Sample complexity of optimal transport barycenters with discrete support

Léo Portales

IRIT, TSE, CNRS and Université de Toulouse

19/06/25

Outline

1. Introduction of the problem
2. The optimal transport divergences we consider
3. Sample complexity and empirical risk minimization
4. The main result

Optimal transport barycenters with sparse support

$\mathbb{D} : \mathcal{M}_1(\mathbb{R}^d) \times \mathcal{M}_1(\mathbb{R}^d) \longrightarrow \mathbb{R}^+$ optimal transport divergence (ex: W_p^p, $W_{\epsilon,p}$, SW_p^p, ...)

Optimal transport barycenters with sparse support

$\mathbb{D} : \mathcal{M}_1(\mathbb{R}^d) \times \mathcal{M}_1(\mathbb{R}^d) \longrightarrow \mathbb{R}^+$ optimal transport divergence (ex: W_p^p, $W_{\epsilon,p}$, SW_p^p, ...) Let $\mu^1, ..., \mu^L \in \mathcal{M}_1(\mathbb{R}^d)$

Optimal transport barycenters with sparse support

$\mathbb{D} : \mathcal{M}_1(\mathbb{R}^d) \times \mathcal{M}_1(\mathbb{R}^d) \longrightarrow \mathbb{R}^+$ optimal transport divergence (ex: W_p^p, $W_{\epsilon,p}$, SW_p^p, ...) Let $\mu^1, ..., \mu^L \in \mathcal{M}_1(\mathbb{R}^d)$

Optimal transport barycenter:

$$\mu^* \in \underset{\mu \in \mathcal{M}_1(\mathbb{R}^d)}{\text{argmin}} \; \mathsf{F}_{\mathbb{D}}(\mu^1, ..., \mu^L, \mu) := \frac{1}{L} \sum_{i=1}^{L} \mathbb{D}(\mu^i, \mu)$$

Optimal transport barycenters with sparse support

$\mathbb{D} : \mathcal{M}_1(\mathbb{R}^d) \times \mathcal{M}_1(\mathbb{R}^d) \longrightarrow \mathbb{R}^+$ optimal transport divergence (ex: W_p^p, $W_{\epsilon,p}$, SW_p^p, ...) Let $\mu^1, ..., \mu^L \in \mathcal{M}_1(\mathbb{R}^d)$

Optimal transport barycenter:

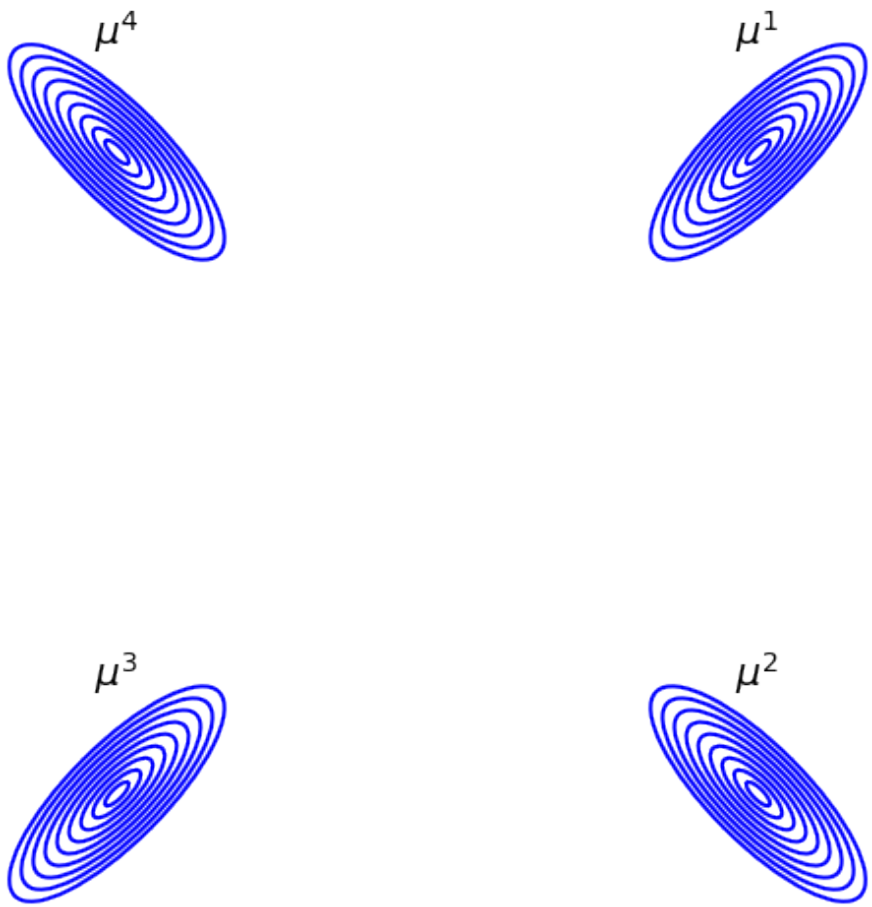
$$\mu^* \in \underset{\mu \in \mathcal{M}_1(\mathbb{R}^d)}{\operatorname{argmin}} \; \mathsf{F}_{\mathbb{D}}(\mu^1, ..., \mu^L, \mu) := \frac{1}{L} \sum_{i=1}^{L} \mathbb{D}(\mu^i, \mu)$$

Optimal transport barycenter with sparse support:

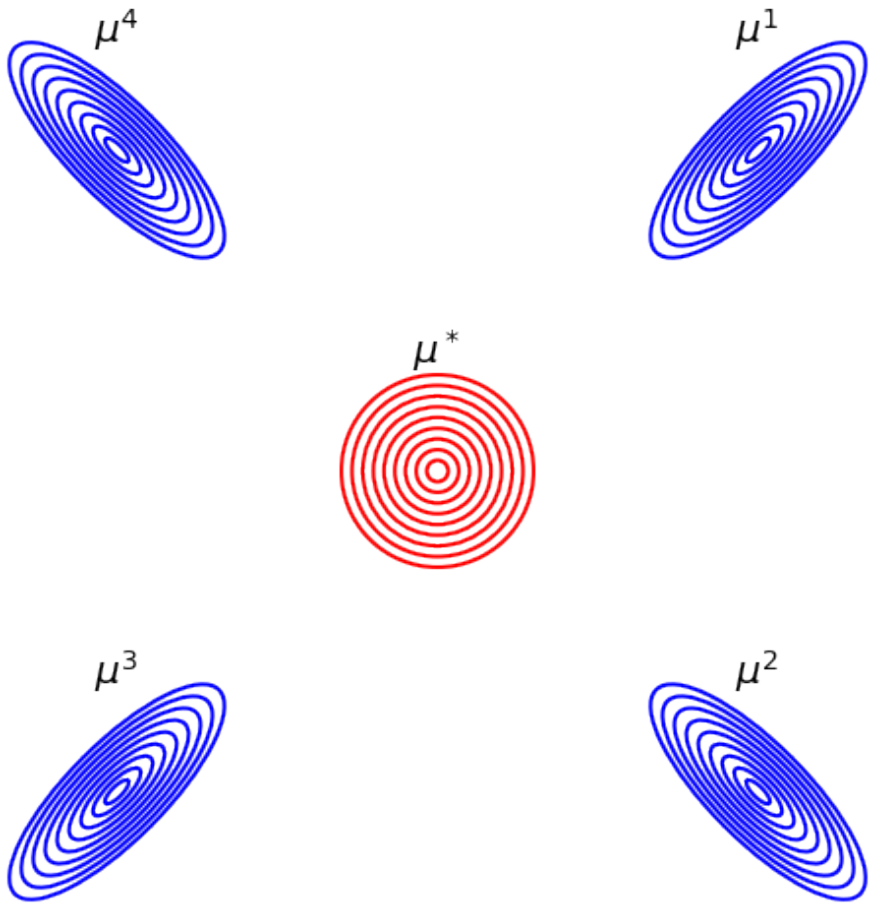
$$\nu^* \in \underset{\mu \in \overline{\mathcal{M}_1^N}}{\operatorname{argmin}} \; \mathsf{F}_{\mathbb{D}}(\mu^1, ..., \mu^L, \mu) := \frac{1}{L} \sum_{i=1}^{L} \mathbb{D}(\mu^i, \mu)$$

where $\overline{\mathcal{M}_1^N}$ is a subset of $\mathcal{M}_1(\mathbb{R}^d)$ composed of discrete measures supported on at most N points.

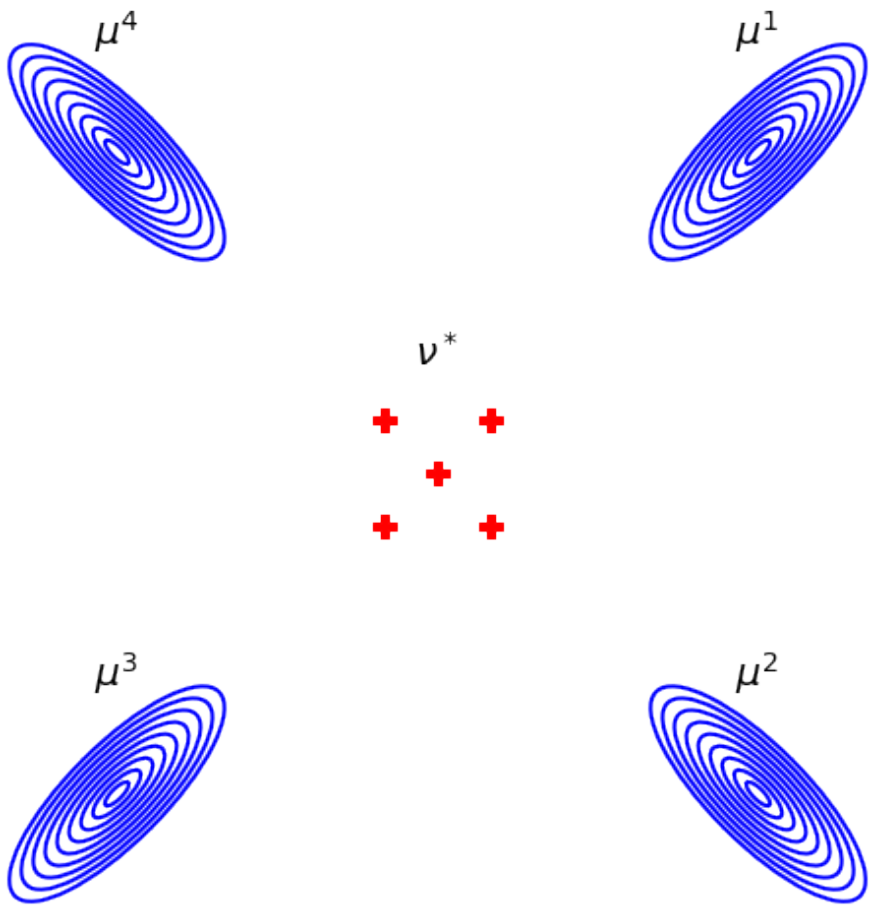
Optimal transport barycenters with sparse support



Optimal transport barycenters with sparse support



Optimal transport barycenters with sparse support



Our problem

- Continuous or large support size target measures $\mu^1, \dots, \mu^L \in \mathcal{M}_1(\mathbb{R}^d)$ need a discrete approximation to be computationally tractable, even when they are known.

Our problem

- Continuous or large support size target measures $\mu^1, \dots, \mu^L \in \mathcal{M}_1(\mathbb{R}^d)$ need a discrete approximation to be computationally tractable, even when they are known.
- This can be done through histogram, online stochastic or sample average aproximation. We focus on the latter.

Our problem

- Continuous or large support size target measures $\mu^1, \dots, \mu^L \in \mathcal{M}_1(\mathbb{R}^d)$ need a discrete approximation to be computationally tractable, even when they are known.
- This can be done through histogram, online stochastic or sample average aproximation. We focus on the latter.
- Moreover most algorithms include a support cardinality constraint on the OT barycenter $|\operatorname{supp}(\nu^*)| \leq N$, a constraint which we include in our analysis.

Our problem

- Continuous or large support size target measures $\mu^1, \dots, \mu^L \in \mathcal{M}_1(\mathbb{R}^d)$ need a discrete approximation to be computationally tractable, even when they are known.
- This can be done through histogram, online stochastic or sample average aproximation. We focus on the latter.
- Moreover most algorithms include a support cardinality constraint on the OT barycenter $|\operatorname{supp}(\nu^*)| \leq N$, a constraint which we include in our analysis.

More specifically we aim at giving statistical guarantees for the sample average approximation of the following problem

$$\min_{(Y,\pi)\in(\mathbb{R}^d)^N\times\Delta_N} \mathsf{F}_{\mathbb{D}}\left(\mu^1, \dots, \mu^L, \sum_{i=1}^N \pi_i \delta_{y_i}\right) := \frac{1}{L}\sum_{\ell=1}^L \mathbb{D}\left(\mu^\ell, \sum_{i=1}^N \pi_i \delta_{y_i}\right).$$

Our problem

More specifically if $\forall \ell \in [\![1, L]\!]$ ${\mu_n}^\ell := \frac{1}{n} \sum_{i=1}^n \delta_{X_i^\ell}$, $X_1^\ell, \ldots, {X_n}^\ell \sim \mu^\ell$

Our problem

More specifically if $\forall \ell \in [\![1, L]\!]$ ${\mu_n}^\ell := \frac{1}{n} \sum_{i=1}^n \delta_{X_i^\ell}$, $X_1^\ell, \ldots, {X_n}^\ell \sim \mu^\ell$ and if

$$ {\nu_n}^* \in \underset{\nu \in \mathcal{M}_1^N(\mathbb{R}^d)}{\operatorname{argmin}} \; \mathsf{F}_{\mathbb{D}}({\mu_n}^1, \ldots, {\mu_n}^L, \nu) $$

Our problem

More specifically if $\forall \ell \in [\![1, L]\!]$ ${\mu_n}^\ell := \frac{1}{n} \sum_{i=1}^n \delta_{X_i^\ell}$, $X_1^\ell, \ldots, {X_n}^\ell \sim \mu^\ell$ and if

$$ {\nu_n}^* \in \underset{\nu \in \mathcal{M}_1^N(\mathbb{R}^d)}{\operatorname{argmin}} \mathsf{F}_{\mathbb{D}}({\mu_n}^1, \ldots, {\mu_n}^L, \nu) $$

or equivalently:

$$ (Y_n, \pi_n) \in \underset{(Y,\pi) \in (\mathbb{R}^d)^N \times \Delta_N}{\operatorname{argmin}} \mathsf{F}_{\mathbb{D}} \left({\mu_n}^1, \ldots, {\mu_n}^L, \sum_{i=1}^N \pi_i \delta_{y_i} \right), $$

Our problem

More specifically if $\forall \ell \in [\![1, L]\!]$ ${\mu_n}^\ell := \frac{1}{n} \sum_{i=1}^n \delta_{X_i^\ell}$, $X_1^\ell, \ldots, {X_n}^\ell \sim \mu^\ell$ and if

$$ {\nu_n}^* \in \underset{\nu \in \mathcal{M}_1^N(\mathbb{R}^d)}{\operatorname{argmin}} \mathsf{F}_{\mathbb{D}}({\mu_n}^1, \ldots, {\mu_n}^L, \nu) $$

or equivalently:

$$ (Y_n, \pi_n) \in \underset{(Y,\pi) \in (\mathbb{R}^d)^N \times \Delta_N}{\operatorname{argmin}} \mathsf{F}_{\mathbb{D}} \left({\mu_n}^1, \ldots, {\mu_n}^L, \sum_{i=1}^N \pi_i \delta_{y_i} \right), $$

then what are statistical upper bounds for

$$ \mathbb{E} \left[\mathsf{F}_{\mathbb{D}} \left(\mu^1, \ldots, \mu^L, \sum_{i=1}^N \pi_i^n \delta_{y_i^n} \right) - \min_{(Y^*,\pi^*)} \mathsf{F}_{\mathbb{D}} \left(\mu^1, \ldots, \mu^L, \sum_{i=1}^N \pi_i^* \delta_{y_i^*} \right) \right] $$

?

Various interesting subcases

Let A be some closed subset of $(\mathbb{R}^d)^N \times \Delta_N$ and $\mathbb{D}$ an OT divergence.

Various interesting subcases

Let A be some closed subset of $(\mathbb{R}^d)^N \times \Delta_N$ and $\mathbb{D}$ an OT divergence. Recall our object of interest:

$$\min_{(Y,\pi)\in A} \frac{1}{L} \sum_{l=1}^{L} \mathbb{D}\left(\mu^l, \sum_{i=1}^{N} \pi_i \delta_{y_i}\right).$$

Depending on A and L, this problem admits various subcases.

Various interesting subcases

Let A be some closed subset of $(\mathbb{R}^d)^N \times \Delta_N$ and $\mathbb{D}$ an OT divergence. Recall our object of interest:

$$\min_{(Y,\pi)\in A} \frac{1}{L}\sum_{l=1}^{L} \mathbb{D}\left(\mu^l, \sum_{i=1}^{N} \pi_i \delta_{y_i}\right).$$

Depending on A and L, this problem admits various subcases.

A	$L=1$	$L>1$
$(\mathbb{R}^d)^N \times \Delta_N$	Optimal quantization	OT barycenter with sparse support
$(\mathbb{R}^d)^N \times \{\bar{\pi}\}$	Constrained quantization	Free support OT barycenter
$\{\bar{Y}\} \times \Delta_N$	-	Fixed support OT barycenter

Various interesting subcases

Let A be some closed subset of $(\mathbb{R}^d)^N \times \Delta_N$ and $\mathbb{D}$ an OT divergence. Recall our object of interest:

$$\min_{(Y,\pi)\in A} \frac{1}{L}\sum_{l=1}^{L} \mathbb{D}\left(\mu^l, \sum_{i=1}^{N} \pi_i \delta_{y_i}\right).$$

Depending on A and L, this problem admits various subcases.

A	$L=1$	$L>1$
$(\mathbb{R}^d)^N \times \Delta_N$	Optimal quantization	OT barycenter with sparse support
$(\mathbb{R}^d)^N \times \{\bar{\pi}\}$	Constrained quantization	Free support OT barycenter
$\{\bar{Y}\} \times \Delta_N$	-	Fixed support OT barycenter

Note: if $\mathbb{D} = W_2^2$, $L=1$, μ discrete and $A = (\mathbb{R}^d)^N \times \Delta_N$ this is the **K-means** problem.

Outline

1. Introduction of the problem
2. The optimal transport divergences we consider
3. Sample complexity and empirical risk minimization
4. The main result

The Wasserstein distance

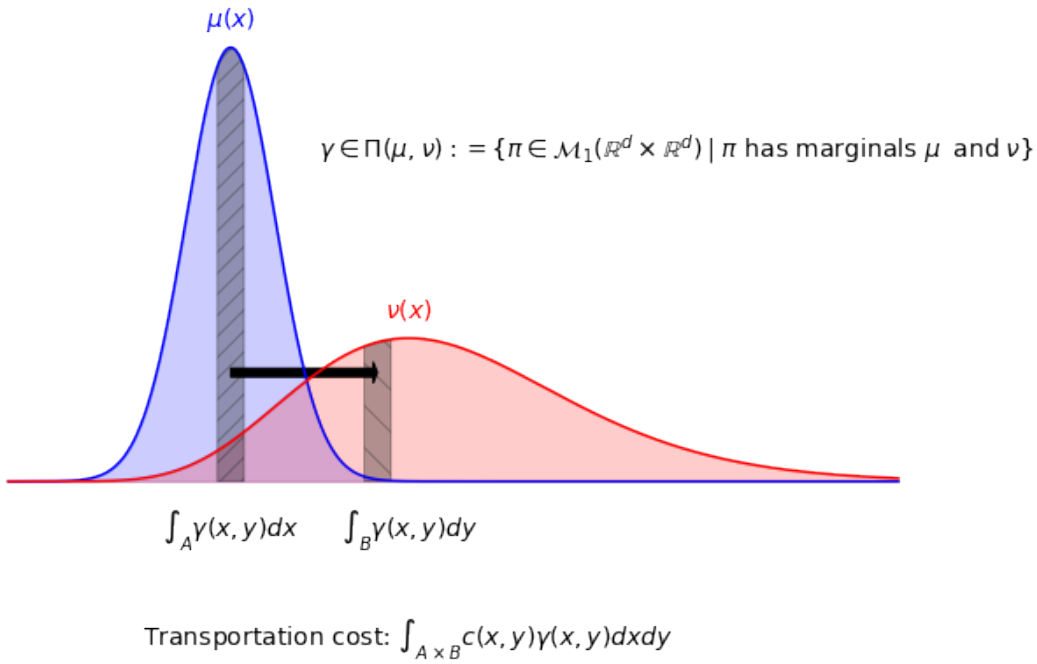
The Wasserstein distance $W_p(\mu, \nu)$ measures the minimal cost of transporting μ onto ν with couplings γ.

$$\boxed{W_p^p(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\gamma(x, y).}$$

The Wasserstein distance

The Wasserstein distance $W_p(\mu, \nu)$ measures the minimal cost of transporting μ onto ν with couplings γ.

$$W_p^p(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\gamma(x, y).$$



Entropic Optimal transport

$$W_{\epsilon,p}(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p d\gamma(x,y) + \epsilon \mathrm{KL}(\gamma, \mu\bigotimes\nu)$$

Entropic Optimal transport

$W_{\epsilon,p}(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p d\gamma(x,y) + \epsilon \mathrm{KL}(\gamma, \mu\bigotimes\nu)$ (where $\mathrm{KL}(\alpha,\beta) = \int_{\mathbb{R}^d} \log\left(\frac{d\alpha(x)}{d\beta(x)}\right) d\alpha(x)$).

Entropic Optimal transport

$W_{\epsilon,p}(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p d\gamma(x,y) + \epsilon \mathrm{KL}(\gamma, \mu\bigotimes\nu)$ (where $\mathrm{KL}(\alpha,\beta) = \int_{\mathbb{R}^d} \log\left(\frac{d\alpha(x)}{d\beta(x)}\right) d\alpha(x)$).

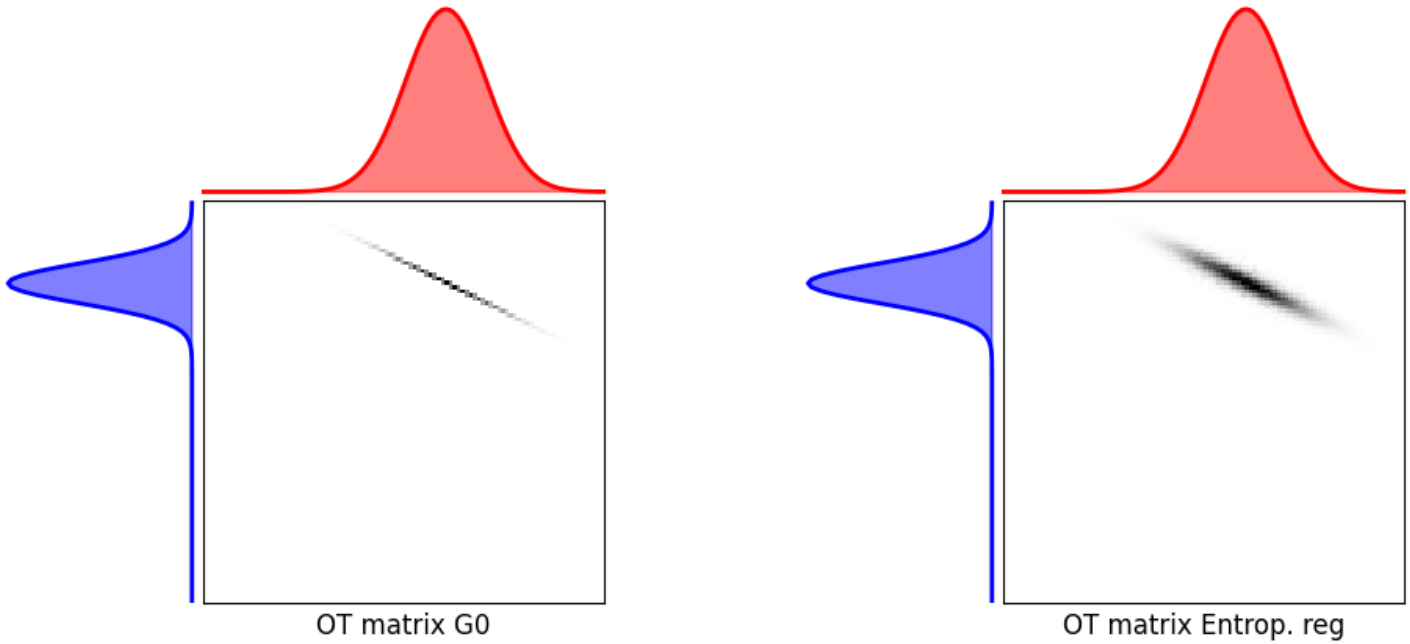


Figure: *POT library example of a $W_{\epsilon,2}$-transport plan*

Wider because it penalizes couplings too far from the product measure.

Sliced optimal transport

1-D optimal transport: if $\mu, \nu \in \mathcal{M}_1(\mathbb{R})$:

$$W_p^p(\mu, \nu) = \int_0^1 \mid F_\mu^{-1}(t) - F_\nu^{-1}(t) \mid^p \, dt$$

where F_α^{-1} denotes the generalized inverse of the c.d.f of α.

Sliced optimal transport

1-D optimal transport: if $\mu, \nu \in \mathcal{M}_1(\mathbb{R})$:

$$W_p^p(\mu, \nu) = \int_0^1 \mid F_\mu^{-1}(t) - F_\nu^{-1}(t) \mid^p \, dt$$

where F_α^{-1} denotes the generalized inverse of the c.d.f of α.

Sliced optimal transport:

The sliced and max-sliced Wasserstein distances leverages the closed form of 1-d OT through projections on the line with directions θ.

Sliced optimal transport

1-D optimal transport: if $\mu, \nu \in \mathcal{M}_1(\mathbb{R})$:

$$W_p^p(\mu, \nu) = \int_0^1 \mid F_\mu^{-1}(t) - F_\nu^{-1}(t) \mid^p dt$$

where F_α^{-1} denotes the generalized inverse of the c.d.f of α.

Sliced optimal transport:

The sliced and max-sliced Wasserstein distances leverages the closed form of 1-d OT through projections on the line with directions θ.

$$SW_p^p(\mu, \nu) = \int_{\mathbb{S}^{d-1}} W_p^p(P_\theta \sharp \mu, P_\theta \sharp \nu) d\sigma(\theta),\ \sigma \sim \mathcal{U}_{\mathbb{S}^{d-1}}$$

$$\text{max-}SW_p^p(\mu, \nu) = \max_{\theta \in \mathbb{S}^{d-1}} W_p^p(P_\theta \sharp \mu, P_\theta \sharp \nu).$$

Outline

1. Introduction of the problem
2. The optimal transport divergences we consider
3. Sample complexity and empirical risk minimization
4. The main result

Statistical learning theory

Empirical risk minimization:

- $X \sim \mu$ (only known through sample)

Sample complexity:

Statistical learning theory

Empirical risk minimization:

- $X \sim \mu$ (only known through sample)
- $\mathcal{F}$ functional class

Sample complexity:

Statistical learning theory

Empirical risk minimization:

- $X \sim \mu$ (only known through sample)
- $\mathcal{F}$ functional class
- L loss function.

Sample complexity:

Statistical learning theory

Empirical risk minimization:

- $X \sim \mu$ (only known through sample)
- $\mathcal{F}$ functional class
- L loss function.

Problem of finding how much information is lost when performing $\min_{f \in \mathcal{F}} \mathbb{P}_n L(f)$ (empirical risk) insted of $\min_{f \in \mathcal{F}} \mathbb{E}[L(f(X))]$ (exact risk).

Sample complexity:

Statistical learning theory

Empirical risk minimization:

- $X \sim \mu$ (only known through sample)
- $\mathcal{F}$ functional class
- L loss function.

Problem of finding how much information is lost when performing $\min_{f \in \mathcal{F}} \mathbb{P}_n L(f)$ (empirical risk) insted of $\min_{f \in \mathcal{F}} \mathbb{E}[L(f(X))]$ (exact risk).

Sample complexity:

Minimum n such that desired precision is reached for the actual risk only minimizing over the empirical risk.

Statistical learning theory

Generalization error:

$$\mathbb{E}\left[L(f_n(X)) - \min_{f\in\mathcal{F}} \mathbb{E}[L(f(X))]\right] \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} \mid \mathbb{P}_n L(f) - \mathbb{E}[L(f(X))] \mid\right]$$
$$\lesssim \frac{\mathcal{C}(n)}{\sqrt{n}}$$

where f_n is a minimizer of $\mathbb{P}_n L(f)$ and $\mathcal{C}(n)$ is a measure of complexity (Rademacher, VC dimension, log-entropy, ...)

Statistical aspect of optimal transport

Study of asymptotic behavior of $\mathbb{D}(\mu, \mu_n)$ and $\mathbb{D}(\mu, \nu_n)$ $\nu \neq \mu$.

[1] On the rate of convergence in Wasserstein distance of the empirical measure, N.Fournier and A.Guillin. 2013

Statistical aspect of optimal transport

Study of asymptotic behavior of $\mathbb{D}(\mu, \mu_n)$ and $\mathbb{D}(\mu, \nu_n)$ $\nu \neq \mu$. In all generality it depends on dimensionality. For instance if μ's support is compact[1]:

$$\mathbb{E}\left[W_p(\mu_n, \mu)\right] \lesssim \begin{cases} n^{-1/d} & \text{if } d > 2p \\ n^{-1/2p} \log(n)^{1/p} & \text{if } d = 2p \\ n^{-1/2p} & \text{if } d < 2p \end{cases}$$

[1] On the rate of convergence in Wasserstein distance of the empirical measure, N.Fournier and A.Guillin. 2013

Statistical aspect of optimal transport

Study of asymptotic behavior of $\mathbb{D}(\mu, \mu_n)$ and $\mathbb{D}(\mu, \nu_n)$ $\nu \neq \mu$. In all generality it depends on dimensionality. For instance if μ's support is compact[1]:

$$\mathbb{E}\left[W_p(\mu_n, \mu)\right] \lesssim \begin{cases} n^{-1/d} & \text{if } d > 2p \\ n^{-1/2p} \log(n)^{1/p} & \text{if } d = 2p \\ n^{-1/2p} & \text{if } d < 2p \end{cases}$$

and by triangle inequality this is also true for for $\mid W_p(\mu_n, \nu) - W_p(\mu, \nu) \mid$.

[1] On the rate of convergence in Wasserstein distance of the empirical measure, N.Fournier and A.Guillin. 2013

Semi-discrete OT: fundamental $1/\sqrt{n}$ rate of convergence

Sharper results by investigating $\mathbb{D}(\mu_n, \nu)$ $(\mu \neq \nu)$ insted. This is because the sample complexity of OT adapts to the "less complex" measure[2].

[2]Empirical optimal transport between different measures adapts to lower complexity, S.Hundrieser, T.Staudt and A.Munk.2022.

Semi-discrete OT: fundamental $1/\sqrt{n}$ rate of convergence

Sharper results by investigating $\mathbb{D}(\mu_n, \nu)$ $(\mu \neq \nu)$ insted. This is because the sample complexity of OT adapts to the "less complex" measure[2]. In particular if ν is discrete then

$$\mathbb{E}[\mid W_p^p(\mu_n, \nu) - W_p^p(\mu, \nu) \mid] = O(1/\sqrt{n}).$$

[2]Empirical optimal transport between different measures adapts to lower complexity, S.Hundrieser, T.Staudt and A.Munk.2022.

Statistical aspect of barycenters

Empirical Wasserstein barycenters are solutions of $\min_{\mu \in \mathcal{M}_1(\mathbb{R}^d)} \frac{1}{L} \sum_{l=1}^{L} W_p^p(\mu_n^l, \mu)$ with for all $l = 1, \dots, L$ μ_n^l empirical measures over i.i.d realizations of μ^l.

[3] Quantitative stability of barycenters in the Wasserstein space, G.Carlier, A.Delalande and Q.Mérigot. 2022.

[4] Randomized Wasserstein barycenter computation: Resampling with statistical guarantees, F.Heinemann, A.Munk and Y.Zemel. 2023.

Statistical aspect of barycenters

Empirical Wasserstein barycenters are solutions of $\min_{\mu \in \mathcal{M}_1(\mathbb{R}^d)} \frac{1}{L} \sum_{l=1}^{L} W_p^p(\mu_n^l, \mu)$ with for all $l = 1, \dots, L$ μ_n^l empirical measures over i.i.d realizations of μ^l. They converge to the true barycenter, furthermore we have[3]

$$\mathbb{E}\left[W_2^2\left(\mu_n^*, \mu^*\right)\right] \lesssim \begin{cases} n^{-1/12} \text{ if } d < 4 \\ n^{-1/12} \log(n)^{1/6} \text{ if } d = 4 \\ n^{-1/3d} \text{ if } d > 4. \end{cases}$$

[3] Quantitative stability of barycenters in the Wasserstein space, G.Carlier, A.Delalande and Q.Mérigot. 2022.

[4] Randomized Wasserstein barycenter computation: Resampling with statistical guarantees, F.Heinemann, A.Munk and Y.Zemel. 2023.

Statistical aspect of barycenters

Empirical Wasserstein barycenters are solutions of $\min_{\mu \in \mathcal{M}_1(\mathbb{R}^d)} \frac{1}{L} \sum_{l=1}^{L} W_p^p(\mu_n^l, \mu)$ with for all $l = 1, \dots, L$ μ_n^l empirical measures over i.i.d realizations of μ^l. They converge to the true barycenter, furthermore we have[3]

$$\mathbb{E}\left[W_2^2\left(\mu_n^*, \mu^*\right)\right] \lesssim \begin{cases} n^{-1/12} \text{ if } d < 4 \\ n^{-1/12} \log(n)^{1/6} \text{ if } d = 4 \\ n^{-1/3d} \text{ if } d > 4. \end{cases}$$

In the discrete setting we have[4]

$$\mathbb{E}\left[\frac{1}{L}\sum_{l=1}^{L} W_p^p(\mu^l, \nu_n) - \min_{\nu \in \mathcal{M}_1(\mathbb{R}^d)} \frac{1}{L}\sum_{l=1}^{L} W_p^p(\mu^l, \nu)\right] \lesssim \frac{1}{\sqrt{n}}.$$

Where ν_n is an empirical barycenter.

[3] Quantitative stability of barycenters in the Wasserstein space, G.Carlier, A.Delalande and Q.Mérigot. 2022.

[4] Randomized Wasserstein barycenter computation: Resampling with statistical guarantees, F.Heinemann, A.Munk and Y.Zemel. 2023.

The K-means as an empirical risk minimization scheme

It was shown[5] that empirical optimal quantization (or K-means) behaves as

$$\mathbb{E}\left[D_2(\mu, Y_n) - \min_{Y \in \overline{\mathrm{B}_d(0,R)}^N} D_2(\mu, Y)\right] \lesssim \sqrt{\frac{N \log(N)}{n}}$$

where $D_2(\mu, Y) = \int_{\mathbb{R}^d} \min_{i=1,\dots,N} \|x - y_i\|^2 d\mu(x)$.

[5] Principles of nonparametric learning, Section 4.4, L.Gyorfi. 2002.

The K-means as an empirical risk minimization scheme

It was shown[5] that empirical optimal quantization (or K-means) behaves as

$$\mathbb{E}\left[D_2(\mu, Y_n) - \min_{Y \in \overline{\mathrm{B}_d(0,R)}^N} D_2(\mu, Y)\right] \lesssim \sqrt{\frac{N \log(N)}{n}}$$

where $D_2(\mu, Y) = \int_{\mathbb{R}^d} \min_{i=1,\dots,N} \|x - y_i\|^2 d\mu(x)$.

(Here $\mathcal{F} = \left\{ x \mapsto \min_{i=1,\dots,N} \|x - y_i\|^2 \mid Y \in \overline{\mathrm{B}_d(0,R)}^N \right\}$).

[5] Principles of nonparametric learning, Section 4.4, L.Gyorfi. 2002.

The K-means as an empirical risk minimization scheme

It was shown[5] that empirical optimal quantization (or K-means) behaves as

$$\mathbb{E}\left[D_2(\mu, Y_n) - \min_{Y \in \overline{\mathrm{B}_d(0,R)}^N} D_2(\mu, Y)\right] \lesssim \sqrt{\frac{N \log(N)}{n}}$$

where $D_2(\mu, Y) = \int_{\mathbb{R}^d} \min_{i=1,\dots,N} \|x - y_i\|^2 d\mu(x)$.

(Here $\mathcal{F} = \left\{ x \mapsto \min_{i=1,\dots,N} \|x - y_i\|^2 \mid Y \in \overline{\mathrm{B}_d(0,R)}^N \right\}$).

Question:

Does this result also hold for the more general problem

$$\min_{(Y,\pi) \in A} \frac{1}{L} \sum_{l=1}^{L} \mathbb{D}\left(\mu^l, \sum_{i=1}^{N} \pi_i \delta_{y_i}\right) \quad A \subset (\mathbb{R}^d)^N \times \Delta_N \, ?$$

[5] Principles of nonparametric learning, Section 4.4, L.Gyorfi. 2002.

Outline

1. Introduction of the problem
2. The optimal transport divergences we consider
3. Sample complexity and empirical risk minimization
4. The main result

Main result

- $\mu^1, ..., \mu^L \in \mathcal{M}_1\left(\overline{\mathrm{B}_d(0,R)}\right)$.

Main result

- $\mu^1, ..., \mu^L \in \mathcal{M}_1\left(\overline{\mathrm{B}_d(0,R)}\right)$.
- $\mu_n^1, \ldots, \mu_n^L$ empirical measures over i.i.d r.v of respective law $\mu^1, \ldots \mu^L$.

Main result

- $\mu^1, ..., \mu^L \in \mathcal{M}_1\left(\overline{\mathrm{B}_d(0,R)}\right)$.
- $\mu_n^1, \ldots, \mu_n^L$ empirical measures over i.i.d r.v of respective law $\mu^1, \ldots \mu^L$.
- $\mathbb{D} = W_p^p$, $W_{p,\epsilon}$, SW_p^p or max-SW_p^p.

Main result

- $\mu^1, ..., \mu^L \in \mathcal{M}_1\left(\overline{\mathrm{B}_d(0,R)}\right)$.
- $\mu_n^1, \ldots, \mu_n^L$ empirical measures over i.i.d r.v of respective law $\mu^1, \ldots \mu^L$.
- $\mathbb{D} = W_p^p$, $W_{p,\epsilon}$, SW_p^p or max-SW_p^p.

Let $\nu_n^* := \sum_{i=1}^N \pi_i^n \delta_{y_i^n}$ be an empirical minimizer of $\mathrm{F}_{\mathbb{D}}(\mu^1, ..., \mu^L, \nu)$ over $A \subset (\mathbb{R}^d)^N \times \Delta_N$. Then

Main result

- $\mu^1, ..., \mu^L \in \mathcal{M}_1\left(\overline{\mathrm{B}_d(0,R)}\right)$.
- $\mu_n^1, \ldots, \mu_n^L$ empirical measures over i.i.d r.v of respective law $\mu^1, \ldots \mu^L$.
- $\mathbb{D} = W_p^p$, $W_{p,\epsilon}$, SW_p^p or max-SW_p^p.

Let $\nu_n^* := \sum_{i=1}^N \pi_i^n \delta_{y_i^n}$ be an empirical minimizer of $\mathrm{F}_{\mathbb{D}}(\mu^1, ..., \mu^L, \nu)$ over $A \subset (\mathbb{R}^d)^N \times \Delta_N$. Then

$$\mathbb{E}\left[\mathrm{F}_{\mathbb{D}}\left(\mu^1, \ldots, \mu^L, \sum_{i=1}^N \pi_i^n \delta_{y_i^n}\right) - \min_{(Y,\pi)\in A} \mathrm{F}_{\mathbb{D}}\left(\mu^1, \ldots, \mu^L, \sum_{i=1}^N \pi_i \delta_{y_i}\right)\right] \lesssim \sqrt{\frac{N}{n}}.$$

A very quick sketch of proof

- As it is typically done in statistical learning theory bound the quantity of interest by twice the generalization error:

$$\sup_{(\pi, Y) \in \Delta_N \times \mathbb{B}_R^N} \left| \mathsf{F}_{\mathbb{D}} \left(\mu_n^1, \ldots, \mu_n^L, \sum_{i=1}^N \pi_i \delta_{Y_i} \right) - \mathsf{F}_{\mathbb{D}} \left(\mu^1, \ldots, \mu^L, \sum_{i=1}^N \pi_i \delta_{y_i} \right) \right| . \qquad (1)$$

A very quick sketch of proof

- As it is typically done in statistical learning theory bound the quantity of interest by twice the generalization error:

$$\sup_{(\pi,Y)\in\Delta_N\times\mathbb{B}_R^N}\left|\mathsf{F}_{\mathbb{D}}\left(\mu_n^1,\ldots,\mu_n^L,\sum_{i=1}^N\pi_i\delta_{Y_i}\right)-\mathsf{F}_{\mathbb{D}}\left(\mu^1,\ldots,\mu^L,\sum_{i=1}^N\pi_i\delta_{y_i}\right)\right|. \tag{1}$$

- Leveraging the dual formulation of $\mathbb{D}$ upper bound and rewrite (1) as a sum of terms of the form

$$\sup_{w,Y,\pi}\left|\mathbb{P}_n^l(f_{Y,\pi,w})-\mathbb{E}_{X\sim\mu^l}[f_{Y,\pi,w}(X)]\right| \tag{2}$$

  where f belongs to a class of function that depends on the divergence used.

A very quick sketch of proof

- As it is typically done in statistical learning theory bound the quantity of interest by twice the generalization error:

$$\sup_{(\pi,Y)\in\Delta_N\times\mathbb{B}_R^N} \left| \mathsf{F}_{\mathbb{D}}\left(\mu_n^1,\dots,\mu_n^L,\sum_{i=1}^N \pi_i\delta_{Y_i}\right) - \mathsf{F}_{\mathbb{D}}\left(\mu^1,\dots,\mu^L,\sum_{i=1}^N \pi_i\delta_{y_i}\right)\right|. \tag{1}$$

- Leveraging the dual formulation of $\mathbb{D}$ upper bound and rewrite (1) as a sum of terms of the form

$$\sup_{w,Y,\pi} \left|\mathbb{P}_n^l(f_{Y,\pi,w}) - \mathbb{E}_{X\sim\mu^l}[f_{Y,\pi,w}(X)]\right| \tag{2}$$

  where f belongs to a class of function that depends on the divergence used.
- Use a measure of complexity to upper bound (2) in expectation (in our case log-entropy).

Thank you for your attention