

Computational and statistical optimal transport

Jiaxi Zhao

26th May, 2021

Brief history of optimal transport

Toward High-dimensional OT

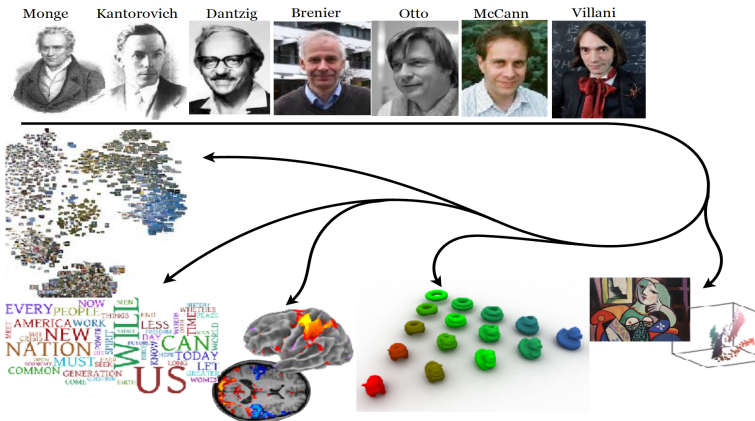


Figure: Copied from Prof. S. Justin's lecture note.

Optimal transport: Monge formulation

Given two target distribution μ, ν in space X, Y and a cost function $c(\cdot, \cdot) : X \times Y \rightarrow \mathbb{R}_{\geq 0}$, we consider to find a mapping $f : X \rightarrow Y$ such that $f_*\mu = \nu$ and minimize the total cost

$$C = \int_X c(x, f(x)) \mu(x) dx.$$

Usual types of cost:

L^p cost.

0-1 cost, total variation.

Related topics: Monge-Ampère equation.

Optimal transport: Kantorovich formulation

Kantorovich provided a relaxed version of this problem, relax from deterministic coupling to stochastic coupling.

$$W_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int c(x, y) d\pi(x, y),$$

$$W_p^p(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int |x - y|^p d\pi(x, y) \text{ (Wasserstein-p distance)}.$$

Denote $X_0 \sim \mu_0 = \delta_{x_0}$, $X_1 \sim \mu_1 = \delta_{x_1}$. Compare

$$W(\mu_0, \mu_1) = \inf_{\pi \in \Pi(\mu_0, \mu_1)} \mathbb{E}_{(X_0, X_1) \sim \pi} c(X_0, X_1) = c(x_0, x_1);$$

V_S

$$\text{TV}(\mu_0, \mu_1) = \int_{\Omega} |\mu_0(x) - \mu_1(x)| dx = 2;$$

V_S

$$\text{KL}(\mu_0 \| \mu_1) = \int_{\Omega} \mu_0(x) \log \frac{\mu_0(x)}{\mu_1(x)} dx = \infty.$$

Linear programming formulation

If we use histograms to replace the distributions in previous formula, i.e. $\mu \in \mathbb{R}^n, \nu \in \mathbb{R}^m$ with $\mathbf{1}^T \mu = \mathbf{1}^T \nu = 1$, we will obtain linear programming (matrix) formulation of the optimal transport, i.e.

$$\min_{\mathbf{P} \in \Pi(\mu, \nu)} \langle \mathbf{C}, \mathbf{P} \rangle, \quad (1)$$
$$\Pi(\mu, \nu) := \left\{ \mathbf{P} \in \mathbb{R}^{n \times m} \mid \mu = \mathbf{P} \mathbf{1}, \mathbf{1}^T \mathbf{P} = \nu^T \right\},$$

here $\mathbf{C} \in \mathbb{R}^{n \times m}$ is the elementwise cost function.

A LP with mn unknown variables and $m + n$ constraints, which is extremely hard to calculate when m, n is large.

Classifications of computational OT

There exist several different classifications of OT:

1. Classification by the target and source distribution:

Discrete OT (LP)

Semi-discrete OT (Voronoi diagram)

Continuous OT

2. Classification by algorithms:

Sinkhorn distance

Semi-dual OT, only applied to Wasserstein-2

Projected Wasserstein distance

3. Different goal:

Calculate distance

Calculate mapping, vector field

Sinkhorn's algorithm and Sinkhorn divergence

One famous in the computational OT is the Sinkhorn algorithm. In the original paper¹, the Sinkhorn distance is defined as

$$d_{\alpha}(\mu, \nu) = \min_{\pi \in \Pi_{\alpha}(\mu, \nu)} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) \right\} = \min_{\mathbf{P} \in \Pi_{\alpha}(\mu, \nu)} \langle \mathbf{C}, \mathbf{P} \rangle, \quad (2)$$

where the set $\Pi_{\alpha}(\mu, \nu) := \{\pi \in \Pi(\mu, \nu) | H(\pi | \mu \otimes \nu) \leq \alpha\}$ with $H(\cdot | \cdot)$ the relative entropy. Entropic OT appearing as the Lagrangian function of the Sinkhorn distance, i.e.

$$\begin{aligned} W_{\epsilon}(\mu, \nu) &= \min_{\pi \in \Pi(\mu, \nu)} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) + \epsilon H(\pi | \mu \otimes \nu) \right\} \\ &= \min_{\mathbf{P} \in \Pi(\mu, \nu)} \langle \mathbf{C}, \mathbf{P} \rangle - \epsilon H(\mathbf{P}). \end{aligned} \quad (3)$$

¹M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems, 26:2292–2300, 2013.

Sinkhorn's algorithm

Let $u = e^{-\frac{\alpha}{\epsilon}}$, $v = e^{-\frac{\beta}{\epsilon}}$ and $\mathbf{K} = e^{-\mathbf{C}/\epsilon}$. We again state the KKT system of (21):


$$\begin{aligned}\mathbf{P}_\epsilon &= \mathbf{diag}(u)\mathbf{K}\mathbf{diag}(v), \\ a &= \mathbf{diag}(u)\mathbf{K}v, \\ b &= \mathbf{diag}(v)\mathbf{K}^\top u.\end{aligned}\tag{31}$$

Then the Sinkhorn's algorithm amounts to alternating updates in the form of

$$\begin{aligned}u^{(k+1)} &= \mathbf{diag}(\mathbf{K}v^{(k)})^{-1}a, \\ v^{(k+1)} &= \mathbf{diag}(\mathbf{K}^\top u^{(k+1)})^{-1}b.\end{aligned}\tag{32}$$

Figure: Sinkhorn's algorithm

It is proved in² that the Sinkhorn's algorithm converges in linear rate with an ϵ depending constant $e^{\frac{1}{\epsilon}}$.

²C. Brauer, C. Clason, D. Lorenz, and B. Wirth. A sinkhorn-newton method for entropic optimal transport. arXiv preprint arXiv:1710.06635, 2017. 

Properties of the Sinkhorn divergence

Approximation power of Sinkhorn divergence

Theorem (A. Genevay et. al, 2019)

Let μ, ν be probability measures on X, Y respectively, subsets of \mathbb{R}^d such that $|X|, |Y| \leq D$ and assume that $c(x, y)$ is L -Lipschitz w.r.t. x, y . It holds

$$0 \leq W_\epsilon(\mu, \nu) - W(\mu, \nu) \leq 2\epsilon d \log \frac{e^2 LD}{\epsilon \sqrt{d}} \sim 2\epsilon d \log \frac{1}{\epsilon}. \quad (4)$$

$$W_\epsilon(\mu, \nu) = \inf_{\rho, \nu} \int_0^1 \int_\Omega \left(\|v(x, t)\|^2 + \frac{\epsilon^2}{4} \|\nabla \log \rho(x, t)\|^2 \right) dx dt.$$

$$\left| \mathbb{E}_{\pi_\epsilon} \|x - y\|^2 - W_2^2(\mu, \nu) \right| \leq \begin{cases} O(\exp(-1/\epsilon)), & \text{discrete OT (LP),} \\ O(\epsilon^2), & \text{semi-discrete OT,} \\ O(\epsilon), & \text{continuous OT.} \end{cases} \quad (5)$$

Properties of the Sinkhorn divergence

Sample Complexity of Sinkhorn divergence

Theorem (A. Genevay et. al, 2019)

Let μ, ν be probability measures on X, Y respectively, subsets of \mathbb{R}^d such that $|X|, |Y| \leq D$ and assume that $c(x, y)$ is C^∞ , L -Lipschitz w.r.t. x, y . One has

$$\mathbb{E} |W_\epsilon(\mu, \nu) - W_\epsilon(\hat{\mu}_N, \hat{\nu}_N)| = O\left(\frac{e^{\frac{\kappa}{\epsilon}}}{\sqrt{n}} \left(1 + \frac{1}{\epsilon^{\lfloor \frac{d}{2} \rfloor}}\right)\right), \quad (6)$$

with $\kappa = 2L|X| + \|c\|_\infty$.

Other regularization

Entropic OT has the following drawback that it breaks the sparsity in the intrinsic LP formulation. Can we use other regularization to obtain better approximate solution?

CoD in OT computation

Theorem (Boissard et. al, 2014.)

Let μ be a probability measure on $[-1, 1]^d$. If μ_n is an empirical measure comprising n i.i.d. samples from μ , then for any $p \in [1, \infty)$,

$$\mathbb{E}W_p(\mu, \mu_n) \leq r_{p,d}(n) := c_p \sqrt{d} \begin{cases} n^{-\frac{1}{2p}}, & d < 2p, \\ n^{-\frac{1}{2p}} (\log n)^{\frac{1}{p}}, & d = 2p, \\ n^{-\frac{1}{d}}, & d > 2p. \end{cases} \quad (7)$$

More on OT

Theorem 3.2 ([Vil03, Theorem 2.12]). *Let P, Q be any two probability distributions on \mathbb{R}^d with finite second order moments. Then,*

1. **(Knott-Smith optimality criterion)** *A coupling $\pi \in \Pi(P, Q)$ is optimal for the primal (2) if and only if there exists a convex function $f \in \text{CVX}(\mathbb{R}^d)$ such that $\text{Supp}(\pi) \subset \text{Graph}(\partial f)$. Or equivalently, for all $d\pi$ -almost (x, y) , $y \in \partial f(x)$. Moreover, the pair (f, f^*) achieves the minimum in the dual form (5).*
2. **(Brenier's theorem)** *If Q admits a density with respect to the Lebesgue measure on \mathbb{R}^d , then there is a unique optimal coupling π for the primal problem. In particular, the optimal coupling satisfies that*

$$d\pi(x, y) = dQ(y)\delta_{x=\nabla f^*(y)},$$

where the convex pair (f, f^) achieves the minimum in the dual problem (5). Equivalently, $\pi = (\nabla f^* \times \text{Id})_{\#} Q$.*

3. *Under the above assumptions of Brenier's theorem, ∇f^* is the unique solution to Monge transportation problem from Q to P , i.e.*

$$\mathbb{E}_Q \|\nabla f^*(Y) - Y\|^2 = \inf_{T: T_{\#} Q = P} \mathbb{E}_Q \|T(Y) - Y\|^2.$$

Figure: Strutture theorem for L^2 OT.

Semidual formulation

$$\begin{aligned} W_2(\mu, \nu) &= \min_{X, Y \sim \mu, \nu} \mathbb{E} \|X - Y\|_2^2 \\ &= \min_{X, Y \sim \mu, \nu} \left(\mathbb{E} \|X\|_2^2 + \mathbb{E} \|Y\|_2^2 - 2\mathbb{E} \langle X, Y \rangle \right) \\ &= \mathbb{E} \|X\|_2^2 + \mathbb{E} \|Y\|_2^2 - 2 \max_{X, Y \sim \mu, \nu} \mathbb{E} \langle X, Y \rangle. \end{aligned}$$

Use Lagrangian multiplier on this maximum term, we obtain for any two probability distributions μ and ν on \mathbb{R}^d with finite second order moments, we have that

$$\max_{X, Y \sim \mu, \nu} \mathbb{E} \langle X, Y \rangle = \inf_{f \text{ cvx}} \mathbb{E}_\mu f(X) + \mathbb{E}_\nu f^*(Y),$$

where f^* is convex conjugate.

Semidual formulation: semidiscrete OT

In semidiscrete case, the potential function has special form, it is the maximum of finite many linear function

$$f(x) = \max_i (y_i^t x + b_i).$$

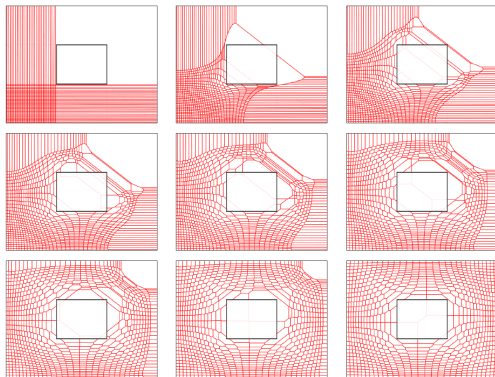


Figure: Numerical experiment




Semidual formulation: Wavelet estimator³

Use M-estimator in truncated wavelet space to estimate the mapping f :

- Minimax rate

- Local Rademacher complexity

Problem: calculating the convex conjugate is very difficult, especially at high dimension.

²Jan-Christian Hutter and Philippe Rigollet. Minimax rates of estimation for smooth optimal transport maps. arXiv preprint arXiv:1905.05828, 2019.   

Semidual formulation: GAN⁴

Instead of calculating convex conjugate, one can directly use NN to parametrize function f, f^* , i.e. solve

$$\inf_{f \in ICNN, g \in ICNN} \mathbb{E} [f(\nabla g(Y)) - f(X) + \langle Y, \nabla g(Y) \rangle].$$

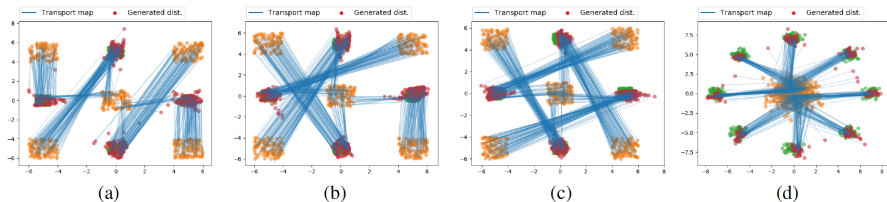


Figure: Numerical experiment

³A. Makkuva et. al, Optimal transport mapping via input convex neural networks, 2019

OT in the Spiked Transport Model

Spiked model:

$$\mu^{(1)} := \text{Law}(X^{(1)} + Z),$$

$$\mu^{(2)} := \text{Law}(X^{(2)} + Z)$$

$X^{(1)}, X^{(2)}$ support on $U \in \mathbb{R}^d$ with intrinsic dimension $k \ll d$.
Common part Z supports on U^\perp .

WPP (Wasserstein projection pursuit) estimator:

$$\widetilde{W}_{p,k}(\mu, \nu) = \max_{U \in V_k(\mathbb{R}^d)} W_p(\mu_U, \nu_U),$$

$$\widehat{W}_{p,k} = \widetilde{W}_{p,k}(\mu_n, \nu_n).$$

μ_U is

OT in the Spiked Transport Model

Good news: overcome CoD

Theorem

Let $(\mu^{(1)}, \mu^{(2)})$ satisfy the spiked transport model, for any $p \in [1, 2]$, if $(\mu^{(1)}, \mu^{(2)})$ satisfy the $T_p(\sigma^2)$ transport inequality, then the WPP estimator $\widehat{W}_{p,k}$ satisfy

$$\mathbb{E} \left| \widehat{W}_{p,k} - W_p(\mu^{(1)}, \mu^{(2)}) \right| \leq c_k \left(r_{p,k}(n) + \sqrt{\frac{d \log n}{n}} \right).$$

Bad news: There is not a polynomial time algorithm to calculate the WPP.

Projected Wasserstein distance

Other interesting topics

Multi-marginal OT and density functional theory (DFT)
OT and GMM