A Particle-Evolving Method for Approximating the Optimal Transport Plan

Shu Liu, Haodong Sun, and Hongyuan Zha

2021.9.16

Outlines

Optimal Transport (OT)

- Optimal transport (OT) provides powerful tools for comparing probability measures in various types.
- In spite of elegant theoretical results, generally computing Wasserstein distance is not an easy task, especially for the continuous case.

Solving the Optimal Coupling to the OT Problem

- Traditional methods:
  - Sinkhorn iteration.
  - linear programming; Monge-Ampère equation; dynamical scheme; or methods involving neural network optimizations.
- This paper proposes a method to directly compute the sample approximation of the optimal coupling between two density functions.

Contributions

- Analyze the theoretical properties of the Entropy Transport problem constrained on probability space and derive its Wasserstein gradient flow.
- Propose an innovative particle-evolving algorithm for obtaining the sample approximation of the optimal transport plan.

Optimal Transport Problem

- We mainly focus on Euclidean Space $\mathbb{R}^d$.
- $\mathcal{P}(E)$: the probability space defined on the given measurable set E.
- The Optimal Transport problem is usually formulated as

$$\inf_{\gamma \in \mathcal{P}\left(\mathbb{R}^d \times \mathbb{R}^d\right),} \iint c(x, y) d\gamma(x, y) \tag{1}$$

  - $\mu, \nu \in \mathcal{P}\left(\mathbb{R}^d\right)$
  - γ_1, γ_2: the marginal distribution of γ w.r.t. component x and y
  - γ_{OT}: the optimizer of (1) as Optimal Transport plan

Reformulation

We reformulate (1) as $\min_{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)} \{\mathcal{E}_\iota(\gamma \mid \mu, \nu)\}$ where

$$\mathcal{E}_\iota(\gamma \mid \mu, \nu) = \iint c(x, y)\, d\gamma(x, y) + \int \iota\left(\frac{d\gamma_1}{d\mu}\right) d\mu + \int \iota\left(\frac{d\gamma_2}{d\nu}\right) d\nu \tag{2}$$

Here ι is defined as $\iota(1) = 0$ and $\iota(s) = +\infty$ when $s \neq 1$.

Relaxation

We relax (2) by replacing $\iota(\cdot)$ with $\Lambda F(\cdot)$.
We focus on $F(s) = s\log s - s + 1$ and $c(x, y) = h(x - y)$ with h as a strictly convex function, and enforce the marginal constraints by KL-divergence:

$$\mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma \mid \mu, \nu) = \iint_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y)\, d\gamma(x, y) + \Lambda D_{\mathrm{KL}}(\gamma_1 \| \mu) + \Lambda D_{\mathrm{KL}}(\gamma_2 \| \nu)$$

Constrained Entropy Transport Problem

$$\mathcal{E}_{\min} = \inf_{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)} \{\mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma \mid \mu, \nu)\} \tag{3}$$

Existence and Uniqueness of the Optimal Solution

- Theorem 1. Suppose $\tilde{\gamma}$ is the solution to original Entropy Transport problem. Then we have $\tilde{\gamma} = Z\gamma$, here $Z = e^{-\frac{\varepsilon_{\min}}{2\Lambda}}$ and $\gamma \in \mathcal{P}\left(\mathbb{R}^d \times \mathbb{R}^d\right)$ is the solution to constrained Entropy Transport problem (3).
- Corollary 1. The constrained ET problem (3) admits a unique optimal solution.

Asymptotic Convergence

Theorem 2. Suppose $c(x,y) = |x-y|^2$, let us assume $\mu, \nu \in \mathcal{P}_2\left(\mathbb{R}^d\right), \mu, \nu \ll \mathscr{L}^d$ where

$$\mathcal{P}_2(E) = \left\{ \gamma \middle| \gamma \in \mathcal{P}(E), \gamma \ll \mathscr{L}^d, \int_E \left| x \right|^2 d\gamma(x) < +\infty \right\} \quad E \text{ measurable},$$

and both μ, ν satisfy the Logarithmic Sobolev inequality with constants $K_\mu, K_\nu > 0$. Let $\{\Lambda_n\}$ be a positive increasing sequence with $\lim_{n\to\infty} \Lambda_n = +\infty$. We consider the sequence of functionals $\{\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot \mid \mu, \nu)\}$. $\{\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\cdot \mid \mu, \nu)\}$ Γ -converges to $\mathcal{E}_\iota(\cdot \mid \mu, \nu)$ on $\mathcal{P}_2\left(\mathbb{R}^d \times \mathbb{R}^d\right)$.
Furthermore, $\min_{\gamma \in \mathcal{P}\left(\mathbb{R}^d \times \mathbb{R}^d\right)} \{\mathcal{E}_{\Lambda_n,\mathrm{KL}}(\gamma \mid \mu, \nu)\}$ admits a unique optimal solution γ_n. At the same time, the Optimal Transport problem (1) also admits a unique optimal solution, we denote it as γ_{OT}. Then

$$\lim_{n\to\infty} \gamma_n = \gamma_{OT} \in \mathcal{P}_2\left(\mathbb{R}^d \times \mathbb{R}^d\right)$$

Wasserstein Gradient Flow of Entropy Transport Functional

The Wasserstein gradient flow of $\mathcal{E}_{\Lambda,\mathrm{KL}}(\cdot \mid \mu, \nu)$:

$$\frac{\partial \gamma_t}{\partial t} = -\operatorname{grad}_W \mathcal{E}_{\Lambda,\mathrm{KL}}(\gamma_t \mid \mu, \nu), \quad \gamma_t|_{t=0} = \gamma_0 \tag{4}$$

We denote $\rho(\cdot, t) = \frac{d\gamma_t}{d\mathscr{L}^{2d}}, \varrho_1 = \frac{d\mu}{d\mathscr{L}^d}, \varrho_2 = \frac{d\nu}{d\mathscr{L}^d}$. Then (4) can be written as:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot \left(\rho \nabla \left(c(x,y) + \Lambda \log\left(\frac{\rho_1(x,t)}{\varrho_1(x)}\right) + \Lambda \log\left(\frac{\rho_2(y,t)}{\varrho_2(y)}\right)\right)\right) \tag{5}$$

Here $\rho_1(\cdot, t) = \int \rho(\cdot, y, t)\, dy$ and $\rho_2(\cdot, t) = \int \rho(x, \cdot, t)\, dx$ are density functions of marginals of γ_t.

Relating the Wasserstein Gradient Flow to a Particle System

- Wasserstein gradient flows can be viewed as a time evolution PDE describing the density evolution of a stochastic process.
- The vector field that drives the random particles at time t should be $-\nabla\left(c(x,y)+\Lambda\log\left(\frac{\rho_1(x,t)}{\varrho_1(x)}\right)+\ \Lambda\log\left(\frac{\rho_2(y,t)}{\varrho_2(y)}\right)\right)$.

Dynamics of the Partical

- The dynamics of $\{(X_t, Y_t)\}_{t\geq 0}$: (here $\dot{X}_t$ denotes the time derivative $\frac{dX_t}{dt}$)

$$\begin{cases} \dot{X}_t = -\nabla_x c\,(X_t, Y_t) + \Lambda\,(\nabla \log \varrho_1\,(X_t) - \nabla \log \rho_1\,(X_t, t)) \\ \dot{Y}_t = -\nabla_y c\,(X_t, Y_t) + \Lambda\,(\nabla \log \varrho_2\,(Y_t) - \nabla \log \rho_2\,(Y_t, t)) \end{cases}$$

  Here $\rho_1(\cdot, t)$ denotes the probability density of random variable X_t and $\rho_2(\cdot, t)$ denotes the density of Y_t. The density $\rho_t(x, y)$ of (X_t, Y_t) solves the PDE (5).

- Transport plan $\leftarrow$ Movement of the particle (X_t, Y_t) at $t \leftarrow \rho\,(X_t, Y_t, t)$.

KDE

We apply the Kernel Density Estimation to approximate the gradient log function $\nabla \log \rho(x)$ by convolving it with kernel $K(x, \xi)$

$$\nabla \log \rho(x) \approx \nabla \log(K * \rho)(x) = \frac{(\nabla_x K) * \rho(x)}{K * \rho(x)}$$

Here $K * \rho(x) = \int K(x, \xi)\rho(\xi)\, d\xi = \mathbb{E}_{\xi \sim \rho} K(x, \xi)$,
$(\nabla_x K) * \rho(x) = \int \nabla_x K(x, \xi)\rho(\xi)\, d\xi = \mathbb{E}_{\xi \sim \rho} \nabla_x K(x, \xi)$.

Blobing Method

$\nabla \log \rho(x)$ is evaluated based on the locations of the particles:

$$\frac{\mathbb{E}_{\xi\sim\rho}\nabla_x K(x,\xi)}{\mathbb{E}_{\xi\sim\rho}K(x,\xi)} \approx \frac{\sum_{k=1}^N \nabla_x K(x,\xi_k)}{\sum_{k=1}^N K(x,\xi_k)} \quad \xi_1,\ldots,\xi_N, \text{ i.i.d. } \sim \rho$$

Simulating the Dynamics of the Partical

In the interacting particle system involving N particles $\{(X_i, Y_i)\}_{i=1,\dots,N}$, for the i-th particle:

$$\begin{cases} \dot{X}_i(t) = -\nabla_x c\,(X_i(t),\, Y_i(t)) - \Lambda \left(\nabla V_1\,(X_i(t)) + \frac{\sum_{k=1}^N \nabla_x K(X_i(t), X_k(t))}{\sum_{k=1}^N K(X_i(t), X_k(t))} \right) \\ \dot{Y}_i(t) = -\nabla_y c\,(X_i(t),\, Y_i(t)) - \Lambda \left(\nabla V_2\,(Y_i(t)) + \frac{\sum_{k=1}^N \nabla_x K(Y_i(t), Y_k(t))}{\sum_{k=1}^N K(Y_i(t), Y_k(t))} \right) \end{cases}$$

where $V_1 = -\log \varrho_1$, $V_2 = -\log \varrho_2$.

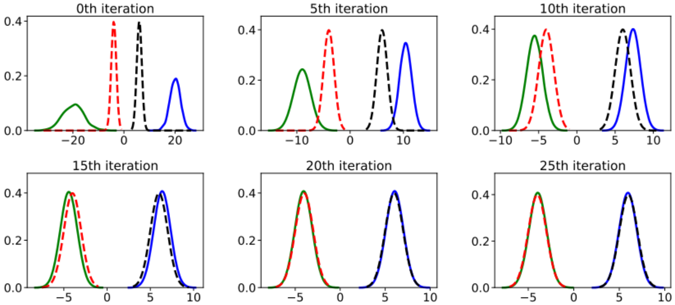
Algorithm Scheme

Algorithm 1. Random Batch Particle Evolution Algorithm

Input: The density functions of the marginals ϱ_1, ϱ_2, timestep Δt, total number of iterations T, parameters of the chosen kernel K
Initialize: The initial locations of all particles $X_i(0)$ and $Y_i(0)$ where $i = 1, 2, \cdots, n$,
for t = 1,2,$\cdots$,T **do**
  Shuffle the particles and divide them into m batches: $\mathcal{C}_1, \cdots, \mathcal{C}_m$
  for each batch $\mathcal{C}_q$ **do**
    Update the location of each particle (X_i, Y_i) $(i \in \mathcal{C}_q)$ according to (11)
  end for
end for
Output: A sample approximation of the optimal coupling: $X_i(T), Y_i(T)$ for $i = 1, 2, \cdots, n$

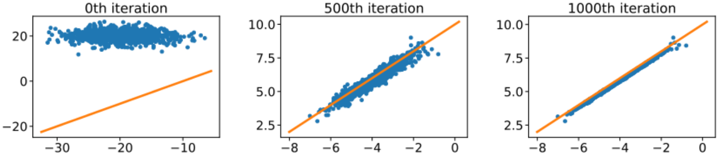
1D Gaussian

- $\varrho_1 = \mathcal{N}(-4, 1), \varrho_2 = \mathcal{N}(6, 1)$
- Red and black dashed lines: two marginal distribution.
- Solid blue and green lines: the kernel estimated density functions of particles at certain iterations.



After first 25 iterations, the particles have matched the marginal distributions very well.

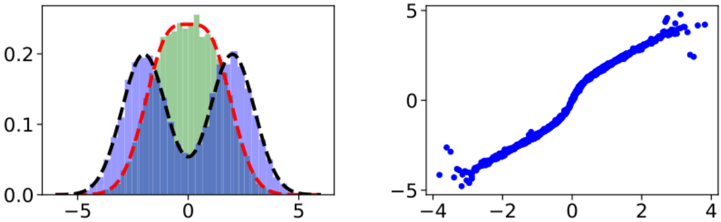
Sample Approximation for 1D Gaussian



The orange dash line corresponds to the Optimal Transport map $T(x) = x + 10$.

Gaussian Mixture

$\varrho_1 = \frac{1}{2}\mathcal{N}(-1,1) + \frac{1}{2}\mathcal{N}(1,1), \varrho_2 = \frac{1}{2}\mathcal{N}(-2,1) + \frac{1}{2}\mathcal{N}(2,1).$



Left:

- The dash lines: two marginal distributions.
- The histogram: the distribution of particles after 5000 iterations.

Right: Sample approximation for the optimal coupling.

Wasserstein Barycenters

Aim: to solve the Wasserstein barycenter

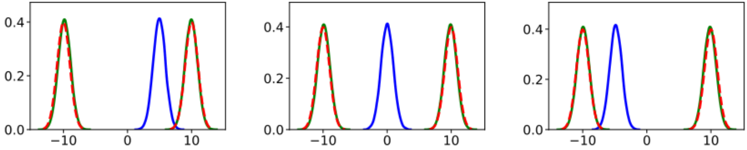
$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^{m} \lambda_i W_2^2(\mu, \mu_i)$$

where $\lambda_i > 0$ are the weights.
We relax the marginal constraints

$$\min_{\gamma \in \mathcal{P}\left(\mathbb{R}^{(m+1)d}\right)} \int_{\mathbb{R}^{(m+1)d}} \sum_{j=1}^{m} \lambda_j \left|x - x_j\right|^2 d\gamma(x, x_1, \ldots, x_m) + \sum_{j=1}^{m} \Lambda_j D_{\mathrm{KL}}\left(\gamma_j \| \mu_j\right)$$

Wasserstein Barycenters

- cost function $c(x, x_1, x_2) = w_1 |x - x_1|^2 + w_2 |x - x_2|^2$
- Experiments are under different weights: $[w_1, w_2] = [0.25, 0.75], [0.5, 0.5], [0.75, 0.25]$.



- Red dashed lines: two marginal distributions
- Solid green lines: kernel estimated density functions of the particles X_1' s and X_2' s.
- Solid blue line: kernel estimated density function of the particles corresponding to the barycenter.

Summary

- This paper proposes a novel algorithm that computes for the sample-wised optimal transport plan by evolving an interacting particle system.
- Existing numerical experiments are low dimensional cases and lack of comparisons with other methods.