Sample Complexity for Regularized Optimal Transport

Valeria Ambrosio, Bjarne Bergh, Tobias Freidling and Lauritz Streck

Supervised by Marcello Carioni and Carola-Bibiane Schönlieb

DPMMS and DAMPT, University of Cambridge

March 30, 2021

Optimal Transport

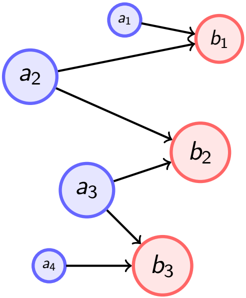


Figure: Discrete Optimal Transport

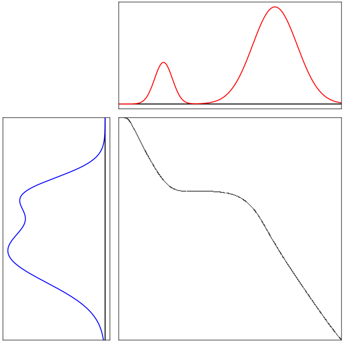


Figure: Continuous Optimal Transport

Optimal Transport

Definition

Let α and β be probability measures on the metric spaces $\mathcal{X}$ and $\mathcal{Y}$. The space of couplings is given by

$$\begin{aligned}\Pi(\alpha, \beta) := \{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})\colon \pi(A \times \mathcal{Y}) = \alpha(A),\\ \pi(\mathcal{X} \times B) = \beta(B) \ \ \forall A, B \subset \mathcal{X}, \mathcal{Y}\}.\end{aligned}$$

Let $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a cost function. Then the *optimal transport problem* is

$$\mathrm{OT}(\alpha, \beta) = \min_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, \mathrm{d}\pi(x, y).$$

Optimal Transport

Definition

Let α and β be probability measures on the metric spaces $\mathcal{X}$ and $\mathcal{Y}$. The space of couplings is given by

$$\begin{aligned}\Pi(\alpha, \beta) := \{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})\colon \pi(A \times \mathcal{Y}) &= \alpha(A), \\ \pi(\mathcal{X} \times B) &= \beta(B) \ \ \forall A, B \subset \mathcal{X}, \mathcal{Y}\}.\end{aligned}$$

Let $c\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a cost function. Then the *optimal transport problem* is

$$\mathrm{OT}(\alpha, \beta) = \min_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, \mathrm{d}\pi(x, y).$$

For the dual formulation

$$\mathrm{D}(\alpha, \beta) = \sup_{\substack{u \in C(\mathcal{X}), v \in C(\mathcal{Y}) \\ u + v \leq c}} \int_{\mathcal{X}} u \, \mathrm{d}\alpha + \int_{\mathcal{Y}} v \, \mathrm{d}\beta$$

we have $\mathrm{D}(\alpha, \beta) = \mathrm{OT}(\alpha, \beta)$.

Applications

If $\mathcal{X} = \mathcal{Y}$ is endowed with distance d, setting $c(x, y) = d(x, y)^p, p \geq 1$, yields the *p-Wasserstein distance*: $W_p(\alpha, \beta) = \left(\mathrm{OT}(\alpha, \beta)\right)^{1/p}$.

Applications

If $\mathcal{X} = \mathcal{Y}$ is endowed with distance d, setting $c(x, y) = d(x, y)^p, p \geq 1$, yields the *p-Wasserstein distance*: $W_p(\alpha, \beta) = \left(\mathrm{OT}(\alpha, \beta)\right)^{1/p}$.

- Image comparison (Earth Mover's Distance)
- Pattern recognition
- Domain adaptation (transfer learning)
- Wasserstein-GAN
- Text retrieval (word embeddings)
- etc.

Regularized Optimal Transport

Definition

Let α and β be probability measures on $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$ bounded, let $c \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a bounded cost function and $\varepsilon > 0$. Assume that $\Phi \colon [0, \infty] \to [0, \infty]$ is convex. The *optimal transport problem with convex regularization* is given by

$$\mathrm{OT}_\varepsilon(\alpha, \beta) = \inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y)\, \mathrm{d}\pi(x, y) + \varepsilon\, G(\pi|\alpha \otimes \beta),$$

where G is defined as

$$G(\pi|\alpha \otimes \beta) = \begin{cases} \int_{\mathcal{X} \times \mathcal{Y}} \Phi\left(\frac{\mathrm{d}\pi}{\mathrm{d}(\alpha \otimes \beta)}\right) \mathrm{d}(\alpha \otimes \beta), & \text{if } \pi \ll \alpha \otimes \beta, \\ +\infty, & \text{otherwise.} \end{cases}$$

Regularization functions

- Shannon entropy: $\Phi(z) = z(\log(z) - 1)$. This yields $G(\pi|\rho_1 \otimes \rho_2) = \mathrm{KL}(\pi \parallel \rho_1 \otimes \rho_2))$

- Tsallis entropy: For $q > 1$,
$$\Phi(z) = \frac{1}{q-1} z^q.$$

- Quadratic regularization: $\Phi(z) = \frac{1}{2} z^2$

Dual formulation

Theorem (Di Marino-Gerolin '20)

The dual formulation of the regularized optimal transport problem relative to the probability measures α, β, bounded cost function c, and entropy function Φ, is given by

$$D_\varepsilon(\alpha,\beta) = \sup_{\substack{u\in\mathcal{C}(\mathcal{X}),\\ v\in\mathcal{C}(\mathcal{Y})}} \left\{ \int_{\mathcal{X}} u d\alpha + \int_{\mathcal{Y}} v d\beta \right.$$
$$\left. - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \Psi\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right) d(\alpha\otimes\beta) \right\},$$

where Ψ is the Legendre conjugate of Φ $(\Psi' = (\Phi')^{-1})$. It holds that the supremum is achieved by some (u^*, v^*),

$$OT_\varepsilon(\alpha,\beta) = D_\varepsilon(\alpha,\beta).$$

(And one can recover π^* from (u^*, v^*)).

Sinkhorn's algorithm

- $D_\varepsilon(\alpha, \beta)$ is a concave optimization problem.
- We can use a fixed point iteration algorithm to solve it, called Sinkhorn's algorithm.

  It starts from fixed v^0 and alternatively updates u^k or v^k as

  $$u^k := \mathrm{argmax}_u\, F_\varepsilon(u, v^{k-1}),$$

  $$v^k := \mathrm{argmax}_v\, F_\varepsilon(u^k, v).$$

  Then u^k and v^k converge to a pair of maximizers (u^*, v^*) for the dual problem.

Sample complexity

In practice:

Let $(X_1, \ldots, X_n) \sim \alpha^n$, $(Y_1, \ldots, Y_n) \sim \beta^n$, and

$$\hat{\alpha}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}, \quad \hat{\beta}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i},$$

be the empirical measures. Then

$$\mathrm{OT}_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n) = \max_{u,v} \left\{ \sum_{i=1}^{n} u(X_i) + \sum_{i=1}^{n} v(Y_i) \right. \\ \left. - \varepsilon \sum_{i=1}^{n} \Psi\left(\frac{u(X_i) + v(Y_i) - c(X_i, Y_i)}{\varepsilon} \right) \right\}.$$

Sample complexity

Our goal is to estimate the *sample complexity*:

$$\mathbb{E}\left[\left|\mathsf{OT}_\varepsilon(\alpha,\beta) - \mathsf{OT}_\varepsilon(\hat{\alpha}_n,\hat{\beta}_n)\right|\right].$$

Sample complexity

Our goal is to estimate the *sample complexity*:

$$\mathbb{E}\left[\left|\mathrm{OT}_\varepsilon(\alpha,\beta) - \mathrm{OT}_\varepsilon(\hat{\alpha}_n,\hat{\beta}_n)\right|\right].$$

It is well known that

$$\begin{aligned} \mathrm{OT} \quad & (\varepsilon = 0) \text{ has sample complexity } \mathcal{O}(n^{-1/d}), \\ \mathrm{MMD} \quad & (\varepsilon = \infty) \text{ has sample complexity } \mathcal{O}(1/\sqrt{n}). \end{aligned}$$

Genevay et al. (2019) showed that

OT_ε with entropic regularization has sample complexity $\mathcal{O}(1/\sqrt{n})$ (with constants depending on ε).

We generalize this to a wider class of regularizations.

Sample Complexity for the Classical Entropy

Theorem (Genevay et al. '19)

Let α and β be probability measures with support on bounded $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$. Let as before X_i be iid distributed wrt α, Y_i wrt β. Let c be a $\mathcal{C}^\infty$ cost function. One has for regularization $\Psi = \exp$,

$$\mathbb{E}|OT_\varepsilon(\alpha, \beta) - OT_\varepsilon(\alpha_n, \beta_n)| = O\left(\frac{\left(1 + \varepsilon^{-\lfloor d/2 \rfloor}\right) \exp\left(O\left(\varepsilon^{-1}\right)\right)}{\sqrt{n}}\right)$$

where the constants only depend on $\mathcal{X}, \mathcal{Y}$ and $\|c^{(s)}\|_\infty$ for $s = 0 \dots \lfloor d/2 \rfloor$.

Sample Complexity for General Regularization Function

Theorem

Same setting, but this time with regularization with Legendre conjugate $\Psi : [0, \infty] \to [0, \infty]$ *convex such that*

$$C_\varepsilon = \frac{\max_{s \leq \frac{d}{2}} \sup_{z \in \frac{1}{\varepsilon} K} |\psi^{(s)}(z)|}{\inf_{z \in \frac{1}{\varepsilon} K} |\Psi''(z)|} < \infty,$$

where the compact set $K \subset \mathbb{R}$ *depends only on* $\mathcal{X}, \mathcal{Y}$ *and* $\|c^{(s)}\|_\infty$ *for* $s = 0, \ldots, \lceil d/2 \rceil$. *Then*

$$\mathbb{E}|OT_\varepsilon(\alpha, \beta) - OT_\varepsilon(\alpha_n, \beta_n)| = O\left(\frac{(1 + (\varepsilon^{-1} C_\varepsilon)^{\lfloor d/2 \rfloor})\psi'(O\left(\varepsilon^{-1}\right))}{\sqrt{n}}\right).$$

Sample Complexity for General Regularization Function

Theorem

Same setting, but this time with regularization with Legendre conjugate $\Psi : [0, \infty] \to [0, \infty]$ *convex such that*

$$C_\varepsilon = \frac{\max_{s \leq \frac{d}{2}} \sup_{z \in \frac{1}{\varepsilon} K} |\psi^{(s)}(z)|}{\inf_{z \in \frac{1}{\varepsilon} K} |\Psi''(z)|} < \infty,$$

where the compact set $K \subset \mathbb{R}$ *depends only on* $\mathcal{X}, \mathcal{Y}$ *and* $\|c^{(s)}\|_\infty$ *for* $s = 0, \ldots, \lceil d/2 \rceil$. *Then*

$$\mathbb{E}|OT_\varepsilon(\alpha, \beta) - OT_\varepsilon(\alpha_n, \beta_n)| = O\left(\frac{(1 + (\varepsilon^{-1} C_\varepsilon)^{\lfloor d/2 \rfloor})\psi'(O(\varepsilon^{-1}))}{\sqrt{n}}\right).$$

- For $\Psi = \exp$, recover previous result up to a factor of d in the exponent.

Idea of the Proof

We want to bound

$$\mathbb{E}|OT_\varepsilon(\alpha_n, \beta_n) - OT_\varepsilon(\alpha, \beta)| = \mathbb{E}\left|\max_{u,v} \frac{1}{n}\sum F_\varepsilon^{X_i Y_i}(u, v) - \max_{u,v} \mathbb{E}\left[F_\varepsilon^{X_1 Y_1}(u, v)\right]\right|$$

where

$$F_\varepsilon^{XY}(u, v) = u(X) - v(Y) - \varepsilon\psi\left(\frac{u(X) + v(Y) - c(X, Y)}{\varepsilon}\right).$$

Idea of the Proof

We want to bound

$$\mathbb{E}|OT_\varepsilon(\alpha_n,\beta_n) - OT_\varepsilon(\alpha,\beta)| = \mathbb{E}\left|\max_{u,v} \frac{1}{n}\sum F_\varepsilon^{X_i Y_i}(u,v) - \max_{u,v} \mathbb{E}\left[F_\varepsilon^{X_1 Y_1}(u,v)\right]\right|$$

where

$$F_\varepsilon^{XY}(u,v) = u(X) - v(Y) - \varepsilon\psi\left(\frac{u(X)+v(Y)-c(X,Y)}{\varepsilon}\right).$$

Assume we can show $\|u\|_{\mathcal{H}^s}, \|v\|_{\mathcal{H}^s} \leq \lambda$ for (u,v) who are optimizers with respect to any (α,β). Then

$$\begin{aligned}
&\mathbb{E}|OT_\varepsilon(\alpha_n,\beta_n) - OT_\varepsilon(\alpha,\beta)| \\
&\leq 3 \sup_{\|u\|_{\mathcal{H}^s},\|v\|_{\mathcal{H}^s}\leq\lambda} \mathbb{E}\left|\frac{1}{n}\sum F_\varepsilon^{X_i Y_i}(u,v) - \mathbb{E}\left[F_\varepsilon^{X_1 Y_1}(u,v)\right]\right|.
\end{aligned}$$

Standard result, e. g. Bartlett-Mendelson '02:
If $s > \lfloor d/2 \rfloor$, so that $\mathcal{H}^s$ consists of continuous functions,

$$\sup_{\|u\|_{\mathcal{H}^s}, \|v\|_{\mathcal{H}^s} \leq \lambda} \mathbb{E} \left| \frac{1}{n} \sum F_{\varepsilon}^{X_i Y_i}(u, v) - \mathbb{E}\left[F_{\varepsilon}^{X_1 Y_1}(u, v) \right] \right| = O_{\lambda, \varepsilon}\left(\frac{1}{\sqrt{n}} \right).$$

Regularity of the Optimizers

Still need to show that $\|u\|_{\mathcal{H}^s}$ is bounded in terms of $\Psi, c, \mathcal{X}, \mathcal{Y}$.

To get such a bound, use maximality for optimizers u, v:

- Pick any φ in $C(X)$ and look at $g(t) := \mathbb{E}F_\varepsilon^{XY}(u + t\varphi, v)$.
- Must have $g'(0) = 0$ by optimality.
- Derive that for any x,

$$1 = \int \psi' \left(\frac{u(x) + v(y) - c(x, y)}{\varepsilon} \right) d\beta(y).$$

- Differentiate this equation in x.

Demonstrating Regularized Optimal Transport

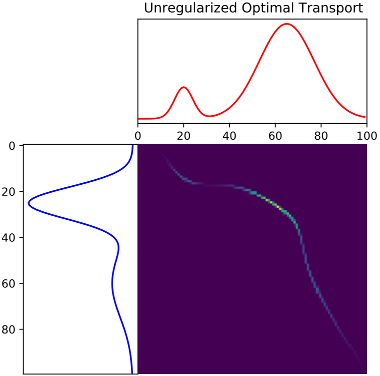


Figure: Unregularized 1D Optimal Transport

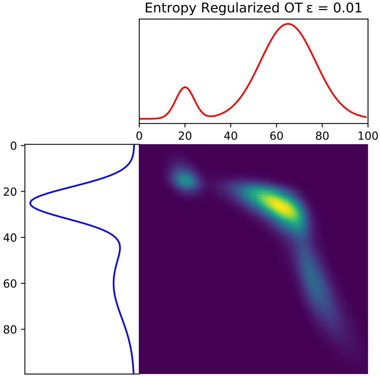


Figure: Entropy Regularized 1D Optimal Transport

Tsallis Entropies

$$H_q(\alpha \| \beta) = \frac{1}{q-1} \int \left[\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta} \right)^{q-1} - 1 \right] \mathrm{d}\alpha \tag{1}$$

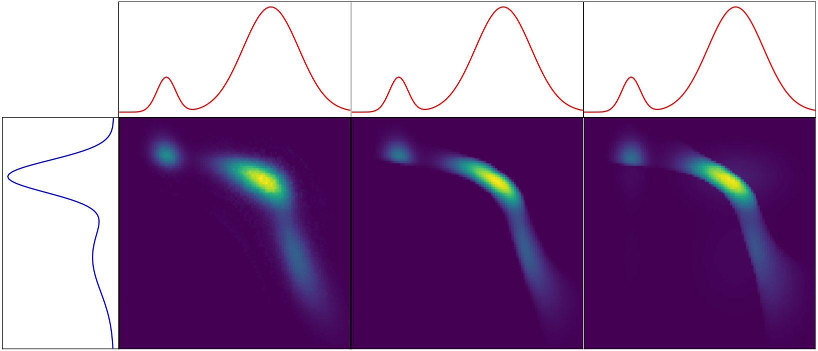
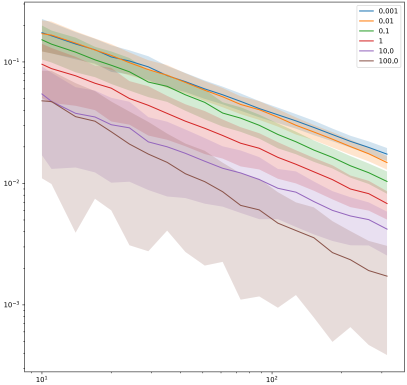


Figure: OT plans with relative Tsallis entropy for $q = (1.1, 2, 3)$

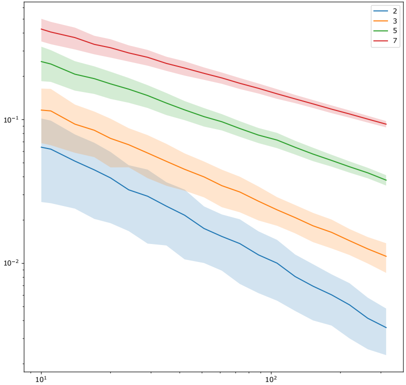
Demonstrating Sample Complexity

- We do not know how to compute $\mathbb{E}|OT_\varepsilon(\alpha, \hat{\alpha}_n)|$ numerically
- Instead compute $\mathbb{E}|OT_\varepsilon(\hat{\alpha}_n, \hat{\alpha}'_n)| \leq 2\mathbb{E}|OT_\varepsilon(\alpha, \hat{\alpha}_n)|$
- For our experiments we choose:
  - α the uniform distribution on a d-dimensional unit hypercube.
  - $c(x, y) = \|x - y\|^2$
  - Average over 300 pairs of samples for each n

Sample Complexity



(a) $d = 3,\ q = 2$, different ε



(b) $\varepsilon = 1, q = 3$, different d

Figure: Optimal Transport between two independent sampled distributions as a function of samples n.

Sample Complexity

Fit functions of the form

$$OT_\varepsilon(\hat{\alpha}_n, \hat{\alpha}'_n) \approx An^t \qquad (2)$$

Fit results (d = 3):

ε \ q	0.5	0.9	1.1	1.5	2.0	3.0	5.0	10.0
0.001	-0.689	-0.689	-0.678	-0.676	-0.675	-0.674	-0.67	-0.672
0.01	-0.819	-0.765	-0.751	-0.733	-0.72	-0.706	-0.689	-0.678
0.1	-1.24	-0.982	-0.89	-0.829	-0.783	-0.747	-0.721	-0.704
1.0	-1.05	-1.03	-1.02	-0.813	-0.766	-0.689	-0.72	-0.701
10.0	-0.999	-0.988	-0.984	-0.938	-0.717	-0.686	-0.69	-0.681
100.0	-1.01	-0.996	-0.998	-1.01	-0.995	-0.63	-0.667	-0.684

Figure: Fit parameter t for the Tsallis relative entropy regularization (d = 3).