Breaking the Curse of Dimension in Smooth Optimal Transport Estimation

Adrien Vacher
LIGM, UGE and INRIA

Boris Muzellec
INRIA Paris

Alessandro Rudi
INRIA Paris

Francis Bach
INRIA Paris

François-Xavier Vialard
LIGM, UGE

Overall motivation

Optimal transport is

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

Statement of the problem

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

An elementary Wasserstein estimation problem

Estimation of a shift

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

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

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

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

Dudley, 1969

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

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

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

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

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

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

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

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

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

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

$$\min_P L(P) \text{ subject to} \tag{9}$$
$$A(P) = b \tag{10}$$
$$P(x) = \sigma_0(x) + \sum_{i=1}^{d} \sigma_i(x) g_i(x) \qquad \text{where } \sigma_i(x) = \sum_j q_j(x)^2\,. \tag{11}$$

SOS in RKHS

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

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

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*

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

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*

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

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 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

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

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

Theorem

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

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

Proof.

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

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

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.