S. degree in electrical engineering with a focus on control systems from the Sharif University of Technology, Tehran, Iran, in 1995 the M.S. degree in electrical and computer engineering from the University of New Mexico, Albuquerque, NM, USA, in 1997 and the Ph.D. degree in control and dynamical systems from the California Institute of Technology, Pasadena, CA, USA, in 2000.

He is currently the JR East Professor and Head of the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, where he is also a core Faculty Member with the Institute for Data, Systems, and Society (IDSS) and a Principal Investigator with the Laboratory for Information and Decision Systems. Previously, he served as the Director of the Sociotechnical Systems Research Center and as the Associate Director of the IDSS, MIT, which he helped found in 2015. He was a Postdoctoral Scholar with Yale University before joining the faculty at the University of Pennsylvania, where he was subsequently promoted through the ranks and held the Alfred Fitler Moore Professorship in network science with the Department of Electrical and Systems Engineering with secondary appointments in computer and information science and operations, information, and decisions in the Wharton School. As a member of the General Robotics, Automation, Sensing and Perception Lab, University of Pennsylvania, he was also the co-founder and Director of the Raj and Neera Singh Program in Networked and Social Systems Engineering, a new undergraduate interdisciplinary degree program. His current research interests include the interplay of dynamic systems and networks with specific emphasis on multiagent coordination and control, distributed optimization, network science, and network economics. Prof. Jadbabaie was the inaugural Editor-in-Chief of the IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, a new interdisciplinary journal sponsored by several IEEE societies. He is a recipient of a National Science Foundation Career Award, an Office of Naval Research Young Investigator Award, the O. Hugo Schuck Best Paper Award from the American Automatic Control Council, and the George S. Axelby Best Paper Award from the IEEE Control Systems Society. He is a senior author of several student best paper awards at IEEE CDC, ACC, and ICASSP conferences. In 2016, He received a Vannevar Bush Fellowship from the Office of Secretary of Defense.