Federated Calculation of the Transportation Barycenter by a Dual Subgradient Method

Zhengqi Lin and Andrzej Ruszczyński

Department of Management Science and Information Systems
Rutgers University, Piscataway, NJ 08854, USA
zhengqi.lin@rutgers.edu; rusz@rutgers.edu

Abstract

We propose an efficient federated dual decomposition algorithm for calculating the free-support Wasserstein barycenter of several distributions. The algorithm does not have access to local data and uses only highly aggregated information. It avoids repeated solutions of mass transportation problems. Owing to the absence of any matrix-vector operations, the algorithm exhibits very low complexity of each iteration and significant scalability. We illustrate its virtues on mixture models.

1 Introduction

The transportation distance, often called the Wasserstein metric, is a powerful tool for comparing probability measures [15]. For $p \geq 1$, the p-th order distance between two measures $\mu, \nu \in \mathcal{P}_p(\mathcal{Y})$ is:

$$W_p(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{Y}\times\mathcal{Y}} d(y,z)^p \, \pi(\mathrm{d}y, \mathrm{d}z) \right)^{1/p}, \tag{1}$$

where $\Pi(\mu,\nu)$ is the set of all joint measures with marginals μ and ν. Its applications are widespread, from generative models [2] to clustering [7] and finance [12, 13].

In this work, we consider a family of probability measures $\{Q(x)\}_{x\in\mathcal{X}}$, parameterized by x from a metric space $\mathcal{X}$. Given a distribution λ on $\mathcal{X}$, we can define an average transportation distance between the entire family Q and a single measure $q \in \mathcal{P}_p(\mathcal{Y})$ as [11]:

$$\mathcal{W}_p^\lambda(Q,q) = \left(\int_{\mathcal{X}} \left[W_p(Q(x),q) \right]^p \, \lambda(\mathrm{d}x) \right)^{1/p}. \tag{2}$$

Our central problem is to find the *transportation barycenter*, which is the measure q that minimizes this average distance, that is, solves the following problem:

$$\min_{q\in\mathcal{P}_p(\mathcal{Y})} \mathcal{W}_p^\lambda(Q,q). \tag{3}$$

This generalizes the Wasserstein barycenter problem introduced in [1]. Such a problem is highly relevant in Federated Learning settings [14], where x might represent a client (e.g., a hospital, a regional office) and $Q(x)$ is the local data distribution. Due to privacy and confidentiality constraints, the local distributions cannot be shared with a central server. This prevents the direct calculation of a simple mixture distribution and motivates the need for algorithms and frameworks that preserve privacy [10]. Barycenters play a key role in fairness in data analysis and learning [6, 4].

Existing algorithms for computing Wasserstein barycenters [3, 8, 5, 9] are not designed for the federated setting, as they typically require direct access to all distributions. They also require repeated and time-consuming solutions of transportation problems. We propose a novel federated dual subgradient algorithm to solve the barycenter problem (3) without accessing local data, ensuring privacy while maintaining computational efficiency and scalability.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ML×OR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

2 Discretization of the Barycenter Problem

We assume that the spaces $\mathcal{X}$ and $\mathcal{Y}$ are finite. We consider a discrete distribution λ supported on a set of clients $\{x^s\}_{s=1,\dots,N}$. Each client s holds a private local data distribution $Q(x^s)$, supported on particles y^{si} with probabilities p_{si}, $i \in \mathcal{I}^s$.

We seek a barycenter $\bar{q}$ that is a uniform discrete measure supported on M points selected from a large, pre-defined set of candidate locations $\mathcal{Z} = \{\zeta^k\}_{k=1,\dots,K} \subset \mathcal{Y}$. Our goal is to find the optimal cloud of M particles $\widetilde{\mathcal{Z}} \subset \mathcal{Z}$ that solves the discrete approximation of (3).

We introduce binary decision variables $\gamma_k = 1$ if a point ζ^k is selected for the barycenter's support ($\zeta^k \in \widetilde{\mathcal{Z}}$), and $\gamma_k = 0$ otherwise. The barycenter is then $\bar{q} = \frac{1}{M}\sum_{k=1}^{K} \gamma_k \delta_{\zeta^k}$. The pth power of the Wasserstein distance $[W_p(Q(x^s), \bar{q})]^p$ is the optimal value of a linear transportation problem. By rescaling the transportation plan variables, we formulate the search for the optimal set $\widetilde{\mathcal{Z}}$ as the following linear mixed-integer programming problem:

$$\min_{\gamma,\beta} \sum_{s=1}^{N} w_s \sum_{i\in\mathcal{I}^s} \sum_{k=1}^{K} d_{sik}\beta_{sik} \tag{4a}$$

$$\text{s.t.} \sum_{i\in\mathcal{I}^s} \beta_{sik} = \gamma_k, \quad s = 1,\dots,N, \quad k = 1,\dots,K, \tag{4b}$$

$$\sum_{k=1}^{K} \beta_{sik} = p_{si} \sum_{k=1}^{K} \gamma_k, \quad s = 1,\dots,N, \quad i \in \mathcal{I}^s, \tag{4c}$$

$$\sum_{k=1}^{K} \gamma_k = M, \tag{4d}$$

$$\beta_{sik} \geq 0,\ \gamma_k \in \{0,1\}, \quad \forall s,i,k. \tag{4e}$$

Here, $d_{sik} = d(y^{si}, \zeta^k)^p$, $w_s = \lambda_s/M$, and β_{sik} are the mass flows from $y^{s,i}$ to ζ^k multiplied by $\sum_{j=1}^{K} \gamma_j$. This problem is computationally intractable for large N, $|\mathcal{I}^s|$, and K by standard solvers, and it violates privacy constraints, as it requires all d_{sik} values at the central location.

3 A Federated Dual Subgradient Algorithm

To solve (4) in a federated and efficient manner, we develop a dual decomposition method. The key innovation is that we introduce not only *global Lagrange multiplier* θ_0 for the cardinality constraint (4d), but also *local dual variables* $\{\theta_{si}\}_{i\in\mathcal{I}^s}$ for each client s, associated with the local mass conservation constraints (4c). These are kept private to each client.

The Lagrangian function is:

$$L(\gamma,\beta;\theta) = \sum_{s,i,k} w_s d_{sik}\beta_{sik} + \sum_{s,i} \theta_{si}\left(p_{si}\sum_k \gamma_k - \sum_k \beta_{sik}\right) + \theta_0\left(\sum_k \gamma_k - M\right). \tag{5}$$

The calculation of the corresponding dual function $L_D(\theta) = \min_{\gamma,\beta} L(\gamma,\beta;\theta)$ subject to (4b) and (4e) decomposes into K independent subproblems, one for each candidate point ζ^k. It has a closed-form solution, owing to the relaxation of the constraints (4c).

The optimal γ_k for each candidate point is determined by a simple rule:

$$\gamma_k = 1 \quad \text{iff} \quad \sum_{s=1}^{N}\Big(\max_{i\in\mathcal{I}^s}\{\theta_{si} - w_s d_{sik}\} - \sum_{i\in\mathcal{I}^s}\theta_{si}p_{si}\Big) > \theta_0. \tag{6}$$

For each (s,i), $\beta_{sik} = \gamma_k$ only for one $i = i^*(s,k)$ maximizing the expression in braces in (6).

We solve the dual problem $\max_\theta L_D(\theta)$ using a subgradient method tailored for the federated architecture, as detailed in Algorithm 1. The subgradients of $L_D(\theta)$ are given by the expressions in the parentheses in (5), with γ and β specified above.

Algorithm 1 Federated Dual Subgradient Algorithm for the Barycenter Problem

Require: Initial $\theta^{(0)}$, M, step size $\alpha^{(0)}$, maxiter. Let $j = 0$.

while $j <$ maxiter **do**
  Local Devices ($s = 1, \ldots, N$)**:**
  for $k = 1, \ldots, K$ **do**
    $T_{sk} \leftarrow \max_{i \in \mathcal{I}^s}(\theta_{si}^{(j)} - w_s d_{sik}) - \sum_{i \in \mathcal{I}^s} \theta_{si}^{(j)} p_{si}$
    $i_{sk}^* \leftarrow \text{argmax}_{i \in \mathcal{I}^s}(\theta_{si}^{(j)} - w_s d_{sik})$
  end for
  Report the vector $[T_{s1}, \ldots, T_{sK}]$ to the global device.
  Global Device:
  for $k = 1, \ldots, K$ **do**
    $\gamma_k \leftarrow \left\{ \sum_{s=1}^N T_{sk} > \theta_0^{(j)} \right\}$
  end for
  Update the step size: $\alpha^{(j+1)} \leftarrow \alpha^{(0)} / \sqrt{j+1}$
  Compute the subgradient: $g_0 \leftarrow \sum_{k=1}^K \gamma_k - M$
  Update the global dual variable: $\theta_0^{(j+1)} \leftarrow \theta_0^{(j)} + \alpha^{(j+1)} g_0$
  Send the vector $[\gamma_1, \ldots, \gamma_K]$ to all local devices.
  Local Devices ($s = 1, \ldots, N$)**:**
  For each $k = 1, \ldots, K$ set $\beta_{si_{sk}^* k} \leftarrow \gamma_k$, and $\beta_{sik} \leftarrow 0$ otherwise.
  For each $i \in \mathcal{I}^s$, compute the subgradient: $g_{si} \leftarrow p_{si} \sum_k \gamma_k - \sum_k \beta_{sik}$
  Update the local dual variables: $\theta_{si}^{(j+1)} \leftarrow \theta_{si}^{(j)} + \alpha^{(j+1)} g_{si}$
  $j \leftarrow j + 1$
end while

The *only* information transmitted from each client s to the server is the vector $[T_{s1}, \ldots, T_{sK}]$, summarizing the aggregate usefulness of the candidate points. Private data, such as particle locations, distances, or local dual variables, is never shared.

For each client, the communication cost per iteration is only $\mathcal{O}(K)$, and the computational complexity is $\mathcal{O}(|\mathcal{I}_s|K)$, without any matrix-vector operations. The coordinator's cost is negligible.

4 Numerical Illustration

We demonstrate our algorithm on the problem to find the Wasserstein barycenter of several 2D Gaussian components. In Figure 1, the left panel displays the components and the grid of K potential barycenter points (black). The right panel shows the computed barycenter (in black) for the weight vector $w = [0.166, 0.385, 0.063, 0.321, 0.065]$.

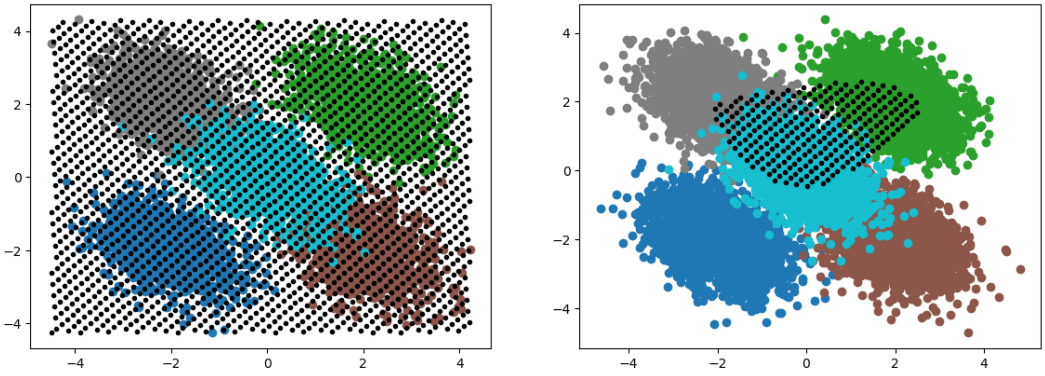

Figure 1: The barycenter selected from a uniform grid.

We also present an experiment to compare two algorithms. The first one is a momentum version of our dual subgradient method, which is a discretized free-support approach, where the barycenter's support is a subset of a larger, randomly selected set of candidate points. The other method is the best free-support algorithm based on alternating Bregman projections, which optimizes the locations of a fixed number of barycenter points. This iterative process computes regularized optimal transport plans to each source measure via the Sinkhorn algorithm at each step, subsequently updating the support locations toward the barycentric mapping of the measures until convergence.[1] Both algorithms are initialized with identical source distributions.

The visualization of barycenters in Figure 2 is similar to the one presented earlier in Figure 1. The weight vector was $w = [0.7, 0.1, 0.05, 0.05, 0.1]$. The $K = 1000$ candidate points for the dual subgradient method were randomly sampled from a normal distribution with independent components having zero means and variances equal to 5. The starting support of cardinality $M = 250$ of the barycenter for the Bregman projection method was sampled from the same distribution. The stopping test for the dual method is whether the relative change in the dual function is smaller than $\varepsilon = 10^{-5}$ and the number of points selected is within $\pm 5\%$ of $M = 250$. For the alternating Bregman projection method, the stopping test is whether the relative change of the barycenter drops below $\varepsilon = 10^{-4}$. The regularization term was 0.05.

Table 1 provides the numbers of iterations, solution time, and the final barycenter value.

The key insight from this analysis is the rapid convergence of the dual method to a high-quality solution, and the high cost of the alternating Bregman projections method. For the nominal regularization 0.05, the latter algorithm did not reach the tolerance threshold ϵ before the maximum iterations. Larger regularization parameters resulted in very low-quality solutions.

The dual subgradient method demonstrates superior computational performance and scalability, and much lower time per iteration. Its primary advantage lies in its architecture, which does not require the expensive computation of optimal transport plans at each iteration. By relying on computationally cheaper dual variable updates, the method is substantially faster per iteration.

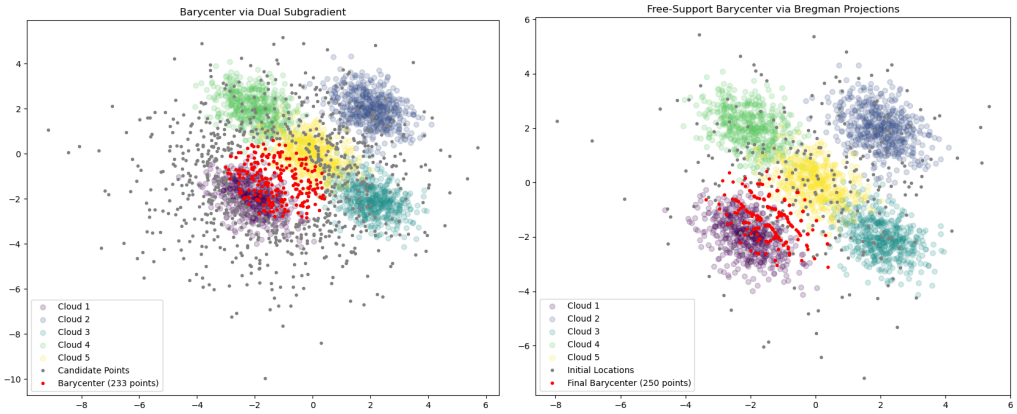

Figure 2: Barycenters by the dual subgradient method and alternating Bregman projections.

Table 1: Comparison of Barycenter Computation Methods

Method	**Converged?**	**Time (s)**	**Iterations**	**Time/Iter. (ms)**	**Barycenter Value**
Subgradient	Yes	23.73	2204	10.77	4.44
Bregman	No	2397.21	1000	2397.21	4.43

[1] Adapted from the Python Optimal Transport library at `https://pythonot.github.io`.

Acknowledgments and Disclosure of Funding

This work was supported by the Office of Naval Research award N00014-21-1-2161 and by the Air Force Office of Scientific Research award FA9550-24-1-0284.

References

[1] Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, January 2011. ISSN 1095-7154. doi: 10.1137/100805741. URL http://dx.doi.org/10.1137/100805741.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN, 2017. arXiv:1701.07875.

[3] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[4] Giulio Vittorio Carassai. Neural optimal transport at scale: Wasserstein barycenters for fair insurance. Master's thesis, ETH Zürich, 2024.

[5] Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. In *Conference on Learning Theory*, pages 1276–1304. PMLR, 2020.

[6] Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020.

[7] Nhat Ho, Xuan Long Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. Multilevel clustering via Wasserstein means, 2017. arXiv:1706.03883.

[8] Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Debiased Sinkhorn barycenters. In *International Conference on Machine Learning*, pages 4692–4701. PMLR, 2020.

[9] Tianyi Lin, Nhat Ho, Xi Chen, Marco Cuturi, and Michael I Jordan. Computational hardness and fast algorithm for fixed-support Wasserstein barycenter. *arXiv preprint arXiv:2002.04783*, 2020.

[10] Zhengqi Lin and Yan Chen. Leveraging optimal transport for distributed two-sample testing: An integrated transportation distance-based framework, 2025. URL https://arxiv.org/abs/2506.16047.

[11] Zhengqi Lin and Andrzej Ruszczyński. An integrated transportation distance between kernels and approximate dynamic risk evaluation in Markov systems. *SIAM Journal on Control and Optimization*, 61(6):3559–3583, 2023. doi: 10.1137/22M1530665. URL https://doi.org/10.1137/22M1530665.

[12] Zhengqi Lin and Andrzej Ruszczynski. Fast dual subgradient optimization of stochastic kernels with application to forward-backward stochastic differential equations. In *IISE Annual Conference. Proceedings*, 2024. doi: 10.21872/2024IISE_5776.

[13] Zhengqi Lin and Andrzej Ruszczynski. Stochastic kernel approximation by transportation distance method. In *IISE Annual Conference. Proceedings*, 2024.

[14] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. *arXiv*, 2016. doi: 10.48550/ARXIV.1602.05629. URL https://arxiv.org/abs/1602.05629.

[15] Cédric Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2009.