FEDDRO: DISTRIBUTIONALLY ROBUST FEDERATED LEARNING WITH WASSERSTEIN BARYCENTER

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Learning (FL) has emerged as a privacy-preserving approach for collaboratively training models without sharing raw data, while a key challenge is that the data across the clients may not be identically distributed. The nominal distribution that the model truly learns is commonly assumed as the Euclidean barycenter. In this paper, we propose **Fed**erated **D**istributionally **R**obust **O**ptimization (FedDRO) that constructs the Wasserstein barycenter among all distributions with a Wasserstein ball as an ambiguity set. We reformulate this paradigm as a min-max optimization problem that trains a robust FL model in an adversarial way and analyze its generalization and optimization properties.

1 PROBLEM FORMULATION

In this paper, we consider a learning scenario where N local clients are connected to a single parameter server. Each client $i \in [1, N]$ observes m_i training samples $\{\mathbf{x}_{i,j}, y_{i,j}\}_{j=1}^{m_i}$ which are independently sampled from distribution P_i. The centralized model is trained to minimize the loss w.r.t. the uniform mixture distribution $\mathcal{U} = \sum_{i=1}^{N} \frac{m_i}{\sum_{i=1}^{N} m_i} P_i$ as FedAvg McMahan et al. (2017). Furthermore, Mohri et al. (2019) proposes AFL to optimize the worst-case w.r.t.the different weight λ_i to construct the weighted average distribution such that the Empirical Risk Problem (ERP) is

$$\min_{h_{\mathbf{w}}} \sup_{\lambda} \mathbb{E}_{(x,y)\sim\mathcal{U}_\lambda}\big[\ell(h_{\mathbf{w}}(x), y\big], \quad \mathcal{U}_\lambda = \sum\nolimits_{i=1}^{N} \lambda_i P_i. \tag{1}$$

The mixture distribution is actually the Euclidean barycenter among all N empirical distributions $P_i, i \in [1, N]$ such that $\mathcal{U}_\lambda = \arg\min_P \sum_{i=1}^{N} \lambda_i ||P_i - P||_2^2$. However, for high-dimensional complex data structures and heterogeneous distributions, thpe Euclidean distance is sensitive to shifted distributions and could not potentially capture the complicated information Cuturi & Doucet (2014).

Considering the limitations of Euclidean barycenter, we choose the Wasserstein distance as a robust measure to quantify the divergence of the distributions Zhu et al. (2023). Following this assumption, the nominal distribution is replaced with the Wasserstein barycenter. Considering the potential mismatch between the nominal distribution and the true distribution, we utilize the distributionally robust optimization (DRO) by introducing an ambiguity set $\mathcal{B}(\mathcal{Q}_{\lambda,p}, \epsilon)$. Therefore, the ERP is formulated as

$$\min_{h_{\mathbf{w}}} \sup_{\mathcal{P}\in\mathcal{B}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}\big[\ell(h_{\mathbf{w}}(\mathbf{x}), y)\big]$$

$$\text{s.t. } \mathcal{B}(\mathcal{Q}_{\lambda,p}, \epsilon) = \big\{\mathcal{P} \in \mathbb{P}(\Xi) : \mathcal{W}_p^p(\mathcal{P}, \mathcal{Q}_{\lambda,p}) \le \epsilon^p\big\}, \quad \mathcal{Q}_{\lambda,p} = \arg\min_{Q} \sum_{k=1}^{N} \lambda_i \mathcal{W}_p^p(P_i, Q), \tag{2}$$

where $\mathcal{W}_p$ is the p-Wasserstein distance. To solve this optimization problem, the first step is to approximate the Wasserstein barycenter $\mathcal{Q}_{\lambda,p}$ among multiple distributions $P_1, \cdots, P_N$ within the federated context. Recently Rakotomamonjy et al. (2023) proposes the interpolating measure to calculate the Wasserstein distance in a Federated scenario and Li et al. (2023) extends this work to approximate the Wasserstein barycenter with the augmentated matrix proposed in Alvarez-Melis & Fusi (2020). However, the augmented matrix is constructed by the features $\mathbf{x}$ and the statistic information of conditional feature distribution $P(\mathbf{x}|Y = y)$ which is assumed to follow the Gaussian distribution $\mathcal{N}(m_y, \Sigma_y)$. In our paper, we need to construct the data clouds $\{\mathbf{x}_\mathcal{B}, y_\mathcal{B}\}$ following

	Test Accuracy			
$\|\|\Lambda\|\|$	0.5	2	3	5
FedAvg	87.4	55.9	48.6	8.3
Ours	**90.9**	**75.1**	**66.2**	**18.5**

Table 1: Test Accuracy on balanced test dataset.

the distributions in $\mathcal{B}(\mathcal{Q}_{\lambda,p}, \epsilon)$. Therefore, we consider two different applications: (1) Class-wise interpolating measures of feature space $\mathcal{X}$, which is applied for the heterogeneous feature space; (2) Data-wise interpolating measure $(\mathcal{X}, \mathcal{Y})$ inspired by the dictionary learning with one-hot encoded labels Fernandes Montesuma et al. (2023). Inspired by Li et al. (2023), Wasserstein distance between P_i and the approximated Wasserstein barycenter $\hat{\mathcal{Q}}$ with uniform λ_i is iteratively optimized by

$$\mathcal{W}_p(P_i, \hat{\mathcal{Q}}) \leq \mathcal{W}_p(P_i, \eta_{P_i}^{(k)}) + \mathcal{W}_p(\eta_{P_i}^{(k)}, \gamma_i^{(k-1)}) + \mathcal{W}_p(\gamma_i^{(k-1)}, \eta_{Q_i}^{(k)}) + \mathcal{W}_p(\eta_{Q_i}^{(k)}, \hat{\mathcal{Q}}^{(k-1)}), \quad (3)$$

where η_{P_i} is the interpolating measure between P_i and γ_i computed by i-th client, η_{Q_i} is the interpolating measure between γ_i and $\hat{\mathcal{Q}}$ computed by the server. Only η_{Q_i} and γ_i are shared for approximations. The server initializes $\gamma_i^{(0)}$ and sends it to i-th client. At each round k, i-th client computes $\mathcal{W}_p(P_i, \gamma_i^{(k-1)})$ and constructs $\eta_{P_i}^{(k)}$. The server computes $\mathcal{W}_p(\gamma_i^{(k-1)}, \hat{\mathcal{Q}}^{(k-1)})$ and shares $\eta_{Q_i}^{(k)}$ with i-th client. Then $\gamma_i^{(k)}$ is updated by i-th client via Rakotomamonjy et al. (2023)

$$\gamma_i^{(k)} \in \arg\min \left[\mathcal{W}_p(\eta_{P_i}^{(k)}, \gamma_i^{(k-1)}) + \mathcal{W}_p(\gamma_i^{(k-1)}, \eta_{Q_i}^{(k)})\right]. \quad (4)$$

Simultaneously, the server updates $\hat{\mathcal{Q}}^{(k)}$ utilizing all $\gamma_i^{(k)}$ based on Cuturi & Doucet (2014). Based on the optimal transport theory, suppose $\mathbf{T}^{\gamma_i}$ is the transportation map between $\hat{\mathcal{Q}}^{(K)}$ and γ_i, then we have $\mathbf{T}_{\#}^{\gamma_i} P_i \overset{\text{dist}}{=} \mathbf{T}_{\#}^{\gamma_j} P_j, \forall i \neq j$. For distributed training, the server could either share the transportation map $\mathbf{T}^{\gamma_i}$ to i-th client or the mapped samples at the last round of Wasserstein barycenter approximation procedure, in which the constructed samples are simply denoted as $\mathcal{Q}_i^{(K)} := \mathbf{T}_i(\mathbf{x}_i)$. Then with Lagrange multiplier $\bar{\lambda} > 0$, the ERM objective in equation 2 is reformulated as follows

$$\min_{\mathbf{w}} \sup_{\theta_i} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[\ell\Big(h_{\mathbf{w}}(\mathbf{T}_i(\mathbf{x}_i) + \theta_i), \mathbf{y}_i \Big) - \bar{\lambda} ||\theta_i||^p \right] \right\}. \quad (5)$$

The reformulation details for the above objective are shown in Appendix B. We summarize our algorithm in Algorithm 1, in which lines 1-7 are approximations of $\mathcal{Q}_{\lambda,p}$, and lines 8-15 are adversarial training in FL.

2 Toy Experiments

In this section, we will show our exploration of the Federated labeled Wasserstein barycenter. Then we conduct a simple comparison to show the validation performance based on the Wasserstein barycenter and Eucludiean barycenter. The technique to solve WDRO is our future exploration.

We simulate the affine transformation $\Lambda\mathbf{x} + \delta$ on the MNIST dataset with 5 clients. For each client, the δ noise is within the range $\{5, 15, 25, 35, 45\}\%$, and the Λ is also random. We calculate the class-wise interpolating measures of feature space $\mathcal{X}$ and approximate the Wasserstein barycener $\mathcal{Q}$ for each class, denoted as $\mathcal{Q} = \{\mathcal{Q}(v)\}_{v=0}^{9}$. We compare the training loss on the $\mathcal{Q}$ with the training loss on the original data via the CNN model in Figure 1 in Appendix. The testing accuracy on the clean MNIST dataset is shown in Table 1.

3 Conclusions

Our paper explores the applications of Wasserstein barycenter to enhance the robustness of Federated Learning (FL) in heterogeneous scenarios. We present `FedDRO`, a framework that leverages the efficient approximation of Wasserstein barycenter within a Federated context based on the advantageous properties of Geodesics in Optimal Transport theory, and adversarial training to solve the WDRO problem during the training procedure.