
Wasserstein Gradient Flows for Scalable and Regularized Barycenter Computation

Eduardo Fernandes Montesuma¹

Yassir Bendou¹

Mike Gartrell^{*2}

¹Sigma Nova, Paris, France

²Rhizome Labs, Paris, France

Abstract

Wasserstein barycenters provide a principled approach for aggregating probability measures, while preserving the geometry of their ambient space. Existing discrete methods are not because as they assume access to the complete set of samples from the input measures. Meanwhile, neural network approaches do scale well, but rely on complex optimization problems and cannot easily incorporate label information. We address these limitations through gradient flows in the space of probability measures. Through time discretization, we achieve a scalable algorithm that i) relies on mini-batch optimal transport, ii) accepts modular regularization through task-aware functions, and iii) seamlessly integrates supervised information into the ground-cost. We empirically validate our approach on domain adaptation benchmarks that span computer vision, neuroscience, and chemical engineering. Our method establishes a new state-of-the-art Wasserstein barycenter solver, with labeled barycenters consistently outperforming unlabeled ones. Our code at <https://github.com/SigmaNova/barycentric-gradient-flows>.

1 INTRODUCTION

Defining the mean or center of a set of probability measures is a fundamental primitive in geometric probability theory [Nielsen, 2020]. Given $K \in \mathbb{N}$ probability measures $\mathcal{Q} = \{Q_k\}_{k=1}^K$ on a metric space (Ω, d) , their *Wasserstein*

barycenter [Agueh and Carlier, 2011] is given by,

$$P^* = \arg \min_{P \in \mathcal{P}_2(\Omega)} \left\{ \mathbb{B}_2(P) = \sum_{k=1}^K \lambda_k \mathbb{W}_2(P, Q_k)^2 \right\}, \quad (1)$$

where $\mathcal{P}_2(\Omega)$ is the set of measures over Ω with finite second moments, and $\lambda = (\lambda_1, \dots, \lambda_K)$, $\lambda_k \geq 0$, $\forall k$, $\sum_{k=1}^K \lambda_k = 1$ is an array of *barycentric coordinates*. Equation 1 defines a notion of the average of the input measures in $\mathcal{Q}$. The main advantage of these barycenters stems from Optimal Transport (OT) [Villani, 2008], that is, they *lift* the geometry of the space (Ω, d) to the space $(\mathcal{P}_2(\Omega), \mathbb{W}_2)$ [Kloeckner, 2010].

Wasserstein barycenters have wide applications to machine learning, such as posterior aggregation in Bayesian inference [Srivastava et al., 2018], model fusion [Singh and Jaggi, 2020], ensembling in time series forecasting [Coz et al., 2023], fairness [Gordaliza et al., 2019, Visentin and Cheridito, 2025], and Domain Adaptation (DA) [Montesuma and Mboula, 2021, Montesuma et al., 2023, 2024b]. In each application, the main advantage is geometric fidelity, that is, the Wasserstein barycenter captures and preserves the underlying structure of samples. Figure 1 shows a conceptual illustration.

Despite their advantages, Wasserstein barycenters remain challenging beyond the Gaussian setting [Altschuler and Boix-Adsera, 2022]. We have identified three gaps in the current literature. First, the seminal algorithm of Cuturi and Doucet [2014] requires having access to the complete set of input measures’ samples at once, making it intractable for large datasets. Neural network methods [Fan et al., 2020, Korotin et al., 2021, Kolesov et al., 2024, Gazdieva et al., 2024] alleviate this issue by operating on mini-batches, but parametrize the barycentric measure with $\mathcal{O}(K)$ neural networks, coupling model complexity with the number of input measures. Second, neural methods cannot seamlessly incorporate label information into their ground-cost, limiting their performance in supervised tasks (cf. Table 2). Third, the objective function in Equation 1 only accounts for distri-

^{*}Work done while at Sigma Nova.

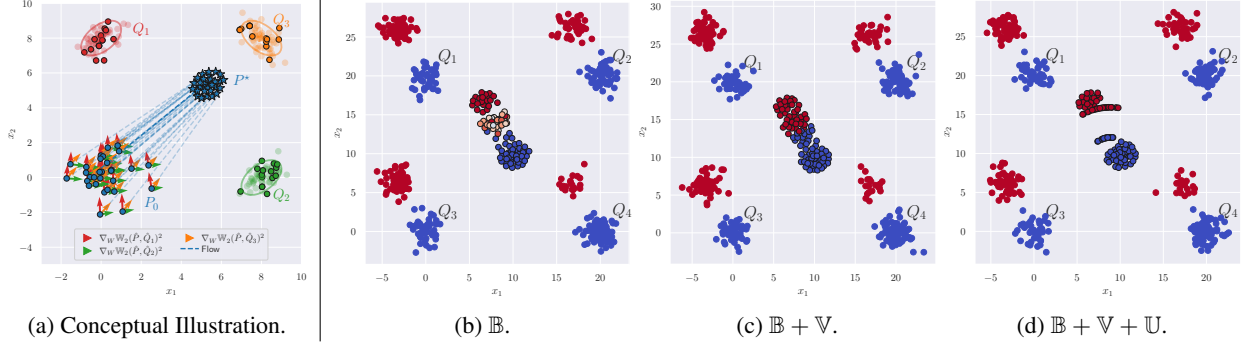


Figure 1: In (a), we show an illustration of our flow. We show the initial barycenter measure P_0 which flow towards the solution P^* . In red, green, and orange, we show the input measures’ mini-batches as solid dots. The flow is given by the combination of gradients $v_{\tau,i} = \sum_{k=1}^K \lambda_k \nabla \mathbb{W}_2(P_{\tau}, Q_k)$. In parallel, (b) shows a pathological barycenter calculation, with fuzzy labels and unclear class boundaries. Adding regularizers $\mathbb{V}$ (b) and $\mathbb{U}$ (c) progressively resolves this issue.

butional fit. In practice, the barycentric measure must satisfy additional structural properties, such as class separation. Current **barycenter solvers** do not offer a principled way of enforcing these properties.

Recent methods study the barycenter problem through the gradient flow perspective. First, Chewi et al. [2020] established a gradient flow algorithm on the Bures-Wasserstein manifold, i.e., on Gaussian measures. Second, Chizat [2025] describes the *doubly regularized* barycenter problem through the same lens. Here, in addition to *entropic regularization* at the level of OT, they proposed an *outer regularization* that penalizes the entropy of the barycentric measure, which leads to a Noisy Particle Gradient Descent (NPGD) algorithm for empirical measures. Although this algorithm incorporates outer regularization, it has the same limitation as Cuturi and Doucet [2014], as it assumes *full-batch* access to the samples of input measures.

We address the aforementioned challenges through the lens of *gradient flows* in the space of probability measures [Ambrosio et al., 2008, Santambrogio, 2017]. We therefore conceptualize the barycenter problem as the flow from an initial measure $P_0 = \mathcal{N}(0, \text{Id})$ following the Wasserstein gradient of the functional,

$$\mathbb{F}(P) = \mathbb{B}(P) + \mathbb{R}(P), \quad (2)$$

where $\mathbb{B}$ denotes the barycenter functional (cf. Equation 1), and $\mathbb{R}$ denotes additional regularizing functionals (see Sections 2.3 and 3.3, and Appendix C.2).

Following the established gradient flow literature [Santambrogio, 2017], we decompose $\mathbb{R}$ into internal, potential, and interaction energies, denoted $\mathbb{G}$, $\mathbb{V}$, and $\mathbb{U}$, respectively. This decomposition generalizes the double regularization strategy of Chizat [2025]. Indeed, setting $\mathbb{V} = \mathbb{U} = 0$ and $\mathbb{G}(P) = \eta_G H(P)$, where $H(P)$ denotes the entropy of P , corresponds to their exact framework. Similarly, Alvarez-Melis and Fusi [2021] devised an approach for flowing a single dataset towards a known target. Our setting differs

structurally from that. First, we go beyond the single measure objective, optimizing the *multi-measure* barycentric objective (cf. Equation 1). Second, while Alvarez-Melis and Fusi [2021] *transforms* an existing dataset towards a known target, ours *synthesizes* a measure (i.e., the barycentric measure) from noise.

Our contributions are as follows. (i) **A mini-batch, time discretized gradient flow algorithm for regularized empirical Wasserstein barycenters (Algorithm 1)**. By randomly sampling the input measures, we reduce the computational complexity of traditional barycenter methods [Cuturi and Doucet, 2014, Montesuma et al., 2023, Chizat, 2025] achieving $2\times$ to $50\times$ speedups with respect to discrete solvers (Figure 5). (ii) **Modular, task-aware regularizing functionals for Wasserstein barycenters (Section 3.3)**. Following previous gradient flow literature [Santambrogio, 2017, Alvarez-Melis and Fusi, 2021], we decompose $\mathbb{R}$ into internal, potential, and interaction energies, allowing incorporation of plug-and-play regularizers into the barycenter problem. This goes beyond Chizat [2025], who considered a particular internal energy as an outer regularization strategy. (iii) **Extensive DA benchmarking (Section 4)**. We empirically show that our algorithm produces barycenters that respect class structure, by incorporating both appropriate functionals (Section 3.3), and labels into the ground-cost of optimal transport (Section 3.2). On five DA benchmarks, spanning computer vision, neuroscience, and chemical engineering, we establish a comparison between empirical and neural barycenter solvers, showing that structured costs are essential for DA performance.

Paper organization. Section 2 reviews OT, Wasserstein barycenters and gradient flows. Section 3 develops our empirical, time-discretized gradient flow framework and algorithms. Section 4 presents our experiments. Finally, Section 5 concludes this paper.

2 BACKGROUND

Throughout this paper (Ω, d) is a Polish metric space, and its elements are denoted $z \in \Omega$. We denote $\mathcal{P}_2(\Omega)$ as the set of measures with finite second moments [Villani, 2008, Definition 6.4]. We work with empirical measures,

$$\text{Emp}_n(\Omega) = \left\{ \hat{P} \in \mathcal{P}_2(\Omega) : \hat{P} = \frac{1}{n} \sum_{i=1}^n \delta_{z_i^{(P)}} \right\}, \quad (3)$$

where $\delta_{z_0}(z) = \delta(z - z_0)$ denotes the Dirac measure centered at $z_0 \in \Omega$. Given a measure $P \in \mathcal{P}_2(\Omega)$, we denote by $\hat{P} \in \text{Emp}_n(\Omega)$ its empirical approximation based on i.i.d. samples $z_1^{(P)}, \dots, z_n^{(P)} \sim P$. The superscript $\cdot^{(P)}$ indicates the measure from which the samples originate.

2.1 OPTIMAL TRANSPORT

OT is a field of mathematics concerned with mass transportation at least effort [Villani, 2008, Peyré and Cuturi, 2019, Montesuma et al., 2025]. It was originally founded by Monge [1781], and seeks a mapping such that,

$$T_{P \rightarrow Q}^* = \arg \inf_{T_{\sharp} P = Q} \mathbb{E}_{z \sim P} [d(z, T(z))^2]. \quad (4)$$

$T_{P \rightarrow Q}^*$ is an OT mapping between P and Q , and $T_{\sharp}$ is the pushforward mapping P to Q . Alternatively, Kantorovich [1942] proposed a formulation in terms of a transport plan $\gamma \in \Gamma(P, Q)$, where $\Gamma(P, Q)$ is the set of all joint measures with marginals P and Q . In this case,

$$\gamma^* = \arg \inf_{\gamma \in \Gamma(P, Q)} \mathbb{E}_{(z, z') \sim \gamma} [d(z, z')^2], \quad (5)$$

where γ^* denotes an OT plan. OT defines the so-called 2-Wasserstein distance, through the infimum value of the optimization problem in Equations 4 and 5,

$$\mathbb{W}_2(P, Q)^2 = \inf_{\gamma \in \Gamma(P, Q)} \mathbb{E}_{(z, z') \sim \gamma} [d(z, z')^2]. \quad (6)$$

Equation 6 defines a metric in $\mathcal{P}_2(\Omega)$ which bridges the notions of Euclidean (i.e., center of mass) and Wasserstein barycenters. An alternative consists of regularizing the OT problem with the entropy of γ , resulting in the Sinkhorn divergence Cuturi [2013],

$$\mathbb{W}_{2,\epsilon}(P, Q)^2 = \inf_{\gamma \in \Gamma(P, Q)} \mathbb{E}[d(z, z')^2] + \lambda H(\gamma), \quad (7)$$

where $H(\gamma) = \text{KL}(\gamma | P \otimes Q)$, for $(P \otimes Q)(z) = P(z)Q(z)$, and KL is the Kullback Leibler divergence. Equation 7 enjoys many desirable properties. First, by relying on the matrix scaling algorithm of Sinkhorn [1964], it has complexity $\mathcal{O}(n^2)$ per iteration. Second, it can be vectorized on modern GPUs, leading to dramatic speed ups. Third, given that the objective in Equation 7 is strictly convex with respect γ , the OT plan γ_ϵ^* is unique.

2.2 WASSERSTEIN BARYCENTER

In the metric setting, the barycenter problem is known as Fréchet [1948] or Karcher [1977] means. In our case, given a finite family of measures $\mathcal{Q} = \{Q_k\}_{k=1}^K$ and a set of barycentric coordinates $\lambda \in \Delta_K = \{a \in \mathbb{R}_{\geq 0}^K : \sum_k a_k = 1\}$, we define the Wasserstein barycenter over $(\mathcal{P}_2(\Omega), \mathbb{W}_2)$ through the following optimization problem,

$$\hat{P}^* \in \arg \min_{\hat{P} \in \text{Emp}_n(\Omega)} \underbrace{\sum_{k=1}^K \lambda_k \mathbb{W}_{2,\epsilon}(\hat{P}, \hat{Q}_k)^2}_{\mathbb{B}_{2,\epsilon}(\hat{P} | \mathcal{Q})}. \quad (8)$$

Here, $\mathbb{B}_{2,\epsilon}(\cdot | \mathcal{Q})$ is a functional mapping $\text{Emp}_n(\Omega)$ to $\mathbb{R}_{\geq 0}$, i.e., the set of non-negative real numbers. Henceforth, we denote $\mathbb{B}_\epsilon(\hat{P})$ for simplicity. We provide a further description of barycenter solvers in Appendix B.

Existence and Uniqueness of Barycenters. The problem in Equation 8 is a minimization over empirical measures. As such, its objective has 3 levels that must be separated. First, each $\mathbb{W}_{2,\epsilon}$ has an associated transport plan, γ_ϵ^* , which is unique by the strict convexity of entropic OT (c.f. Section 2.1). Second, the continuous barycenter (i.e., the minimizer of Equation 1), denoted $P^* \in \mathcal{P}_2(\Omega)$, always exists on $\Omega = B(0, R)$ by the direct method of calculus of variations [Santambrogio, 2015, Box 1.1], and Agueh and Carlier [2011] shows that uniqueness holds when at least one input measure vanishes on small sets. Third, the empirical barycenter $\hat{P}^* \in \text{Emp}_n(\Omega)$ exists, since the objective is continuous on the compact Ω^n . However, the **empirical barycenter** is not unique, as Problem 8 is not convex.

2.3 GRADIENT FLOWS

Our method relies on the gradient flow interpretation of the minimization in Equation 8. Starting from the Euclidean setting, let $F : \mathbb{R}^d \rightarrow \mathbb{R}$ be a functional. A gradient flow is the solution to the Ordinary Differential Equation (ODE),

$$\dot{x}(t) = -\nabla F(x(t)), \text{ subject to } x(0) = x_0. \quad (9)$$

When discretized with the *forward Euler discretization*, one has gradient descent,

$$\dot{x}(t) = \frac{x_{t+1} - x_t}{\alpha} \implies x_{t+1} = x_t - \alpha \nabla F(x_t)$$

We now extend this framework to probability measures. Let $\mathbb{F} : \mathcal{P}_2(\Omega) \rightarrow \mathbb{R}$ be a functional. The evolution of a curve $\{P_t\}_{t \geq 0}$ in $\mathcal{P}_2(\Omega)$ is given by the continuity equation

$$\partial_t P_t = -\text{div}(P_t v_t), \quad (10)$$

where $v_t : \Omega \rightarrow \Omega$ is a velocity field, and div is the divergence operator, understood in a measure theoretic sense. Choosing $v_t = -\nabla_W \mathbb{F}(P_t)$ yields the gradient flow of $\mathbb{F}$,

that is, the measure theoretic analogue of Equation 9. Here, ∇_W denotes the Wasserstein gradient of the functional $\mathbb{F}$ (c.f. [Chewi et al., 2024, Section 5.4] and Appendix C.1).

The structure of $\mathbb{F}$ determines the dynamics of P_t . Following the classical OT theory [Santambrogio, 2015, Chapter 7], we consider the functional $\mathbb{F}(P) = \mathbb{B}(P) + \mathbb{R}(P)$, where,

$$\mathbb{R}(P) = \underbrace{\int G(p(z))dz}_{\text{Internal energy } \mathbb{G}(P)} + \underbrace{\int V(z)dP(z)}_{\text{Potential energy } \mathbb{V}(P)} + \underbrace{\int \int U(z, z')dP(z)dP(z')}_{\text{Interaction energy } \mathbb{U}(P)} \quad (11)$$

where $p : \Omega \rightarrow \mathbb{R}$ denotes the density of P , $G : \mathbb{R} \rightarrow \mathbb{R}$, $V : \Omega \rightarrow \mathbb{R}$, and $U : \Omega \times \Omega \rightarrow \mathbb{R}$. For example, when $G(s) = s \log s$, $\mathbb{G}$ is the *entropy functional* of [Chizat, 2025], associated with diffusion dynamics. $\mathbb{G}$, $\mathbb{V}$, and $\mathbb{U}$ introduce diffusion, drift, and pairwise behaviors in the flow, respectively. See Section 3.3 for further examples.

3 WASSERSTEIN BARYCENTERS AS GRADIENT FLOWS

In this section, we describe a new method for computing empirical OT barycenters. We use, $\mathbb{F}(\hat{P}) = \mathbb{B}_\epsilon(\hat{P}) + \mathbb{R}(\hat{P})$, where $\mathbb{B}_\epsilon$ is the barycenter objective defined in Equation 8, and $\mathbb{R}$ is a combination of different energies defined in Equation 11. We implement a flow from a prior measure $P_0 = \mathcal{N}(0, \text{Id})$ to an empirical minimizer of $\mathbb{F}(\hat{P})$ through a time-discretized version of the continuity equation. We proceed in two steps. **First**, we derive the empirical version of the continuity equation (c.f., Equation 10). For $\hat{P}_t$,

$$\begin{aligned} \partial_t \hat{P}_t &= \frac{1}{n} \sum_{i=1}^n \partial_t \delta_{z_{t,i}^{(P)}} = -\frac{1}{n} \sum_{i=1}^n \text{div}(\dot{z}_{t,i}^{(P)} \delta_{z_{t,i}^{(P)}}), \quad (12) \\ &= -\text{div}(\hat{P}_t v_t), \end{aligned}$$

where we identified $v_t(z_{t,i}^{(P)}) = \dot{z}_{t,i}^{(P)}$, i.e., the velocity field displacing samples of $\hat{P}_t$. **Second**, we discretize the gradient flow in time, through an index $\tau = \alpha t$, $\alpha > 0$. Again, using the forward Euler scheme,

$$\dot{z}_{t,i}^{(P)} = \frac{z_{\tau+1,i}^{(P)} - z_{\tau,i}^{(P)}}{\alpha} \implies z_{\tau+1,i}^{(P)} = z_{\tau,i}^{(P)} + \alpha v_{\tau,i}, \quad (13)$$

where $v_{\tau,i} = v_\tau(z_{\tau,i}^{(P)})$. We can obtain different update strategies based on the velocity field,

$$v_{\tau,i} = -\nabla_{z_{\tau,i}^{(P)}} \mathbb{F}(\hat{P}_\tau), \quad (14)$$

$$v_{\tau,i} = -\nabla_{z_{\tau,i}^{(P)}} \mathbb{F}(\hat{P}_\tau) + \sqrt{2\alpha\eta_G} \xi_{\tau,i}, \quad (15)$$

$$v_{\tau,i} = u_{\tau,i}, \quad u_{\tau+1,i} = \beta u_{\tau,i} - \alpha \nabla_{z_{\tau,i}^{(P)}} \mathbb{F}(\hat{P}_\tau). \quad (16)$$

where $\xi_{\tau,i} \sim \mathcal{N}(0, \text{Id})$. Equations 14, 15, and 16 correspond to the steepest descent, Langevin dynamics, and momentum descent strategies, respectively. $\eta_G \geq 0$ denotes the diffusion coefficient. The Langevin dynamics version of the flow corresponds to the NPGD dynamics of Chizat [2025]. Furthermore, for empirical measures, the functionals in $\mathbb{F}$ are tractable, and take the following form,

$$\begin{aligned} \mathbb{B}_\epsilon(\hat{P}) &= \sum_{k=1}^K \lambda_k \sum_{i=1}^n \sum_{j=1}^{n_k} \gamma_{\epsilon,k,i,j}^* d(z_i^{(P)}, z_j^{(Q_k)})^2, \\ \mathbb{G}(\hat{P}) &= \frac{1}{n} \sum_{i=1}^n G(\hat{p}(z_i^{(P)})), \\ \mathbb{V}(\hat{P}) &= \frac{1}{n} \sum_{i=1}^n V(z_i^{(P)}), \\ \mathbb{U}(\hat{P}) &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n U(z_i^{(P)}, z_j^{(P)}), \end{aligned} \quad (17)$$

where $\hat{p}$ denotes the *estimated* density of $\hat{P}$. Since $\hat{P}$ is empirical it is, in general, difficult to compute $\mathbb{G}$. A notable exception is the entropy functional (i.e., $G(s) = s \log s$), which adds a diffusion term to the gradient flow dynamics. This corresponds to Langevin dynamics [Chizat, 2025] (Equation 15). For general choices of G , one needs to estimate the density $\hat{p}$ at each iteration of the empirical flow.

Algorithm 1 Empirical Time Discretized Wasserstein Gradient Flow

Input: $\lambda \in \Delta_K$, $n \in \mathbb{N}$, $m \in \mathbb{N}$, $\alpha \in \mathbb{R}_+$, $\epsilon \in [0, +\infty)$

Output: Barycenter support $\{z_i^{(P)}\}_{i=1}^n$

- 1: $\{z_{0,i}^{(P)}\}_{i=1}^n$, such that $z_{0,i}^{(P)} \sim \mathcal{N}(0, \text{Id})$.
 - 2: **for** $\tau \leftarrow 1$ to n_{iter} **do**
 - 3: $\mathbb{F}(\hat{P}_\tau) \leftarrow \mathbb{V}(\hat{P}_\tau) + \mathbb{U}(\hat{P}_\tau) + \mathbb{G}(\hat{P}_\tau)$
 - 4: **for** $k \leftarrow 1$ to K **do**
 - 5: Sample $\{z_i^{(Q_k)}\}_{i=1}^m$ i.i.d. from Q_k .
 - 6: $\mathbb{F}(\hat{P}_\tau) \leftarrow \mathbb{F}(\hat{P}_\tau) + \lambda_k \mathbb{W}_{2,\epsilon}(\hat{P}_\tau, \hat{Q}_k)^2$
 - 7: **end for**
 - 8: $z_{\tau+1,i}^{(P)} \leftarrow z_{\tau,i}^{(P)} + \alpha v_{\tau,i}$
 - 9: **end for**
-

An immediate consequence of Equation 17 is that our gradient flow formulation encompasses several previous works. For instance, with $G = U = V = 0$, we have the classical free-support barycenters of Cuturi and Doucet [2014]. Furthermore, for the entropy functional and $\epsilon > 0$, we have the NPGD algorithm of Chizat [2025]. Finally, as we show in Section 3.2 below, with a specific metric on $\Omega = \mathcal{X} \times \mathcal{Y}$, we retrieve the algorithm of Montesuma et al. [2023].

3.1 SCALING WITH MINI-BATCH OT

The computation of $\mathbb{B}_\epsilon$ in the discrete approach [Cuturi and Doucet, 2014] involves an OT problem between the

complete set of samples in the input measures. Therefore, when the input measures are large scale, the barycenter problem becomes infeasible. We circumvent this issue with mini-batch OT [Fratras et al., 2021].

The main idea behind our scalable barycenter algorithm is accessing $\hat{Q}_k \in \mathcal{Q}$ through sampling. At each iteration of Algorithm 1, we sample a mini-batch $\{z_i^{(Q_k)}\}_{i=1}^m$ of m samples from each input measure (cf. line 5 in Algorithm 1). Since we sample the same number of samples from each measure, we can actually vectorize the computation of the K OT problems in the evaluation of $\mathbb{W}_{2,\epsilon}$.

Let $\{z_{\tau,i}^{(P)}\}_{i=1}^n$ be the barycenter support, and $\{z_j^{(Q_k)}\}_{j=1}^m$ be a mini-batch from Q_k . The Gibbs kernel $E_{k,i,j} = e^{-(C_{k,i,j}/\epsilon)}$, and $C_{k,i,j} = d(z_{\tau,i}^{(P)}, z_j^{(Q_k)})^2$ are arrays with shape (K, n, m) . The Sinkhorn iterations become,

$$\psi_{k,j}^{(\ell+1)} = \frac{b_{k,j}}{\sum_{i=1}^n E_{k,i,j} \phi_{k,i}^{(\ell)}}, \phi_{k,i}^{(\ell+1)} = \frac{a_{k,i}}{\sum_{j=1}^m E_{k,i,j} \psi_{k,j}^{(\ell+1)}}, \quad (18)$$

where the division operation should be understood coordinate-wise. Furthermore, $b_j = n^{-1}$ and $a_{k,i} = m^{-1}$ is the importance of each sample in the barycenter, and input measure support, respectively. For simplicity, we assume uniform weights. For $L \in \mathbb{N}$ iterations, we can retrieve $\gamma_{\epsilon,k}^* = \text{diag}(\phi_k^{(L)}) E_k \text{diag}(\psi_k^{(L)})$. Since each $\gamma_{\epsilon,k}^*$ has shape (n, m) , we can actually stack them in an array of shape (K, n, m) , similarly to E and C .

The key insight is that the contractions $\sum_i E_{k,i,j} \phi_{k,i}^{(\ell)}$ and $\sum_j E_{k,i,j} \psi_{k,j}^{(\ell)}$ are batched matrix-vector products across the leading K axis, meaning that **the K OT problems can be vectorized**. This is possible because we draw the same number of samples from each input measure, making it possible to vectorize the Sinkhorn iterations. For instance, a similar strategy is possible for measures over a fixed grid [Peyré and Cuturi, 2019, Remark 4.16]. The OT plan γ_ϵ , obtained at the end of the Sinkhorn iterations, is used to compute $\mathbb{B}_\epsilon$, in lines 4 through 7 in Algorithm 1.

3.2 FLOWS OVER JOINT MEASURES

An interesting advantage of the gradient flow formalism is that the only necessary ingredient is a differentiable metric over Ω . Therefore, we generalize the setting of Cuturi and Doucet [2014], who assumed $\Omega = \mathbb{R}^d$, and $d(z, z') = \|z - z'\|_2$. As we show in our experiments, in machine learning applications one has $\Omega = \mathcal{X} \times \mathcal{Y}$, i.e., measures over the joint space of features and labels. Let $z = (x, y)$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$,

$$d(z, z') = \sqrt{\|x - x'\|_2^2 + \beta \|y - y'\|_2^2}, \quad (19)$$

where $\beta \geq 0$ is a parameter that balances the feature distance terms and the label distance terms.

For regression applications (i.e., $\mathcal{Y} = \mathbb{R}$), this distance is quite natural. However, for classification, $\mathcal{Y}$ is categorical (e.g., $\mathcal{Y} = \{1, \dots, n_{\text{classes}}\}$). One possible strategy, used in Montesuma and Mboula [2021] and Alvarez-Melis and Fusi [2021], is fixing the labels and flowing only the features. In contrast to these works, we embed $\mathcal{Y}$ into the compact continuous space $\Delta_{n_{\text{classes}}}$, through a one-hot encoding operation. For our flow, we parametrize labels through a change of variables,

$$y_{i,c}^{(P)} = \text{softmax}(\ell_{i,1}^{(P)}, \dots, \ell_{i,n_c}^{(P)}) = \frac{\exp(\ell_{i,c}^{(P)})}{\sum_{c=1}^{n_c} \exp(\ell_{i,c}^{(P)})},$$

thus, instead of optimizing over $z = (x, y)$, we optimize over $z = (x, \ell)$, $x \in \mathbb{R}^d$, $\ell \in \mathbb{R}^{n_{\text{classes}}}$. From the soft probabilities, we can retrieve the actual discrete labels with an argmax , $y_{\text{hard},i}^{(P)} = \text{argmax}_{c=1, \dots, n_c} y_{i,c}^{(P)}$.

3.3 TASK-AWARE REGULARIZING FUNCTIONALS

One of the advantages of our proposed method is regularizing the barycenter calculation with internal, interaction, and potential energy functionals. This idea was already used in practice by Alvarez-Melis and Fusi [2021] for transfer learning problems. Here we propose the following functionals,

$$V_E(z) = -\eta_V \sum_{c=1}^{n_{\text{classes}}} y_c \log y_c, \quad (20)$$

$$U_R(z, z') = \begin{cases} \eta_U h(d(x, x')) & \text{if } y_{\text{hard}} \neq y'_{\text{hard}}, \\ 0 & \text{otherwise} \end{cases}, \quad (21)$$

where $h : \mathbb{R} \rightarrow \mathbb{R}$ is lower semi-continuous and bounded from below, and $\eta_V > 0$ and $\eta_U > 0$ are scaling parameters. For instance, in our experiments we use the hinge loss, $h(u) = \max(0, \text{margin} - d)$, where $\text{margin} \geq 0$ is a margin parameter. Equations 20 and 21 correspond to entropy and repulsion, respectively. The first functional penalizes barycenters that have fuzzy labels. The second functional encourages classes to be well separated.

3.4 CONVERGENCE

One of the difficulties in analyzing the gradient flow of the barycenter functional comes from the fact that $P \mapsto \mathbb{W}_2(P, Q)^2$ is not geodesically convex in $\mathcal{P}_2(\Omega)$ [Santambrogio, 2017, Section 4.4]. As a result, we are minimizing a non-convex functional. **Explicit constants and proofs are available in Appendix D.** Our analysis relies on a measure-theoretic version of the Polyak Łojasiewicz (PL) inequality [Polyak, 1964, Karimi et al., 2016],

$$\|\nabla_W \mathbb{B}(P)\|_{L_2(P)}^2 \geq C_{\text{PL}} (\mathbb{B}(P) - \mathbb{B}^*). \quad (22)$$

When this inequality holds, we have the following convergence result.

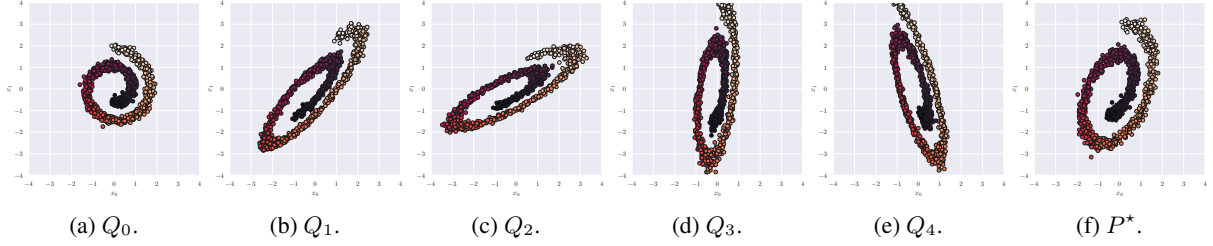


Figure 2: Swiss roll measures in a location-scatter family, where a source measure is transformed under an affine function $T_k(x) = A_k x + b_k$. Colors reflect the relative position of points in the manifold.

Theorem 3.1. Let $\Omega = \mathcal{B}(0, R)$ be the closed ball in $\mathbb{R}^d$ with radius $R > 0$. Let

$$P^* = \arg \min_{P \in \mathcal{P}_2(\Omega)} \sum_{k=1}^K \lambda_k \mathbb{W}_2(P, Q_k)^2$$

be the continuous barycenter of $\mathcal{Q} = \{Q_k\}_{k=1}^K$, $Q_k \in \mathcal{P}_2(\Omega)$ $k = 1, \dots, K$, with barycentric coordinates $\lambda = (\lambda_1, \dots, \lambda_K) \in \Delta_K$. Approximate each Q_k with m i.i.d. samples $\{z_i^{(Q_k)}\}_{i=1}^m$, $z_i^{(Q_k)} \sim Q_k$ and $\{z_{\tau,i}^{(P)}\}_{i=1}^n$ be the barycenter support. Let $\hat{P}^*$ be a minimizer of

$$\hat{P}^* \in \arg \min_{\hat{P} \in \text{Emp}_n(\Omega)} \sum_{k=1}^K \lambda_k \mathbb{W}_{2,\epsilon}(\hat{P}, \hat{Q}_k)^2.$$

Under the PL inequality in Equation 22 the following holds,

$$\mathbb{E}[\hat{\mathbb{B}}(\hat{P}_\tau) - \hat{\mathbb{B}}^*] \leq e^{-C_{PL}\tau} (\mathbb{B}(\hat{P}_0) - \mathbb{B}^*) + C_R \sqrt{\frac{C_{d,n}}{m}}, \quad (23)$$

where the expectation is taken with respect to samples from Q_k . The constants C_R and $C_{d,n}$ depend on the radius R , number of barycenter samples n , and dimensions d . m denotes the number of samples drawn from each Q_k .

The first term in the r.h.s. of Equation 23 decays exponentially with the iterations τ , and is a direct consequence of the PL inequality in Equation 22. The second term captures the empirical approximation error of the involved measures, that is, $Q_k \in \mathcal{Q}$ and P_τ . In practice, $\tau \rightarrow +\infty$ leads to an error governed by the empirical approximations.

On the PL Inequality. Equation 22 is a standard tool used in proving the convergence of non-convex optimization algorithms, however, it is an open question whether it holds in general settings. For instance, Chewi et al. [2020] shows that this inequality holds for Gaussian measures under specific spectral conditions. In the Appendix, we show that the same result holds for *location-scatter families of measures* (Appendix D.2), based on the isometry between these sets, and the Bures-Wasserstein manifold established by [Álvarez-Esteban et al., 2016, Theorem 2.3]. On top of this theoretical justification, we provide an experiment validating the bound on the Swiss roll measure (Appendix D.1).

4 EXPERIMENTS

In the following, all experiments were conducted on a virtual machine with 8 CPUs (AMD EPYC 7413, 48GB of RAM) and an NVIDIA L4 GPU (24GB of VRAM). Appendix F includes additional details, fine-grained results, complexity analysis, and ablations.

4.1 SWISS ROLL MEASURES

In Figure 2, we show the Swiss roll measure [Korotin et al., 2021] Q_0 , alongside four variations obtained via $T_{k,\#}Q_0$, $k = 1, \dots, 4$, $T_k(x) = A_k x + b_k$, for which the ground-truth barycenter is known in closed-form [Álvarez-Esteban et al., 2016], that is, $P^* = (\sum_k \lambda_k T_k)_{\#}Q_0$. These measures are shown in Figure 2.

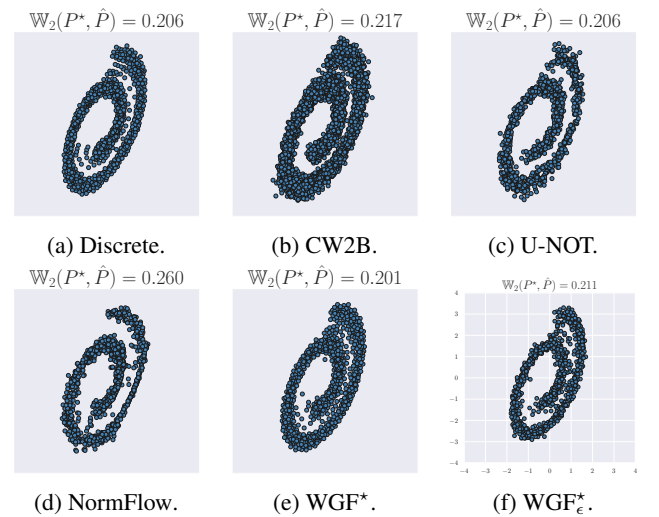


Figure 3: Comparison of unsupervised barycenter methods. Solvers (b, c, d) are neural, and (e, f) are ours.

We compare five barycenter solvers, including discrete solvers [Cuturi and Doucet, 2014, Montesuma et al., 2023], and neural solvers: CW2B [Korotin et al., 2021], U-NOT Gazdieva et al. [2024], and NormFlow Visentin and Cheridito [2025]. First, we benchmark unsupervised solvers in Figures 3 (a) through (f). Quantitatively, empirical meth-

ods, notably [Cuturi and Doucet, 2014, Algorithm 2] and our Wasserstein Gradient Flow (WGF) algorithm (cf. Algorithm 1) achieve the lowest Wasserstein distance to the ground truth in Figure 2. Overall, while the neural network solvers are usually more scalable than discrete methods in terms of number of samples, their optimization is more complicated and very sensitive to hyper-parameters.

Second, we experiment with integrating labels in the ground cost as described in Section 3.2. These methods are shown in Figures 4 (a), (b) and (c). In all cases, integrating labels produces barycenters that are closer to the ground-truth. We conclude that using the labels gives a strong inductive bias in the barycenter computation, which explains the gain in performance of labeled barycenters in the next section.

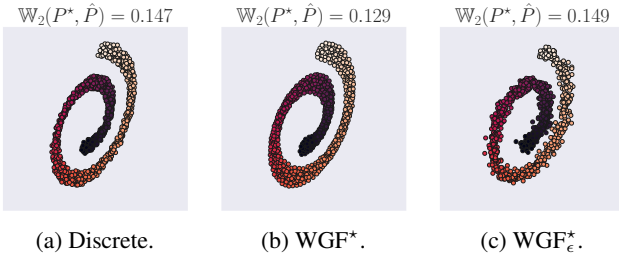


Figure 4: Comparison of supervised barycenter methods.

Next, we compare the running time of our WGF solver with the discrete solver of Cuturi and Doucet [2014], for an increasing barycenter support size $n \in \{2^{10}, \dots, 2^{16}\}$ and batch size $m \in \{2^8, 2^9, 2^{10}\}$ for 200 iterations. The discrete solver is run for $n \in \{2^{10}, \dots, 2^{14}\}$. For this algorithm, larger support sizes **result in memory overflow**. Our WGF is able to compute larger barycenters via mini-batching.

We summarize our results in Figure 5. First, in Figure 5 (a), our WGF leads to speedups ranging from $2\times$ ($n = 2^{12}, m = 2^{10}$) to $50\times$ ($n = 2^{14}, m = 2^8$). Similar gains were obtained using Exact OT on CPU (Appendix F.3). Second, in Figure 5 (b), we isolate the gain of using GPU acceleration and our vectorization strategy (Section 3.1), showing consistent gains ranging from $2.1\times$ to $8.3\times$. Overall, our WGF strategy provides compounding speedups from three sources: (i) minibatching, (ii) entropic regularization, and (iii) GPU parallelism.

4.2 MULTI-SOURCE DOMAIN ADAPTATION

One of the main applications of Wasserstein barycenters is multi-source domain adaptation (MSDA) [Kouw and Loog, 2019, Sun et al., 2015, Pan and Yang, 2009]. In this setting one needs to adapt multiple labeled source measures $Q_1, \dots, Q_K$ to a single unlabeled target measure Q_T . The goal is to learn, from samples $\{\{x_i^{(Q_k)}, y_i^{(Q_k)}\}_{i=1}^{n_k}\}_{k=1}^K$ and $\{x_i^{(Q_T)}\}_{i=1}^{n_T}$, a classifier h that achieves low risk or error in

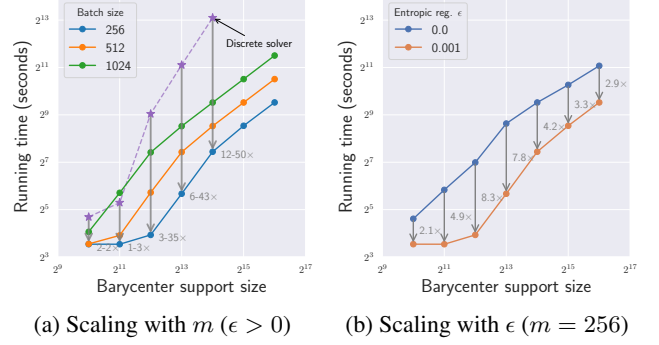


Figure 5: Running time analysis for our WGF algorithm. In (a), we compare the running time of our technique to that of the discrete solver of Cuturi and Doucet [2014]. In (b), we isolate the effect of GPU acceleration, showing consistent gains with increasing support size.

the target domain measure,

$$\mathcal{R}_{Q_T}(h) = \mathbb{E}_{(x,y) \sim Q_T} [\mathcal{L}(y, h(x))],$$

for a loss function $\mathcal{L}$ (e.g., cross-entropy loss). We isolate the quality of barycenters by doing adaptation at the level of embeddings. This approach allows us to perform domain adaptation in a higher semantic space, where distributions are more meaningful and comparable across domains. Thus, we assume that a meaningful feature extractor ϕ , called the backbone, has been previously learned. We obtain the feature extractor by fine-tuning a neural network on the labeled source domain data (Appendix E.2).

Experimental Setup. The unsupervised WGF is fit with *unlabeled* measures, i.e., $Q_k = n_k^{-1} \sum_i \delta_{x_i^{(Q_k)}}$, where $x_i^{(Q_k)}$ are embedding vectors. The supervised WGF is fit with *labeled* measures, that is, $Q_k = n_k^{-1} \sum_{i=1}^{n_k} \delta_{(x_i^{(Q_k)}, y_i^{(Q_k)})}$. In both cases, $\hat{P} = n^{-1} \sum_{i=1}^n \delta_{(x_i^{(P)}, y_i^{(P)})}$ which allows us to apply the regularizing functionals defined in Section 3.3.

Wasserstein Barycenter Transport. Montesuma and Mboula [2021] proposed a technique based on Wasserstein barycenters for domain adaptation. The idea is to map the synthesized barycenter support to the target domain through the barycentric mapping [Courty et al., 2016]. Then, a classifier can be learned on the target domain based on the transported data. We adapt this idea to unsupervised barycenter algorithms by using the computed barycenter as a pivot domain. See Appendix E.4 for more details.

Benchmarks. We run our experiments on five benchmarks: Office 31 [Saenko et al., 2010], Office Home [Venkateswara et al., 2017], BCI-CIV-2a [Brunner et al., 2008], IS-RUC [Khalighi et al., 2013], and TEP [Montesuma et al., 2024c]. The first two, second two, and last benchmarks correspond to computer vision, neuroscience, and chemical engineering benchmarks, respectively. We show in Table 1 an overview of our experimental setting.

Benchmark	Backbone	# Samples	# Domains	# Dim.	# Classes
Office31	ResNet50	3287	3	2048	31
BCI-CIV-2a	CBraMod	5184	10	200	4
TEP	CNN	17289	6	128	29
Office Home	ResNet101	15500	4	2048	65
ISRUC	CBraMod	89240	100	512	5

Table 1: Overview of benchmarks used for domain adaptation, sorted by number of samples.

Compared Methods. Overall, we compare seven Wasserstein barycenter strategies with ours. Besides those used in the previous section, we include the NOT approach of Kolesov et al. [2024], and the Gaussian Mixture Model (GMM) barycenter of Montesuma et al. [2024b]. For completeness, we include four other state-of-the-art methods in domain adaptation over embedding vectors. Those methods are: WJDOT Turrisi et al. [2022], DaDiL-R, and E Montesuma et al. [2023], and GMM-DaDiL Montesuma et al. [2024b]. For each benchmark, we use one domain as the target domain (e.g., Amazon vs. {dSLR, Webcam} in the Office 31 benchmark), and we measure the classification accuracy, i.e., the percentage of correct predictions.

Main Results. We present our main results in Table 2, which reports the average performance per domain on each benchmark. We provide fine-grained results in Appendix F. In general, labeled barycenter methods (the 3 last rows in Table 2) have a clear advantage over unsupervised methods. We argue that label information is essential in domain adaptation success, which is consistent with previous research [Courty et al., 2016, Montesuma and Mboula, 2021, Montesuma et al., 2023, 2024b]. Among barycenter methods, our WGF method achieves the best performance in all benchmarks, outperforming previous methods in MSDA, such as WJDOT and DaDiL in the ISRUC, BCI-CIV, and TEP benchmarks.

Unsupervised vs. Supervised Barycenters. An intriguing finding of Table 2 is that, on some benchmarks, unsupervised WGF achieves decent performance, even outperforming supervised baselines (e.g., WJDOT of Turrisi et al. [2022]). We probe the reason why these methods have unreasonably good performance, by performing a t-SNE [Maaten and Hinton, 2008] visualization of the barycenters against

Benchmark	$\mathcal{X} \times \mathcal{Y}$	Office31	OfficeHome	BCI-CIV-2a	ISRUC	TEP	Avg. Rank
Backbone	-	ResNet-50	ResNet-101	CBraMod	CBraMod	CNN	-
Source-Only	-	86.40	75.95	50.30	76.63	78.48	-
WJDOT	-	86.80	76.59	N/A	76.95	86.13	7.8
DaDiL-R	-	89.91	77.86	53.41	74.68	86.14	6.4
DaDiL-E	-	89.79	78.14	N/A	75.89	85.87	5.0
GMM-DaDiL	-	90.63	78.81	57.10	75.47	86.85	4.4
Discrete	$\times$	87.82	77.17	57.38	70.67	83.81	8.8
NormFlow	$\times$	85.91	76.73	55.45	78.04	82.89	10.0
CW2B	$\times$	86.37	76.44	57.25	75.84	85.83	9.0
NOT	$\times$	86.22	75.36	57.29	76.39	84.93	9.4
U-NOT	$\times$	86.97	76.85	57.17	77.06	85.43	7.8
WGF (ours)	$\times$	88.27	77.56	57.68	78.16	85.51	4.9
Discrete	$\checkmark$	87.93	77.09	57.43	78.20	86.09	5.0
GMM	$\checkmark$	88.54	77.87	57.04	74.58	84.67	8.2
WGF (ours)	$\checkmark$	89.52	78.42	57.68	80.02	86.87	1.9

Table 2: Average classification accuracy on target domains. The top group (WJDOT through GMM-DaDiL) include **general domain adaptation methods**. The middle and bottom groups show **barycenter solvers** without and with access to source labels ($\mathcal{X} \times \mathcal{Y} = \times$, $\mathcal{X} \times \mathcal{Y} = \checkmark$), respectively. Bold indicates best among barycenter methods.

the source domain samples, shown in Figure 6. We see that even without labels, the barycenter samples cluster around the classes. This effect stems from the geometry of the embeddings, which were optimized (through supervision on source domains) to separate classes. As a result, unsupervised OT already captures, to some extent, the cluster structure of the barycentric measure. Figures 6 (b) and (c) investigate the remaining performance gap, as we transport the measures towards the target. Figure 6 (b) shows the transported sources to the barycentric domain (left) and the synthesized barycenter (right), and Figure 6 shows the final transported samples in the target domain for the unsupervised strategy (left) and supervised one (right). We conclude that using a synthesized supervised barycenter is better than using an unsupervised barycenter as pivot domain (Appendix E.4). Therefore, *the inductive bias of using labels in the ground cost is essential for domain adaptation*. This further explains the performance of DaDiL methods, which directly synthesize the target domain through an approximate barycenter.

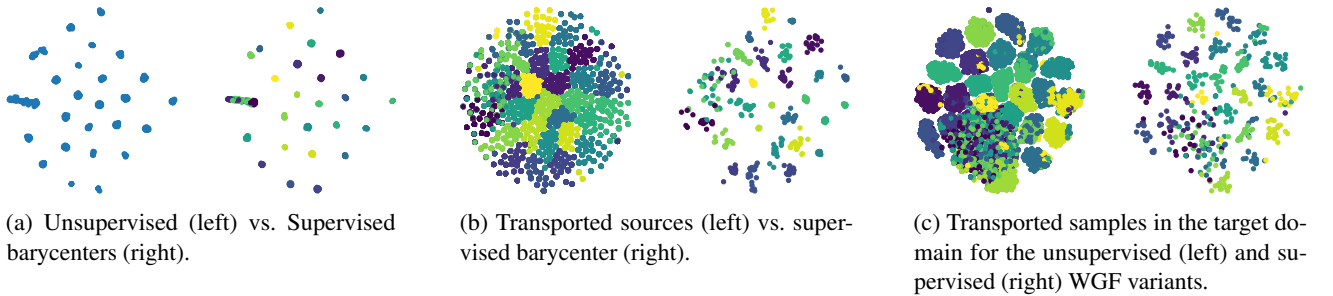


Figure 6: Comparison of unsupervised vs. supervised barycenter transports on the TEP benchmark.

Dataset	Metric	Baseline	WGF ($\mathcal{X} \times \mathcal{Y}$)	WGF _u ($\mathcal{X}$)	NormFlow	NOT	U-NOT	CW2B
Adult ^{Gender}	Acc	0.808 $\pm$ 0.002	0.793 ⁽⁵⁾ $\pm$ 0.034	0.792 ⁽⁶⁾ $\pm$ 0.004	0.806 ⁽¹⁾ $\pm$ 0.003	0.794 ⁽³⁾ $\pm$ 0.004	0.804 ⁽²⁾ $\pm$ 0.002	0.794 ⁽⁴⁾ $\pm$ 0.004
	DI	1.955 $\pm$ 0.083	0.990 ⁽¹⁾ $\pm$ 0.014	1.011 ⁽²⁾ $\pm$ 0.013	1.462 ⁽⁶⁾ $\pm$ 0.041	0.975 ⁽³⁾ $\pm$ 0.157	1.253 ⁽⁵⁾ $\pm$ 0.090	0.926 ⁽⁴⁾ $\pm$ 0.039
Adult ^{Race}	Acc	0.808 $\pm$ 0.002	0.833 ⁽¹⁾ $\pm$ 0.020	0.799 ⁽⁶⁾ $\pm$ 0.006	0.807 ⁽²⁾ $\pm$ 0.002	0.807 ⁽³⁾ $\pm$ 0.001	0.807 ⁽⁴⁾ $\pm$ 0.002	0.806 ⁽⁵⁾ $\pm$ 0.001
	DI	0.613 $\pm$ 0.015	0.977 ⁽³⁾ $\pm$ 0.007	0.985 ⁽²⁾ $\pm$ 0.015	0.741 ⁽⁶⁾ $\pm$ 0.034	0.997 ⁽¹⁾ $\pm$ 0.051	0.865 ⁽⁵⁾ $\pm$ 0.026	1.041 ⁽⁴⁾ $\pm$ 0.019
Adult ^{G+R}	Acc	0.808 $\pm$ 0.002	0.848 ⁽¹⁾ $\pm$ 0.019	0.792 ⁽⁵⁾ $\pm$ 0.002	0.802 ⁽²⁾ $\pm$ 0.003	0.791 ⁽⁶⁾ $\pm$ 0.002	0.801 ⁽³⁾ $\pm$ 0.002	0.792 ⁽⁴⁾ $\pm$ 0.005
	DI	2.130 $\pm$ 0.180	0.905 ⁽³⁾ $\pm$ 0.061	1.031 ⁽²⁾ $\pm$ 0.047	1.859 ⁽⁶⁾ $\pm$ 0.287	0.990 ⁽¹⁾ $\pm$ 0.265	1.201 ⁽⁵⁾ $\pm$ 0.254	0.843 ⁽⁴⁾ $\pm$ 0.076
COMPAS ^{Race}	Acc	0.674 $\pm$ 0.009	0.666 ⁽⁵⁾ $\pm$ 0.019	0.659 ⁽⁶⁾ $\pm$ 0.007	0.672 ⁽¹⁾ $\pm$ 0.008	0.668 ⁽⁴⁾ $\pm$ 0.006	0.670 ⁽³⁾ $\pm$ 0.003	0.671 ⁽²⁾ $\pm$ 0.006
	DI	1.973 $\pm$ 0.176	1.050 ⁽³⁾ $\pm$ 0.078	1.018 ⁽¹⁾ $\pm$ 0.020	1.278 ⁽⁶⁾ $\pm$ 0.178	1.071 ⁽⁴⁾ $\pm$ 0.022	1.244 ⁽⁵⁾ $\pm$ 0.060	1.036 ⁽²⁾ $\pm$ 0.066
Credit ^{Gender}	Acc	0.812 $\pm$ 0.002	0.813 ⁽¹⁾ $\pm$ 0.012	0.802 ⁽⁶⁾ $\pm$ 0.001	0.812 ⁽⁴⁾ $\pm$ 0.002	0.811 ⁽⁵⁾ $\pm$ 0.001	0.813 ⁽¹⁾ $\pm$ 0.002	0.812 ⁽³⁾ $\pm$ 0.001
	DI	0.922 $\pm$ 0.113	1.087 ⁽⁴⁾ $\pm$ 0.068	1.044 ⁽²⁾ $\pm$ 0.019	1.067 ⁽³⁾ $\pm$ 0.066	1.094 ⁽⁵⁾ $\pm$ 0.135	1.001 ⁽¹⁾ $\pm$ 0.084	1.104 ⁽⁶⁾ $\pm$ 0.025
Law ^{Gender}	Acc	0.631 $\pm$ 0.006	0.618 ⁽⁶⁾ $\pm$ 0.014	0.627 ⁽⁵⁾ $\pm$ 0.007	0.631 ⁽⁴⁾ $\pm$ 0.001	0.632 ⁽³⁾ $\pm$ 0.003	0.635 ⁽¹⁾ $\pm$ 0.005	0.635 ⁽²⁾ $\pm$ 0.001
	DI	0.938 $\pm$ 0.012	1.018 ⁽³⁾ $\pm$ 0.018	0.996 ⁽²⁾ $\pm$ 0.004	1.001 ⁽¹⁾ $\pm$ 0.009	1.042 ⁽⁶⁾ $\pm$ 0.106	0.962 ⁽⁵⁾ $\pm$ 0.030	1.022 ⁽⁴⁾ $\pm$ 0.026
Crime ^{Race}	MSE	0.0182 $\pm$ 0.0010	0.0424 ⁽⁶⁾ $\pm$ 0.0045	0.0370 ⁽⁵⁾ $\pm$ 0.0021	0.0182 ⁽¹⁾ $\pm$ 0.0008	0.0306 ⁽⁴⁾ $\pm$ 0.0046	0.0238 ⁽³⁾ $\pm$ 0.0023	0.0192 ⁽²⁾ $\pm$ 0.0014
	MG	0.2122 $\pm$ 0.0207	0.0043 ⁽²⁾ $\pm$ 0.0019	0.0000 ⁽¹⁾ $\pm$ 0.0000	0.2111 ⁽⁶⁾ $\pm$ 0.0217	0.1107 ⁽³⁾ $\pm$ 0.0143	0.1287 ⁽⁴⁾ $\pm$ 0.0311	0.2053 ⁽⁵⁾ $\pm$ 0.0275
Mean rank (14 metric cells)			3.14	3.64	3.50	3.64	3.36	3.64

Table 3: Fairness-repair results across seven tabular benchmarks (mean $\pm$ std over 3 seeds; superscript in each cell is the per-dataset rank among the six methods, 1 = best). Classification reports **Acc** (higher is better) and **DI** (disparate impact, target 1; distance is $|\text{DI} - 1|$). Crime (regression) reports **MSE** (lower is better) and **MG**, the group mean-prediction gap (target 0). Rank-1 means are **bold**. WGF repairs the joint $\mathcal{X} \times \mathcal{Y}$; WGF_u is the feature-only ablation on $\mathcal{X}$.

4.3 FAIRNESS

In this section, we examine the performance of our WGF against four neural solvers (NF, NOT, U-NOT and CW2B) in the context of algorithmic fairness. Briefly, the problem consists of learning a model $h \in \mathcal{H}$ that is fair against a protected variable s (e.g., gender). Gordaliza et al. [2019] formalized this task, by considering probability measures $P(X|S = s)$. These authors proposed a method called *repairing*, in which (1) they compute the barycenter of $\{P(X|S = s)\}_{s=1}^S$ and (2) they align each $P(X|S = s)$ with the barycenter. Inspired by our domain adaptation results, we also propose a new method based on the joint alignment, meaning, computing the barycenter of $P(X, Y|S = s)$.

We test the two versions of our WGF algorithm (marginal, $\mathcal{X}$, and joint, $\mathcal{X} \times \mathcal{Y}$ repairs) against four neural barycenter solvers (NormFlow, NOT, U-NOT and CW2B) over five datasets: Adult income [Kohavi et al., 1996], COMPAS [Angwin et al., 2022], Law school bar exam [Wightman, 1998], Communities and crime [Redmond and Baveja, 2002]. All algorithms are evaluated against two axes: (i) downstream performance (e.g., accuracy for classification, mean squared error for regression); (ii) fairness, that is, the disparate impact metric,

$$\text{DI} = \frac{\Pr(\hat{Y} = 1|S = \text{unprivileged})}{\Pr(\hat{Y} = 1|S = \text{privileged})},$$

which measures the ratio, in probability, of picking the unprivileged group over the privileged group (the closer to 1,

the better), and for regression we use the maximum gap,

$$\text{MG} = |\mathbb{E}[\hat{Y}|S = \text{unprivileged}] - \mathbb{E}[\hat{Y}|S = \text{privileged}]|.$$

Our results are summarized in Table 3. Overall, similarly to domain adaptation, repairing the joint measures has the best average performance.

5 CONCLUSION

We have presented a new Wasserstein barycenter algorithm (Algorithm 1), inspired by the gradient flow literature [Santambrogio, 2015], that addresses two persistent challenges: scalability and regularization. We have derived a new algorithm that is amenable to mini-batch OT [Fratras et al., 2021] (Section 3.1). Under entropic OT, we further demonstrate that OT computations can be vectorized across the K input measures, leading to further speedups (Figure 5).

Our framework is flexible, and accommodates modular plug-in regularization from internal, potential, and interaction energies (Section 3.3). Extensive experiments on domain adaptation demonstrate the value of our approach and that incorporating labels in the OT ground-cost is necessary for state-of-the-art performance (Table 2).

Finally, our work highlights several promising avenues for future research. Specifically, our framework could be extended to more complex differentiable structures, including Riemannian manifolds. Additionally, it prompts further investigation into how to effectively integrate labels and other regularization schemes into neural network solvers.

References

- Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.
- Jason M Altschuler and Enric Boix-Adsera. Wasserstein barycenters are np-hard to compute. *SIAM Journal on Mathematics of Data Science*, 4(1):179–203, 2022.
- 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.
- David Alvarez-Melis and Nicolò Fusi. Dataset dynamics via gradient flows in probability space. In *International conference on machine learning*, pages 219–230. PMLR, 2021.
- Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.
- Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.
- Julio Backhoff-Veraguas, Mathias Beiglböck, and Gudmun Pammer. Existence, duality, and cyclical monotonicity for weak transport costs. *Calculus of Variations and Partial Differential Equations*, 58(6):203, 2019.
- Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pages 1–12, 2011.
- Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.
- Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. Bci competition 2008–graz data set a. *Institute for knowledge discovery (laboratory of brain-computer interfaces)*, Graz University of Technology, 16(1-6):34, 2008.
- Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. Gradient descent algorithms for bures-wasserstein barycenters. In *Conference on Learning Theory*, pages 1276–1304. PMLR, 2020.
- Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 3, 2024.
- Lénaïc Chizat. Doubly regularized entropic wasserstein barycenter. *Foundations of Computational Mathematics*, pages 1–38, 2025.
- Samuel Cohen, Michael Arbel, and Marc Peter Deisenroth. Estimating barycenters of measures in high dimensions. *arXiv preprint arXiv:2007.07105*, 2020.
- Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.
- Camille Le Coz, Alexis Tantet, Rémi Flamary, and Riwal Plougonven. A barycenter-based approach for the multi-model ensembling of subseasonal forecasts. *arXiv preprint arXiv:2310.17933*, 2023.
- Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.
- Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.
- Nabarun Deb, Promit Ghosal, and Bodhisattva Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Advances in Neural Information Processing Systems*, 34:29736–29753, 2021.
- Jiaojiao Fan, Amirhossein Taghvaei, and Yongxin Chen. Scalable computations of wasserstein barycenter via input convex neural networks. *arXiv preprint arXiv:2007.04462*, 2020.
- Kilian Fatras, Younes Zine, Szymon Majewski, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Minibatch optimal transport distances; analysis and applications. *arXiv preprint arXiv:2101.01792*, 2021.
- Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.
- Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l’institut Henri Poincaré*, volume 10, pages 215–310, 1948.
- Wilfrid Gangbo and Robert J McCann. The geometry of optimal transportation. 1996.
- Milena Gazdieva, Jaemoo Choi, Alexander Kolesov, Jaewoong Choi, Petr Mokrov, and Alexander Korotin. Robust barycenter estimation using semi-unbalanced neural optimal transport. *arXiv preprint arXiv:2410.03974*, 2024.
- Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

- Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International conference on machine learning*, pages 2357–2365. PMLR, 2019.
- Nathael Gozlan, Cyril Roberto, Paul-Marie Samson, and Prasad Tetali. Kantorovich duality for general transport costs and applications. *Journal of Functional Analysis*, 273(11):3327–3405, 2017.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- L Kantorovich. On the transfer of masses (in russian). In *Doklady Akademii Nauk*, volume 37, pages 227–229, 1942.
- Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977.
- Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.
- Sirvan Khalighi, Teresa Sousa, Gabriel Pires, and Urbano Nunes. Automatic sleep staging: A computer assisted approach for optimal combination of features and polysomnographic channels. *Expert Systems with Applications*, 40(17):7046–7059, 2013.
- Benoît Kloeckner. A geometric study of wasserstein spaces: Euclidean spaces. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 9(2):297–323, 2010.
- Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.
- Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.
- Alexander Kolesov, Petr Mokrov, Igor Udovichenko, Milena Gazdieva, Gudmund Pammer, Evgeny Burnaev, and Alexander Korotin. Estimating barycenters of distributions with neural optimal transport. *arXiv preprint arXiv:2402.03828*, 2024.
- Alexander Korotin, Lingxiao Li, Justin Solomon, and Evgeny Burnaev. Continuous wasserstein-2 barycenter estimation without minimax optimization. *arXiv preprint arXiv:2102.01752*, 2021.
- Alexander Korotin, Vage Egiazarian, Lingxiao Li, and Evgeny Burnaev. Wasserstein iterative networks for barycenter estimation. *Advances in Neural Information Processing Systems*, 35:15672–15686, 2022a.
- Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. *arXiv preprint arXiv:2201.12220*, 2022b.
- Wouter M Kouw and Marco Loog. A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):766–785, 2019.
- Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger-kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.
- Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008.
- Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR, 2020.
- Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l’Académie Royale des Sciences de Paris*, 1781.
- Eduardo Montesuma, Fred Maurice Ngole Mboula, and Antoine Souloumiac. Multi-source domain adaptation through dataset dictionary learning in wasserstein space. In *ECAI 2023*, pages 1739–1746. IOS Press, 2023.
- Eduardo Fernandes Montesuma and Fred Maurice Ngole Mboula. Wasserstein barycenter for multi-source domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16785–16793, 2021.
- Eduardo Fernandes Montesuma, Fred Ngolè Mboula, and Antoine Souloumiac. Multi-source domain adaptation meets dataset distillation through dataset dictionary learning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5620–5624. IEEE, 2024a.
- Eduardo Fernandes Montesuma, Fred Ngolè Mboula, and Antoine Souloumiac. Lighter, better, faster multi-source domain adaptation with gaussian mixture models and optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 21–38. Springer, 2024b.
- Eduardo Fernandes Montesuma, Michela Mulas, Fred Ngolè Mboula, Francesco Corona, and Antoine Souloumiac.

- Benchmarking domain adaptation for chemical processes on the tennessee eastman process. In *MLACCE Workshop*, 2024c.
- Eduardo Fernandes Montesuma, Fred Maurice Ngolè Mboula, and Antoine Souloumiac. Recent advances in optimal transport for machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2): 1161–1180, 2025. doi: 10.1109/TPAMI.2024.3489030.
- Frank Nielsen. An elementary introduction to information geometry. *Entropy*, 22(10), 2020. ISSN 1099-4300. doi: 10.3390/e22101100. URL <https://www.mdpi.com/1099-4300/22/10/1100>.
- Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.
- Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.
- Boris T Polyak. Gradient methods for solving equations and inequalities. *USSR Computational Mathematics and Mathematical Physics*, 4(6):17–32, 1964.
- Léo Portales, Edouard Pauwels, and Elsa Cazelles. Sample complexity of optimal transport barycenters with discrete support. *arXiv preprint arXiv:2505.21274*, 2025.
- Joaquin Quiñero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2022.
- Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In *The 22nd International Conference on artificial intelligence and statistics*, pages 849–858. PMLR, 2019.
- Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.
- Christopher Reinartz, Murat Kulahci, and Ole Ravn. An extended tennessee eastman simulation dataset for fault-detection and decision support systems. *Computers & chemical engineering*, 149:107281, 2021.
- Tobias Ringwald and Rainer Stiefelhausen. Adaptiope: A modern benchmark for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 101–110, 2021.
- Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.
- Filippo Santambrogio. *Optimal transport for applied mathematicians*, volume 87. Springer, 2015.
- Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7:87–154, 2017.
- Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. *Advances in Neural Information Processing Systems*, 33:22045–22055, 2020.
- Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.
- Sanvesh Srivastava, Cheng Li, and David B Dunson. Scalable bayes via barycenter in wasserstein space. *Journal of Machine Learning Research*, 19(8):1–35, 2018.
- Shiliang Sun, Honglei Shi, and Yuanbin Wu. A survey of multi-source domain adaptation. *Information Fusion*, 24: 84–92, 2015.
- Asuka Takatsu. Wasserstein geometry of gaussian measures. *Osaka Journal of Mathematics*, 2011.
- Rosanna Turrise, Rémi Flamary, Alain Rakotomamonjy, and Massimiliano Pontil. Multi-source domain adaptation via weighted joint distributions optimal transport. In *Uncertainty in artificial intelligence*, pages 1970–1980. PMLR, 2022.
- Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5): 988–999, 1999.
- Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.
- Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2008.
- Gabriele Visentin and Patrick Cheridito. Computing optimal transport maps and wasserstein barycenters using conditional normalizing flows. *arXiv preprint arXiv:2505.22364*, 2025.
- Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. *arXiv preprint arXiv:2412.07236*, 2024.
- Linda F Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.