

Sobolev Gradient Ascent for Optimal Transport : Barycenter Optimization and Convergence Analysis

Kaheon Kim[1] Bohan Zhou[2] Changbo Zhu[1] Xiaohui Chen[3]

April 16, 2026

[1] Department of ACMS, University of Notre Dame

[2] Deparment of Mathematics, University of California, Santa Barbara

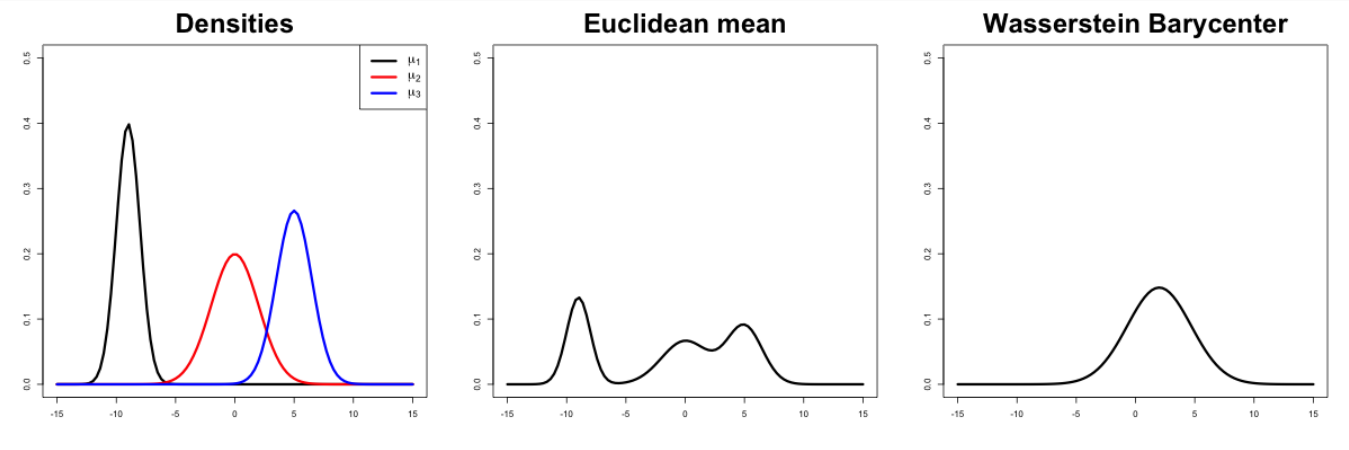
[3] Deparment of Mathematics, University of Southern California

ICLR

International Conference On Learning Representations

Introduction

Wasserstein Barycenter : Average of probability distributions.



Wasserstein Barycenter for $\mu_1, \cdots, \mu_n \in \mathcal{P}(\Omega)$ is formulated as

$$\bar{\mu} = \arg \min_{\nu \in \mathcal{P}(\Omega)} \mathcal{B}(\nu) := \sum_{i=1}^{m} \frac{\alpha_i}{2} W_2^2(\mu_i, \nu).$$

where α_i's are positive weights such that $\sum_{i=1}^{m} \alpha_i = 1$.

Existing Methodologies

With Regularization

- **CWB** : Convolutional Wasserstein Barycenter (Cuturi and Doucet (2014))
- **DSB** : Debiased Sinkhorn Barycenter (Janati et al. (2020))

Limitation

- Blurriness due to regularization
- High computational complexity: $O(m^2)$

  (m: number of grid points)

Without Regularization

- **WDHA** : Wasserstein Descent $\dot{\mathbb{H}}^1$ Ascent (Kim et al. (2025))

Limitation

- Gap between the theoretically analyzable algorithm and the computationally feasible implementation.

Contribution

We will introduce Sobolev Gradient Ascent(SGA) Algorithm that achieves

- A global convergence rate comparable to classical subgradient methods
- Improved computational efficiency by avoiding c-concavity projections
- Strong empirical performance compared with existing OT barycenter solvers

Formulation

Constraint-free concave dual formulation for the Wasserstein barycenter problem:

$$\max_{f_1,\dots,f_{m-1}} \Big\{ \mathcal{D}(f_1,\dots,f_{m-1}) := \sum_{i=1}^{m-1} \alpha_i \int_\Omega f_i^c \, d\mu_i + \alpha_m \int_\Omega f_{\text{mix}}^c \, d\mu_m \Big\},$$

where $f_{\text{mix}} = -\sum_{i=1}^{m-1} \frac{\alpha_i}{\alpha_m} f_i$ and the supremum over $f_1,\dots,f_{m-1}$ is unconstrained.

Theorem 4 (Strong duality and barycenter characterization)

1. *If $(\tilde{f}_1,\dots,\tilde{f}_{m-1})$ maximizes $\mathcal{D}(f_1,\dots,f_{m-1})$, then for each $i = 1,\dots,m-1$, $\tilde{f}_i$ coincides μ_i-a.e. with a c-concave function. With $\tilde{f}_{\text{mix}} := -\sum_{i=1}^{m-1} \frac{\alpha_i}{\alpha_m} \tilde{f}_i$, $\tilde{f}_{\text{mix}}$ coincides μ_m-a.e. with a c-concave function.*
2. *We have*

$$\min_{\nu \in \mathcal{P}_2(\Omega)} \mathcal{B}(\nu) \;=\; \max_{f_1,\dots,f_{m-1}} \mathcal{D}(f_1,\dots,f_{m-1}).$$

Let $(\tilde{f}_1,\dots,\tilde{f}_{m-1})$ be a c-concave maximizer of $\mathcal{D}$, and define $\tilde{f}_{\text{mix}}$ as above. Then the (unique) Wasserstein barycenter satisfies

$$\tilde{\nu} = (T_{\tilde{f}_i^c})_{\#}\mu_i \;=\; (T_{\tilde{f}_{\text{mix}}^c})_{\#}\mu_m,$$

where $T_h = id - \nabla h$.

Sobolev Gradient Ascent (SGA) Algorithm

The $\dot{\mathbb{H}}^1$-gradient of $\mathcal{D}$ at f_i, denoted as $\nabla_{f_i}\mathcal{D}$, can be expressed as

$$\nabla_{f_i}\mathcal{D} = (-\Delta)^{-1}\delta\mathcal{D}_{f_i} = (-\Delta)^{-1}\Big(- \alpha_i\big((T_{f_i^c})_{\#}\mu_i - (T_{f^c_{\text{mix}}})_{\#}\mu_m\big)\Big),$$

where $(-\Delta)^{-1}$ is the negative inverse Laplacian operator.

Algorithm 4 SGA Algorithm for Barycenter Computation

Initialize $f_i^{(0)}$ for $i = 1, 2, \ldots, m-1$
for $t = 1, 2, \ldots, T$ **do**
  for $i = 1, 2, \ldots, m-1$ **do**
    $f_i^{(t)} = f_i^{(t-1)} + \eta_{t-1}\,\nabla_{f_i}\mathcal{D}\big(f_1^{(t-1)}, \ldots, f_{m-1}^{(t-1)}\big)$
  end for
end for

Theoretical Result : Convergence rate for barycenter optimization with SGA

Theorem 6

(i) **Constant step size.** If the total number of iterations T is fixed a priori, then by setting $M = \max_{t=1}^{T} \left(\sum_{i=1}^{m-1} \alpha_i^2 \left\| (T_{(f_i^{(t-1)})^c})_{\#}\mu_i - (T_{(f_{\text{mix}}^{(t-1)})^c})_{\#}\mu_m \right\|_{\dot{\mathbb{H}}^{-1}}^2 \right)^{1/2}$ and $\eta_{t-1} = \left(\sum_{i=1}^{m-1} \|f_i^{(0)} - \tilde{f}_i\|_{\dot{\mathbb{H}}^1}^2 \right)^{1/2} / (M\sqrt{T})$, we have

$$\mathcal{D}(\tilde{f}_1, \ldots, \tilde{f}_{m-1}) - \mathcal{D}(f_1^{(\text{best})}, \ldots, f_{m-1}^{(\text{best})}) \leqslant M \left(\sum_{i=1}^{m-1} \|f_i^{(0)} - \tilde{f}_i\|_{\dot{\mathbb{H}}^1}^2 \right)^{1/2} \frac{1}{\sqrt{T}},$$

where $f_{\text{mix}}^{(t-1)} = -\sum_{i=1}^{m-1} \frac{\alpha_i}{\alpha_m} f_i^{(t-1)}$, $(\tilde{f}_1, \ldots, \tilde{f}_{m-1})$ is any c-concave maximizer of $\mathcal{D}$, and

$$\mathcal{D}(f_1^{(\text{best})}, \ldots, f_{m-1}^{(\text{best})}) \geqslant \max_{t=1}^{T} \mathcal{D}(f_1^{(t)}, \ldots, f_{m-1}^{(t)}).$$

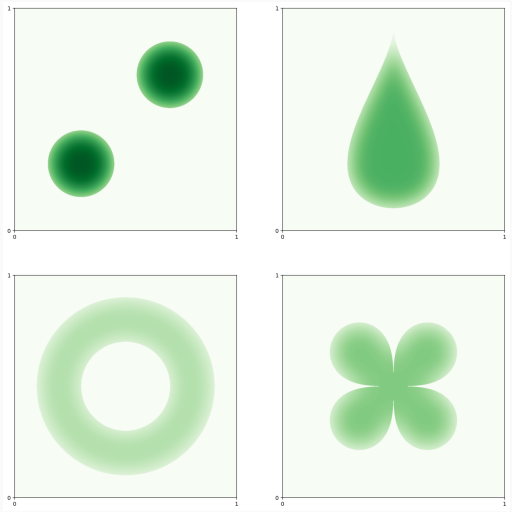
(ii) **Annealing step size.** If the total number of iterations T is not fixed a priori, then by setting $\eta_{t-1} = \left(\sum_{i=1}^{m-1} \|f_i^{(0)} - \tilde{f}_i\|_{\dot{\mathbb{H}}^1}^2 \right)^{1/2} / (M\sqrt{t})$, we have

$$\mathcal{D}(\tilde{f}_1, \ldots, \tilde{f}_{m-1}) - \mathcal{D}(f_1^{(\text{best})}, \ldots, f_{m-1}^{(\text{best})}) \leqslant M \left(\sum_{i=1}^{m-1} \|f_i^{(0)} - \tilde{f}_i\|_{\dot{\mathbb{H}}^1}^2 \right)^{1/2} \frac{\ln(T) + 2}{\sqrt{T}}.$$

Numerical Studies : 2D Simulations

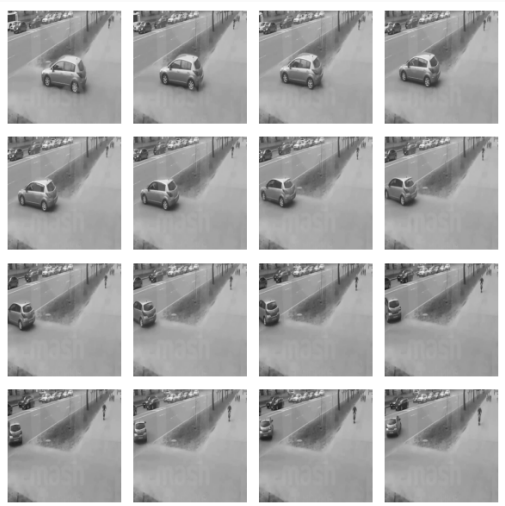
Experiment 1: Synthetic Distributions

- Data: 4 Shape data
- Grid Size: $m = 1024 \times 1024$



Experiment 2: Real Data

- Data: 16 surveillance video frames
- Grid Size: $m = 400 \times 400$



Comparisons

- Convolutional Wasserstein Barycenter (CWB)
- Debiased Sinkhorn Barycenter (DSB)
- Wasserstein Descent $\dot{\mathbb{H}}^1$ Ascent (WDHA)

Numerical Studies : 2D Synthetic Distributions

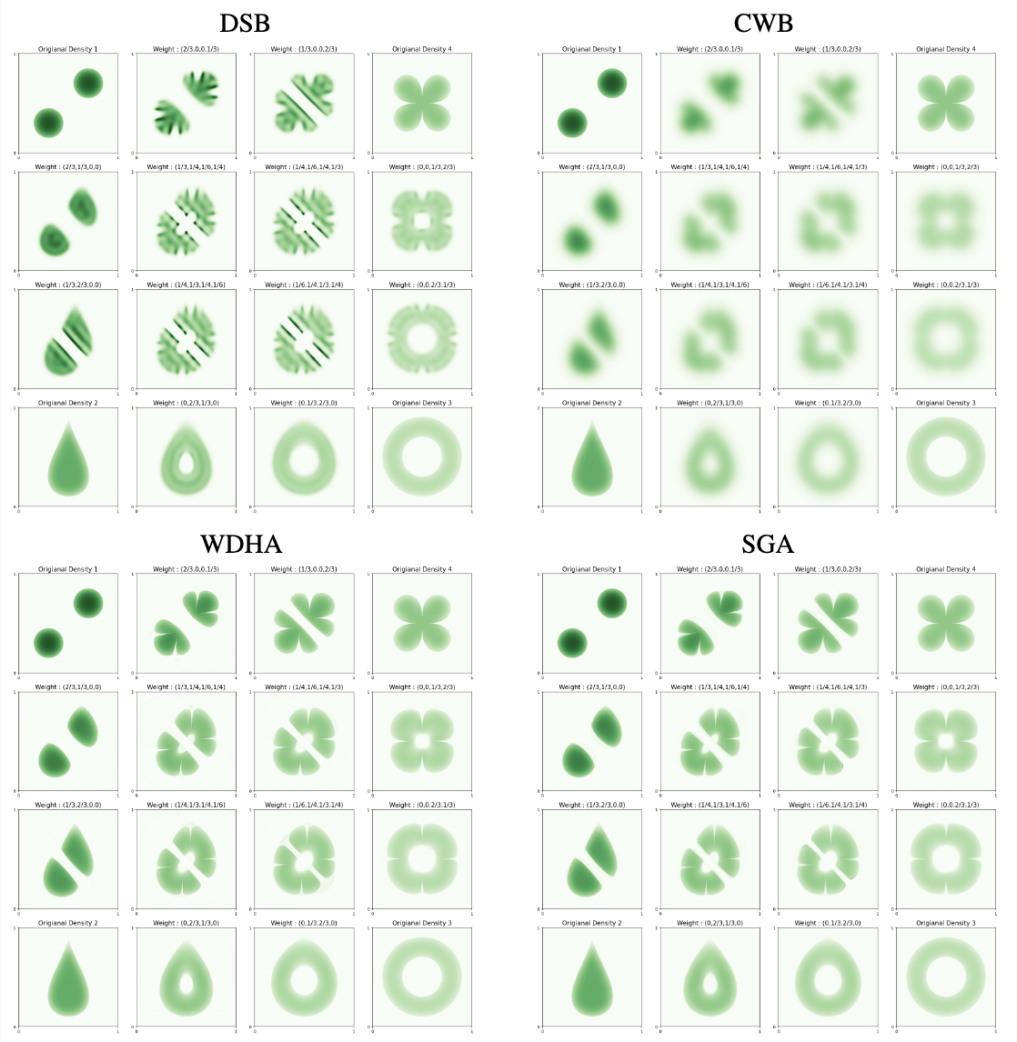


Figure 1: Comparison of weighted Wasserstein barycenter of densities supported on different shapes.

Numerical Studies : 2D Synthetic Distributions

Table 1: Barycenter functional value and running time for SGA, WDHA, CWB, and DSB for the 2D synthetic distribution example. [↓] means the smaller number, the better performer among the four methods under comparison.

Weights	Barycenter Functional Value ($\times 10^3$) [↓]				Time (in sec) [↓]			
	SGA	WDHA	CWB	DSB	SGA	WDHA	CWB	DSB
$(\frac{2}{3}, 0, 0, \frac{1}{3})$	**3.917**	3.940	4.202	3.926	**298**	625	1346	3643
$(\frac{1}{3}, 0, 0, \frac{2}{3})$	**3.917**	3.923	4.160	3.925	**221**	569	1328	3679
$(\frac{2}{3}, \frac{1}{3}, 0, 0)$	**5.257**	5.275	5.559	5.260	**257**	602	1145	3728
$(\frac{1}{3}, \frac{1}{4}, \frac{1}{6}, \frac{1}{4})$	**5.923**	5.952	6.153	5.936	**700**	1064	2290	5003
$(\frac{1}{4}, \frac{1}{6}, \frac{1}{4}, \frac{1}{3})$	**5.299**	5.313	5.513	5.314	**640**	1049	2464	5192
$(0, 0, \frac{1}{3}, \frac{2}{3})$	**1.646**	1.649	1.821	1.648	**202**	550	1262	3798
$(\frac{1}{3}, \frac{2}{3}, 0, 0)$	**5.257**	5.268	5.513	5.263	**214**	537	1256	3885
$(\frac{1}{4}, \frac{1}{3}, \frac{1}{4}, \frac{1}{6})$	**5.964**	5.987	6.182	5.977	**672**	984	2520	5062
$(\frac{1}{6}, \frac{1}{4}, \frac{1}{3}, \frac{1}{4})$	**5.219**	5.232	5.424	5.239	**626**	961	2515	5076
$(0, 0, \frac{2}{3}, \frac{1}{3})$	**1.645**	1.647	1.841	1.648	**206**	498	1255	3664
$(0, \frac{2}{3}, \frac{1}{3}, 0)$	**3.812**	3.833	4.006	3.814	**211**	550	1247	3675
$(0, \frac{1}{3}, \frac{2}{3}, 0)$	**3.813**	3.814	4.023	3.814	**209**	510	1251	3648

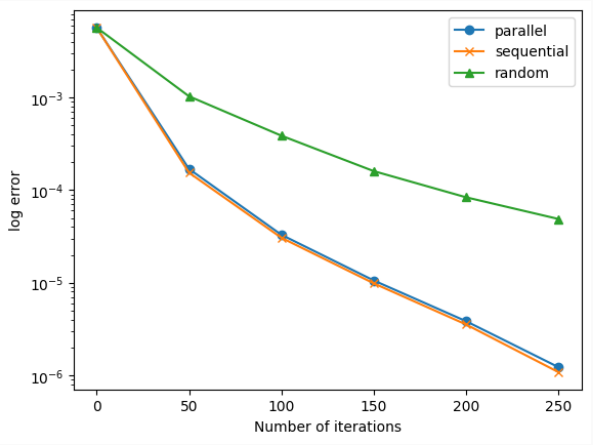


Figure 2: Empirical convergence rates of parallel, sequential, and random SGA algorithms.

Numerical Studies : 2D Real Data

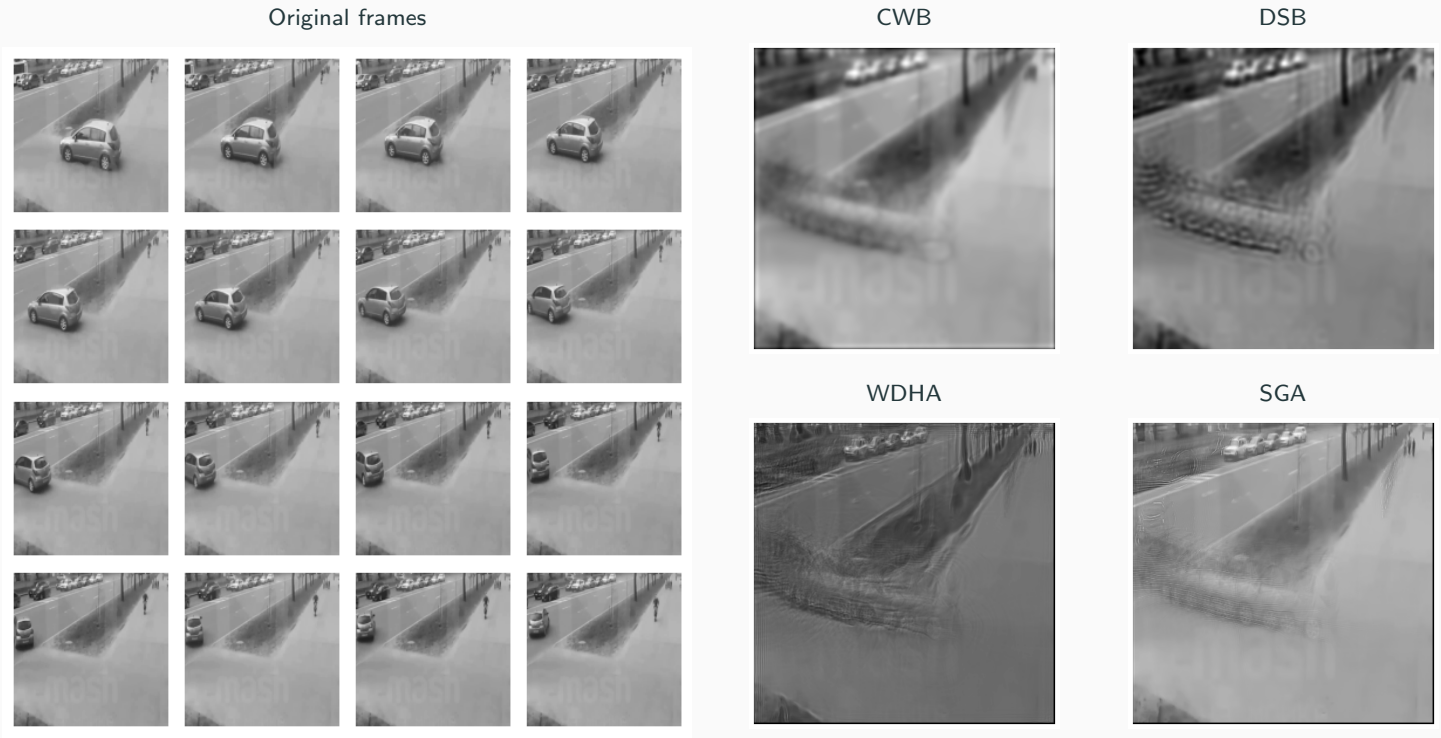


Figure 3: Left half displays 16 surveillance video frames. Right half displays their barycenters computed by CWB (top left), DSB (top right), WDHA (bottom left), and SGA (bottom right).

SGA	WDHA	CWB	DSB
3.11×10^{-5}	3.68×10^{-5}	3.22×10^{-5}	3.21×10^{-5}

Table 2: Barycenter functional value for SGA, WDHA, CWB, and DSB for 16 video surveillance data.

Numerical Studies : 3D Simulation

Experiment 3: Interpolation between synthetic 3D Distributions

- Data: 2 uniform distribution data (ball and cube)
- Grid Size: $m = 200 \times 200 \times 200$

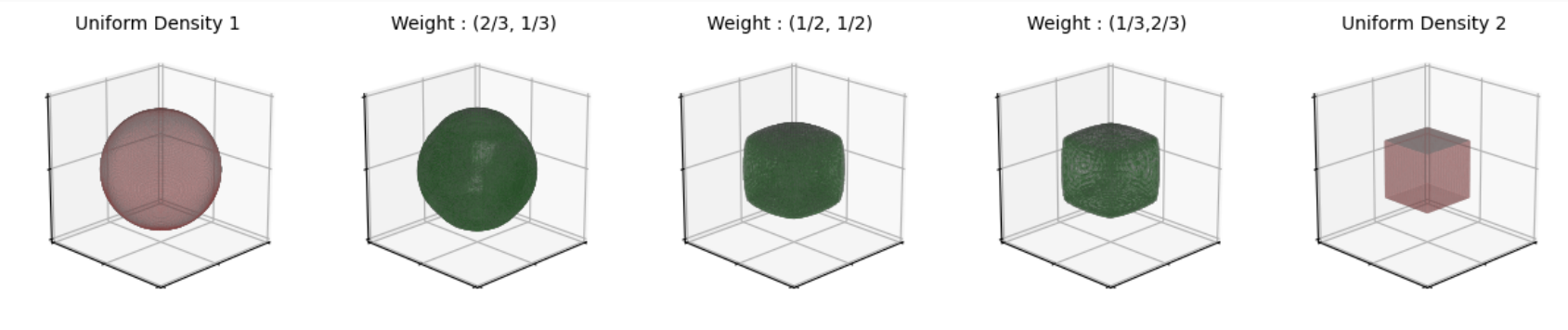


Figure 4: Interpolation using SGA between a 3D ball and a 3D cube.

Thank you!

References

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

Janati, H., Cuturi, M., and Gramfort, A. (2020). Debiased Sinkhorn barycenters. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4692–4701.

Kim, K., Yao, R., Zhu, C., and Chen, X. (2025). Optimal transport barycenter via nonconvex concave minimax optimization. In *International Conference on Machine Learning (ICML)*.