# A Unified Approach for Computing Wasserstein Barycenters of Discrete and Continuous Measures

**Peng Xu**
Department of Statistics
University of Illinois Urbana-Champaign
pengxu1@illinois.edu

**Changbo Zhu**
Department of ACMS
University of Notre Dame
czhu4@nd.edu

**Xiaohui Chen**
Department of Mathematics
Thomas Lord Department of Computer Science
University of Southern California
xiaohuic@usc.edu

## Abstract

Computing the unregularized Wasserstein barycenter for measure-valued data is a challenging optimization task. Recent algorithms have been tailored to either discrete measures as point clouds or continuous measures discretized on regular grids. In this work, we propose a primal mirror descent algorithm for computing the exact Wasserstein barycenter in the Fisher-Rao geometry. Our algorithm is a unified approach that is flexible enough to simultaneously cover discrete and absolutely continuous input measures, with best-iterate convergence guarantees in both settings. In particular, when all input measures are discrete, our algorithm solves a sequence of semi-discrete optimal transport subproblems and produces absolutely continuous iterates whose best objective value converges to the optimal value under suitable initialization and support conditions. We use synthetic and real data examples to demonstrate promising results in terms of accuracy and computational cost.

## 1 Introduction

The Wasserstein barycenter extends Euclidean averaging to probability measures by minimizing the weighted sum of squared 2-Wasserstein distances to a collection of input distributions. Since its inception in the optimal transport (OT) literature [Agueh and Carlier, 2011], it has drawn increasing attention for a wide range of geometric and statistical learning methods on distribution data [Rabin et al., 2012, Solomon et al., 2015, Srivastava et al., 2018, Ho et al., 2017, Zhuang et al., 2022, Zhu and Müller, 2025]. The characterization of Wasserstein barycenters via expectations of optimal transport maps has been investigated by Bigot and Klein [2018].

Despite the progress on the applications, practical algorithms for computing the exact Wasserstein barycenter largely bifurcate into two discretization schemes of the input distributions. For *discrete probability measures* as point cloud data, early work showed that the barycenter problem can be formulated as a large-scale linear programming (LP) problem [Carlier et al., 2015, Ge et al., 2019]. While this formulation is exact, its computational cost grows rapidly with the number of support points, making it impractical for high-resolution grids or dense point cloud data. Álvarez-Esteban et al. [2016] proposed an iterative fixed-point algorithm in which the weighted average of Monge transport maps is used to update the barycenter. This update coincides with a unit-step Wasserstein gradient descent for the barycenter functional [Chewi et al., 2020]. Explicit convergence rates are

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

available in the Gaussian setting, where a Polyak–Łojasiewicz (PŁ) inequality holds [Chewi et al., 2020]. In general, however, the Wasserstein barycenter functional is not geodesically convex in the $W_2$ geometry [Ambrosio et al., 2008], and establishing convergence guarantees for gradient-based algorithms with general measure-valued inputs remains challenging.

Complementary to the above primal and entropic approaches, recent works have established strong duality and utilized gradient flows in the potential function space. Kim et al. [2025, 2026] proposed Sobolev gradient based algorithms that optimize dual or semi-dual potentials on *continuous probability densities* discretized on regular grids. These methods produce high-accuracy solutions when sufficiently fine discretization is available, but their reliance on finite-difference approximations makes them unsuitable for coarse or irregular domains and limits their ability to handle discrete measures directly, even in the two-marginal case; see also [Jacobs and Léger, 2020]. To overcome the curse of dimensionality, recent works have explored neural-network parameterizations of Kantorovich potentials or transport maps [Fan et al., 2021, Kolesov et al., 2024]. Such approaches scale favorably with dimension but necessarily trade off accuracy for representational flexibility and training stability. In addition, neural network based methods typically require large training sets, careful architecture choices, and do not guarantee exact satisfaction of optimality conditions.

### 1.1 Our Contribution

In this work, we propose **FRBary**, a mirror-descent algorithm in the Fisher–Rao geometry for computing the exact Wasserstein barycenter of a collection of probability measures $\{\mu_1, \dots, \mu_n\}$. Our method has the following key advantages:

- The proposed algorithm applies to discrete and absolutely continuous measures in a unified and adaptive manner, allowing the input collection to contain measures of heterogeneous types, including point-cloud and absolutely continuous distributions. While discrete optimal transport methods can, in principle, be applied to continuous measures through discretization, their computational cost grows rapidly as the grid resolution is refined. In contrast, our approach adapts to the intrinsic structure of each input measure through semi-discrete and absolutely continuous optimal transport subproblems, leading to substantially improved computational efficiency.
- Our method is a primal approach for computing the exact Wasserstein barycenter. Under the assumptions stated in Section 4, we prove that the minimum barycenter objective value along the iterates converges to the optimal value at rate $O(\log T/\sqrt{T})$ after $T$ iterations, with guarantees covering fully continuous, fully discrete, and mixed discrete–continuous inputs. When the barycenter has finite KL divergence relative to the initialization, a constant step size chosen for the iteration horizon improves the best-iterate objective rate to $O(T^{-1/2})$.
- The algorithm is a semi-discrete approach. For discrete input measures (point clouds), it solves a sequence of semi-discrete optimal transport subproblems while maintaining absolutely continuous barycenter approximations. In many downstream statistical models for distribution-valued data, such as distribution-on-distribution regression [Chen et al., 2023, Zhu and Müller, 2023] and Wasserstein PCA [Bigot et al., 2017, Xu et al., 2026], the barycenter serves as an anchor measure and is typically required to be absolutely continuous. This feature of our algorithm therefore facilitates statistical modeling and inference for discrete or sample-based distributions.

For limitations, the current implementation primarily focuses on the computation of barycenters for 2D and 3D measures.

**Notations.** We write $\mathcal{P}(\Omega)$ for the set of probability measures supported on a compact and convex set $\Omega \subset \mathbb{R}^d$, and $\mathcal{P}_{\mathrm{ac}}(\Omega)$ for the subset of measures that are absolutely continuous with respect to the Lebesgue measure. We denote by $\mathrm{id}$ the identity map on $\mathbb{R}^d$, by $\langle x, y\rangle$ the Euclidean inner product of vectors $x, y \in \mathbb{R}^d$, and by $\|x\| = \sqrt{\langle x, x\rangle}$ the associated Euclidean norm. For sequences of real numbers $a_T, b_T$ depending on $T$, we write $a_T \asymp b_T$ if there exist constants $0 < c_1 \le c_2 < \infty$, independent of $T$, such that $c_1 b_T \le a_T \le c_2 b_T$ for all sufficiently large $T$.

## 2 Preliminaries

We classify optimal transport problems into three settings: continuous, discrete and semi-discrete, due to their fundamental differences in algorithmic structure and theoretical properties. These correspond to the cases where both probability measures are absolutely continuous, both are discrete, or one is discrete and the other is absolutely continuous. In the following, we review optimal transport in each of these three settings and discuss the corresponding Wasserstein barycenter problem.

**Continuous optimal transport.** In a nutshell, the optimal transport (OT) problem aims to find the most cost-efficient way to transport a pile of sand to a collection of holes; cf. [Villani, 2003, 2009, Santambrogio, 2015] as excellent references on OT theory. For any $\mu, \nu \in \mathcal{P}_{\text{ac}}(\Omega)$, the quadratic (2-Wasserstein) distance between them is defined as

$$W_2^2(\mu, \nu) \coloneqq \inf_{T:T_{\#}\mu=\nu} \int_\Omega \|x - T(x)\|^2 \, \mathrm{d}\mu(x),$$

where the infimum is taken over all measurable transport maps $T : \Omega \to \Omega$ that push forward $\mu$ to $\nu$ (i.e., $T_{\#}\mu = \nu$). This is the Monge formulation [Monge, 1781], where the minimizer (whenever it exists) is called the optimal transport from $\mu$ to $\nu$, denoted $T_{\mu\to\nu}$. For discrete measures, i.e., $\mu, \nu$ lie in $\mathcal{P}(\Omega)$, but not in $\mathcal{P}_{\text{ac}}(\Omega)$, the Monge problem often has no solution because each atom is indivisible and cannot be split to match the target weights. In contrast, the Kantorovich formulation always has a solution, and defines $W_2^2(\mu, \nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\Omega\times\Omega} \|x - y\|^2 \, \mathrm{d}\pi(x, y)$, where $\Pi(\mu, \nu)$ denotes the set of all couplings on $\Omega \times \Omega$ with marginals $\mu$ and $\nu$. For absolutely continuous measures, the two formulations coincide. The Kantorovich dual formulation takes the form

$$\frac{1}{2} W_2^2(\mu, \nu) = \sup_\phi \int_\Omega \phi(x) \, \mathrm{d}\mu(x) + \int_\Omega \phi^c(y) \, \mathrm{d}\nu(y),$$

where $\phi^c(y) \coloneqq \inf_{x\in\Omega}\{\|x - y\|^2/2 - \phi(x)\}$ is the $c$-transform of $\phi$. The maximizer to the dual is called the Kantorovich potential from $\mu$ to $\nu$, denoted as $\phi_{\mu\to\nu}$. Brenier's theorem implies that $x \mapsto \|x\|^2/2 - \phi_{\mu\to\nu}(x)$ is always convex on $\Omega$ and the optimal transport map can be obtained via $T_{\mu\to\nu} = \mathrm{id} - \nabla\phi_{\mu\to\nu}$ [Brenier, 1991]. By discretizing $\mu$ and $\nu$ over a regular grid of size $M$, the Sobolev gradient based approach can be applied to compute the Kantorovich potential with per-iteration time complexity $O(M \log(M))$ and space complexity $O(M)$ [Kim et al., 2026, Jacobs and Léger, 2020].

**Discrete optimal transport.** Let $\widehat{\mu}_1 = \sum_{i=1}^{m_1} u_{1,i} \cdot \delta_{x_{1,i}}$ and $\widehat{\mu}_2 = \sum_{j=1}^{m_2} u_{2,j} \cdot \delta_{x_{2,j}}$ be two discrete probability measures on $\mathbb{R}^d$, where for $k = 1, 2$, $u_{k,i} \geq 0$ and $\sum_i u_{k,i} = 1$. The Kantorovich dual problem reduces to

$$\frac{W_2^2(\widehat{\mu}_1, \widehat{\mu}_2)}{2} = \max_{\boldsymbol{\phi}\in\mathbb{R}^{m_1}} \left\{ \sum_{i=1}^{m_1} u_{1,i} \, \phi_i + \sum_{j=1}^{m_2} u_{2,j} \, \phi_j^c \right\},$$

where the maximization is over vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{m_1})^\top \in \mathbb{R}^{m_1}$, and $\boldsymbol{\phi}^c = (\phi_1^c, \ldots, \phi_{m_2}^c)^\top \in \mathbb{R}^{m_2}$ is the discrete $c$-transform of $\boldsymbol{\phi}$, defined by

$$\phi_j^c \coloneqq \min_{1\leq i\leq m_1} \left\{ \frac{1}{2}\|x_{1,i} - x_{2,j}\|^2 - \phi_i \right\},$$

where $j \in [m_2]$. Equivalently, this formulation enforces the dual feasibility condition $\phi_i + \phi_j^c \leq \|x_{1,i} - x_{2,j}\|^2/2$ for all $i, j$. The maximizer $\boldsymbol{\phi}$ is referred to as the discrete Kantorovich potential from $\widehat{\mu}_1$ to $\widehat{\mu}_2$, and $\boldsymbol{\phi}^c$ is the associated dual potential in the reverse direction.

**Semi-discrete optimal transport.** Many statistical and data-analytic problems involve computing the Wasserstein distance between a known, absolutely continuous reference measure and an empirical distribution constructed from observed data. This motivates the semi-discrete optimal transport setting, which has received a lot of attention; see [Sadhu et al., 2024, Taşkesen et al., 2023, Agarwal et al., 2024], among others. Formally, let $\nu \in \mathcal{P}_{\text{ac}}(\Omega)$ be a known reference measure, and suppose we observe i.i.d. samples $x_1, \ldots, x_m$ from another absolutely continuous measure $\mu$. Define the

weighted empirical (discrete) approximation $\widehat{\mu} \coloneqq \sum_{i=1}^{m} u_i \delta_{x_i}$. The 2-Wasserstein distance between the empirical measure $\widehat{\mu}$ and the continuous measure $\nu$ in this semi-discrete setting admits the Kantorovich dual formulation

$$\frac{W_2^2(\widehat{\mu}, \nu)}{2} = \max_{\boldsymbol{\phi} \in \mathbb{R}^m} \left\{ \sum_{i=1}^{m} u_i \phi_i + \int_{\Omega} \phi^c(y) \, \mathrm{d}\nu(y) \right\},$$

where the maximization is over the potential vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)^\top \in \mathbb{R}^m$, which plays the role of the discrete dual potential corresponding to the empirical measure $\widehat{\mu}$, and $\phi^c : \Omega \to \mathbb{R}$, defined as $\phi^c(y) \coloneqq \min_{1 \le i \le m} \{\|y - x_i\|^2/2 - \phi_i\}$, is the quadratic $c$-transform of $\boldsymbol{\phi} \in \mathbb{R}^m$. For $i \in [m]$, the partial derivative of the dual objective

$$I_{\widehat{\mu},\nu}(\boldsymbol{\phi}) \coloneqq \sum_{i=1}^{m} u_i \phi_i + \int_{\Omega} \phi^c(y) \, \mathrm{d}\nu(y) \tag{1}$$

with respect to $\phi_i$ is $\partial I_{\widehat{\mu},\nu} / \partial \phi_i \, (\boldsymbol{\phi}) \coloneqq u_i - \nu(L_i(\boldsymbol{\phi}))$, where

$$L_i(\boldsymbol{\phi}) = \left\{ y \;\middle|\; \frac{1}{2}\|y - x_i\|_2^2 - \phi_i \le \frac{1}{2}\|y - x_k\|_2^2 - \phi_k, \forall k \right\}$$

is the Laguerre cell associated with $x_i$. Thus, by discretizing $\nu$ over a regular grid of size $M$, the concave objective $I_{\widehat{\mu},\nu}$ can still be solved by a first-order gradient-based approach with per-iteration time complexity of $O(mM)$ and a space complexity of $O(m + M)$. By contrast, a direct LP formulation involves $mM$ transport variables and requires $O(mM)$ storage when represented explicitly. We use $\boldsymbol{\phi}_{\widehat{\mu} \to \nu}$ to represent the maximizing vector of $I_{\widehat{\mu},\nu}$ and $\phi_{\nu \to \widehat{\mu}} : \Omega \to \mathbb{R}$ to denote the corresponding $c$-transform. While both the discrete and semi-discrete settings admit finite-dimensional dual formulations, they differ substantially in structure. In the discrete case, optimal transport reduces to a linear program over a coupling matrix subject to marginal constraints. In contrast, the semi-discrete setting induces a partition of $\Omega$ into Laguerre cells, yielding an explicit transport map that assigns each location to a support point.

**Wasserstein barycenter.** The Wasserstein barycenter, introduced by [Agueh and Carlier, 2011], defines a notion of averaging for a collection of probability measures. Given weights $w_1, w_2, \ldots, w_n > 0$ with $\sum_{i=1}^{n} w_i = 1$ and measures $\mu_1, \mu_2, \ldots, \mu_n \in \mathcal{P}(\Omega)$, a *Wasserstein barycenter* is any minimizer

$$\lambda^\star \in \operatorname*{arg\,min}_{\lambda \in \mathcal{P}(\Omega)} \left\{ \mathcal{E}(\lambda) \coloneqq \sum_{i=1}^{n} \frac{w_i}{2} W_2^2(\mu_i, \lambda) \right\}. \tag{2}$$

Intuitively, $\lambda^\star$ is the distribution that minimizes the weighted sum of squared Wasserstein distances to all input measures. When at least one $\mu_i$ is absolutely continuous with respect to the Lebesgue measure and the supports are not confined to lower-dimensional affine subspaces, the minimizer exists, and is unique and absolutely continuous [Kim and Pass, 2017, Le Gouic and Loubes, 2017]. If all measures are discrete, it has been shown that Wasserstein barycenters of discrete measures are themselves discrete, and that one can always choose a barycenter that is provably sparse [Anderes et al., 2016]. In addition, every discrete barycenter admits a non–mass splitting optimal transport to each of the input discrete measures. Such non–mass splitting maps rarely exist between two arbitrary discrete measures unless special mass-balance conditions happen to hold, which highlights a distinctive structural property of Wasserstein barycenters in the discrete case.

## 3 Mirror Descent for Wasserstein Barycenter

To motivate our proposed algorithm in the Wasserstein space, we first briefly review the mirror descent algorithm in the Euclidean setting. It is well known that gradient descent for minimizing a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ can be interpreted as a linearization of the proximal point method of the form [Beck and Teboulle, 2003]:

$$x^{k+1} = \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \left\{ \left\langle x, \nabla f(x^k) \right\rangle + \frac{1}{2\eta_k} \|x - x^k\|^2 \right\},$$

where $\eta_k > 0$ is the step size or learning rate. Mirror descent generalizes this update by replacing the Euclidean squared norm with a more general Bregman divergence induced by a strictly convex mirror map $\psi : \mathbb{R}^d \to \mathbb{R}$. Specifically, the mirror descent update takes the form

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^d} \left\{ \langle x, \nabla f(x^k) \rangle + \frac{1}{\eta_k} \mathrm{D}_\psi(x, x^k) \right\},$$

where the Bregman divergence associated with $\psi$ is defined by $\mathrm{D}_\psi(x, y) := \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle$. By choosing an appropriate mirror map, mirror descent adapts the geometry of the optimization procedure to the structure of the objective function. For a convex nonsmooth objective, the same update is defined by replacing $\nabla f(x^k)$ with any subgradient $g^k \in \partial f(x^k)$. This is the form of mirror descent used below for the Wasserstein barycenter objective.

The fundamental roadblock for deriving a primal and convergent gradient-based algorithm for solving the Wasserstein barycenter problem (2) is that the Wasserstein barycenter functional is not geodesically convex in $W_2$-geometry. Therefore in the existing literature, deriving principled convergence analysis for gradient-based methods to compute the barycenter exists only in highly specialized setups, e.g. Gaussian measures [Chewi et al., 2020], or directly assumes a Wasserstein PŁ inequality that is difficult to verify [Boufadene and Vialard, 2025, Zhu and Chen, 2025].

Our key starting observation is that the objective functional $\mathcal{E}$ in (2) is convex under ordinary mixtures of measures, since $W_2^2(\cdot, \mu)$ is convex under such mixtures for every $\mu \in \mathcal{P}(\Omega)$; cf. [Santambrogio, 2015, Proposition 7.19]. In Wasserstein space, the first variation is an extension of gradient in Euclidean space to a general metric space, when $\lambda$ is absolutely continuous, a subgradient of $\mathcal{E}$ at $\lambda$ is represented by the bounded function $\delta\mathcal{E}_\lambda : \Omega \to \mathbb{R}$, and has the expression $\delta\mathcal{E}_\lambda := \sum_{i=1}^n w_i \phi_{\lambda \to \mu_i}$. Using the first variation formula and picking the mirror map as negative relative entropy (reflecting the linear or $L^2$ geometry), we arrive at the following mirror descent algorithm for minimizing $\mathcal{E}$:

$$\lambda^{k+1} = \arg\min_\lambda \int \delta\mathcal{E}_{\lambda^k} \, \mathrm{d}\lambda + \frac{\mathrm{D}_{\mathrm{KL}}(\lambda \parallel \lambda^k)}{\eta_k} = \arg\min_\lambda \sum_{i=1}^n w_i \int \phi_{\lambda^k \to \mu_i} \, \mathrm{d}\lambda + \frac{\mathrm{D}_{\mathrm{KL}}(\lambda \parallel \lambda^k)}{\eta_k},$$

where $\eta_k$ is the user-selected learning rate and $\mathrm{D}_{\mathrm{KL}}$ is the KL-divergence. Let $\rho^k$ denote the density of $\lambda^k$, the above problem can then be written as

$$\arg\min_\rho \sum_{i=1}^n w_i \, \mathbb{E}_\rho[\phi_{\rho^k \to \mu_i}] + \frac{\mathrm{D}_{\mathrm{KL}}(\rho \parallel \rho^k)}{\eta_k}, \tag{3}$$

with the corresponding first-order condition $\sum_{i=1}^n w_i \phi_{\rho^k \to \mu_i} + \eta_k^{-1} \ln(\rho^{k+1}/\rho^k) + \text{const.} = 0$. This implies that $\rho^{k+1} \propto \rho^k \exp\left(-\eta_k \sum_{i=1}^n w_i \phi_{\rho^k \to \mu_i}\right)$.

Now, we summarize our mirror descent algorithm for exact Wasserstein barycenter computation in Algorithm 1. Our mirror descent algorithm, FRBary, can be interpreted as a discrete Fisher-Rao gradient flow [Bauer et al., 2016, Yan et al., 2024] of the Wasserstein barycenter functional, in the sense that we evolve the barycenter measure along the Fisher-Rao Riemannian geometry, i.e., steepest descent under the Fisher information metric, while the objective itself is defined via Wasserstein distances to the input measures.

**Algorithm 1** Mirror descent for minimizing $\mathcal{E}(\lambda)$

**Require:** Initial measure $\lambda^0 \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ with density $\rho^0$ such that $0 < a \le \rho^0(x) \le b < \infty$ for any $x \in \Omega$;
Step sizes $\{\eta_k\}_{k \ge 0}$;
Number of iterations $T$.

1: **for** $k = 0, 1, \ldots, T$ **do**
2: Compute Kantorovich potentials $\phi_{\rho^k \to \mu_i}$ for $i \in [n]$.
3: Update the density by

$$\rho^{k+1} \propto \rho^k \exp\left(-\eta_k \sum_{i=1}^n w_i \phi_{\rho^k \to \mu_i}\right). \tag{4}$$

Algorithm 1 remains applicable when some input measures $\mu_i$ are replaced by their empirical discrete counterpart $\widehat{\mu}_i = \sum_{j=1}^{m_i} u_{i,j} \, \delta_{x_{i,j}}$. In this case, the update involves the Kantorovich potential $\phi_{\rho^k \to \widehat{\mu}_i}$

arising from a semi-discrete optimal transport problem. As discussed earlier, let $\boldsymbol{\phi}_{\widehat{\mu}_i \to \rho^k} \in \mathbb{R}^{m_i}$ denote the discrete optimal potential vector from $\widehat{\mu}_i$ to $\rho^k$. Its $c$-transform, $\phi_{\rho^k \to \widehat{\mu}_i}$, can be interpreted as the Kantorovich potential from the absolutely continuous measure $\rho^k$ to the discrete measure $\widehat{\mu}_i$, in the sense that the map $(\mathrm{id} - \nabla \phi_{\rho^k \to \widehat{\mu}_i})$ pushes forward $\rho^k$ to $\widehat{\mu}_i$. More precisely, the proposed method can be interpreted as a semi-discrete barycenter computation: the input measures are discrete, while the mirror-descent iterates remain absolutely continuous and belong to $\mathcal{P}_{\mathrm{ac}}(\Omega)$. Since $\mathcal{P}_{\mathrm{ac}}(\Omega)$ is dense in $\mathcal{P}(\Omega)$ under the $W_2$ metric, a discrete Wasserstein barycenter can be approximated arbitrarily well by absolutely continuous measures. Theorem 4.2 establishes convergence of the best objective value attained by our iterates to the discrete barycenter's objective value.

**Relation to mirror descent.** Cuturi and Doucet [2014] studied KL mirror descent for fixed-support barycenter weights, recovering the same multiplicative update. Yao et al. [2024] developed implicit KL proximal schemes, including smoothing arguments for singular minimizers, while Bonet et al. [2024] studied mirror descent in Wasserstein transport geometry. Our explicit Fisher–Rao scheme addresses unregularized barycenters with heterogeneous inputs and provides convergence guarantees for discrete barycenters.

**Comparison with entropic Wasserstein barycenter.** Our method is different from existing entropic OT approaches for Wasserstein barycenters. Entropic methods replace each Wasserstein term $W_2^2(\mu_i, \lambda)$ by an entropically regularized distance

$$W_{2,\varepsilon}^2(\mu_i, \lambda) \coloneqq \min_{\pi \in \Pi(\mu_i, \lambda)} \left\{ \int \|x - y\|_2^2 \, \mathrm{d}\pi(x, y) + \varepsilon \, \mathrm{D}_{\mathrm{KL}}(\pi \parallel \mu_i \otimes \lambda) \right\},$$

and compute a barycenter of the modified objective $\mathcal{E}_\varepsilon(\lambda) = \sum_{i=1}^n w_i W_{2,\varepsilon}^2(\mu_i, \lambda)/2$, typically on a fixed grid using Sinkhorn iterations [Janati et al., 2020, Solomon et al., 2015]. The resulting "entropic barycenter" is therefore biased, depends on the choice of $\varepsilon$, and is restricted to the prescribed grid in these common-grid implementations. In contrast, we keep the *original* barycenter functional $\mathcal{E}(\lambda) = \sum_{i=1}^n w_i W_2^2(\mu_i, \lambda)/2$ unchanged and use negative entropy solely as the mirror map in a mirror-descent scheme on $\mathcal{P}_{\mathrm{ac}}(\Omega)$. At each iteration, we solve (semi-discrete) OT problems between a continuous candidate $\lambda$ and the inputs $\mu_i$, and update $\lambda$ by a KL–Bregman step $\lambda^{k+1} = \arg\min_\lambda \langle \delta \mathcal{E}_{\lambda^k}, \lambda \rangle + \eta_k^{-1} \, \mathrm{D}_{\mathrm{KL}}(\lambda \parallel \lambda^k)$. The KL divergence here plays the role of a geometry (the Bregman divergence of the mirror map), not a regularizer of the OT cost. Consequently, our method targets the unregularized Wasserstein barycenter objective without introducing entropic regularization bias, while maintaining absolutely continuous approximations even when all input measures are discrete.

**Implementation.** Our algorithm can handle inputs of mixed types and at each iteration, the main task is the computation of the Kantorovich potentials $\{\phi_{\rho^k \to \mu_i}\}$ between the current iterate $\rho^k$ and the input measures $\{\mu_i\}$. The computational complexity for each $\phi_{\rho^k \to \mu_i}$ is determined by the type of the input measure $\mu_i$, and our algorithm adapts accordingly by invoking the appropriate optimal transport solver. When the input $\mu_i$ is discrete (point cloud), the resulting subproblem is a semi-discrete optimal transport problem, for which Laguerre cell–based methods are more efficient. When the input $\mu_i$ is absolutely continuous and supported on a high-resolution grid, Sobolev gradient–based approaches [Kim et al., 2026, Jacobs et al., 2021] provide an effective solver. In practice, the above methods often involve discretizing the iterate $\rho^k$ over a high-resolution regular grid; see Section 2 for discussions of their computational complexities. Alternatively, the iterate $\rho^k$ may be parameterized by a neural network, leading to a grid-free implementation. However, for moderate-dimensional problems (e.g., 2D or 3D), discretization-based approaches often provide higher accuracy and more stable computation. A special case arises when all measures are histogram-type inputs, viewed as piecewise-constant densities on a common low-resolution grid. In this setting, it is natural to represent the iterate using the same grid, yielding a finite-dimensional discretization of the barycenter problem. When the grid size is moderate, the resulting discrete optimal transport subproblems can be solved using LP-based solvers. Additional implementation details are deferred to Appendix C.

## 4 Convergence Analysis

The Kantorovich potential $\phi_{\lambda^k \to \mu_i}$ is unique only up to an additive constant, and the update rule (4) is invariant under such shifts. Since $\Omega$ is compact, define $R \coloneqq \sup_{x \in \Omega} \|x\|$. Because the cost

function is finite on the compact set $\Omega \times \Omega$, an optimal Kantorovich potential $\phi_{\lambda^k \to \mu_i}$ can be chosen to be finite on $\Omega$. We fix the normalization $\inf_{x\in\Omega} \phi_{\lambda^k\to\mu_i}(x) = 0$. Then, by Lemma A.1 in the Appendix, the potential is uniformly bounded, and its $L^\infty$-norm satisfies $\|\phi_{\lambda^k\to\mu_i}\|_{L^\infty(\Omega)} \le 2R^2$. As a consequence, the first variation of $\mathcal{E}$ at $\lambda^k$ is also uniformly bounded in $L^\infty(\Omega)$ by $2R^2$. Combining this bound with Lemma A.2 in the Appendix yields $\mathrm{D_{KL}}(\rho^k \parallel \rho^{k+1}) \le 2\eta_k^2 R^4$. If at least one of the measures $\{\mu_i\}$ is absolutely continuous, then the barycenter $\lambda^\star$ is absolutely continuous. Combining the above inequalities with standard mirror-descent arguments yields the theorem stated below. A complete proof is deferred to the Appendix.

**Theorem 4.1.** *Assume that at least one of $\{\mu_1, \dots, \mu_n\}$ is absolutely continuous and that $\mathrm{D_{KL}}(\lambda^\star \parallel \lambda^0) < \infty$. For the iterates generated by Algorithm 1, we have*

$$\min_{0\le k\le T}\big(\mathcal{E}(\lambda^k) - \mathcal{E}(\lambda^\star)\big) \le \frac{\mathrm{D_{KL}}(\lambda^\star \parallel \lambda^0)}{\sum_{k=0}^T \eta_k} + 2R^4 \frac{\sum_{k=0}^T \eta_k^2}{\sum_{k=0}^T \eta_k}.$$

The above rate matches the classical mirror-descent guarantee for convex nonsmooth optimization in Euclidean space, which establishes best-iterate convergence, namely convergence of the best objective value among the iterates. Since the Wasserstein barycenter functional is convex but generally nonsmooth, our analysis naturally yields the same type of guarantee. If we choose a constant step size $\eta_k = T^{-1/2}$, then $\sum_{k=0}^T \eta_k \asymp \sqrt{T}$ and $\sum_{k=0}^T \eta_k^2 \asymp 1$, so the right-hand side is of order $O(T^{-1/2})$. If instead we use the diminishing step size $\eta_k = (k+1)^{-1/2}$, then $\sum_{k=0}^T \eta_k \asymp \sqrt{T}$ and $\sum_{k=0}^T \eta_k^2 \asymp \log T$, yielding the rate $O(\log T/\sqrt{T})$.

Whenever the best objective value converges to the optimum, compactness of $\Omega$ and continuity of $\mathcal{E}$ imply that every accumulation point of a sequence of selected best iterates is a Wasserstein barycenter. If the barycenter is unique, this selected sequence converges to it in $W_2$. This does not establish convergence of the full iterate sequence or a convergence rate in $W_2$.

We note that if each $\mu_i$ is discrete, then the barycenter $\lambda^\star$ is also discrete, and hence $\mathrm{D_{KL}}(\lambda^\star \parallel \lambda^0) = +\infty$. In this case the preceding guarantee is vacuous, since the right-hand side diverges. To establish convergence in the discrete setting, we instead compare $\lambda^\star$ to a smoothed approximation $\lambda_\varepsilon$ with density $\rho_\varepsilon$ such that

$$W_2(\lambda_\varepsilon, \lambda^\star) = O(\varepsilon), \tag{5}$$

$$\mathrm{D_{KL}}(\rho_\varepsilon \parallel \rho^0) = O(\log(1/\varepsilon)), \tag{6}$$

where $\rho^0$ denotes the initial density in the mirror-descent algorithm. In Lemmas A.4 and A.3 (Appendix), we show that such a $\lambda_\varepsilon$ can be constructed via a truncated Gaussian smoothing procedure. Then, for any fixed $T$, choosing $\varepsilon$ appropriately allows us to control both terms in the decomposition $\mathcal{E}(\lambda^k) - \mathcal{E}(\lambda^\star) = (\mathcal{E}(\lambda^k) - \mathcal{E}(\lambda_\varepsilon)) + (\mathcal{E}(\lambda_\varepsilon) - \mathcal{E}(\lambda^\star))$. This leads to the following theorem, where we let $\partial\Omega$ be the boundary of $\Omega$ and $\mathrm{dist}(x, \partial\Omega) = \inf_{y\in\partial\Omega}\|x-y\|_2$. The complete proof is deferred to the Appendix.

**Theorem 4.2.** *Let $\{\mu_i = \sum_{j=1}^{m_i} u_{i,j}\delta_{x_{i,j}}\}_{i=1}^n$ be a set of discrete probability measures on $\Omega$. Suppose that there exists $r > 0$ such that $\mathrm{dist}(x_{i,j}, \partial\Omega) \ge r$ for all $i \in [n]$ and $j \in [m_i]$. For $T \ge 4d/r^2$, running Algorithm 1 with step size $\eta_k = (k+1)^{-1/2}$ or $\eta_k = T^{-1/2}$ yields*

$$\min_{0\le k\le T}\big(\mathcal{E}(\lambda^k) - \mathcal{E}(\lambda^\star)\big) = O\left(\frac{\log T}{\sqrt{T}}\right).$$

The condition $\mathrm{dist}(x_{i,j}, \partial\Omega) \ge r$ requires that all support points $\{x_{i,j}\}$ lie strictly in the interior of $\Omega$, and is used to derive the bounds (5) and (6). In practice, $\Omega$ can always be chosen as a sufficiently large bounding box so that this condition holds.

Unlike Theorem 4.1, the convergence analysis in Theorem 4.2 does not require the discrete barycenter to have finite KL divergence relative to the initialization. Instead, the proof compares the iterates with a Gaussian-smoothed approximation of the barycenter whose KL divergence is explicitly controlled. This approximation introduces the logarithmic term in (6), resulting in an additional $\log T$ factor in the convergence rate even when the optimal constant step size $\eta_k = 1/\sqrt{T}$ is used. The assumptions required for Theorems 4.1 and 4.2 remain substantially weaker than those imposed by the dual approach of Kim et al. [2026]. In that work, the algorithm iteratively updates $n-1$ Kantorovich

potentials $\{\phi_i\}_{i=1}^{n-1}$ and defines the remaining one as $\phi_{\text{mix}} := -\sum_{i=1}^{n-1}(w_i/w_n)\phi_i$. To establish convergence, they require that the potentials remain continuous throughout the optimization process and that $\|(\text{id} - \nabla\phi_i^c)_{\#}\mu_i - (\text{id} - \nabla\phi_{\text{mix}}^c)_{\#}\mu_n\|_{\dot{H}^{-1}}^2$ is uniformly bounded, where $\|\cdot\|_{\dot{H}^{-1}}$ denotes the dual norm of the homogeneous Sobolev space.

# 5 Experiments

We present a series of experiments to evaluate the performance of the proposed method and compare it with benchmark methods from the literature. Due to space constraints, additional simulation results and details, including hardware specifications, are provided in Appendix D.

## 5.1 Heterogeneous Input

In this experiment, we consider heterogeneous inputs consisting of: (i) a $28 \times 28$ digit from the MNIST dataset [LeCun et al., 1998], treated as a piecewise constant density; (ii) a uniform density supported on a heart-shaped domain, discretized over a high-resolution grid of size $256 \times 256$; and (iii) a point cloud (discrete measure) of $10^4$ samples drawn from the Swiss roll distribution. To the best of our knowledge, only free-support barycenter methods are applicable in this setting. We therefore compare our approach with the exact and Sinkhorn free-support barycenters implemented in the Python Optimal Transport (POT) library [Flamary et al., 2024]. The convolutional Wasserstein barycenter [Solomon et al., 2015] and the debiased Sinkhorn barycenter [Janati et al., 2020] rely on a common discretization of the underlying domain (i.e., a shared grid or support) and are therefore not directly applicable in our setting with heterogeneous inputs and irregular supports. Similarly, Sobolev gradient-based approaches [Kim et al., 2026, 2025] are not applicable, as they require all input measures to be supported on high-resolution grids.

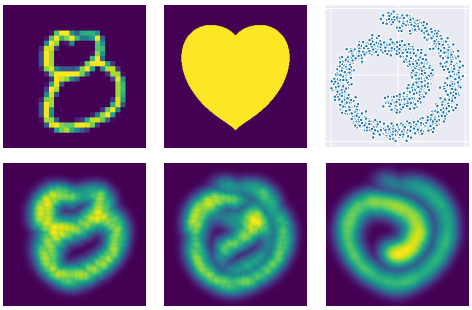
The results from our method and the free-support approaches are presented in Figure 1a and Figure 1b, respectively. The free-support methods produce point clouds with a pre-specified sample size, set to 1000 in this simulation study. In this setting, all three true barycenters are absolutely continuous and admit densities. Our method directly estimates each barycenter density, which we discretize on a $100 \times 100$ grid, enabling the generation of an arbitrary number of samples. In contrast, both exact and Sinkhorn free-support approaches produce only a fixed set of samples for each case and must be rerun to generate additional points. While densities can be estimated from these samples, this requires additional smoothing and the introduction of tuning parameters, such as the bandwidth. Moreover, all methods are run for up to 250 iterations and as shown in Table 1, our method requires less computational time compared with the exact free-support approach despite producing a larger output, namely density values on $10^4$ grid points compared to $10^3$ sample points. This is expected, as the exact free-support method solves linear programs and scales poorly with the number of input points. The Sinkhorn free-support method uses less computation time for the left barycenter, as it converges before reaching the full 250 iterations. However, its overall performance is unsatisfactory, as it tends to push support points toward the boundary of the output shape.

| Method | Barycenter (left) | Barycenter (middle) | Barycenter (right) |
|---|---|---|---|
| FRBary | 275 | 357 | 286 |
| Exact Free-support | 609 | 798 | 793 |
| Sinkhorn Free-support | 84 | 834 | 468 |

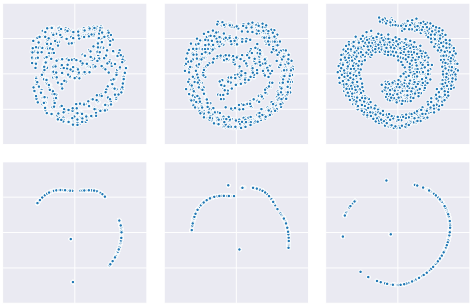
Table 1: Comparison of computation time (in seconds).

## 5.2 MNIST

We use the MNIST dataset [LeCun et al., 1998] as a simple real-world benchmark. The goal is to compute the barycenter of $n = 10$ images of the digit 3, where each image has size $28 \times 28$ pixels and is normalized and treated as a 784-dimensional histogram. Since the ground truth barycenter can be computed via linear programming (LP), we assess accuracy by measuring the $W_2$ and $L^2$ distances between the estimated barycenter densities and the ground truth. We compare our method with the entropically regularized Wasserstein barycenter (EWB) [Benamou et al., 2015] and the debiased



(a) (**Top row**): input distributions. (**Bottom row**): barycenters computed by FRBary. (**left**): from the first two inputs; (**middle**): from all three inputs; (**right**): from the last two inputs.



(b) (**Top row**): barycenters computed by the exact free-support approach. (**Bottom row**): barycenters computed by the Sinkhorn free-support approach.

Figure 1: Barycenters from heterogeneous input distributions.

| | $W_2$ | $L^2$ |
|---|---|---|
| FRBary | $1.360 \times 10^{-4}$ | $1.573 \times 10^{-4}$ |
| EWB | $1.904 \times 10^{-4}$ | $2.875 \times 10^{-4}$ |
| DSB | $3.237 \times 10^{-4}$ | $5.582 \times 10^{-4}$ |

Table 2: Quantifying the difference between the estimated barycenter and the ground truth.

Sinkhorn barycenter (DSB) [Janati et al., 2020], both of which are well suited for this setting and are implemented in the POT library.

Figure 2 displays the ground truth barycenter along with the estimated barycenters. Our result is obtained after $T = 500$ iterations. EWB and DSB, with regularization parameter set to $0.0004$, are run until convergence; smaller values (e.g., $0.0003$) lead to runtime errors. While our method is not the most computationally efficient in this setting, it yields the most accurate barycenter. Visually, our estimate is close to the exact solution. Quantitatively, Table 2 reports the $W_2$ and $L^2$ distances to the ground truth, showing that our method achieves lower error than both EWB and DSB.

### 5.3 2D/3D Gaussian Point Clouds

We evaluate our method on point clouds generated from Gaussian distributions. Since the Wasserstein barycenter of Gaussian measures is Gaussian, its mean and covariance can be computed accurately using the Gaussian barycenter characterization. This provides a population reference for assessing our estimates from finite point clouds. We first consider four two-dimensional Gaussian measures with randomly generated means and covariance matrices. Each input point cloud contains $10{,}000$ samples. Our method outputs a barycenter density discretized over a $100 \times 100$ grid. The learning rate is set to $0.1 \times (k+1)^{-0.3}$, and the algorithm is run for $T = 125$ iterations. Samples from the estimated barycenter are then obtained via rejection sampling. We also repeat this experiment for three three-dimensional Gaussian measures. In this case, the barycenter density is discretized over a

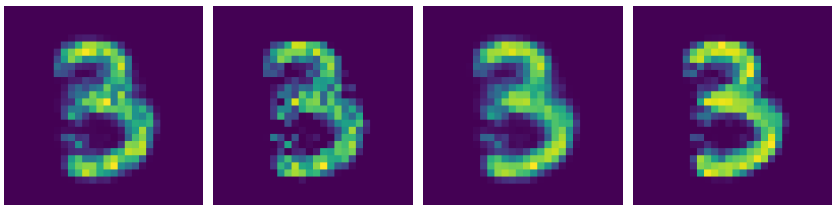


Figure 2: **Left to right**: ground truth (LP), mirror descent barycenter (FRBary), entropic barycenter (EWB), and debiased entropic barycenter (DSB), where $\varepsilon = 0.0004$ for the latter two.

$50 \times 50 \times 50$ grid, and 3,000 samples are drawn from each input distribution. The learning rate is set to $0.2 \times (k+1)^{-0.2}$, and the algorithm is run for $T = 100$ iterations. Figure 3 compares samples generated from the estimated barycenter with those from the true barycenter. We estimate the mean vectors and covariance matrices from these samples and report their distances to the ground truth in Figure 4. The results indicate that the learned densities closely approximate the true barycenters. More discussions for Gaussian distributions can be found in Appendix B.

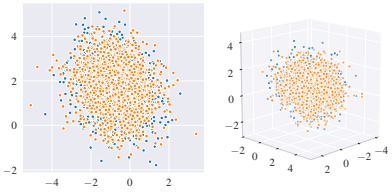


Figure 3: 2D (left) and 3D (right) samples drawn from the true (blue) and estimated (orange) barycenter.

| $d$ | $n$ | $m$ | distance |
|---|---|---|---|
| 2 | 4 | 10000 | Mean: $4.184 \times 10^{-3}$<br>Cov: $1.746 \times 10^{-1}$ |
| 3 | 3 | 3000 | Mean: $1.226 \times 10^{-2}$<br>Cov: $7.007 \times 10^{-2}$ |

Figure 4: Euclidean distances between the true and estimated mean vectors, and Frobenius distances between the true and estimated covariance matrices.

### 5.4 Color Palette Averaging

A 3-channel image can be viewed as samples from a color palette distribution on $[0, 1]^3$ after channel normalization. The palette barycenter across images can therefore be computed directly from the pixel samples. Figure 5 shows the color averaging result for $n = 2$ images. The image resolutions are $3160 \times 1846$ and $2560 \times 1573$. We employ the neural network implementation to compute the barycenter, where the barycenter log-density is represented by a 3-layer MLP with SiLU activations. During training, the OT potential functions are estimated via stochastic semi-discrete OT using the `OTT` package [Cuturi et al., 2022]. A quadrature-based alternative is described in Appendix C.2 The algorithm is run for 200 iterations using a learning rate schedule of $3 \times (k+1)^{-0.1}$. After convergence, samples are drawn from the barycenter using underdamped Langevin dynamics. We then apply a discrete OT map to match the color histogram of the input images. The original images exhibit strong red–blue contrast; the averaged images show more similar color tones, with darker blue/green on the left, and an orange tint on the right.



(a) Input images. (b) Images with averaged color palette.

Figure 5: Results of color averaging with stochastic palette barycenter.

## 6 Conclusion

We introduced FRBary, a Fisher–Rao mirror-descent framework for computing unregularized Wasserstein barycenters of discrete, continuous, and heterogeneous input measures. We established best-iterate objective convergence guarantees and demonstrated competitive accuracy across several numerical examples. Future work includes extending the analysis to inexact OT potentials and developing efficient implementations for higher-dimensional distributions.

## Acknowledgments and Disclosure of Funding

X. Chen was partially supported by NSF DMS-2413404 and a gift from the Simons Foundation. C. Zhu was supported by NSF DMS-2412832.

## References


Pankaj Agarwal, Sharath Raghvendra, Pouyan Shirzadian, and Keegan Yao. A combinatorial algorithm for the semi-discrete optimal transport problem. *Advances in Neural Information Processing Systems*, 37:25857–25887, 2024.

M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.*, 43(2):904–924, 2011. ISSN 0036-1410. doi: 10.1137/100805741.

Pedro C Álvarez-Esteban, E Del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2): 744–762, 2016.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Ethan Anderes, Steffen Borgwardt, and Jacob Miller. Discrete Wasserstein barycenters: Optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84(2):389–409, 2016.

Martin Bauer, Martins Bruveris, and Peter W Michor. Uniqueness of the Fisher–Rao metric on the space of smooth densities. *Bulletin of the London Mathematical Society*, 48(3):499–506, 2016.

Cédric Beaulac and Jeffrey S Rosenthal. Introducing a new high-resolution handwritten digits data set with writer characteristics. *SN Computer Science*, 4(1):66, 2022.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003. ISSN 0167-6377. doi: https://doi.org/10.1016/S0167-6377(02)00231-6. URL `https://www.sciencedirect.com/science/article/pii/S0167637702002316`.

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. doi: 10.1137/141000439.

Jérémie Bigot and Thierry Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *ESAIM: Probability and Statistics*, 22:35–57, 2018.

Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Geodesic PCA in the Wasserstein space by convex PCA. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53(1):1 – 26, 2017. doi: 10.1214/15-AIHP706.

Clément Bonet, Théo Uscidda, Adam David, Pierre-Cyril Aubin-Frankowski, and Anna Korba. Mirror and preconditioned gradient descent in Wasserstein space. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 25311–25374. Curran Associates, Inc., 2024. doi: 10.52202/079017-0798. URL `https://proceedings.neurips.cc/paper_files/paper/2024/hash/2cf153951b5e9b39564fc4a0ef6adc1a-Abstract-Conference.html`.

Siwan Boufadene and François-Xavier Vialard. On the global convergence of Wasserstein gradient flow of the Coulomb discrepancy. *SIAM Journal on Mathematical Analysis*, 57(4):4556–4587, 2025. doi: 10.1137/24M1647631.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991. doi: https://doi.org/10.1002/cpa.3160440402. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160440402`.

Guillaume Carlier, Adam Oberman, and Edouard Oudet. Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM: M2AN*, 49(6):1621–1642, 11 2015. doi: 10.1051/m2an/2015033. URL `https://doi.org/10.1051/m2an/2015033`.

Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. Wasserstein regression. *Journal of the American Statistical Association*, 118(542):869–882, 2023.

Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J. Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1276–1304. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/chewi20a.html.

Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In *Proceedings of the 31st International Conference on Machine Learning*, pages 685–693, 2014.

Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal Transport Tools (OTT): A JAX Toolbox for all things Wasserstein, 2022. URL https://arxiv.org/abs/2201.12324. arXiv:2201.12324.

Jiaojiao Fan, Amirhossein Taghvaei, and Yongxin Chen. Scalable computations of Wasserstein barycenter via input convex neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1571–1581. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/fan21d.html.

Rémi Flamary, Cédric Vincent-Cuaz, Nicolas Courty, Alexandre Gramfort, Oleksii Kachaiev, Huy Quang Tran, Laurène David, Clément Bonet, Nathan Cassereau, Théo Gnassounou, Eloi Tanguy, Julie Delon, Antoine Collas, Sonia Mazelet, Laetitia Chapel, Tanguy Kerdoncuff, Xizheng Yu, Matthew Feickert, Paul Krzakala, Tianlin Liu, and Eduardo Fernandes Montesuma. POT Python Optimal Transport (version 0.9.5), 2024. URL https://github.com/PythonOT/POT.

Dongdong Ge, Haoyue Wang, Zikai Xiong, and Yinyu Ye. Interior-point methods strike back: Solving the Wasserstein barycenter problem. *Advances in Neural Information Processing Systems*, 32, 2019.

Ferdinand Genans, Antoine Godichon-Baggioni, François-Xavier Vialard, and Olivier Wintenberger. Stochastic optimization in semi-discrete optimal transport: Convergence analysis and minimax rate. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 51490–51527. Curran Associates, Inc., 2025. doi: 10.52202/085713-1716. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/49f7b16c2799e4f12feb1d037a39fd37-Paper-Conference.pdf.

Nhat Ho, Xuan Long Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. Multilevel clustering via Wasserstein means. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1501–1509. JMLR.org, 2017.

M. Jacobs and F. Léger. A fast approach to optimal transport: the back-and-forth method. *Numer. Math.*, 146(3):513–544, 2020. ISSN 0029-599X. doi: 10.1007/s00211-020-01154-8.

Matt Jacobs, Wonjun Lee, and Flavien Léger. The back-and-forth method for Wasserstein gradient flows. *ESAIM: Control, Optimisation and Calculus of Variations*, 27:28, 2021.

Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Debiased Sinkhorn barycenters. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4692–4701. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/janati20a.html.

Kaheon Kim, Rentian Yao, Changbo Zhu, and Xiaohui Chen. Optimal transport barycenter via nonconvex concave minimax optimization. In *International Conference on Machine Learning (ICML)*, July 2025.

Kaheon Kim, Bohan Zhou, Changbo Zhu, and Xiaohui Chen. Sobolev gradient ascent for optimal transport: Barycenter optimization and convergence analysis. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IjL1xEoxXi.

Young-Heon Kim and Brendan Pass. Wasserstein barycenters over Riemannian manifolds. *Advances in Mathematics*, 307:640–683, 2017. ISSN 0001-8708. doi: 10.1016/j.aim.2016.11.026. URL https://www.sciencedirect.com/science/article/pii/S0001870815304643.

Alexander Kolesov, Petr Mokrov, Igor Udovichenko, Milena Gazdieva, Gudmund Pammer, Evgeny Burnaev, and Alexander Korotin. Estimating barycenters of distributions with neural optimal transport. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 25016–25041. PMLR, 21–27 Jul 2024.

Alexander Korotin, Lingxiao Li, Justin Solomon, and Evgeny Burnaev. Continuous Wasserstein-2 barycenter estimation without minimax optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3tFAs5E-Pe.

Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168(3–4):901–917, 2017. doi: 10.1007/s00440-016-0727-z.

Hugo Leclerc and Quentin Mérigot. Semi-discret optimal transportation tools, 2025. URL https://github.com/sd-ot.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.

Eduardo Fernandes Montesuma, Yassir Bendou, and Mike Gartrell. Wasserstein gradient flows for scalable and regularized barycenter computation, 2025. URL https://arxiv.org/abs/2510.04602. arXiv:2510.04602.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision*, pages 435–446, 2012.

Ritwik Sadhu, Ziv Goldfeld, and Kengo Kato. Stability and statistical inference for semidiscrete optimal transport maps. *The Annals of Applied Probability*, 34(6):5694–5736, 2024.

F. Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser, 2015. ISBN 978-3-319-20827-5; 978-3-319-20828-2. doi: 10.1007/978-3-319-20828-2. Calculus of variations, PDEs, and modeling.

Justin Solomon, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics*, 34(4), 2015.

Sanvesh Srivastava, Cheng Li, and David B. Dunson. Scalable Bayes via barycenter in Wasserstein space. *The Journal of Machine Learning Research*, 19(1):312–346, 2018.

Bahar Taşkesen, Soroosh Shafieezadeh-Abadeh, and Daniel Kuhn. Semi-discrete optimal transport: Hardness, regularization and numerical solution. *Mathematical Programming*, 199(1):1033–1106, 2023.

Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003. ISBN 9780821833124. URL https://books.google.com/books?id=R_nWqjq89oEC.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Gabriele Visentin and Patrick Cheridito. Computing optimal transport maps and Wasserstein barycenters using conditional normalizing flows. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=WMIueIRcAm.

WallpaperAccess.com. WallpaperAccess, 2026. URL https://wallpaperaccess.com/.

Peng Xu, Changbo Zhu, Young-Heon Kim, and Xiaohui Chen. Another Look at Log-PCA for Probability Measures: A Dynamical Formulation and Statistical Convergence, 2026. URL https://arxiv.org/abs/2606.17196. arXiv:2606.17196.

Yuling Yan, Kaizheng Wang, and Philippe Rigollet. Learning Gaussian mixtures using the Wasserstein–Fisher–Rao gradient flow. *The Annals of Statistics*, 52(4):1774–1795, 2024.

Lei Yang, Jia Li, Defeng Sun, and Kim-Chuan Toh. A fast globally linearly convergent algorithm for the computation of Wasserstein barycenters. *Journal of Machine Learning Research*, 22(21): 1–37, 2021.

Rentian Yao, Linjun Huang, and Yun Yang. Minimizing convex functionals over space of probability measures via KL divergence gradient flow. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2530–2538. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/yao24a.html.

Rentian Yao, Atsushi Nitanda, Xiaohui Chen, and Yun Yang. Learning density evolution from snapshot data, 2025. URL https://arxiv.org/abs/2502.17738. arXiv:2502.17738.

Changbo Zhu and Hans-Georg Müller. Autoregressive optimal transport models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):1012–1033, 2023.

Changbo Zhu and Hans-Georg Müller. Geodesic optimal transport regression. *Biometrika*, page asaf086, 12 2025. ISSN 1464-3510. doi: 10.1093/biomet/asaf086.

Shuailong Zhu and Xiaohui Chen. Convergence analysis of the Wasserstein proximal algorithm beyond geodesic convexity, 2025. URL https://arxiv.org/abs/2501.14993. arXiv:2501.14993.

Yubo Zhuang, Xiaohui Chen, and Yun Yang. Wasserstein $K$-means for clustering probability distributions. In *Proceedings of 36th Conference on Neural Information Processing Systems*, 2022.

# A Technical Details

## A.1 Technical Lemmas

**Lemma A.1.** *Let $\Omega \subset \mathbb{R}^d$ be compact with*

$$R \coloneqq \sup_{x\in\Omega} \|x\|.$$

*Fix the quadratic cost $c(x,y) = \|x-y\|^2/2$. Let $\lambda \in \mathcal{P}_{\text{ac}}(\Omega)$ and let $\mu \in \mathcal{P}(\Omega)$ be arbitrary (in particular, $\mu$ may be absolutely continuous or discrete). Let $\phi_{\lambda\to\mu}$ be a $c$-concave Kantorovich potential that is finite on $\Omega$ and normalized so that $\inf_{x\in\Omega} \phi_{\lambda\to\mu}(x) = 0$. Then, we have $\phi_{\lambda\to\mu} \in L^\infty(\Omega)$ and*

$$\|\phi_{\lambda\to\mu}\|_{L^\infty(\Omega)} \le 2R^2.$$

*Proof.* Since $\phi_{\lambda\to\mu}$ is $c$-concave and finite on $\Omega$, there exists an extended-real function $\psi : \Omega \to \mathbb{R}\cup\{-\infty\}$ such that $\phi_{\lambda\to\mu} = \psi^c$, i.e.

$$\phi_{\lambda\to\mu}(x) = \inf_{y\in\Omega}\left\{\frac{1}{2}\|x-y\|_2^2 - \psi(y)\right\}, \qquad x\in\Omega.$$

If $\mu = \sum_{i=1}^m u_i\delta_{x_i}$ is discrete, one may take $\psi(y) = \psi_i$ for $y = x_i$ and $\psi(y) = -\infty$ otherwise, so that the infimum reduces to a minimum over $i$. Fix $x, x' \in \Omega$ and $\varepsilon > 0$. Since $\phi_{\lambda\to\mu}(x')$ is finite, there exists $y_\varepsilon \in \Omega$ with $\psi(y_\varepsilon) \in \mathbb{R}$ such that

$$\phi_{\lambda\to\mu}(x') \ge \frac{1}{2}\|x'-y_\varepsilon\|_2^2 - \psi(y_\varepsilon) - \varepsilon.$$

By definition of the $c$-transform, for the same $y_\varepsilon$,

$$\phi_{\lambda\to\mu}(x) \le \frac{1}{2}\|x-y_\varepsilon\|_2^2 - \psi(y_\varepsilon).$$

Subtracting yields

$$\begin{aligned}
\phi_{\lambda\to\mu}(x) - \phi_{\lambda\to\mu}(x') &\le \frac{1}{2}\big(\|x-y_\varepsilon\|_2^2 - \|x'-y_\varepsilon\|_2^2\big) + \varepsilon \\
&\le \sup_{y\in\Omega}\frac{1}{2}\big(\|x-y\|_2^2 - \|x'-y\|_2^2\big) + \varepsilon \\
&\le \sup_{y\in\Omega}\frac{1}{2}\|x-y\|_2^2 + \varepsilon \\
&\le 2R^2 + \varepsilon,
\end{aligned}$$

since $\|x-y\|_2 \le 2R$ for all $x, y \in \Omega$. Letting $\varepsilon \downarrow 0$ gives $\phi_{\lambda\to\mu}(x) - \phi_{\lambda\to\mu}(x') \le 2R^2$. Swapping $(x, x')$ yields

$$|\phi_{\lambda\to\mu}(x) - \phi_{\lambda\to\mu}(x')| \le 2R^2 \qquad \forall x, x' \in \Omega.$$

Normalize $\phi_{\lambda\to\mu}$ so that $\inf_{x\in\Omega}\phi_{\lambda\to\mu}(x) = 0$. Then for any $x \in \Omega$,

$$0 \le \phi_{\lambda\to\mu}(x) \le \inf_{x'\in\Omega}\phi_{\lambda\to\mu}(x') + 2R^2 = 2R^2,$$

hence $\|\phi_{\lambda\to\mu}\|_{L^\infty(\Omega)} \le 2R^2$. □

The following Lemma follows from the proof of Theorem 5 in Yao et al. [2025].

**Lemma A.2.** *Let $\rho^k, \rho^{k+1}$ satisfy*

$$\rho^{k+1}(x) = \frac{1}{Z_k}\cdot\rho^k(x)\exp\big(-\eta g_k(x)\big),$$

*where $Z_k \coloneqq \int_\Omega \rho^k(y)\exp\big(-\eta g_k(y)\big)\,\mathrm{d}y$ and $g_k \in L^\infty(\Omega)$. Then*

$$\mathrm{D}_{\mathrm{KL}}(\rho^k \parallel \rho^{k+1}) \le \frac{\eta^2}{2}\|g_k\|_{L^\infty(\Omega)}^2.$$

*Proof.* We let $c_k \coloneqq \log Z_k$, then it holds that $\log[\rho^k(x)/\rho^{k+1}(x)] = \eta g_k(x) + c_k$. Using the standard identity for symmetrized KL,

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(\rho^k \parallel \rho^{k+1}) + \mathrm{D}_{\mathrm{KL}}(\rho^{k+1} \parallel \rho^k) &= \int_\Omega \log \frac{\rho^k}{\rho^{k+1}} \,\mathrm{d}(\rho^k - \rho^{k+1}) \\ &= \int_\Omega \big(\eta g_k + c_k\big) \,\mathrm{d}(\rho^k - \rho^{k+1}) \\ &= \eta \int_\Omega g_k \,\mathrm{d}(\rho^k - \rho^{k+1}),\end{aligned}$$

where the constant term vanishes since $\int_\Omega \mathrm{d}(\rho^k - \rho^{k+1}) = 0$. Hence, by $L^\infty$–$L^1$ duality,

$$\mathrm{D}_{\mathrm{KL}}(\rho^k \parallel \rho^{k+1}) + \mathrm{D}_{\mathrm{KL}}(\rho^{k+1} \parallel \rho^k) \le \eta \|g_k\|_{L^\infty(\Omega)} \|\rho^k - \rho^{k+1}\|_{L^1(\Omega)}. \tag{7}$$

Next, by Pinsker's inequality,

$$\mathrm{D}_{\mathrm{KL}}(\rho^{k+1} \parallel \rho^k) \ge \frac{1}{2} \|\rho^{k+1} - \rho^k\|^2_{L^1(\Omega)}. \tag{8}$$

Combining (7)–(8) yields

$$\mathrm{D}_{\mathrm{KL}}(\rho^k \parallel \rho^{k+1}) \le \eta \|g_k\|_{L^\infty(\Omega)} \cdot a - \frac{1}{2} a^2, \qquad a \coloneqq \|\rho^k - \rho^{k+1}\|_{L^1(\Omega)}.$$

Finally, maximizing the right-hand side over $a \ge 0$ gives $\eta \|g_k\|_\infty \cdot a - a^2/2 \le (\eta^2/2)\|g_k\|^2_\infty$, which completes the proof. □

**Gaussian smoothing.** Here, we describe the procedure of Gaussian smoothing and state two lemmas related to it. Let $\widehat{\mu} = \sum_{j=1}^m u_j \delta_{x_j}$ be a discrete probability measure on $\Omega \subset \mathbb{R}^d$, and let $X \sim \widehat{\mu}$ and $Z \sim \mathcal{N}(0, I_d)$ be independent, where $\mathcal{N}(0, I_d)$ denotes the standard $d$-dimensional normal distribution. For $\varepsilon > 0$, define $\widetilde{\mu}_\varepsilon$ to be the distribution of $Y \coloneqq X + \varepsilon Z$. Then $\widetilde{\mu}_\varepsilon$ is absolutely continuous with density $\widetilde{\rho}_\varepsilon$ given by the Gaussian mixture

$$\widetilde{\rho}_\varepsilon(y) = \sum_{j=1}^m u_j \widetilde{f}_\varepsilon(y - x_j),$$

where

$$\widetilde{f}_\varepsilon(x) = (2\pi\varepsilon^2)^{-d/2} \exp\bigg(-\frac{\|x\|_2^2}{2\varepsilon^2}\bigg).$$

Let $\mu_\varepsilon$ denote the distribution obtained by truncating $\widetilde{\mu}_\varepsilon$ to $\Omega$ and renormalizing: $\mu_\varepsilon(A) \coloneqq \widetilde{\mu}_\varepsilon(A \cap \Omega)/\widetilde{\mu}_\varepsilon(\Omega)$ for any $A \in \mathcal{B}(\mathbb{R}^d)$. Then, $\mu_\varepsilon$ admits a density on $\Omega$, denoted by $\rho_\varepsilon$, of the form $\rho_\varepsilon(x) = \widetilde{\rho}_\varepsilon(x)/\widetilde{\mu}_\varepsilon(\Omega), x \in \Omega$.

Let $\partial\Omega$ be the boundary of $\Omega$ and $\mathrm{dist}(x, \partial\Omega) = \inf_{y \in \partial\Omega} \|x - y\|_2$. The next lemma bounds the 2-Wasserstein distance between $\mu_\varepsilon$ and $\widehat{\mu}$.

**Lemma A.3.** *Assuming that $\Omega$ is compact and there exists $r > 0$ such that for any $j$,*

$$\mathrm{dist}(x_j, \partial\Omega) \ge r.$$

*Then, for any $\varepsilon \in \big(0, r/(2\sqrt{d})\big]$, we have*

$$W_2(\mu_\varepsilon, \widehat{\mu}) \le \sqrt{d}\,\varepsilon + C_{R,d} \exp\bigg(-\frac{r^2}{32\,\varepsilon^2}\bigg),$$

*where $C_{R,d} > 0$ is a constant depending on $d$ and $R = \sup_{x \in \Omega} \|x\|_2$.*

*Proof.* Let $X \sim \widehat{\mu}$ and $Z \sim \mathcal{N}(0, I_d)$ be independent, and define $Y \coloneqq X + \varepsilon Z$. Then $Y \sim \widetilde{\mu}_\varepsilon$. Consider the coupling between $\widehat{\mu}$ and $\widetilde{\mu}_\varepsilon$ given by the joint law of $(X, Y)$. By definition of $W_2$,

$$W_2^2(\widetilde{\mu}_\varepsilon, \widehat{\mu}) \le \mathbb{E}\|Y - X\|_2^2 = \varepsilon^2\, \mathbb{E}\|Z\|_2^2 = d\varepsilon^2.$$

Next, let $\mu_\varepsilon$ be the truncation and renormalization of $\widetilde{\mu}_\varepsilon$ to $\Omega$. By the triangle inequality,

$$W_2(\mu_\varepsilon, \widehat{\mu}) \le W_2(\mu_\varepsilon, \widetilde{\mu}_\varepsilon) + W_2(\widetilde{\mu}_\varepsilon, \widehat{\mu}).$$

We then bound $W_2(\mu_\varepsilon, \widetilde{\mu}_\varepsilon)$. Let $X' \sim \mu_\varepsilon$ be independent of $(X, Z)$, and define

$$\widehat{Y} = \begin{cases} Y, & Y \in \Omega, \\ X', & Y \notin \Omega. \end{cases}$$

Since the conditional distribution of $Y$ given $Y \in \Omega$ is $\mu_\varepsilon$, we have $\widehat{Y} \sim \mu_\varepsilon$. Thus, $(Y, \widehat{Y})$ is a coupling of $\widetilde{\mu}_\varepsilon$ and $\mu_\varepsilon$, entailing

$$W_2^2(\mu_\varepsilon, \widetilde{\mu}_\varepsilon) \le \mathbb{E}\|Y - \widehat{Y}\|_2^2 = \mathbb{E}\left[\|Y - X'\|_2^2 \mathbf{1}_{\{Y \notin \Omega\}}\right].$$

Since $X' \in \Omega$, we have $\|X'\|_2 \le R$. Moreover, writing $Y = X + \varepsilon Z$ with $\|X\|_2 \le R$, we obtain $\|Y\|_2 \le R + \varepsilon\|Z\|_2$. Hence, by the triangle inequality,

$$\|Y - X'\|_2 \le \|Y\|_2 + \|X'\|_2 \le 2R + \varepsilon\|Z\|_2.$$

Therefore, we have $W_2^2(\mu_\varepsilon, \widetilde{\mu}_\varepsilon) \le \mathbb{E}\left[(2R + \varepsilon\|Z\|_2)^2 \mathbf{1}_{\{Y \notin \Omega\}}\right]$. Finally, since $\{Y \notin \Omega\} \subseteq \{\varepsilon\|Z\|_2 \ge r\}$, we obtain

$$\begin{aligned} W_2^2(\mu_\varepsilon, \widetilde{\mu}_\varepsilon) &\le \mathbb{E}\big[(2R + \varepsilon\|Z\|_2)^2 \mathbf{1}_{\{\|Z\|_2 \ge r/\varepsilon\}}\big] \\ &\le \left\{8(2R)^4 + 8\varepsilon^4\,\mathbb{E}\|Z\|_2^4\right\}^{1/2} \mathbb{P}(\|Z\|_2 \ge r/\varepsilon)^{1/2} \\ &\le C_{R,d}\,\mathbb{P}(\|Z\|_2 \ge r/\varepsilon)^{1/2} \\ &\le C_{R,d} \exp\left(-\frac{r^2}{16\varepsilon^2}\right). \end{aligned}$$

The lemma is concluded by combining the results. □

**Lemma A.4.** *Let $\widehat{\mu} = \sum_{j=1}^m u_j \delta_{x_j}$ and $\rho^0$ be a density on $\Omega$ satisfying $0 < a \le \rho^0(x) \le b < \infty$ for all $x \in \Omega$. Assuming that $\Omega \subset \mathbb{R}^d$ is compact, and there exists $r > 0$ such that for any $j$,*

$$\operatorname{dist}(x_j, \partial\Omega) \ge r.$$

*Then, for any $\varepsilon \in \big(0, r/(2\sqrt{d})\big]$, we have*

$$\mathrm{D}_{\mathrm{KL}}(\rho_\varepsilon \parallel \rho^0) \le C_{\mathrm{KL}} + d\log(1/\varepsilon),$$

*where $C_{\mathrm{KL}} \coloneqq -(d/2)\log(2\pi) - \log(1 - \exp(-d/2)) - \log a$.*

*Proof.* Recall that $\rho_\varepsilon(x) = \widetilde{\rho}_\varepsilon(x)/\widetilde{\mu}_\varepsilon(\Omega)$ for $x \in \Omega$, where

$$\widetilde{\rho}_\varepsilon(x) = \sum_{j=1}^m u_j\, \widetilde{f}_\varepsilon(x - x_j), \quad \text{and} \quad \widetilde{f}_\varepsilon(z) = (2\pi\varepsilon^2)^{-d/2} \exp\left(-\frac{\|z\|_2^2}{2\varepsilon^2}\right).$$

We decompose

$$\mathrm{D}_{\mathrm{KL}}(\rho_\varepsilon \parallel \rho^0) = \int_\Omega \rho_\varepsilon(x) \log \rho_\varepsilon(x)\,\mathrm{d}x - \int_\Omega \rho_\varepsilon(x) \log \rho^0(x)\,\mathrm{d}x.$$

Since $\widetilde{\rho}_\varepsilon(x) \le (2\pi\varepsilon^2)^{-d/2}$ for all $x$, we have

$$\rho_\varepsilon(x) \le \frac{(2\pi\varepsilon^2)^{-d/2}}{\widetilde{\mu}_\varepsilon(\Omega)}.$$

Let $\alpha_\varepsilon \coloneqq \widetilde{\mu}_\varepsilon(\Omega) = \mathbb{P}(Y \in \Omega)$ and $\delta_\varepsilon \coloneqq 1 - \alpha_\varepsilon = \mathbb{P}(Y \notin \Omega)$. Since $\operatorname{dist}(X, \partial\Omega) \ge r$ a.s., $\{Y \notin \Omega\} \subseteq \{\|Y - X\|_2 \ge r\} \subseteq \{\varepsilon\|Z\|_2 \ge r\}$. Hence, for $\varepsilon \le r/(2\sqrt{d})$,

$$\delta_\varepsilon \le \mathbb{P}(\|Z\|_2 \ge r/\varepsilon) \le \exp\left(-\frac{(r/\varepsilon - \sqrt{d})^2}{2}\right) \le \exp\left(-\frac{r^2}{8\,\varepsilon^2}\right),$$

where we used the Gaussian norm tail bound $\mathbb{P}(\|Z\|_2 \ge t) \le \exp(-(t - \sqrt{d})^2/2)$ for $t \ge \sqrt{d}$. This also implies that

$$\alpha_\varepsilon \ge 1 - \exp\left(-\frac{r^2}{8\,\varepsilon^2}\right) \ge 1 - \exp\left(-\frac{d}{2}\right). \tag{9}$$

Using $\int_\Omega \rho_\varepsilon = 1$ and $\log \rho_\varepsilon(x) \le \log\|\rho_\varepsilon\|_\infty$,

$$\int_\Omega \rho_\varepsilon(x) \log \rho_\varepsilon(x)\,\mathrm{d}x \le -\frac{d}{2}\log(2\pi) - d\log\varepsilon - \log\alpha_\varepsilon.$$

Since $\rho^0(x) \ge a$ on $\Omega$, we have $-\log\rho^0(x) \le -\log a$, hence

$$-\int_\Omega \rho_\varepsilon(x) \log \rho^0(x)\,\mathrm{d}x \le -\log a.$$

Combining the above bounds yields

$$\mathrm{D_{KL}}(\rho_\varepsilon \parallel \rho^0) \le \left(-\frac{d}{2}\log(2\pi) - \log\alpha_\varepsilon - \log a\right) + d\log(1/\varepsilon).$$

□

### A.2 Proof of Theorem 4.1

*Proof.* If at least one of $\{\mu_1, \dots, \mu_n\}$ is absolutely continuous, then the minimizer $\lambda^\star$ is absolutely continuous. By assumption, $\mathrm{D_{KL}}(\lambda^\star \parallel \lambda^0) < \infty$. Let $\rho^k$ and $\rho^\star$ denote the densities of $\lambda^k$ and $\lambda^\star$, respectively. The KL-mirror descent update in Algorithm 1 takes the form

$$\rho^{k+1}(x) = \frac{1}{Z^k} \cdot \rho^k(x) \exp\{-\eta_k g^k(x)\},$$

where $Z^k := \int_\Omega \rho^k(y)\exp\{-\eta_k g^k(y)\}\,dy$ and $g^k(x) = \frac{\delta\mathcal{E}}{\delta\rho}(\rho^k)(x) = \sum_{i=1}^n w_i\,\phi_{\rho^k\to\mu_i}(x)$. By Lemma A.1, each normalized Kantorovich potential satisfies $\|\phi_{\rho^k\to\mu_i}\|_{L^\infty(\Omega)} \le 2R^2$. Therefore,

$$\|g^k\|_{L^\infty(\Omega)} \le \sum_{i=1}^n w_i \|\phi_{\rho^k\to\mu_i}\|_{L^\infty(\Omega)} \le 2R^2 =: G.$$

Let $\rho$ be any probability density on $\Omega$ satisfying $D_{\mathrm{KL}}(\rho\|\rho^0) < \infty$. From the update formula, we have $\log(\rho^k(x)/\rho^{k+1}(x)) = \eta_k g^k(x) + c^k$, where $c_k := \log Z_k$. Consequently,

$$\begin{aligned}\int_\Omega g^k(x)\big(\rho^k(x) - \rho(x)\big)\,\mathrm{d}x &= \frac{1}{\eta_k}\int_\Omega \log\frac{\rho^k(x)}{\rho^{k+1}(x)}\big(\rho^k(x) - \rho(x)\big)\,\mathrm{d}x \\ &= \frac{1}{\eta_k}\Big(\mathrm{D_{KL}}(\rho^k \parallel \rho^{k+1}) + \mathrm{D_{KL}}(\rho \parallel \rho^k) - \mathrm{D_{KL}}(\rho \parallel \rho^{k+1})\Big).\end{aligned}$$

By Lemma A.2, $\mathrm{D_{KL}}(\rho^k \parallel \rho^{k+1}) \le (\eta_k^2/2)\|g^k\|^2_{L^\infty(\Omega)}$, and hence

$$\int_\Omega g^k(x)\big(\rho^k(x) - \rho(x)\big)\,\mathrm{d}x \le \frac{1}{\eta_k}\Big(\mathrm{D_{KL}}(\rho \parallel \rho^k) - \mathrm{D_{KL}}(\rho \parallel \rho^{k+1})\Big) + \frac{\eta_k}{2}\|g^k\|^2_{L^\infty(\Omega)}.$$

Since $\mathcal{E}$ is convex along linear mixtures of densities, $g^k$ is a subgradient of $\mathcal{E}$ at $\rho^k$, we have for any $\nu \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ with density $\rho$,

$$\mathcal{E}(\rho^k) - \mathcal{E}(\rho) \le \int_\Omega g^k(x)\big(\rho^k(x) - \rho(x)\big)\,\mathrm{d}x.$$

Combining the above inequality with the previous bound yields

$$\eta_k(\mathcal{E}(\rho^k) - \mathcal{E}(\rho)) \le \mathrm{D_{KL}}(\rho \parallel \rho^k) - \mathrm{D_{KL}}(\rho \parallel \rho^{k+1}) + 2R^4\eta_k^2.$$

Summing the above inequality for $k = 0, \dots, T$ gives

$$\sum_{k=0}^T \eta_k\big(\mathcal{E}(\rho^k) - \mathcal{E}(\rho)\big) \le \mathrm{D_{KL}}(\rho \parallel \rho^0) + 2R^4\sum_{k=0}^T \eta_k^2.$$

By taking minimization over $k$, we have

$$\min_{0\le k\le T}\big(\mathcal{E}(\rho^k) - \mathcal{E}(\rho)\big) \le \frac{\mathrm{D_{KL}}(\rho \parallel \rho^0)}{\sum_{k=0}^T \eta_k} + 2R^4\frac{\sum_{k=0}^T \eta_k^2}{\sum_{k=0}^T \eta_k}. \tag{10}$$

Since the above inequality holds for any $\rho$, setting $\rho = \rho^*$ proves the theorem. □

### A.3 Proof of Theorem 4.2

*Proof.* We follow the Gaussian smoothing procedure described above Lemma A.4. Let $\lambda^\star = \sum_{j=1}^m u_j \delta_{x_j}$ be discrete and $X \sim \lambda^\star$. For $\varepsilon > 0$, define $\widetilde{\lambda}_\varepsilon$ to be the distribution of $Y := X + \varepsilon Z$ with density $\widetilde{\rho}_\varepsilon$. Let $\lambda_\varepsilon$ denote the truncated distribution of $\widetilde{\lambda}_\varepsilon$ to $\Omega$. The density of $\lambda_\varepsilon$ is denoted as $\rho_\varepsilon$. By letting $\rho = \rho_\varepsilon$ in Equation 10, we have

$$\min_{0\le k\le T} \big(\mathcal{E}(\lambda^k) - \mathcal{E}(\lambda_\varepsilon)\big) \le \frac{\mathrm{D_{KL}}(\rho_\varepsilon \parallel \rho^0)}{\sum_{k=0}^T \eta_k} + 2R^4 \frac{\sum_{k=0}^T \eta_k^2}{\sum_{k=0}^T \eta_k}.$$

On the other hand, for each $i$, by the triangle inequality, $\big|W_2(\lambda_\varepsilon, \mu_i) - W_2(\lambda^\star, \mu_i)\big| \le W_2(\lambda_\varepsilon, \lambda^\star)$, and since $\Omega$ is compact, $W_2(\cdot,\cdot) \le 2R$ on $\mathcal{P}(\Omega)$, hence $W_2(\lambda_\varepsilon, \mu_i) + W_2(\lambda^\star, \mu_i) \le 4R$. Therefore,

$$W_2^2(\lambda_\varepsilon, \mu_i) - W_2^2(\lambda^\star, \mu_i) \le 4R \cdot W_2(\lambda_\varepsilon, \lambda^\star).$$

Summing over $i$ with weights gives $\mathcal{E}(\lambda_\varepsilon) - \mathcal{E}(\lambda^\star) \le 4R \cdot W_2(\lambda_\varepsilon, \lambda^\star)$. Finally, taking $\min_{0\le k\le T}$ and combining the bounds yields

$$\begin{aligned}\min_{0\le k\le T} \mathcal{E}(\lambda^k) - \mathcal{E}(\lambda^\star) &= \min_{0\le k\le T} \big(\mathcal{E}(\lambda^k) - \mathcal{E}(\lambda_\varepsilon)\big) + \big(\mathcal{E}(\lambda_\varepsilon) - \mathcal{E}(\lambda^\star)\big) \\ &\le \frac{\mathrm{D_{KL}}(\rho_\varepsilon \parallel \rho^0)}{\sum_{k=0}^T \eta_k} + 2R^4 \frac{\sum_{k=0}^T \eta_k^2}{\sum_{k=0}^T \eta_k} + 4R \cdot W_2(\lambda_\varepsilon, \lambda^\star).\end{aligned}$$

Since every support point of $\lambda^\star$ is a weighted average of input support points and $\Omega$ is convex, $\mathrm{dist}(z, \partial\Omega) \ge r$ for every $z \in \mathrm{supp}(\lambda^\star)$. Using Lemmas A.4 and A.3, we obtain

$$\begin{aligned}&\min_{0\le k\le T} \big(\mathcal{E}(\lambda^k) - \mathcal{E}(\lambda^\star)\big) \\ &\le \frac{C_{\mathrm{KL}}}{\sum_{k=0}^T \eta_k} + \frac{d \log(1/\varepsilon)}{\sum_{k=0}^T \eta_k} + 2R^4 \frac{\sum_{k=0}^T \eta_k^2}{\sum_{k=0}^T \eta_k} + 4R\bigg(\sqrt{d}\,\varepsilon + C_{R,d}\,\exp\bigg(-\frac{r^2}{32\,\varepsilon^2}\bigg)\bigg).\end{aligned}$$

Choosing $\varepsilon = T^{-1/2} \le \frac{r}{2\sqrt{d}}$ yields

$$\begin{aligned}&\min_{0\le k\le T} \big(\mathcal{E}(\lambda^k) - \mathcal{E}(\lambda^\star)\big) \\ &\le \frac{C_{\mathrm{KL}}}{\sum_{k=0}^T \eta_k} + \frac{d}{2}\frac{\log T}{\sum_{k=0}^T \eta_k} + 2R^4 \frac{\sum_{k=0}^T \eta_k^2}{\sum_{k=0}^T \eta_k} + 4R\bigg(\sqrt{d}\,T^{-1/2} + C_{R,d}\exp\bigg(-\frac{r^2 T}{32}\bigg)\bigg) \\ &= O\bigg(\frac{\log T}{\sum_{k=0}^T \eta_k} + \frac{\sum_{k=0}^T \eta_k^2}{\sum_{k=0}^T \eta_k} + \frac{1}{\sqrt{T}}\bigg).\end{aligned}$$

For either choice of step size, $\eta_k = 1/\sqrt{k+1}$ or $\eta_k = 1/\sqrt{T}$, the right-hand side is of order $O\Big(\log T/\sqrt{T}\Big)$. □

## B Explicit Mirror Descent for Gaussian Measures

Consider centered Gaussian inputs $\mu_i = \mathcal{N}(0, \Sigma_i)$ with $\Sigma_i \succ 0$. Their Wasserstein barycenter is also a centered Gaussian [Agueh and Carlier, 2011]. If the current mirror-descent iterate is $\lambda^k = \mathcal{N}(0, S_k)$ with $S_k \succ 0$, the associated Kantorovich potentials and KL divergence admit the following closed-form expressions:

$$\phi_{\lambda^k \to \mu_i}(x) = \frac{1}{2} x^\top (I - A_{i,k}) x, \qquad A_{i,k} = S_k^{-1/2} (S_k^{1/2} \Sigma_i S_k^{1/2})^{1/2} S_k^{-1/2},$$

and

$$\mathrm{D_{KL}}(\lambda \parallel \lambda^k) = \frac{1}{2} \left( \mathrm{Tr}\big(S_k^{-1} S\big) - \ln \frac{|S|}{|S_k|} - d \right).$$

By noting $\mathbb{E}_\lambda[\phi_{\lambda^k \to \mu_i}] = [\mathrm{Tr}(S) - \mathrm{Tr}(A_{i,k} S)]/2$, the mirror descent objective (3) can be written explicitly as

$$\mathcal{L}(S) = \frac{1}{2} \left[ \mathrm{Tr}(S) - \sum_{i=1}^{n} w_i \, \mathrm{Tr}(A_{i,k} S) + \frac{2}{\eta_k} \, \mathrm{D_{KL}}(\lambda \parallel \lambda^k) \right],$$

along with its gradient

$$\frac{\partial \mathcal{L}}{\partial S} = \frac{1}{2} \left( I - \sum_{i=1}^{n} w_i A_{i,k} + \frac{1}{\eta_k} (S_k^{-1} - S^{-1}) \right).$$

The first-order optimality condition then yields an explicit update rule for the precision matrix:

$$S_{k+1}^{-1} = S_k^{-1} + \eta_k \left( I - \sum_{i=1}^{n} w_i A_{i,k} \right).$$

This update is well defined provided that

$$S_k^{-1} + \eta_k \left( I - \sum_{i=1}^{n} w_i A_{i,k} \right) \succ 0,$$

which holds for sufficiently small $\eta_k > 0$. This approach requires only the computation of geometric means of covariance matrices and therefore its per-iteration computational cost is comparable to that of Bures–Wasserstein gradient descent.

We evaluate this explicit mirror descent method by computing the barycenter of $100$ Gaussian measures on $\mathbb{R}^{50}$ with randomly generated covariance matrices. As shown in Figure 6, the performance of our method closely matches that of Bures–Wasserstein gradient descent, converging only a few iterations after the latter.

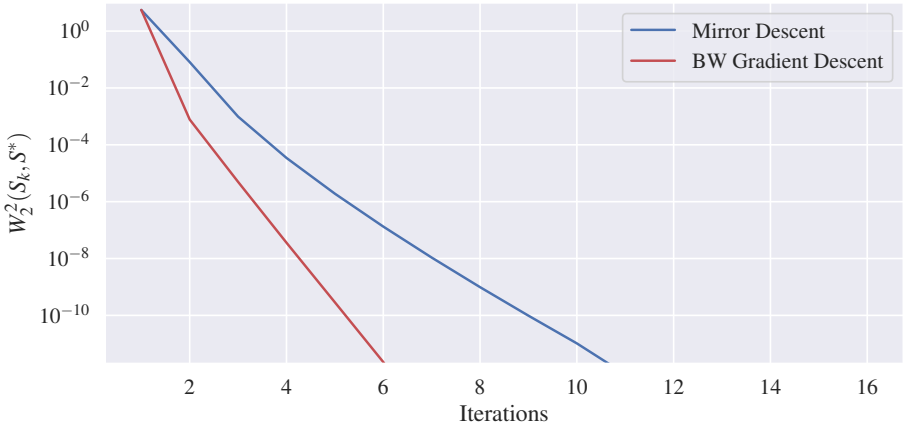



Figure 6: Comparison with Bures–Wasserstein gradient descent. Performance is measured by the Bures–Wasserstein distance to the ground truth.

# C Implementation Details

The collection of input measures $\mu_1, \ldots, \mu_n$ in Algorithm 1 may be of mixed types, with some measures being discrete and others absolutely continuous. In this section, we provide detailed implementation of Algorithm 1 adapted to different input types.

For Gaussian measures, the Kantorovich potential admits a closed-form expression. Outside this special case, the potential needs to be computed using iterative algorithms. Absolutely continuous measures are infinite-dimensional objects and must be approximated by finite-dimensional representations for computational purposes, typically through parameterizations via neural networks or grid-based discretizations. In the following, we divide the discussion into two parts: (1) discretization-based methods, and (2) neural network-based methods.

## C.1 Discretization-based Method

For discrete inputs, each measure $\mu_i$ is represented by a sample matrix $X_i \in \mathbb{R}^{m_i \times d}$, whose rows correspond to i.i.d. samples drawn from $\mu_i$. The optimization variable $\lambda^k$ is discretized over a regular grid of size $M$ and represented by the vector $\boldsymbol{\rho}^k := (\rho_1^k, \ldots, \rho_M^k)^\top$. We solve a semi-discrete optimal transport problem between $\boldsymbol{\rho}^k$ and each $X_i$ at every iteration, and update $\boldsymbol{\rho}^k$ using the resulting Kantorovich potentials. Algorithm 2 provides the full implementation details for this case. The density update in Algorithm 2 is performed in log-space for numerical stability. The $\mathrm{logsumexp}$ function used in Algorithm 2 is defined as

$$\mathrm{logsumexp}(v, \Delta) = \log \sum_{i=1}^{M} \Delta \cdot \exp(v_i),$$

where $\Delta$ denotes the volume of a single grid cell. This operation normalizes the updated density so that it integrates to one over the grid. In Algorithm 2, the semi-discrete optimal transport problems are solved using the `sdot` package [Leclerc and Mérigot, 2025].

For absolutely continuous inputs or histogram inputs (e.g., Section 5.2), each measure is represented by a probability vector $p_i$ supported on the grid $\boldsymbol{y}$, satisfying $p_i \geq 0$ and $p_i^\top \mathbf{1}_M = 1$. The distinction between absolutely continuous measures and histograms lies primarily in the grid resolution: the grid size $M$ is typically large for absolutely continuous measures and relatively small for histogram representations. Accordingly, the Kantorovich potentials are also represented as vectors in $\mathbb{R}^M$, and the objective of the inner optimization (3) reduces to

$$\mathcal{L}(\boldsymbol{\rho}) = \sum_{i=1}^{n} w_i \sum_{j=1}^{M} \boldsymbol{\rho}_j \phi_{\boldsymbol{\rho}^k \to \mu_i}(y_j) + \frac{1}{\eta_k} \sum_{j=1}^{M} \boldsymbol{\rho}_j \ln \frac{\boldsymbol{\rho}_j}{\boldsymbol{\rho}_j^k} = \sum_{i=1}^{n} w_i \boldsymbol{\rho}^\top \boldsymbol{\phi}_{i,k} + \frac{1}{\eta_k} \boldsymbol{\rho}^\top \ln(\boldsymbol{\rho} \oslash \boldsymbol{\rho}_k),$$

where $\oslash$ denotes element-wise division. Differentiating with respect to $\boldsymbol{\rho}$ yields

$$\nabla_{\boldsymbol{\rho}}\mathcal{L} = \sum_{i=1}^{n} w_i \phi_{i,k} + \frac{1}{\eta_k}(\mathbf{1}_M + \ln \boldsymbol{\rho} - \ln \boldsymbol{\rho}_k).$$

All implementation details for this case are provided in Algorithm 3.

**Algorithm 2** Barycenter Algorithm for Point Clouds

**Require:** Initial barycenter $\boldsymbol{\rho}^0$;
Data matrices $X_1, \ldots, X_n$;
Step sizes $\{\eta_k\}_{k\geq 0}$;
Number of iterations $T$.

**for** $k = 0, 1, \ldots, T$ **do**
  **for** $i = 1, \ldots, n$ **do**
    Compute semi-discrete OT between $\boldsymbol{\rho}^k$ and $X_i$;
    Obtain the Kantorovich potentials $\phi_{i,k} \in \mathbb{R}^{m_i}$;
    Compute the $c$-transform of $\phi_{i,k}$ on $\boldsymbol{y}$: $\boldsymbol{\phi}^c_{i,k} = \big(\phi^c_{i,k}(y_1), \ldots, \phi^c_{i,k}(y_M)\big)$;
  Compute the average potential $\bar{\boldsymbol{\phi}}^c_k = \sum_{i=1}^n w_i \boldsymbol{\phi}^c_{i,k}$;
  Update the density by

$$\log \boldsymbol{\rho}^{k+1} \leftarrow \log \boldsymbol{\rho}^k - \eta_k \bar{\boldsymbol{\phi}}^c_k;$$
$$\log \boldsymbol{\rho}^{k+1} \leftarrow \log \boldsymbol{\rho}^{k+1} - \mathrm{logsumexp}(\log \boldsymbol{\rho}^{k+1}, \Delta);$$

**Algorithm 3** Barycenter Algorithm for Histograms

**Require:** Initial barycenter $\boldsymbol{\rho}^0$;
Input histograms $p_1, \ldots, p_n$;
Step sizes $\{\eta_k\}_{k\geq 0}$;
Number of iterations $T$.

**for** $k = 0, 1, \ldots, T$ **do**
  **for** $i = 1, \ldots, n$ **do**
    Compute OT between $\boldsymbol{\rho}^k$ and $p_i$;
    Obtain discrete Kantorovich potentials $\phi_{i,k} \in \mathbb{R}^M$;
  Compute the average potential $\bar{\boldsymbol{\phi}}_k = \sum_{i=1}^n w_i \phi_{i,k}$;
  Update the density by

$$\log \boldsymbol{\rho}^{k+1} \leftarrow \log \boldsymbol{\rho}^k - \eta_k \left(\mathbf{1}_M + \bar{\boldsymbol{\phi}}_k\right);$$
$$\log \boldsymbol{\rho}^{k+1} \leftarrow \log \boldsymbol{\rho}^{k+1} - \mathrm{logsumexp}(\log \boldsymbol{\rho}^{k+1}, 1);$$

### C.2 Neural Network-based Method

Our proposed algorithm provides a general framework for computing barycenters and is not restricted to a fixed grid. In particular, we can parameterize the barycenter log-density by a neural network $h$. Semi-discrete OT potentials can be estimated using stochastic optimization [Genans et al., 2025]. Here, we instead approximate integration using fixed Sobol quadrature nodes on a bounded domain $\Omega$, avoiding sampling from the evolving barycenter density during training.

At each iteration, the normalized values of $\exp(h)$ at these nodes define quadrature masses for computing the OT potentials against each input measure, represented by a weighted point cloud. We evaluate the corresponding $c$-transforms at the quadrature nodes, average them across inputs, and apply the multiplicative mirror-descent update in log-space. The network is then fitted to the updated log-density values using squared-error regression. The quadrature nodes and pairwise costs remain fixed throughout optimization, allowing the cost matrices to be cached. Algorithm 4 summarizes this implementation. With fixed quadrature nodes, the inner problem is a discrete approximation of the semi-discrete OT dual.

This procedure approximates the functional update through quadrature and neural-network fitting. The convergence guarantees in Section 4 concern exact functional updates and do not directly account for these approximation errors.

**Algorithm 4** Barycenter Algorithm with Neural Network and Quadrature

**Require:** Neural network $h$;
Input measures $\mu_i = \sum_{j=1}^{m_i} u_{i,j}\delta_{y_{i,j}}$, $i = 1, \ldots, n$;
Bounded integration domain (box) $\Omega$;
Step sizes $\{\eta_k\}_{k\geq 0}$;
Number of quadrature nodes $m$ and iterations $T$.

Initialize $h$ to match the log-density of a chosen distribution;
Generate Sobol quadrature nodes $B_x = (x_1, \ldots, x_m)^\top \in \mathbb{R}^{m\times d}$ on $\Omega$;
Set quadrature weights $\omega_q = |\Omega|/m$ for $q = 1, \ldots, m$;
Define $c(x, y) = \|x - y\|_2^2/2$ and cache $c(x_q, y_{i,j})$ for all $i, q, j$;
Initialize buffer array $B_h \leftarrow h(B_x)$;
**for** $k = 0, 1, \ldots, T-1$ **do**
  Compute normalized quadrature masses $p_q^k \leftarrow \omega_q e^{h(x_q)} / \sum_{r=1}^m \omega_r e^{h(x_r)}$;
  **for** $i = 1, \ldots, n$ **do**
    Compute the discrete-side Kantorovich potential

$$\phi_{i,k} \leftarrow \operatorname*{arg\,max}_{\phi\in\mathbb{R}^{m_i}} \left\{ \sum_{j=1}^{m_i} u_{i,j}\phi_j + \sum_{q=1}^{m} p_q^k \min_{1\leq j\leq m_i} \{c(x_q, y_{i,j}) - \phi_j\} \right\};$$

    Compute $\phi_{i,k}^c(x_q) \leftarrow \min_{1\leq j\leq m_i}\{c(x_q, y_{i,j}) - (\phi_{i,k})_j\}$ for $q = 1, \ldots, m$;
  Compute the average Kantorovich potential $\bar{\phi}_k^c(B_x) \leftarrow \sum_{i=1}^n w_i \phi_{i,k}^c(B_x)$;
  Update buffer: $B_h \leftarrow B_h - \eta_k \bar{\phi}_k^c(B_x)$;
  Normalize buffer: $B_h \leftarrow B_h - \operatorname{logsumexp}\big((B_h)_q + \log\omega_q\big)\mathbf{1}_m$;
  Update $h$ by minimizing $m^{-1}\|h(B_x) - B_h\|_2^2$.

### C.3 Limitations

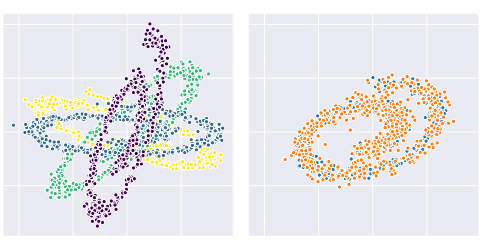
In this work, our implementation primarily focuses on 2D and 3D examples. For the discretization-based approach, we are constrained by the available OT solvers. The grid-based and geometric OT solvers used in this work are particularly effective in 2D and 3D, which cover many classical applications in statistics and imaging.

For the neural network approach, our limited explorations suggest that the method can extend beyond three dimensions, but achieving both accuracy and computational efficiency remains challenging. The quadrature-based implementation avoids sampling from the evolving barycenter during training, but its accuracy depends on the choice of bounded integration domain, quadrature resolution, inner OT solves, and neural-network approximation. Maintaining adequate quadrature accuracy in higher dimensions may require substantially more nodes. Applications requiring sample output also need a separate sampling procedure for the learned density. Developing efficient implementations in higher dimensions is left for future work.

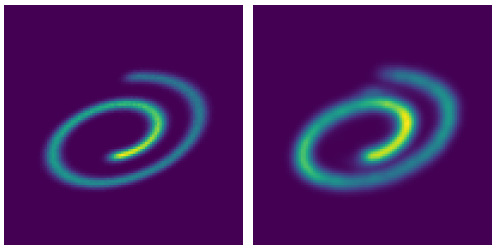
# D Additional Numerical Results

## D.1 Swiss Roll

The two-dimensional Swiss roll dataset from `scikit-learn` [Pedregosa et al., 2011] has been widely used in the Wasserstein barycenter literature as a qualitative benchmark [Korotin et al., 2021, Visentin and Cheridito, 2025, Montesuma et al., 2025]. In the following example, we generate $n = 4$ input Swiss rolls, and sample $10000$ points from each. The input distributions are obtained by applying random linear transformations, for which the true Wasserstein barycenter can be computed. The estimated barycenter density is discretized on a $100 \times 100$ grid. We initialize our algorithm with a uniform density and run it for $T = 150$ iterations, with a decaying learning rate $0.6 \times (k + 1)^{-0.1}$. Afterwards, points are sampled from the estimated density using simple rejection sampling. No additional data are sampled during the optimization. Our results are shown in Figure 7a and Figure 7b.



(a) **Left**: Samples from the input Swiss rolls; **Right**: Samples from the true (blue) and estimated (orange) barycenter.



(b) **Left**: True Swiss roll barycenter density; **Right**: Estimated barycenter density.

Figure 7: Swiss roll.

**Comparison with free-support approaches.** Using the Swiss roll data, we further compared our method (with both discretization and neural network-based method) against several representative free-support barycenter approaches: the neural network–based CW2B [Korotin et al., 2021], as well as exact and entropic solvers from the POT package. We computed the barycenter of 4 Swiss roll distributions, each consisting of 10000 points; the results are shown in Figure 8.

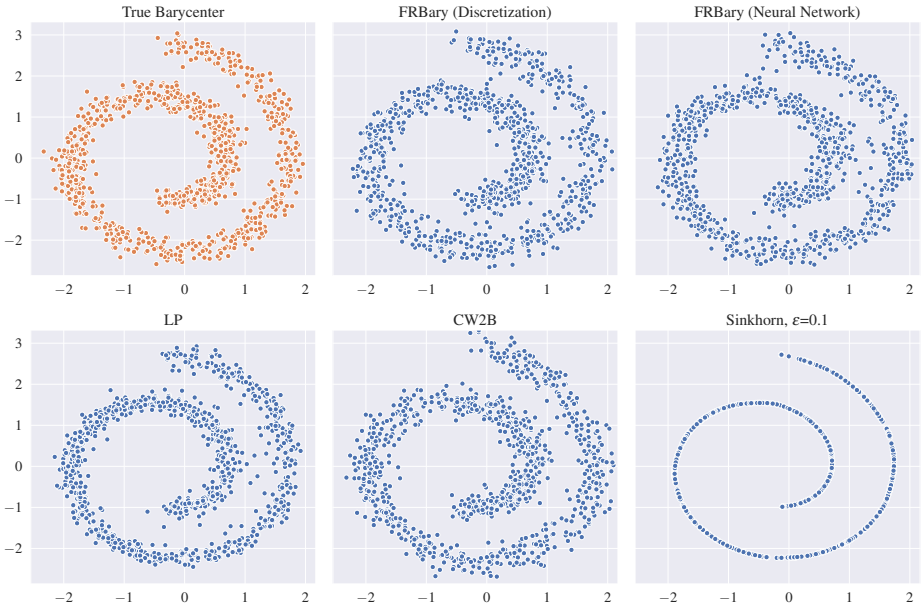



Figure 8: Comparison with various free-support barycenter methods.

For all methods, we used the default parameters provided in their respective implementations. The neural FRBary and the CW2B models were trained on a single NVIDIA 5090 GPU, while all other methods were run on CPUs. Notably, the Sinkhorn approach incurred substantial computational cost and produced results that deviated significantly from the ground truth, consistent with Section 5.1. The computational costs are summarized in Table 3. We also report the sliced Wasserstein distance (SWD) between each converged solution and samples drawn from the true barycenter.

| Solver | Iterations | Time (minutes) | SWD |
|---|---|---|---|
| FRBary (Discretization) | 300 | 11 | 0.0272 |
| FRBary (NN) | 200 | 15 | 0.0635 |
| Exact | 100 | 35 | 0.0255 |
| CW2B | 10000 | 17 | 0.0476 |
| Sinkhorn ($\varepsilon = 0.1$) | 100 | $> 270$ | 0.0516 |

Table 3: Computation costs of free-support solvers.

**Effect of grid resolution.** We investigate the effect of the barycenter grid resolution using the Swiss roll example. After optimization, we sample from the interpolated barycenter density and compute the sliced Wasserstein distance to samples from the true barycenter. As shown in Table 4, a coarse grid substantially degrades accuracy, whereas refinement beyond $30 \times 30$ provides no further improvement in this experiment.

| Grid resolution | SWD |
|---|---|
| $10 \times 10$ | 0.1987 |
| $20 \times 20$ | 0.0748 |
| $30 \times 30$ | 0.0412 |
| $50 \times 50$ | 0.0413 |
| $100 \times 100$ | 0.0445 |

Table 4: Effect of barycenter grid resolution on the Swiss-roll experiment.

### D.2 MNIST

**Scalability with sample size.** To test the scalability of our method with respect to the sample size $n$, we further compute barycenters for all 10 digits using $n = 500$ samples each; the results are shown in Figure 9. The computation of the Kantorovich potentials is parallelized across the input, and each barycenter requires approximately 5 minutes to compute. In contrast, the LP-based implementation we use fails in this setting due to memory limitations.

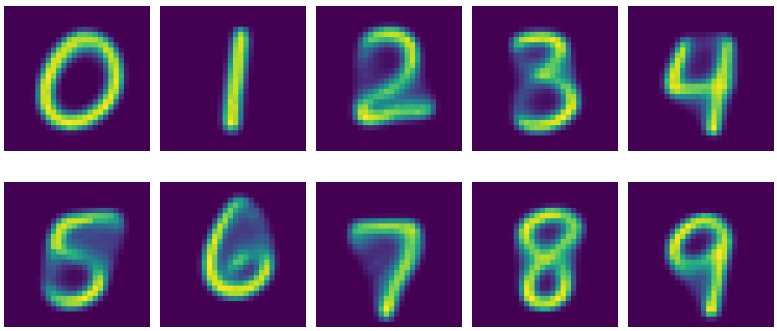


Figure 9: Barycenters computed from $n = 500$ images.

**Effect of the inner OT solver.** We assess the sensitivity of FRBary to the choice of inner OT solver using the MNIST example. Keeping the remaining settings fixed, we compare POT's exact LP solver (POT), the back-and-forth method (BFM), and the semi-discrete solver (SD). We use the LP barycenter $\mu^\star$ as the reference and track $W_2^2(\lambda^k, \mu^\star)$ during optimization. As shown in Figure 10, all three solvers produce visually similar barycenters and substantially reduce the error. POT achieves the lowest error, followed by SD, while BFM exhibits a more pronounced plateau. The higher residual error of BFM may reflect the coarse spatial discretization of the $28 \times 28$ MNIST images. BFM is designed for continuous densities discretized on spatial grids, and its attainable accuracy can depend on grid resolution, particularly when transport maps are discontinuous. Thus, this low-resolution example does not fully reflect its advantages for high-resolution inputs. These results indicate that the qualitative output is robust to the solver choice, although the attainable accuracy and convergence trajectories differ.

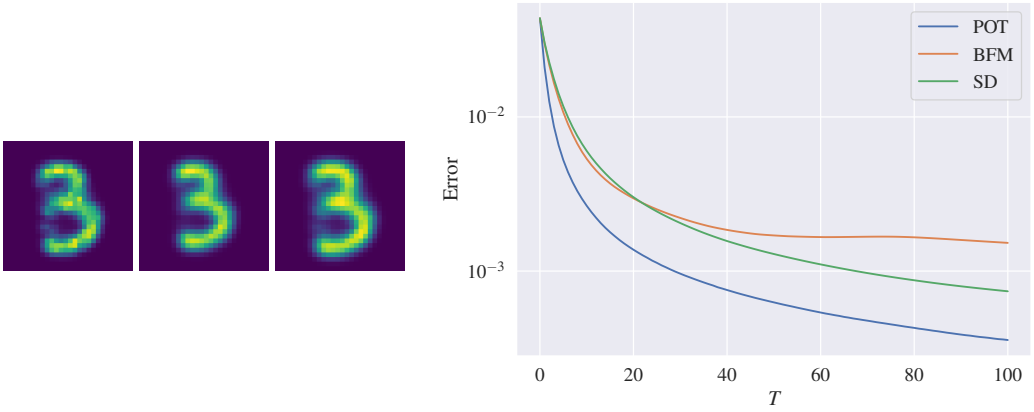



Figure 10: Effect of the inner OT solver on FRBary. **Left**: barycenters obtained using POT, BFM, and SD, respectively. **Right**: $W_2^2$ errors relative to the LP reference barycenter over iterations.

### D.3 High-resolution Digits

In this experiment, we compute the barycenter of 10 high-resolution ($100\times100$ and $200\times200$) images of handwritten digit "3" (Figure 11), obtained by downsampling the HWD+ dataset [Beaulac and Rosenthal, 2022]. We compare against three baselines: sGS-ADMM [Yang et al., 2021], debiased Sinkhorn [Janati et al., 2020], and WDHA [Kim et al., 2025]. Both sGS-ADMM and WDHA are exact barycenter solvers.

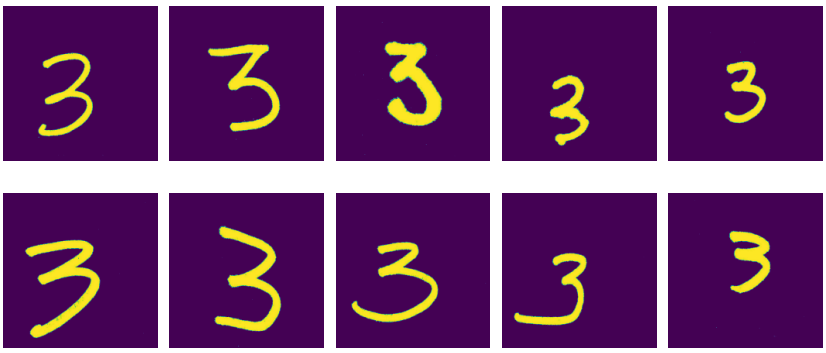
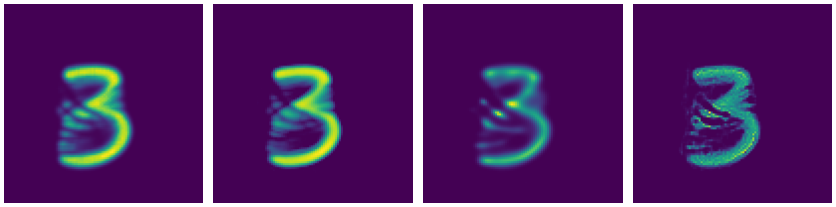
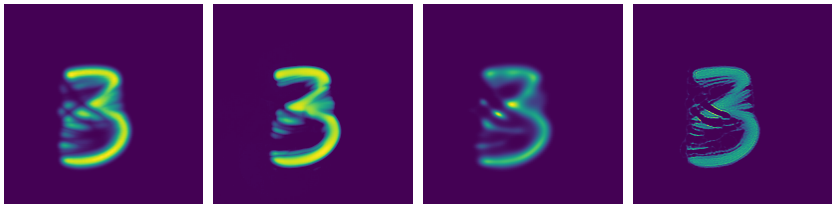


Figure 11: 10 images of the digit "3" from the HWD+ dataset.

The estimated barycenters are shown in Figure 12. In both resolutions, our method and sGS-ADMM produce visually superior barycenters. However, the higher-resolution sGS-ADMM result may exhibit artifacts compared to its lower-resolution counterpart, suggesting slight convergence issues.



(a) Barycenters of $100 \times 100$ images.



(b) Barycenters of $200 \times 200$ images.

Figure 12: **Left to right**: Barycenters computed through FRBary, sGS-ADMM, debiased Sinkhorn ($\varepsilon = 0.005$), and WDHA.

Computation times are summarized in Table 5. For sGS-ADMM, we ran the algorithm until the default convergence threshold ($10^{-5}$) was reached. For debiased Sinkhorn, we set the regularization parameter to $0.005$. For WDHA, we used the default parameters provided by the authors.

**Accuracy versus runtime.** We further compare convergence against wall-clock time for the $100 \times 100$ digit images using the relative objective error $|\mathcal{E}(\lambda^k) - \mathcal{E}^\star|/|\mathcal{E}^\star|$, where $\mathcal{E}^\star$ denotes the reference objective value. As shown in Figure 13, WDHA reaches a low error most quickly, but its trajectory exhibits oscillations. FRBary reduces the error more steadily and reaches a slightly lower final error than WDHA in the reported run, although it requires more time. The sGS-ADMM solver initially progresses more slowly but ultimately attains the lowest error. These results illustrate the trade-off between runtime, convergence behavior, and final accuracy among the tested implementations.

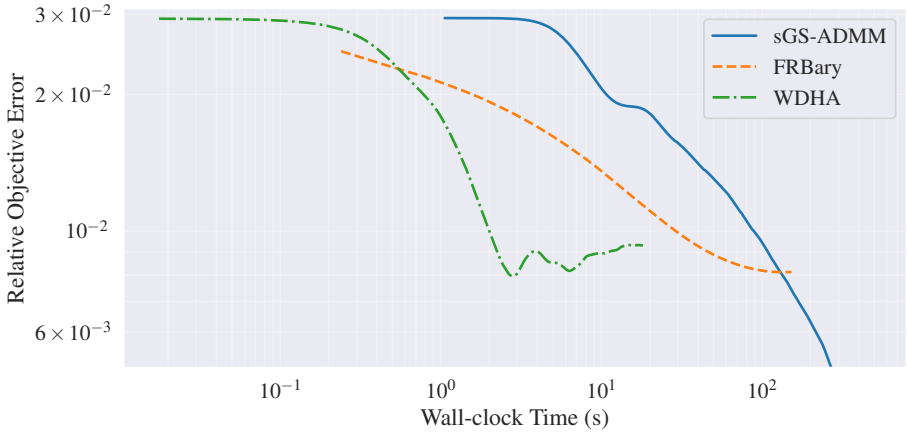



Figure 13: Relative barycenter objective error versus wall-clock time for the $100 \times 100$ digit images. Both axes use logarithmic scales.

Our method exhibits a clear computational advantage over sGS-ADMM: solving the higher-resolution barycenter with sGS-ADMM required more than 4.5 hours. Both our method and WDHA

| Solver | $100 \times 100$ Images | | $200 \times 200$ Images | |
|---|---|---|---|---|
| | Iterations | Time (seconds) | Iterations | Time (seconds) |
| FRBary | 200 | 147 | 200 | 423 |
| sGS-ADMM | 1650 | 473 | 2950 | 16470 |
| Debiased Sinkhorn | 100 | 0.84 | 100 | 6.73 |
| WDHA | 1000 | 5.5 | 1000 | 67 |

Table 5: Computation costs of histogram-based solvers.

use Sobolev gradient-based algorithms [Kim et al., 2026, Jacobs et al., 2021], which provides strong scalability. In particular, when all measures are supported on a regular grid of size $M$, the per-iteration complexity of our method scales as $O(nM \log M)$, with memory complexity $O(M)$. In contrast, sGS-ADMM has per-iteration complexity $O(nM^2)$ and memory complexity $O(M^2)$. As a result, sGS-ADMM becomes impractical for higher-resolution inputs; for $250 \times 250$ images, the provided implementation requires more than 300 GB of memory.

### D.4 Color Palette Averaging with Discretized Barycenter

For the color-averaging experiment, we also explored a discretization-based implementation. The images were resized to $1920 \times 1122$ and $1920 \times 1180$, respectively. The barycenter was discretized on a $50 \times 50 \times 50$ voxel grid. Due to the large input size ($\sim 2 \times 10^6$ pixels per image), the barycenter was estimated stochastically by sampling $1024$ pixels at each iteration. The algorithm was run for 150 iterations using a learning rate schedule of $4 \times (k+1)^{-0.2}$.

After convergence, we computed a semi-discrete OT map between the final estimate and the inputs; pixels were then mapped to the centroids of their corresponding Laguerre cells. The results are shown in Figure 14, while Figure 15 visualizes the color distributions of the input images and the estimated barycenter. Compared to the neural network approach in Figure 5, the transferred images exhibit lower visual quality and noticeable artifacts. Furthermore, because the target points are resampled at every iteration, warm-start strategies for semi-discrete OT cannot be used, resulting in a total runtime of approximately 60 minutes.

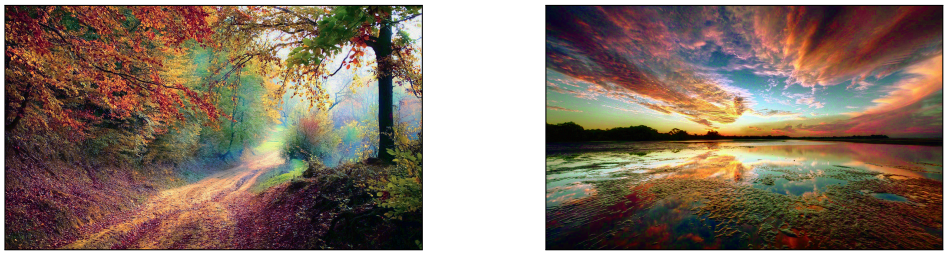


Figure 14: Results of color averaging with discretization-based barycenter.

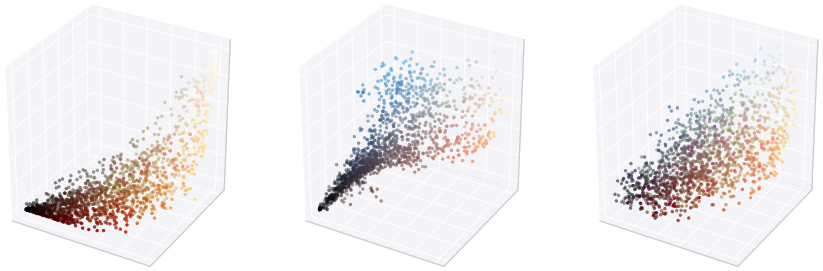


Figure 15: Color palettes of the input images (left and middle) and the barycenter (right).

### D.5 Empirical Convergence Behavior

Figure 16 plots the objective discrepancy against the iteration count for the 2D and 3D Gaussian settings in Section 5.3, using a population reference computed from the Gaussian barycenter characterization. We use a step size decaying as $(k+1)^{-1/2}$ and include $T^{-1/2}$ and $T^{-1}$ reference curves to illustrate the empirical convergence behavior. The observed decay is consistent with the best-iterate upper bound $O(\log T/\sqrt{T})$ in Theorem 4.2.

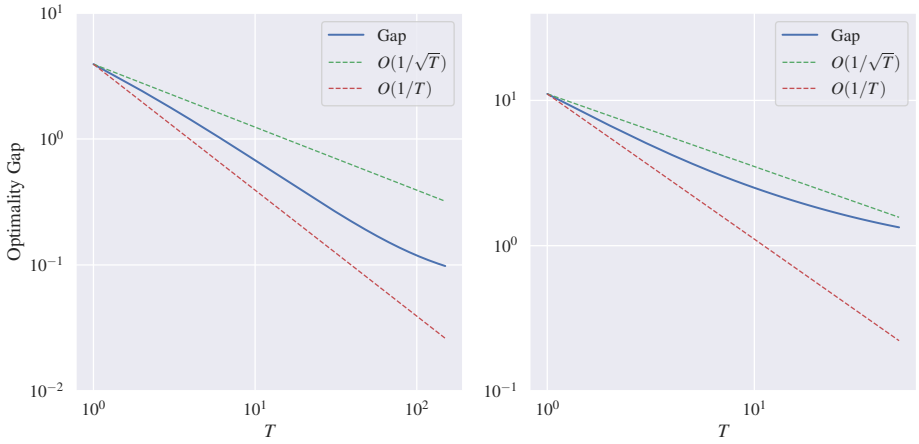



Figure 16: Empirical objective gaps for the 2D and 3D Gaussian experiments, with reference decay curves.

### D.6 Miscellaneous

**Code availability.** Our implementation and notebooks for reproducing the experiments are available at `https://github.com/Francis-Hsu/FRBary`.

**Computation time.** Table 6 reports the total elapsed time for several experiments presented in Section 5 and Appendix D. We observe that the time required to solve the optimal transport subproblems increases as the barycenter approaches convergence. Since the exact solver in the `POT` package does not support warm starts, the computation becomes increasingly expensive in later iterations.

| Experiment | $n$ | $m$ | $M$ | $T$ | Elapsed time |
|---|---|---|---|---|---|
| 2D Gaussian | 4 | 10000 | $100 \times 100$ | 125 | $\sim$ 4 mins 40 secs |
| 3D Gaussian | 3 | 3000 | $50 \times 50 \times 50$ | 100 | $\sim$ 12 mins 6 secs |
| Swiss Rolls | 4 | 10000 | $100 \times 100$ | 150 | $\sim$ 4 mins 50 secs |
| MNIST | 10 | — | $28 \times 28$ | 500 | $\sim$ 9 mins |
| MNIST | 500 | — | $28 \times 28$ | 250 | $\sim$ 5 mins (Parallel) |
| Color Averaging (NN) | 2 | 1024 (Stochastic) | $50 \times 50 \times 50$ | 150 | $\sim$ 11 mins 11 secs |

Table 6: Settings and total runtime for the experiments in Section 5.

**Hardware information.** All experiments in this work, except for one, were conducted on a machine equipped with a single Intel Core i9-14900K CPU, an NVIDIA 5090 GPU, and 32 GB of RAM. Due to its substantial memory requirements, the $200 \times 200$ sGS-ADMM experiment was instead run on a compute node equipped with an AMD EPYC 7713 CPU and 256 GB of RAM.

**License of assets.** The images used in Section 5.4 are obtained from WallpaperAccess.com. The `POT` and `pysdot` packages are distributed under the terms of the MIT license. The `OTT` package is distributed under the Apache-2.0 license. The repository for the back-and-forth OT solver does not explicitly specify a software license.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions, which is a mirror descent algorithm for computing the exact Wasserstein barycenter in the Fisher–Rao geometry.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitations in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions for our main results are discussed in Section 4, and technical details are provided in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The setup for each experiment is described in detail in Section 5 and Appendix D. The provided code covers most of the experiments presented in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide several Jupyter notebooks demonstrating the application of our proposed method across a variety of settings.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The details of the experiments are provided in the descriptions presented in Section 5 and Appendix D.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We do not report error bars, as the experiments are primarily deterministic optimization benchmarks and qualitative evaluations rather than statistical learning experiments. For experiments involving stochastic optimization, the batch sizes are sufficiently large that the resulting variability is minimal.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Hardware information is described in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work proposes an optimization framework for solving the Wasserstein barycenter problem. It does not involve human subjects or participants, nor does it release any datasets. We therefore assume it complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper develops core methodology for computing Wasserstein barycenter. Because the contribution is purely algorithmic and not tied to any specific application or deployment scenario, it introduces no identifiable societal benefits or risks. Broader impacts therefore lie outside the scope of the work.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not release any specific models or datasets, and therefore does not pose such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Software and data mentioned in this paper are properly cited. Their licenses are explicitly mentioned in Appendix D.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.