Statistical estimation of optimal transport distances and an extension of Gromov-Wasserstein distances

A. Vacher — LIGM, UGE and INRIA
B. Muzellec — INRIA Paris
A. Rudi — INRIA Paris
F. Bach — INRIA Paris
G. Peyré — ENS
T. Séjourné — ENS

Contents

Overall motivation

Optimal transport is

- Gaining interest in data science.
- Data distribution $\mathcal{P}$ accessible via samples $x_1, \ldots, x_n \in \mathbb{R}^d, d >> 1$.
- Typical situation: find a parametrized distribution $\mathcal{Q}_\theta$ close to $\mathcal{P}$.

Statement of the problem

Given samples $x_1, \ldots, x_n \sim \mathcal{P}$ and $y_1, \ldots, y_n \sim \mathcal{Q}$,
How to estimate efficiently $W_2(\mathcal{P}, \mathcal{Q})$?

An elementary Wasserstein estimation problem

Estimation of a shift

Consider $x_1, \ldots, x_n \sim \mathcal{N}(\mu, \mathrm{Id}_d)$ and $y_1, \ldots, y_n \sim \mathcal{N}(\mu + \delta, \mathrm{Id}_d)$.

- $\mathbb{E}[|\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i) - \delta|] \lesssim \sqrt{\frac{2d}{n}}$.

Kernel based distances

Reproducing Kernel Hilbert Spaces (RKHS)

Consider $H \subset \mathcal{F}(\Omega, \mathbb{R})$ Hilbert Space such that $H \hookrightarrow C^0(\Omega)$.

- $\delta_x \in H^*$.
- $\langle \delta_x, v \rangle = v(x) =: \langle k(x, \cdot), v \rangle_H$.

Dual norms (a.k.a. Maximum Mean Discrepancy (MMD))

- $\mathcal{M}_1(\Omega) \subset H^*$, $\|\mu\|_{H^*} = \sup_{\|f\|_H \leq 1} \langle f, \mu \rangle$.
- $\|\hat{\mu} - \mu\|_{H^*} \lesssim \sqrt{\frac{2|k|_\infty}{n}}$ where $\hat{\mu} := \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ independent of the dimension.

Why? $\|\mu\|_{H^*}^2 = \|k^{1/2}\mu\|_{L^2}^2$ and Monte-Carlo rate.

W1 optimal tranpsort

Recall that

$$W_1(\mu, \hat{\mu}) = \sup_{f s.t. \|\nabla f\|_\infty \leq 1} \langle f, \mu - \hat{\mu} \rangle \,. \tag{1}$$

Dudley, 1969

If $d > 2$, on a bounded domain for the support of $\mathcal{P}$,

$$\mathbb{E}[|W_1(\mathcal{P}_n, \mathcal{P})|] \lesssim O(n^{-1/d}) \,. \tag{2}$$

Sharp if $\mathcal{P}$ has density w.r.t. Lebesgue.

Compare with kernel norms! $n^{-1/2}$.

Goal: Define Est s.t. $\mathbb{E}[\mathrm{Est}(\mathcal{P}_n, \mathcal{Q}_n) - W_2^2(\mathcal{P}, \mathcal{Q})] \lesssim \frac{1}{\sqrt{n}} (\star)$.

Example: $\mathrm{Est}(\mathcal{P}_n, \mathcal{Q}_n) = W_2^2(\mathcal{P}_n, \mathcal{Q}_n) \implies O(n^{-1/d})$ in $O(n^3 \log(n))$.

Q: Can we design statistical and computational efficient estimators of high-dimensional W_2 in good cases?

Goal: Define Est s.t. $\mathbb{E}[\mathrm{Est}(\mathcal{P}_n, \mathcal{Q}_n) - W_2^2(\mathcal{P}, \mathcal{Q})] \lesssim \frac{1}{\sqrt{n}} (\star)$.

Example: $\mathrm{Est}(\mathcal{P}_n, \mathcal{Q}_n) = W_2^2(\mathcal{P}_n, \mathcal{Q}_n) \implies O(n^{-1/d})$ in $O(n^3 \log(n))$.

Q: Can we design statistical and computational efficient estimators of high-dimensional W_2 in good cases?

A: Yes, in the case of "smooth" W_2 using

Sum of Squares (SOS) approach on RKHS and sampling inequalities.

State of the art

- Entropic optimal transport (EOT) with λ regularization: $O(\frac{1}{\lambda^{\lfloor d/2 \rfloor}\sqrt{n}})$.
- (Chizat et al, 2020), Estimation of $(\star)$ via EOT: $O(\varepsilon^{-d/2+2})$ and $O(\varepsilon^{-(d'+5.5)})$ operations. Curse of dimension.
- (Hütter, Rigollet, 2019), Minimax rates of convergences for smooth OT. No computationally feasible algorithm.
- (Weed, Berthet, 2019), need $O(\varepsilon^{-\frac{d+2s}{1+s}})$ samples and $O(\varepsilon^{-(2d+d/2)})$ Computational time suffers from curse of dimensionality.

Smooth OT

Dual static formulation of OT:

$$\mathrm{OT}(\mu,\nu) = \sup_{u,v\in C(\mathbb{R}^d)} \int u(x)d\mu(x) + \int v(y)d\nu(y) \qquad (3)$$
$$\text{subject to} \quad c(x,y) \geq u(x) + v(y), \ \forall (x,y) \in X \times Y,$$

Theorem

Let X, Y be two bounded open subsets of $\mathbb{R}^d$, let c be the quadratic cost $c(x,y) = \frac{\|x-y\|^2}{2}$ and $k \geq 0$. If (μ,ν) admit densities $(\rho_\mu, \rho_\nu) \in \mathcal{C}^k(X) \times \mathcal{C}^k(Y)$, bounded away from zero and infinity, and Y is convex, then the optimal map $T = \nabla u$ sending μ onto ν is $\mathcal{C}^{k+1}$.

Actually, only need the optimal potentials are

$$(u_*, v_*) \in H^{s+2}(X) \times H^{s+2}(Y) \ \text{ where } s > d+1\,.$$

Contents

Leveraging smoothness

Sampling inequalities:

- $\Omega \subset \mathbb{R}^d$ with interior cone condition: include convex bounded sets.
- $X = \{x_1, \ldots, x_n\}$ the sampling set.
- Define *fill distance* $h = \sup_{y \in \Omega} \min_{x_i \in X} \|x_i - y\|_2$.

Then, it holds (Wendland, Rieger 2005)

$$\|f\|_{\infty(\Omega)} \leq C h^{s-d/2} \|f\|_{H^s(\Omega)} + 2|f|_{\infty(X)} \,. \tag{4}$$

if $h \leq \frac{cste(\Omega)}{\lfloor s \rfloor^2}$ and $s > d/2$.

Sample Ω: $x_1, \ldots, x_n$: $p < 1 - \delta$, if $n \geq n_0(R, d)$, then

$$h \leq C n^{-1/d} \left[\log \left(\frac{n}{\delta} \right) \right]^{2/d} . \tag{5}$$

Main issues to leverage smoothness in dual OT

- How to optimize on the set $\{(u,v)\,;\, c(x,y) - u(x) - v(y) \geq 0\}$, $\|u\|_{H^s}, \|v\|_{H^s} \leq M$?
- Subsampling the inequality: Control $\inf_D f$ if $f_X \geq 0$?
  $\rightarrow$ Only Lipschitz bound can be used.
- Imposing to work on Fenchel-Legendre pairs ?
  $\rightarrow$ Not feasible computationally

Solutions

Replace inequality by equality : represent nonnegative functions using sum of squares (SOS)

Sum of squares relaxation (Lasserre,...)

Optimizing on nonnegative polynomials

$$\min_P L(P) \text{ subject to} \tag{6}$$

$$A(P) = b \tag{7}$$

$$P(x) \geq 0 \text{ for } x \text{ s.t. } Q_i(x) \geq 0\,. \tag{8}$$

Include optimization of polynomials: $\min P(x)$.

Structural result: Positivestellensatz

$$\min_P L(P) \text{ subject to} \tag{9}$$

$$A(P) = b \tag{10}$$

$$P(x) = \sigma_0(x) + \sum_{i=1}^{d} \sigma_i(x) Q_i(x) \qquad \text{where } \sigma_i(x) = \sum_j q_j(x)^2\,. \tag{11}$$

SOS in RKHS

- Finding Global Minima via Kernel Approximations (Rudi, Marteau-Ferrey, Bach, 2020).

$$c(x,y) - u(x) - v(y) = \sum_{i=1}^{k} h_i(x,y)^2 \,. \tag{12}$$

SOS in RKHS

- Finding Global Minima via Kernel Approximations (Rudi, Marteau-Ferrey, Bach, 2020).

$$c(x,y) - u(x) - v(y) = \sum_{i=1}^{k} h_i(x,y)^2 \,. \tag{12}$$

Assume H RKHS with kernel k:

$$c(x,y) - u(x) - v(y) = \sum_{i=1}^{k} \langle h_i, k \rangle_H^2 = \langle k, Ak \rangle_H \,, \tag{13}$$

where A self-adjoint, finite rank: $A = \sum_{i=1}^{k} h_i \otimes h_i$.

Representation result for smooth OT

Theorem

Let $(u_\star, v_\star)$ be Kantorovich potentials such that $u_\star \in H^{s+2}(X)$ and $v_\star \in H^{s+2}(Y)$ for $s > d+1$. There exist functions $w_1, \ldots, w_d \in H^s(X \times Y)$ such that

$$\tfrac{1}{2}\|x-y\|^2 - u_\star(x) - v_\star(y) = \textstyle\sum_{i=1}^d w_i(x,y)^2, \quad \forall (x,y) \in X \times Y.$$

Representation result for smooth OT

Theorem

Let $(u_\star, v_\star)$ be Kantorovich potentials such that $u_\star \in H^{s+2}(X)$ and $v_\star \in H^{s+2}(Y)$ for $s > d+1$. There exist functions $w_1, \dots, w_d \in H^s(X \times Y)$ such that

$$\tfrac{1}{2}\|x-y\|^2 - u_\star(x) - v_\star(y) = \textstyle\sum_{i=1}^d w_i(x,y)^2, \quad \forall (x,y) \in X \times Y.$$

Proof.

Consider $f(x) = \frac{\|x\|^2}{2} - u_\star(x), f^\star(y) = \frac{\|y\|^2}{2} - v_\star(y)$,
$f(x) + f^\star(y) - \langle x, y\rangle = h(x,y) \geq 0\,.$
$\rightarrow$ Second order Taylor expansion on $h(x,y)$ with remainder at points $(x, T(x))$.

$$h(x,y) = \langle y - T(x), \int_0^1 (1-t)\nabla^2_{yy} h dt (y - T(x))\,. \tag{14}$$

Strong convexity of $f^\star$ + square root of $\nabla^2_{yy} h$. □

Soft-penalized OT-SOS formulation

"Continuous formulation"

$$\text{OT-SOS}(\mu,\nu) = \sup_{u,v,A} \int u(x)d\mu(x) + \int v(y)d\nu(y) - \lambda_1 \operatorname{tr}(A) - \lambda_2(\|u\|_H^2 + \|v\|_H^2) \quad (15)$$

such that $c - (u+v) = \langle k, Ak \rangle$.

"Sampled formulation"

$$\widehat{\text{OT-SOS}}(\hat{\mu},\hat{\nu}) = \sup_{u,v,A} \int u(x)d\hat{\mu}(x) + \int v(y)d\hat{\nu}(y) - \lambda_1 \operatorname{tr}(A) - \lambda_2(\|u\|_H^2 + \|v\|_H^2) \quad (16)$$

such that $c(x_k,y_k) - u(x_k) - v(y_k) = \langle k(x_k,y_k), Ak(x_k,y_k) \rangle$.

Approximation result

Theorem

- $\delta \in (0,1]$.
- $(\tilde{x}_j, \tilde{y}_j)$ $j \in [1,\ell]$ *uniform sampling on* $X \times Y$.

There exists $\ell_0(d,m)$ *and* $C_1, C_2(u_\star, v_\star)$ *s.t. if* $\ell \geq \ell_0$ *and if*

$$\lambda_1 \geq C_1 \ell^{-m/2d+1/2} \log \tfrac{\ell}{\delta}, \quad \lambda_2 \geq \|\mu - \hat{\mu}\|_{(H^s)^*} + \|\nu - \hat{\nu}\|_{(H^s)^*} + \lambda_1, \tag{17}$$

then, with probability $1-\delta$, *we have*

$$|\widehat{\mathrm{OT}}(\hat{\mu},\hat{\nu}) - \mathrm{OT}(\mu,\nu)| \leq C_2 \lambda_2 \,.$$

where

$$\widehat{\mathrm{OT}}(\hat{\mu},\hat{\nu}) = \int \hat{u}(x) d\hat{\mu}(x) + \int \hat{v}(y) d\hat{\nu}(y) \tag{18}$$

$\hat{u}, \hat{v}$ maximizers of $\widehat{\text{OT-SOS}}(\hat{\mu},\hat{\nu})$.

Reduction to an SDP problem

- $\mathbf{Q}_{i,j} = k_X(\tilde{x}_i, \tilde{x}_j) + k_Y(\tilde{y}_i, \tilde{y}_j)$
- $z_j = \hat{w}_\mu(\tilde{x}_j) + \hat{w}_\nu(\tilde{y}_j) - \lambda_2 c(\tilde{x}_j, \tilde{y}_j)$
- $q^2 = \|\hat{\mu}\|^2_{(H^s)_*} + \|\hat{\nu}\|^2_{(H^s)^*}$
- $\mathbf{K}_{i,j} = k_{XY}(\tilde{x}_i, \tilde{y}_i, \tilde{x}_j, \tilde{y}_j)$
- $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^\top$ (Cholesky).

The dual problem writes:

$$\begin{aligned} \min_{\gamma \in \mathbb{R}^\ell} \ & \tfrac{1}{4\lambda_2}\gamma^\top \mathbf{Q}\gamma - \tfrac{1}{2\lambda_2}\textstyle\sum_{j=1}^\ell \gamma_j z_j + \tfrac{q^2}{4\lambda_2} \\ \text{such that} \quad & \textstyle\sum_{j=1}^\ell \gamma_j \Phi_j \Phi_j^\top + \lambda_1 \operatorname{Id}_\ell \succeq 0. \end{aligned} \tag{19}$$

$$\widehat{\mathrm{OT}} = \frac{q^2}{2\lambda_2} - \frac{1}{2\lambda_2}\sum_{j=1}^\ell \hat{\gamma}_j(\hat{w}_\mu(\tilde{x}_j) + \hat{w}_\nu(\tilde{y}_j)) \tag{20}$$

Computational complexity

Solving the SDP formulation: IPM

$$O\big(C + E\ell + \ell^{3.5}\log\tfrac{\ell}{\varepsilon}\big) \text{ time}, \qquad O(\ell^2) \text{ memory}, \tag{21}$$

where C is the cost for computing q^2 and E is the cost to compute one z_j.

Theorem

The cost to achieve $|\widehat{\mathrm{OT}} - \mathrm{OT}(\mu,\nu)| \le \varepsilon$*:*

1. *Time:* $\tilde{O}(\varepsilon^{-\max(4,\frac{7d}{m-d})})$.
2. *Space:* $\tilde{O}(\varepsilon^{-\frac{4d}{m-d}})$. *#samples of* μ,ν*:* $\tilde{O}(\varepsilon^{-2})$.

Proof.

$\varepsilon^{-2} = n$, $\varepsilon = \frac{1}{\sqrt{n}}$.

$$\begin{aligned}\tilde{O}(C + E\ell + \ell^{3.5}) &= \tilde{O}(n_\mu^2 + n_\nu^2 + (n_\mu + n_\nu)\ell + \ell^{3.5}) \\ &= \tilde{O}(\varepsilon^{-4} + \varepsilon^{-2-2d/(m-d)} + \varepsilon^{-7d/(m-d)}) = \tilde{O}(\varepsilon^{-\max(4,7d/(m-d))}).\end{aligned}$$

Contents

Summary

- Leverage smoothness via sampling inequalities.
- Remove inequality constraint with equality (SOS).
- Need structural result on the optimum.
- Reduction to SDP formulation.

No free lunch: curse of dimension is in the constants.

What's more ?

- Extension to mixture of gaussians: $\tilde{O}(\varepsilon^{-\frac{7d}{m-d}})$. Space: $\tilde{O}(\varepsilon^{-\frac{4d}{m-d}})$.
- Access to densities: Time, $\tilde{O}(\varepsilon^{-\frac{7d}{m-d}})$. Space: $\tilde{O}(\varepsilon^{-\frac{4d}{m-d}})$. #evaluations of μ, ν: $\tilde{O}(\varepsilon^{-\frac{d}{m+1}})$.
- To come: Almost (statistical) optimal rates for estimation of potentials.
- Open: Efficient implementations/approximations?

More details: "A Dimension-free Computational Upper-bound for Smooth Optimal Transport Estimation". `https://arxiv.org/abs/2101.05380`

What can be said on potentials?

Minimax rate of estimation: Hütter and Rigollet.

Up to log terms (for the upper bound):

$$\mathbb{E}(\|\nabla f_* - \nabla f_n\|^2_{L^2(\mu)}) \sim n^{-\frac{m+1}{m+d/2}} . \tag{22}$$

m is smoothness of data.
Estimator computationally not feasible.

What can be achieved in our framework?
Note that $m = \infty \implies \frac{1}{n}$, we have only estimates in $\frac{1}{\sqrt{n}}$ on W_2.

Stability estimates

Question:

How good performance on the cost gives good performances on the potentials?

Recall OT convex optimization problem: Given $\mu, \nu \in \mathcal{P}_1$,

$$\min_{\gamma \in \mathcal{C}(\mu,\nu)} \langle \gamma(x,y), c(x,y) \rangle = \sup_{f,g \in C(X) \times C(Y)} \{ \langle f, \mu \rangle + \langle g, \nu \rangle \mid f(x) + g(y) \leq c(x,y) \} . \quad (23)$$

Estimate $\|f_\star - f_n\|_{\dot{H}_1} + \|g_\star - g_n\|_{\dot{H}_1}$ with $\mathrm{Dual}(f_\star, g_\star) - \mathrm{Dual}(f_n, g_n)$?

The semi-dual functional: case $c(x, y) = \frac{1}{2}\|x - y\|^2$

Definition (Legendre transform)

Let $f : X \mapsto \mathbb{R}$ and $c(x, y)$ continuous, define

$$f^*(y) = \inf_x y \cdot x - f(x) \,. \tag{24}$$

$$\inf_f J_{\mu,\nu}(f) = \langle f, \mu \rangle + \langle f^*, \nu \rangle \,. \tag{25}$$

Retains more convexity.

Strong convexity estimates

Proposition

Consider potentials γ-strongly convex C^1 with M-Lipschitz gradient

$$\frac{1}{2M}\|\nabla f - \nabla f_\star\|^2_{L^2(\mu)} \leq (J(f) - J(f_\star)) \leq \frac{1}{2\gamma}\|\nabla f - \nabla f_\star\|^2_{L^2(\mu)}\,. \tag{26}$$

Reaching the $\frac{1}{n}$ estimation.

$$\text{We have: } |J_{\mu_n,\nu_n}(f_n) - J_{\mu,\nu}(f_\star)| \sim \frac{1}{\sqrt{n}} \tag{27}$$

$$\forall\, f \text{ smooth } |J_{\mu_n,\nu_n}(f) - J_{\mu,\nu}(f)| \sim \frac{1}{\sqrt{n}}\,. \tag{28}$$

Thus

$$|J_{\mu,\nu}(f_n) - J_{\mu,\nu}(f_\star)| \sim \frac{1}{\sqrt{n}}\,. \tag{29}$$

Reaching the $\frac{1}{n}$ estimation.

$$\text{We have: } |J_{\mu_n,\nu_n}(f_n) - J_{\mu,\nu}(f_\star)| \sim \frac{1}{\sqrt{n}} \tag{27}$$

$$\forall\, f \text{ smooth } |J_{\mu_n,\nu_n}(f) - J_{\mu,\nu}(f)| \sim \frac{1}{\sqrt{n}}\,. \tag{28}$$

Thus

$$|J_{\mu,\nu}(f_n) - J_{\mu,\nu}(f_\star)| \sim \frac{1}{\sqrt{n}}\,. \tag{29}$$

- Suppose $J_{\mu_n,\nu_n}(f_n)$ realizes the minimum then $J_{\mu_n,\nu_n}(f_n) \leq J_{\mu_n,\nu_n}(f_\star)$
- And $J_{\mu,\nu}(f_\star) \leq J_{\mu,\nu}(f_n)$.

$$\begin{aligned} 0 \leq J_{\mu,\nu}(f_n) - J_{\mu,\nu}(f_\star) &\leq J_{\mu,\nu}(f_n) - J_{\mu,\nu}(f_\star) + J_{\mu_n,\nu_n}(f_\star) - J_{\mu_n,\nu_n}(f_n) \\ &= \langle \mu - \mu_n, f_\star - f_n \rangle + \langle \nu - \nu_n, f_\star^* - f_n^* \rangle \sim \frac{1}{n}\,. \end{aligned} \tag{30}$$

+ regularity assumptions (Gagliardo-Nirenberg inequality).

Main result

Theorem

Let $\delta, \varepsilon \in]0,1[^2$ and

$$\lambda_n^1 = \lambda_n^2 = \lambda_n = \left(\frac{\log(\frac{2}{\delta})}{n}\right)^{\frac{m+1}{m+d/2+\varepsilon}} + C_1 \left(\frac{\log(\frac{n}{\delta})}{n}\right)^{\frac{m-d}{2d}}, \tag{31}$$

where C_1 is a constant that does not depend on n and δ. With probability $\geq 1-\delta$ for $n \geq n_0(X, Y, d, m)$,

$$\|\nabla \hat{f}_n - \nabla f_*\|^2_{L^2(\mu)} + \|\nabla \hat{g}_n - \nabla g_*\|^2_{L^2(\nu)} \leq C_2 \lambda_n\,, \tag{32}$$

where C_2 independent from n and δ (but ∞ when $\varepsilon = 0$). The minimax rate is nearly attained:

$$\|\nabla \hat{f}_n - \nabla f_*\|^2_{L^2(\mu)} + \|\nabla \hat{g}_n - \nabla g_*\|^2_{L^2(\nu)} \leq C_2 \left(\frac{\log(\frac{2}{\delta})}{n}\right)^{\frac{m+1}{m+d/2+\varepsilon}}. \tag{33}$$

Contents

Unbalanced optimal transport

Optimal transport applications: Imaging, machine learning, gradient flows, ...

Bottleneck in optimal transport: data has fixed total mass.

- Relax the mass constraint to extend OT distance between positive measures of arbitrary mass.
- Develop associated numerical algorithms.

Unbalanced optimal transport

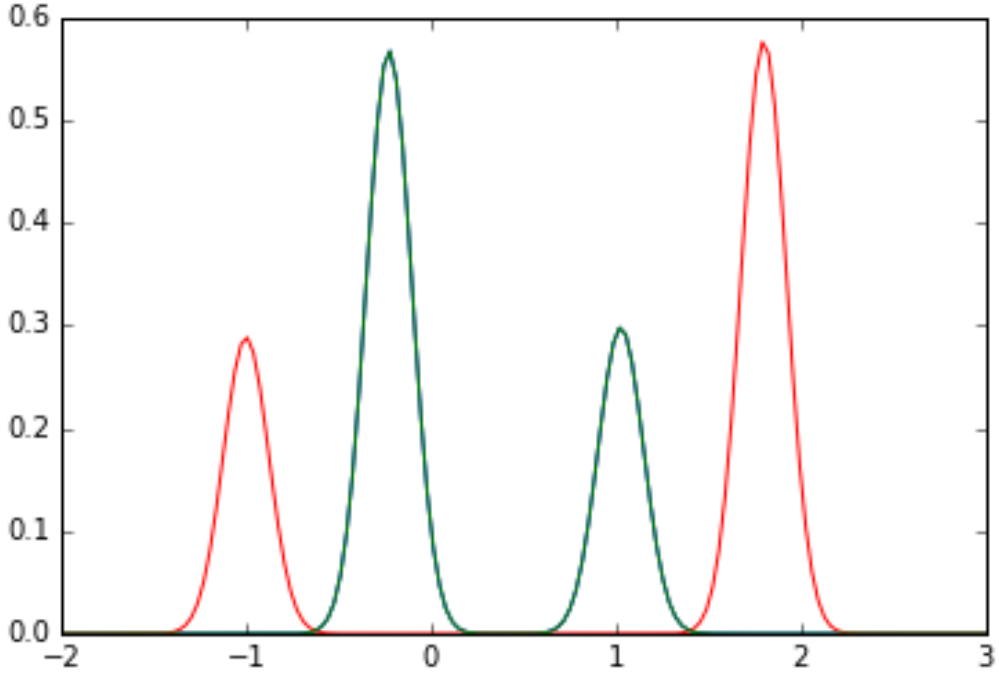


Figure: Optimal transport between bimodal densities

Unbalanced optimal transport

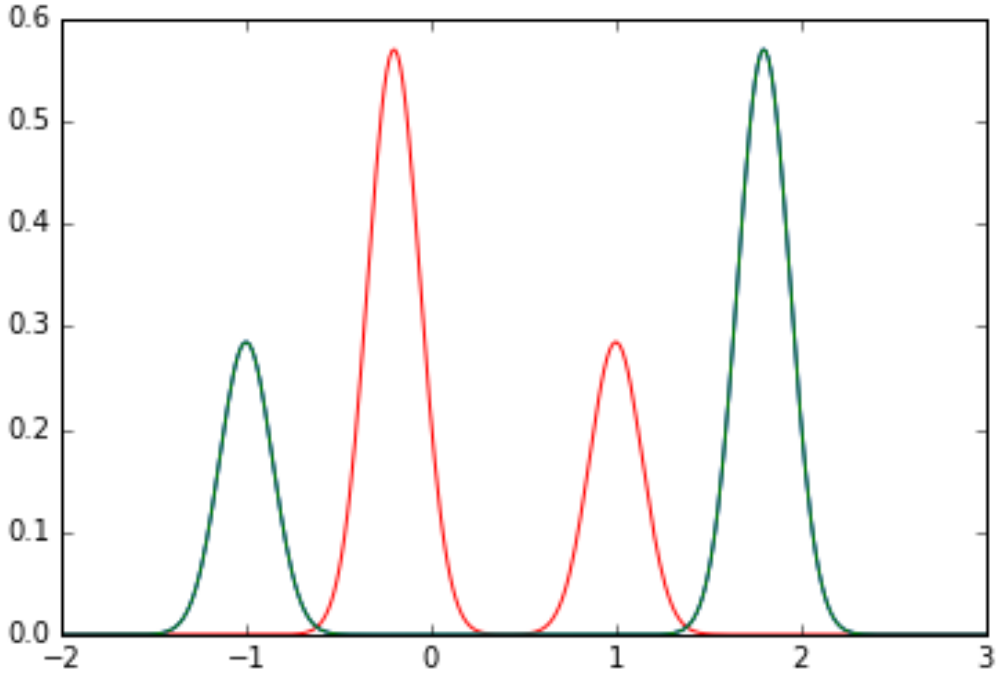


Figure: Another transformation

Two possible directions

Pros and cons:

- Extend static formulation:

$$\min \lambda KL(\mathrm{Proj}^1_* \gamma, \rho_1) + \lambda KL(\mathrm{Proj}^2_* \gamma, \rho_2) + \int_{M^2} \gamma(x,y) d(x,y)^2 \,\mathrm{d}x \,\mathrm{d}y \quad (34)$$

  Good for numerics, but is it a distance ?

- Extend dynamic formulation: on the tangent space of a density, choose a metric on the transverse direction.
  Built-in metric property but does there exist a static formulation ?

An extension of Benamou-Brenier formulation

Add a source term in the constraint: (weak sense)

$$\dot{\rho} = -\nabla \cdot (\rho v) + \alpha\rho \,,$$

where α can be understood as the growth rate.

$$WF^2(\mu,\nu) := \inf_{(v,\alpha)} \frac{1}{2} \int_0^1 \int_M |v(x,t)|^2 \rho(x,t) \, \mathrm{d}x \, \mathrm{d}t + \frac{\delta^2}{2} \int_0^1 \int_M \alpha(x,t)^2 \rho(x,t) \, \mathrm{d}x \, \mathrm{d}t \,.$$

where δ is a length parameter.

A relaxed static OT formulation

Define

$$KL(\gamma, \nu) = \int \frac{\mathrm{d}\gamma}{\mathrm{d}\nu} \log\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\nu}\right) \mathrm{d}\nu + |\nu| - |\gamma|$$

$$WF^2(\rho_1, \rho_2) = \inf_{\gamma} KL(\mathrm{Proj}^1_* \gamma, \rho_1) + KL(\mathrm{Proj}^2_* \gamma, \rho_2) - \int_{M^2} \gamma(x, y) \log(\cos^2(\min(d(x, y)/2, \delta\pi/2))) \, \mathrm{d}x \, \mathrm{d}y$$

Theorem

On a Riemannian manifold (compact without boundary) or Euclidean space, the static and dynamic formulations are equal.

Illustration

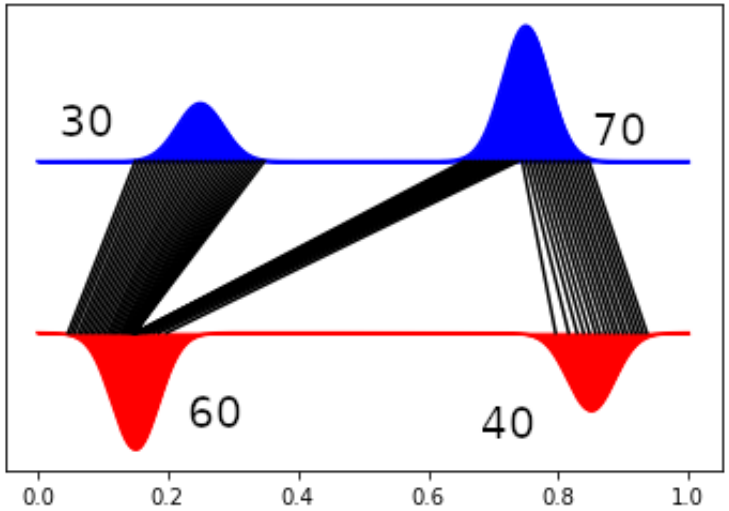


Figure: Standard OT

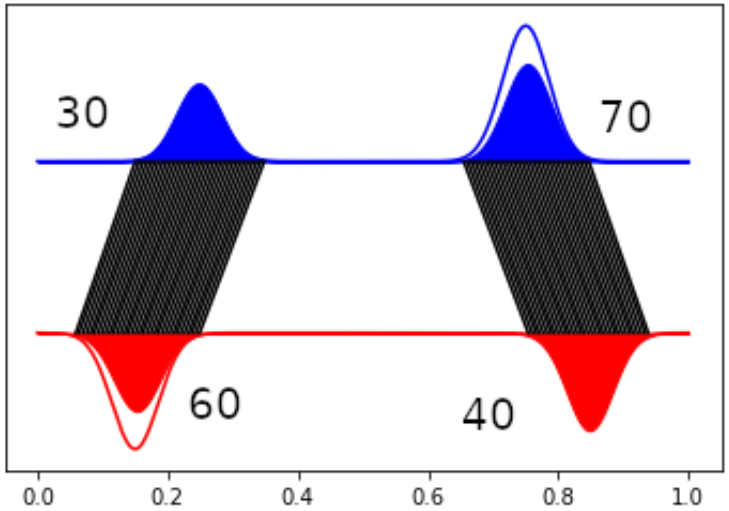


Figure: Wasserstein-Fisher-Rao

Yet another equivalent formulation

Liero, Mielke, Savaré.

WFR is Wasserstein 2 on $\mathcal{P}(M \times \mathbb{R}_+)$ with second moment constraint.

$$WFR(\mu, \nu) = \min_{\tilde{\mu}, \tilde{\nu}} W_2(\tilde{\mu}, \tilde{\nu}) \tag{35}$$

with the constraint:

$$\int_{\mathbb{R}_+} r^2 \, \mathrm{d}\tilde{\mu}(x, r) = \mu(x) \,, \tag{36}$$

and

$$\int_{\mathbb{R}_+} r^2 \, \mathrm{d}\tilde{\nu}(x, r) = \nu(x) \,, \tag{37}$$

Here $r^2 = m$ and on $\mathcal{C}(M \times \mathbb{R}_+)$ cone metric:

$$d((x_1, r_1), (x_2, r_2))^2 \;=\; r_2^2 + r_1^2 - 2 r_1 r_2 r \cos\left(\frac{1}{2} d_M(x_1, x_2) \wedge \pi\right) \,. \tag{38}$$

Introduction to Gromov-Wasserstein

Comparing metric-measure spaces

Let $\mathcal{X} = \{(X, d, \mu), (X, d)$ polish space, μ probability measure$\}$.
How to compare such spaces ?
Isometric mm-spaces: $\phi : X \mapsto Y$, $\phi_*(\mu) = \nu$ and ϕ isometric: $\phi^* d_Y = d_X$.

- Developed by Gromov, Memoli, Sturm.
- Applications in and outside mathematics: graph matching, quantum chemistry, NLP.

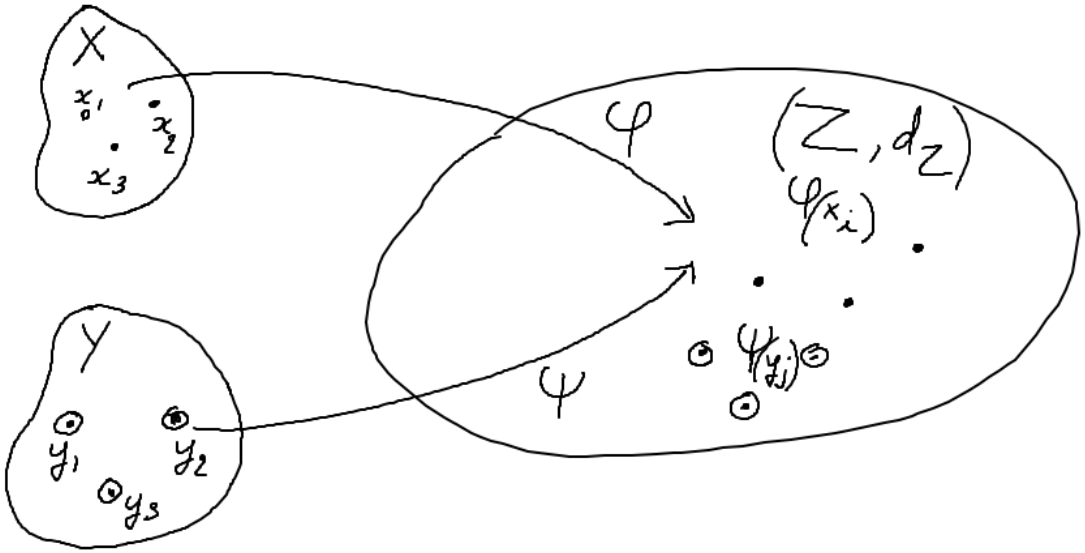
Two different distances

D^2 distance: infimum on the set of embeddings, Sturm, 2006

$$D^2(X,Y) := \inf_{\psi,\phi}\{\inf_{\pi}\langle \pi, d_Z^2(\psi(x),\phi(y))\rangle\,;\,(\psi,\phi):(X,Y)\mapsto Z \text{ and } \pi\in C_{\mu_X,\mu_Y}\}\,,$$

ψ,ϕ being isometric embeddings.

1. Reformulation on minimising a coupling pseudo-metric on $X\times Y$.
2. Non-convex optimization problem.

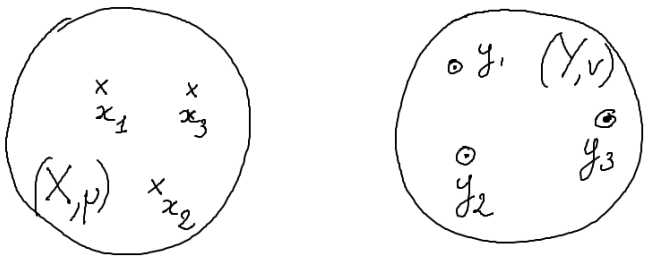


Two different distances

GW^2 distance: comparison of pairwise distances (distortion distances)

$$\mathrm{GW}^2(X,Y) := \inf_{\pi}\{\langle \pi(x,y)\otimes\pi(x',y'), |d_X(x,x')-d_Y(y,y')|^2\rangle\,;\, \pi\in C_{\mu_X,\mu_Y}\}\,.$$

1. Non-convex optimization problem.
2. When X, Y are Euclidean spaces $\implies$ concave optimization problem.



$$(X\times X, \mu\otimes\mu) \qquad (Y\times Y, \nu\otimes\nu)$$

$$\inf_{\tilde{\pi}\in C(\mu\otimes\mu,\nu\otimes\nu)} \langle \tilde{\pi}, |d_X - d_Y|^2\rangle$$

with <u>constraint</u>: $\tilde{\pi} = \pi\otimes\pi$, $\pi\in C(\mu,\nu)$

Properties of D and GW

1. Same topology (on compact spaces with uniformly bounded diameters).
2. D gives complete metric space, not GW.
3. Both are length spaces. E.g. $(X \times Y, (td_Y^2 + (1-t)d_X^2)^{1/2}, \pi)$ for GW.
4. GW has non-negative Alexandrov curvature.

How to solve GW numerically

Developing the squares, problem equivalent to

$$\mathrm{GW}^2(X,Y) = \inf_{\pi}\{-\langle \pi(x,x') \otimes \pi(y,y'), d_X(x,x')d_Y(y,y')\rangle \,;\, \pi \in C_{\mu_X,\mu_Y}\}\,.$$

If X, Y are Euclidean, $-\|x-x'\|\|y-y'\|$ is a <u>negative kernel on C_{μ_X,μ_Y}</u>.

$\implies$ concave *minimization* problem.

Proposition

Konno's result: following relaxation is tight

$$\mathrm{GW}^2(X,Y) = \inf_{\pi,\gamma}\{-\langle \pi \otimes \gamma, d_X(x,x')d_Y(y,y')\rangle \,;\, \pi,\gamma \in C_{\mu_X,\mu_Y}\} \quad (39)$$

Linear w.r.t. to each variable $\implies$ alternate minimization.

In practice, add entropic regularization.

$$\mathrm{GW}^2_\varepsilon(X,Y) = \inf_{\pi,\gamma}\{-\langle \pi \otimes \gamma, d_X(x,x')d_Y(y,y')\rangle + \varepsilon\,\mathrm{KL}(\pi, \mu\otimes\nu) + \varepsilon\,\mathrm{KL}(\gamma, \mu\otimes\nu) \,;\, \pi,\gamma \in C_{\mu_X,\mu_Y}\} \quad (40)$$

Contents

Two possible directions again

Consider mm-spaces: (X, d, μ) with $\mu \in \mathcal{M}_+(X)$.

1. Extend D.
2. Extend GW.

Simple method: use UOT instead of OT where it appears.

We are interested in the extension of GW since the optimization problem seems nicer for numerics.

Two possible directions again

Consider mm-spaces: (X, d, μ) with $\mu \in \mathcal{M}_+(X)$.

1. Extend D. De Ponti, Mondino, 2020.
2. Extend GW. Séjourné, Peyré, Vialard, 2020.

Simple method: use UOT instead of OT where it appears.

We are interested in the extension of GW since the optimization problem seems nicer for numerics.

Don't mess up with homogeneity

First idea that doesn't work

$$\langle |d(x,x') - d(y,y')|^2, \pi \otimes \pi \rangle + \mathrm{KL}(\pi_1, \mu) + \mathrm{KL}(\pi_2, \nu)\,. \tag{41}$$

Don't mess up with homogeneity

First idea that doesn't work

$$\langle |d(x,x') - d(y,y')|^2, \pi \otimes \pi \rangle + \mathrm{KL}(\pi_1, \mu) + \mathrm{KL}(\pi_2, \nu)\,. \tag{41}$$

Better proposal: see it as optimal transport in $\pi \otimes \pi$

$$UGW := \langle \Gamma(|d(x,x') - d(y,y')|), \pi \otimes \pi \rangle + \mathrm{KL}([\pi \otimes \pi]_1, \mu \otimes \mu) + \mathrm{KL}([\pi \otimes \pi]_2, \nu \otimes \nu)\,. \tag{42}$$

1. Good for numerics (similar alternating minimization schemes).
2. Not a distance (as shown by numerics).

The conic distance between unbalanced mm spaces

$$\mathrm{GW}^2(X,Y) := \inf_{\pi} \langle \pi \otimes \pi, d^2((d_X(x,x'), rr'), (d_Y(y,y'), ss')) \rangle \quad \text{s.t.} \int_{r,s} r^2 s^2 \pi((x,r),(y,s)) \in C_{\mu_X,\mu_Y}\,.$$

with $\pi([x,r],[y,s])$ and d the distance on the cone over $\mathbb{R}_+$.

The conic distance between unbalanced mm spaces

$$\mathrm{GW}^2(X,Y) := \inf_{\pi} \langle \pi \otimes \pi, d^2((d_X(x,x'), rr'), (d_Y(y,y'), ss')) \rangle \quad \text{s.t.} \int_{r,s} r^2 s^2 \pi((x,r),(y,s)) \in C_{\mu_X,\mu_Y} .$$

with $\pi([x,r],[y,s])$ and d the distance on the cone over $\mathbb{R}_+$.

- Invariant with respect to dilations: let $v : ([x,r],[y,s]) \mapsto \mathbb{R}_+$, then define $h_v([x,r],[y,s]) = ([x,r/v],[y,s/v])$ and

$$\mathrm{Dil}_v : \mathcal{M}_+(\mathcal{C}(X),\mathcal{C}(Y)) \to \mathcal{M}_+(\mathcal{C}(X),\mathcal{C}(Y)) \tag{43}$$

$$\alpha \mapsto [h_v]_\sharp(v^p \alpha) , \tag{44}$$

- Key point for proving triangle inequality.

Main result:

Theorem

CGW is a distance.

$$UGW(X,Y) = 0 \Leftrightarrow X \text{ and } Y \text{ isomorphic.} \tag{45}$$

$$UGW \leq CGW \tag{46}$$

Proof.

Lift the optimal plan of UGW to the cone to obtain a competitor in CGW. □

Illustration

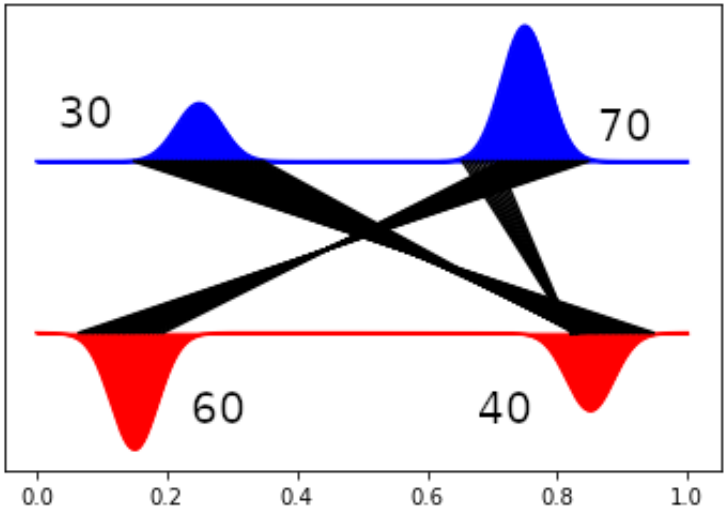


Figure: Standard GW

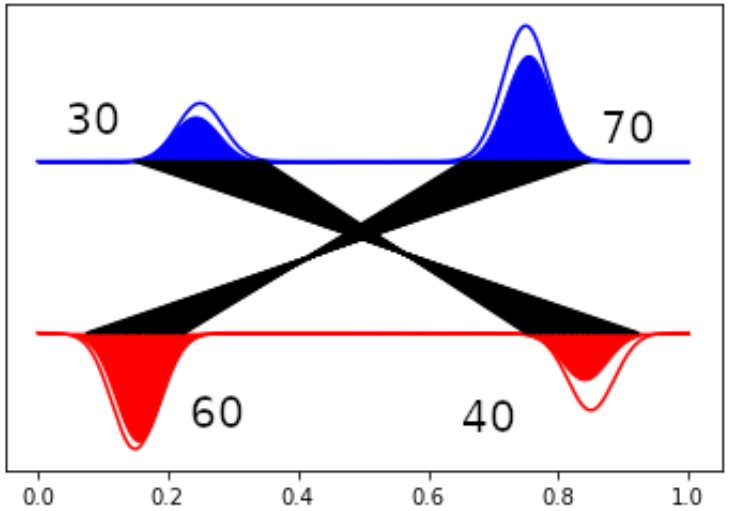


Figure: Unbalanced GW

Perspectives

1. Applications: "Gromov-Wasserstein optimal transport to align single-cell multi-omics data", (Demetci et al.).

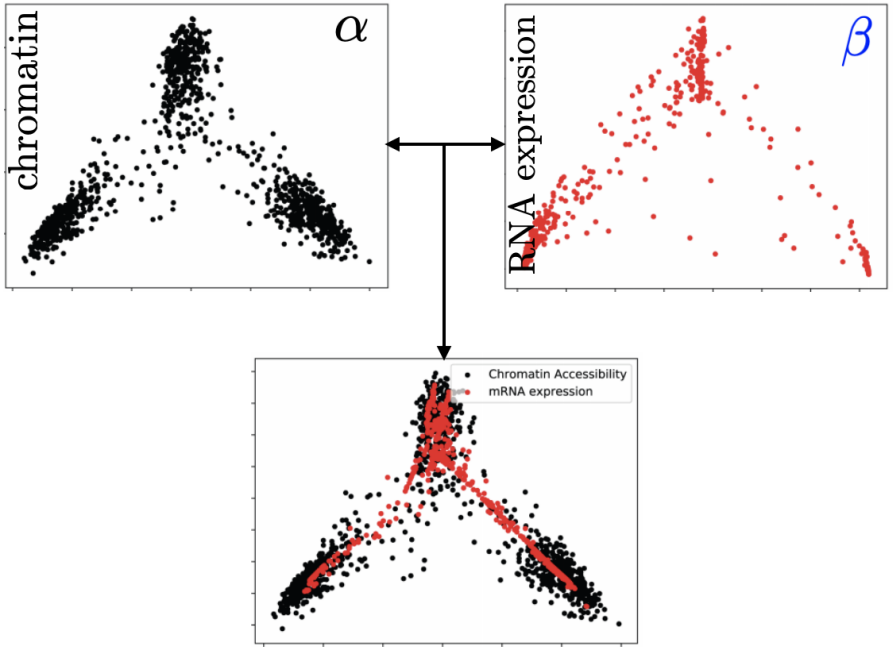


Figure: A

2. Metric and topological properties of CGW.
3. Statistical estimations of GW and UGW.
4. Mitigating the bias of entropy regularization.