Optimal Transport Barycenter via Nonconvex-Concave Minimax Optimization

Kaheon Kim[1] Rentian Yao[2] Changbo Zhu[1] Xiaohui Chen[3]

June 15, 2025

[1]Department of ACMS, University of Notre Dame

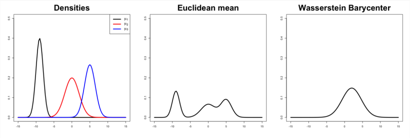
[2]Deparment of Mathematics, University of British Columbia

[3]Deparment of Mathematics, University of Southern California

ICML
International Conference
On Machine Learning

Introduction

Optimal Transport (Wasserstein) Barycenter : Average of probability distributions.



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

$$\bar{\mu} = \arg\min_{\mu \in \mathcal{P}(\Omega)} \frac{1}{n} \sum_{i=1}^{n} W_2^2(\mu, \mu_i)$$

- For 1D distributions, the barycenter $\bar{\mu}$ satisfies $Q_{\bar{\mu}} = \frac{1}{n} \sum_{i=1}^{n} Q_{\mu_i}$.
- No closed-form solution for multivariate distributions

Contributions

Existing Methodologies (available through POT package)

- Convolutional Wasserstein Barycenter(Cuturi et al 2014)
- Debiased Sinkhorn Barycenter(Janati et al 2020)

Those methods have disadvantage under high resolution settings

- Blurriness Issue due to the regularization
- High Computational complexity of $O(m^2)$ (m : # grids)

Our main contribution lies in

- Resolve Blurriness Issue by targeting exact barycenter
- Efficient Computation with $O(m \log m)$ (m : # grids)

Formulation

Our formulation : Substituting Wasserstein metric with Kantorovich dual formulation

$$
\begin{aligned}
\bar{\mu} &= \arg \min_{\nu \in \mathcal{P}(\Omega)} \frac{1}{n} \sum_{i=1}^{n} W_2^2(\nu, \mu_i) \\
&= \arg \min_{\nu \in \mathcal{P}(\Omega)} \frac{1}{n} \sum_{i=1}^{n} \max_{\varphi_i : \text{convex}} \underbrace{\int \left(\frac{\|x\|_2^2}{2} - \varphi_i(x) \right) d\nu(x) + \int \left(\frac{\|y\|_2^2}{2} - \varphi_i^*(y) \right) d\mu_i(y)}_{\mathcal{I}_\nu^{\mu_i}(\varphi_i)}
\end{aligned}
$$

Nonconvex-concave Minimax problem

$$
\min_{\nu \in \mathcal{P}_2^r(\Omega)} \max_{\varphi_i \in \mathbb{F}_{\alpha,\beta}} \mathcal{J}(\nu, \varphi) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}_\nu^{\mu_i}(\varphi_i)
$$

- $\mathcal{P}_2^r$: set of absolutely continuous probability measures whose second order moment is finite
- $\mathbb{F}_{\alpha,\beta}$: a set of $\alpha-$strongly convex and β-smooth functions
  - $\mathbb{F}_{0,\infty}$: Set of convex functions

Wasserstein Descent $\dot{\mathbb{H}}^1$-Ascent

Our Approach : Applying Gradient Descent Acent(Lin et al, 2020)-like algorithm for our Nonconvex-concave Minimax problem

$$\min_{\nu \in \mathcal{P}_2^r(\Omega)} \max_{\varphi_i \in \mathbb{F}_{\alpha,\beta}} \mathcal{J}(\nu, \varphi) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}_\nu^{\mu_i}(\varphi_i)$$

Gradients in two different geometric spaces

- (Descent) Wasserstein Gradient (Zemel et al 2019)

$$\nabla \mathcal{J}(\nu, \boldsymbol{\varphi}) = \text{id} - \nabla \bar{\varphi}, \text{ where } \bar{\varphi} = \frac{1}{n} \sum_{i=1}^{n} \varphi_i$$

- (Ascent) $\dot{\mathbb{H}}^1$ gradient (Jacobs et al 2020)

$$\mathbb{\nabla}_{\varphi_i} \mathcal{J}(\nu, \boldsymbol{\varphi}) = \frac{1}{n} (-\Delta)^{-1} (-\nu + (\nabla \varphi_i^*)_{\#} \mu_i)$$

- $(-\Delta)^{-1}$: Inverse Laplacian Operator
- Solved by Fast Fourier Transform(FFT) : $O(m \log m)$ (m : # grids)

Wasserstein Descent $\dot{\mathbb{H}}^1$-Ascent(WDHA)
Alternating Wasserstein gradient(descent)/$\dot{\mathbb{H}}^1$-gradient(ascent)

Algorithm 3: Wasserstein-Descent $\dot{\mathbb{H}}^1$-Ascent Algorithm

Initialize $\nu^1, \boldsymbol{\varphi}^1$;

for $t = 1, 2, \cdots, T-1$ **do**

  for $i = 1, 2, \ldots, n$ **do**

    $\widehat{\varphi}_i^{t+1} = \varphi_i^t + \eta \boldsymbol{\nabla}_{\varphi_i} \mathcal{J}(\nu^t, \boldsymbol{\varphi}^t)$;

    $\varphi_i^{t+1} = \mathcal{P}_{\mathbb{F}_{\alpha,\beta}}(\widehat{\varphi}_i^{t+1})$;

  $\nu^{t+1} = (\mathrm{id} - \tau \mathbb{\nabla} \mathcal{J}(\nu^t, \boldsymbol{\varphi}^t))_{\#} \nu^t$;

return $\{(\nu^t, \boldsymbol{\varphi}^t)\}_{t=1}^T$;

Figure 1: Wasserstein Descent $\dot{\mathbb{H}}^1$-Ascent

- $\mathcal{P}_{\mathbb{F}_{\alpha,\beta}}$: Projection onto $\mathbb{F}_{\alpha,\beta}$

Theoretical Result

Definition 1 (Stationary point)
We call $\nu \in \mathcal{P}_2^r(\Omega)$ a stationary point of $\mathcal{F}_{\alpha,\beta}$ if $\int_\Omega \|\mathbb{V}\mathcal{F}_{\alpha,\beta}(\nu)\|_2^2 d\nu = 0$

Theorem 2 (Convergence of WDHA)
Assume that there are constant a and b, such that the density functions satisfy $0 < a \leq \mu_i(x) \leq b < \infty$ for all $i = 1, 2 \ldots, n$ and $x \in \Omega$. Recall that $A = a\alpha^d/\beta$ and $B = b\beta^d/\alpha$. If $\max_t \|\nu^t\|_\infty \leq V < \infty$ for some constant $V > 0$, by choosing the step sizes (τ, η) satisfying $\eta < 1/B$ and $\tau < \frac{A^2\eta}{A\eta(A\alpha+A+V)+4V\sqrt{4-2A\eta}}$, we have

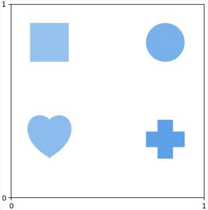
$$\min_{t=1,\ldots,T} \int_\Omega \|\mathbb{V}\mathcal{F}_{\alpha,\beta}\|_2^2 \, d\nu^t \leq \frac{1}{T}\sum_{t=1}^{T} \int_\Omega \left\|\mathbb{V}\mathcal{F}_{\alpha,\beta}(\nu^t)\right\|_2^2 d\nu^t \leq \frac{\frac{4\tau V\bar{\delta}^1}{A\eta} + \mathcal{F}_{\alpha,\beta}(\nu^1) - \mathcal{F}_{\alpha,\beta}(\nu^{T+1})}{T\tau/2}$$

where $\mathcal{F}_{\alpha,\beta}(\nu) := \frac{1}{n}\sum_{i=1}^n \mathcal{I}_\nu^{\mu_i}$, $\mathcal{L}^{\mu_i}(\nu) := \max_{\varphi_i \in \mathbb{F}_{\alpha,\beta}} \mathcal{I}_\nu^{\mu_i}(\varphi_i)$ and $\bar{\delta}^1 = \bar{\delta}^1(\nu^1, \boldsymbol{\varphi}^1, \mu_1, \ldots, \mu_n) > 0$ is a constant.

Numerical Studies : Simulation

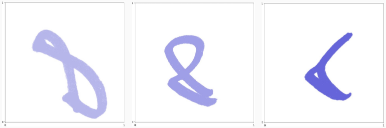
Experiment 1: Uniform Distributions

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



Experiment 2: Hand Digit Data

- Data: 100 hand-written 8 digits
- Grid Size: $m = 500 \times 500$



Comparisons

- Convolutional Wasserstein Barycenter(CWB)
- Debiased Sinkhorn Barycenter (DSB)
- (Experiment 1) CWB and DSB with Thresholding
  - Removing intensities smaller than the threshold so that the removed intensities amount to 10% of the mass

Experimental Setting

Algorithm 4: Wasserstein-Descent $\dot{\mathbb{H}}^1$-Ascent Algorithm

Initialize $\nu^1, \boldsymbol{\varphi}^1$;

for $t = 1, 2, \cdots, T-1$ **do**

  for $i = 1, 2, \ldots, n$ **do**

    $\widehat{\varphi}_i^{t+1} = \varphi_i^t + \eta_i^t \boldsymbol{\nabla}_{\varphi_i} \mathcal{J}(\nu^t, \boldsymbol{\varphi}^t)$;

    $\varphi_i^{t+1} = (\widehat{\varphi}_i^{t+1})^{**}$;

  $\nu^{t+1} = (\mathrm{id} - \tau_t \boldsymbol{\nabla} \mathcal{J}(\nu^t, \boldsymbol{\varphi}^t))_{\#} \nu^t$;

return $\{(\nu^t, \boldsymbol{\varphi}^t)\}_{t=1}^T$;

Figure 2: Wasserstein Descent $\dot{\mathbb{H}}^1$-Ascent(Empirical)

- Number of Iteration : $T = 300$
- Stepsize for t-th iteration
  - Wasserstein Descent : $\exp(-t/T)$
  - $\dot{\mathbb{H}}^1$ Ascent : $\eta_i^1 = 1$ and $\eta_{t+1} = 0.99\eta_t$ if $\mathcal{I}_{\nu^t}^{\mu_i}(\varphi_i^{t+1}) < \mathcal{I}_{\nu^t}^{\mu_i}(\varphi_i^t)$
- Convex hull $(\cdot)^{**}$: Computationally efficient compared to $\mathcal{P}_{\mathbb{F}_{\alpha,\beta}}$.

Numerical Studies : Results

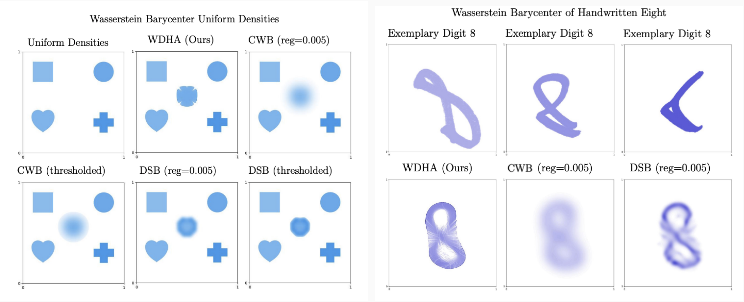


Figure 3: Comparison between methodologies

	CWB	$\text{CWB}_{\text{(thresholded)}}$	DSB	$\text{DSB}_{\text{(Thresholded)}}$	WDHA
$\frac{1}{n}\sum_{i=1}^{n} W_2^2(\nu^{\text{est}}, \mu_i)$	75.0689	74.7346	74.5804	74.7346	**74.5791**
Time(s)	3731	3731	7249	7249	**676**

Table 1: Numerical Comparison between methodologies(Shape Data)

	CWB	DSB	WDHA
Time(s)	10808	11186	**3299**

Table 2: Numerical Comparison between methodologies(Hand-digit Data)

Thank you!