# 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 $\lambda_i$ 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 $\lambda_i$ 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 $\eta_{P_i}$ is the interpolating measure between $P_i$ and $\gamma_i$ computed by $i$-th client, $\eta_{Q_i}$ is the interpolating measure between $\gamma_i$ and $\hat{\mathcal{Q}}$ computed by the server. Only $\eta_{Q_i}$ and $\gamma_i$ 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 $\gamma_i$, 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 $\delta$ noise is within the range $\{5, 15, 25, 35, 45\}\%$, and the $\Lambda$ 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.

## URM STATEMENT

The first two authors are non-white and outside the range of 30-50 years. All authors are not located in North America, Western Europe and UK, or East Asia.

## REFERENCES

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pp. 685–693. PMLR, 2014.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020a.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. *Advances in neural information processing systems*, 33:15111–15122, 2020b.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

Eduardo Fernandes Montesuma, Fred Ngolè Mboula, and Antoine Souloumiac. Multi-source domain adaptation through dataset dictionary learning in wasserstein space. *arXiv e-prints*, pp. arXiv–2307, 2023.

Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 2022.

Joel Goh and Melvyn Sim. Distributionally robust optimization and its tractable approximations. *Operations research*, 58(4-part-1):902–917, 2010.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pp. 130–166. Informs, 2019.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021.

Wenqian Li, Shuran Fu, Fengrui Zhang, and Yan Pang. Data valuation and detections in federated learning. *arXiv preprint arXiv:2311.05304*, 2023.

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*, pp. 1273–1282. PMLR, 2017.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pp. 4615–4625. PMLR, 2019.

Tung-Anh Nguyen, Tuan Dung Nguyen, Long Tan Le, Canh T Dinh, and Nguyen H Tran. On the generalization of wasserstein robust federated learning. *arXiv preprint arXiv:2206.01432*, 2022.

Alain Rakotomamonjy, Kimia Nadjahi, and Liva Ralaivola. Federated wasserstein distance. *arXiv preprint arXiv:2310.01973*, 2023.

Amirhossein Reisizadeh, Farzan Farnia, Ramtin Pedarsani, and Ali Jadbabaie. Robust federated learning: The case of affine distribution shifts. *Advances in Neural Information Processing Systems*, 33:21554–21565, 2020.

Jiacheng Zhu, Jielin Qiu, Aritra Guha, Zhuolin Yang, XuanLong Nguyen, Bo Li, and Ding Zhao. Interpolation for robust learning: Data augmentation on geodesics. *arXiv preprint arXiv:2302.02092*, 2023.

## A Related Work

Distributionally Robust Optimization (DRO) is a powerful framework that explicitly accounts for uncertainty in the underlying probability distributions of the problem parameters (Delage & Ye (2010);Goh & Sim (2010)). Wasserstein distance, a metric measuring the divergence of different distributions, is introduced into the DRO and has been utilized to seek data-driven optimal decisions in recent works (Mohajerin Esfahani & Kuhn (2018);Kuhn et al. (2019);Gao et al. (2022)). The Wasserstein Distributionally Robust Optimization (WDRO) has been theoretically justified by Kuhn et al. (2019) that it has many conceptual and computational benefits, and it has been widely applied to solve certain machine learning tasks such as Blanchet et al. (2019) and Gao & Kleywegt (2023). For a set of probability distributions, the Wasserstein Barycenter represents a distribution that minimizes the sum of the Wasserstein distances to the given distributions.

Federated Learning (FL) has emerged as a privacy-preserving approach for collaboratively training models without sharing raw data, however, robust federated learning remains a challenging problem because of the nature of non-i.i.d. data from clients' devices (Kairouz et al. (2021)). Typical solutions for this problem include personalized federated learning, where a personalized model is adapted to each client (Fallah et al. (2020);Deng et al. (2020a);Li et al. (2021)), and distributionally robust models, where the global model is trained using a worst-case objective over an ambiguity set and can deliver uniformly good performance for all clients (Mohri et al. (2019);Reisizadeh et al. (2020);Deng et al. (2020b)). Nguyen et al. (2022) utilized the WDRO scheme in federated learning and empirically demonstrated its robustness in distribution shift settings. In this work, we explore the application of Wasserstein Barycenter in robust FL.

## B Reformulation of our problem

We follow the techniques to solve the Wasserstein Distributionally Robust Optimization problem in Kuhn et al. (2019). Consider the min-max problem as follows,

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

$$\text{s.t. } \mathcal{B}(\mathcal{Q}_{\lambda,p},\epsilon)=\big\{\mathcal{P}\in\mathbb{P}(\Xi):\mathcal{W}_p^p(\mathcal{P},\mathcal{Q}_{\lambda,p})\le\epsilon^p\big\},\quad \mathcal{Q}_{\lambda,p}=\arg\min_{Q}\sum_{k=1}^{N}\lambda_i\mathcal{W}_p^p(P_i,Q) \tag{6}$$

To fit the distributed training setting, we assume $i$-th client holds $\mathbf{T}^{\eta_i}$ ( transportation map between $P_i$ and $\gamma_i$ ) and $\mathbf{T}^{\gamma_i}$ ( transportation map between $\gamma_i$ and $\mathcal{Q}_{\lambda,p}=\hat{\mathcal{Q}}^{(K)}$). Then the mapped samples are $\mathbf{T}^{\gamma_i}_{\#}(\gamma_i)$, follows $\mathcal{Q}_{\lambda,p}$, abbreviated as $\mathbf{T}_i(x)$. Then we could restrict the original wasserstein ball to a subset that contains only perturbed samples of the form

$$\mathcal{B}(\Theta)=\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\delta_{\mathbf{T}_i(\mathbf{x}_{i,j})+\theta_i},\ \ \mathbf{T}_i(\mathbf{x}_{i,j})+\theta_i\in\Xi \tag{7}$$

where $\theta_i\in\mathbb{R}^m$ is the displacement of samples from $i$-th client. Therefore, all distributions in $\mathcal{B}(\mathcal{Q}_\lambda,\epsilon,p)$ are encoded by a perturbation matrix $\Theta=(\theta_1,\cdots,\theta_N)\in\mathbb{R}^{m\times N}$. Based on the Theorem 6 in Kuhn et al. (2019), the worst case risk of any fixed loss function $\ell\in\mathcal{L}$ is bounded with $L_{p,1}$-norm uncertainty set, that is, the wasserstein constraint $\mathcal{W}_p(\mathcal{P},\mathcal{Q}_{\lambda,p})\le\epsilon$ is equivalent to the inequality $||\theta_i||^p\le\epsilon^p$. Therefore, with the assumption that $\ell(h_{\mathbf{w}})$ is concave, the ERM of the inner sup is defined as

$$\sup_{\theta_i}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\ell\Big(h_{\mathbf{w}}(\mathbf{T}_i(x_{i,j})+\theta_i),y_{i,j}\Big)$$
$$\text{s.t. } \theta_i\in\mathbb{R}^m,\ ||\theta_i||^p\le\epsilon^p,\ \mathbf{T}_i(\mathbf{x}_{i,j})+\theta_i\in\Xi,\ \forall i\in[1,N]. \tag{8}$$

Then given a Lagrange multiplier $\bar{\lambda}>0$, we have:

$$\sup_{\theta_i}\Big\{\frac{1}{N}\sum_{i=1}^{N}\Big[\sum_{j=1}^{M}\ell\Big(h_{\mathbf{w}}(\mathbf{T}_i(\mathbf{x}_{i,j})+\theta_i),y_{i,j}\Big)-\bar{\lambda}||\theta_i||^p\Big]\Big\} \tag{9}$$

Here we use a norm penalty requiring a bounded distance and find the worst-case transformation that results in the maximum loss for the samples of $i$-th client.

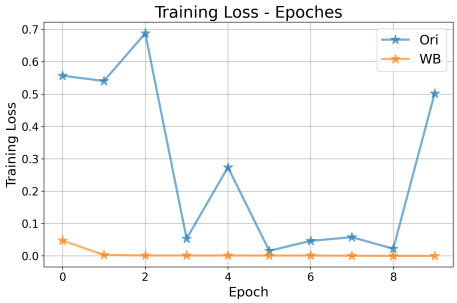


Figure 1: Training loss on Wasserstein barycenter and original data

**Algorithm 1** `FedDRO`

**Input:** Initialisation of $\mathcal{Q}^{(0)}, \gamma_i^{(0)}, \mathbf{w}_i^{(0)}, \forall i = [1, N], \eta_1, \eta_2, \tau, T$

**for** $k = 1$ to $K$ **do**
  Clients compute distance $\mathcal{W}_p(P_i, \gamma_i^{(k-1)})$ and construct $\eta_{P_i}^{(k)}$
  Server computes distance $\mathcal{W}_p(\mathcal{Q}^{(k-1)}, \gamma_i^{(k-1)})$ and sends $\eta_{Q_i}^{(k)}$ to $i$-th client
  The $i$-th client updates $\gamma_i^{(k)}$ based on $\eta_{P_i}^{(k)}$ and $\eta_{Q_i}^{(k)}$
  Server updates $\mathcal{Q}^{(k)}$ based on all $\gamma_i^{(k)}$s
**end for**
Server sends $\mathcal{Q}_i^{(K)}$ or $\mathbf{T}^{\gamma_i}$ to $i$-th client
**for** $t = 0$ to $T - 1$, $i$-th client computes **do**
  $\theta_i^{(t+1)} = \theta_i^{(t)} + \eta_1 \nabla_\theta h_i(\mathbf{w}_i^{(t)}, \theta_i^{(t)})$
  **if** $t$ does not divide $\tau$ **then**
    $\mathbf{w}_i^{(t+1)} = \mathbf{w}_i^{(t)} - \eta_2 \nabla_{\mathbf{w}} h_i(\mathbf{w}_i^{(t)}, \theta_i^{(t)})$
  **else** $i$-th client sends $\mathbf{w}_i^{(t)} - \eta_2 \nabla_{\mathbf{w}} h_i(\mathbf{w}_i^{(t)}, \theta_i^{(t)})$ to the server
    Server updates $\mathbf{w}_i^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \left[\mathbf{w}_i^{(t)} - \eta_2 \nabla_{\mathbf{w}} h_i(\mathbf{w}_i^{(t)}, \theta_i^{(t)})\right]$
  **end if**
**end for**

**Output:** $\bar{\mathbf{w}}^{(T)} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{w}_i^{(T)}$