Tutorial on Optimal Transport Theory

With a machine learning touch

Lénaïc Chizat*

Jul. 24th 2019 - IFCAM summer school - IISc Bangalore

*CNRS and Université Paris-Sud

A Geometric Motivation

Setting: Probability measures $\mathcal{P}(\mathcal{X})$ on a metric space $(\mathcal{X}, \mathrm{dist})$.

Goal

Build a metric on $\mathcal{P}(\mathcal{X})$ consistent with the geometry of $(\mathcal{X}, \mathrm{dist})$.

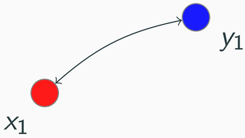
A Geometric Motivation

Setting: Probability measures $\mathcal{P}(\mathcal{X})$ on a metric space $(\mathcal{X}, \text{dist})$.

Goal

Build a metric on $\mathcal{P}(\mathcal{X})$ consistent with the geometry of $(\mathcal{X}, \text{dist})$.

$$\mu = \delta_{x_1}, \qquad \nu = \delta_{y_1}$$



Distance between μ and ν...

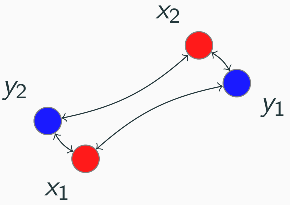
$$\text{dist}(x_1, y_1)$$

A Geometric Motivation

Setting: Probability measures $\mathcal{P}(\mathcal{X})$ on a metric space $(\mathcal{X}, \text{dist})$.

Goal

Build a metric on $\mathcal{P}(\mathcal{X})$ consistent with the geometry of $(\mathcal{X}, \text{dist})$.



$$\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}, \quad \nu = \frac{1}{n}\sum_{j=1}^{n} \delta_{y_j}$$

Distance between μ and ν...

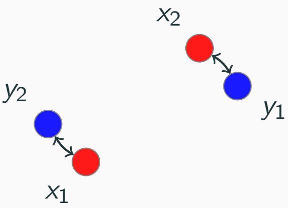
$$\frac{1}{n^2}\sum_{ij} \text{dist}(x_i, y_j)?$$

A Geometric Motivation

Setting: Probability measures $\mathcal{P}(\mathcal{X})$ on a metric space $(\mathcal{X}, \text{dist})$.

Goal

Build a metric on $\mathcal{P}(\mathcal{X})$ consistent with the geometry of $(\mathcal{X}, \text{dist})$.



$$\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}, \quad \nu = \frac{1}{n}\sum_{j=1}^{n} \delta_{y_j}$$

Distance between μ and ν...

$$\min_{\sigma \text{ perm.}} \frac{1}{n}\sum_{i} \text{dist}(x_i, y_{\sigma(i)})?$$

A Geometric Motivation

Setting: Probability measures $\mathcal{P}(\mathcal{X})$ on a metric space $(\mathcal{X}, \mathrm{dist})$.

Goal

Build a metric on $\mathcal{P}(\mathcal{X})$ consistent with the geometry of $(\mathcal{X}, \mathrm{dist})$.

$\mu, \nu \in \mathcal{P}(\mathcal{X})$

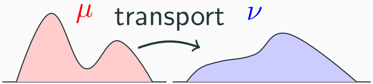
Distance between μ and ν...

?

Origin and Ramifications

Monge Problem (1781)

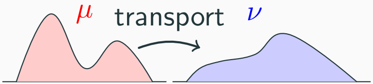
Move dirt from one configuration to another with least effort



Origin and Ramifications

Monge Problem (1781)

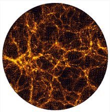
Move dirt from one configuration to another with least effort

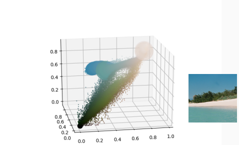


Strong modelization power:

- probability distribution, empirical distribution
- weighted undistinguishable particles
- density of a gas, a crowd, cells...

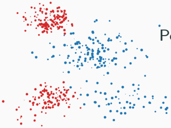
Early universe
(Brenier *et al.* '08)





Color histograms (Delon *et al.*)

Crowd motion
(Roudneff *et al.*, '12)



Point clouds

Part 1: Qualitative Overview

- **classical theory**
- **selection of properties and variants**

Part 2: Algorithms and Approximations

- **entropic regularization**
- **computational aspects**
- **statistical aspects**

Outline

Outline

Main Theoretical Facts

A Glimpse of Applications

Unbalanced Optimal Transport

Differentiability

Computation and Approximation

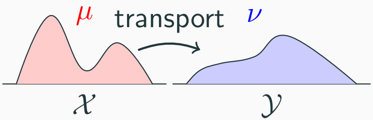
Density Fitting

Losses between Probability Measures

Definition

Ingredients

- Metric spaces $\mathcal{X}$ and $\mathcal{Y}$ (complete, separable)
- Cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ (lower bounded, lsc)
- Probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$



Transport map

Definition (pushforward)

Let $T : \mathcal{X} \to \mathcal{Y}$ be a map. The *pushforward measure* of μ by T is characterized by

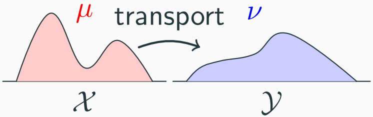
$$T_{\#}\mu(B) = \mu(T^{-1}(B)) \quad \text{for all } B \subset \mathcal{Y}.$$

If X is a random variable such that $\mathrm{Law}(X) = \mu$, then

$$\mathrm{Law}(T(X)) = T_{\#}\mu.$$

Ingredients

- Metric spaces $\mathcal{X}$ and $\mathcal{Y}$ (complete, separable)
- Cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ (lower bounded, lsc)
- Probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$



Definition (Monge problem)

$$\inf_{T:\mathcal{X}\to\mathcal{Y}} \left\{ \int_{\mathcal{X}} c(x, T(x))\mathrm{d}\mu(x) \; ; \; T_{\#}\mu = \nu \right\}$$

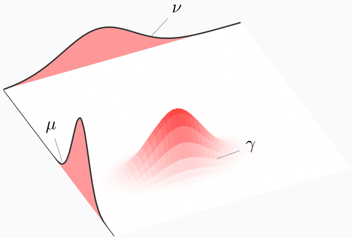
$\rightsquigarrow$ in some cases: no solution, no feasible point...

Transport Plans

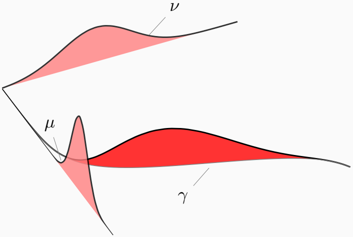
Definition (Set of transport plans)

Positive measures on $\mathcal{X} \times \mathcal{Y}$ with specified marginals :

$$\Pi(\mu, \nu) := \left\{ \gamma \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y}) : \mathrm{proj}^x_\# \gamma = \mu, \ \mathrm{proj}^y_\# \gamma = \nu \right\}$$



Product coupling
$\gamma = \mu \otimes \nu$



Deterministic coupling
$\gamma = (\mathrm{Id} \times T)_\# \mu$

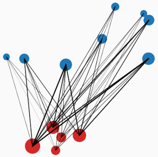
- Generalizes permutations, bistochastic matrices, matchings
- convex, weakly compact

Transport Plans

Definition (Set of transport plans)

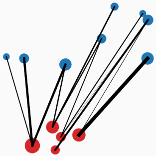
Positive measures on $\mathcal{X} \times \mathcal{Y}$ with specified marginals :

$$\Pi(\mu, \nu) := \left\{ \gamma \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y}) : \mathrm{proj}^x_{\#} \gamma = \mu, \ \mathrm{proj}^y_{\#} \gamma = \nu \right\}$$



Product coupling

$\gamma = \mu \otimes \nu$



Cycle-free coupling

- Generalizes permutations, bistochastic matrices, matchings
- convex, weakly compact

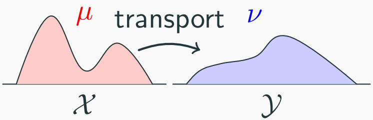
Definition

Ingredients

- Metric spaces $\mathcal{X}$ and $\mathcal{Y}$ (complete, separable)
- Cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ (lower bounded, lsc)
- Probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$

Definition (Optimal transport problem)

$$C(\mu, \nu) := \min_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\gamma(x, y)$$



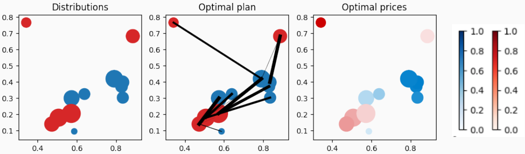
Probabilistic view: $\min_{(X,Y)} \{\mathbb{E}[c(X, Y)] : X \sim \mu \text{ and } Y \sim \nu\}$

Duality

Theorem (Kantorovich duality)

$$\min_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\gamma(x,y) \qquad \text{(Primal)}$$
$$=$$
$$\max_{\substack{\phi \in L^1(\mu) \\ \psi \in L^1(\nu)}} \left\{ \int_{\mathcal{X}} \phi(x) d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y) : \phi(x) + \psi(y) \leq c(x,y) \right\} \qquad \text{(Dual)}$$

Economy: (Primal) centralized vs. (Dual) externalized planification

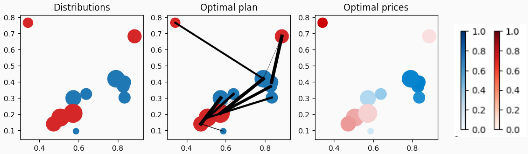


Duality

Theorem (Kantorovich duality)

$$\min_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\gamma(x,y) \qquad \text{(Primal)}$$
$$=$$
$$\max_{\substack{\phi \in L^1(\mu)\\ \psi \in L^1(\nu)}} \left\{ \int_{\mathcal{X}} \phi(x)d\mu(x) + \int_{\mathcal{Y}} \psi(y)d\nu(y) : \phi(x)+\psi(y) \leq c(x,y) \right\} \qquad \text{(Dual)}$$

Economy: (Primal) centralized vs. (Dual) externalized planification



At optimality

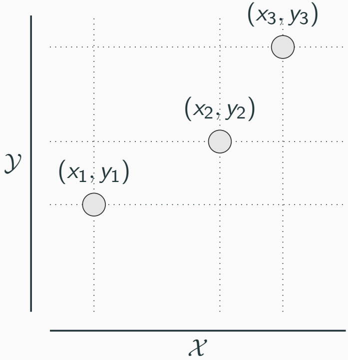
- $\phi(x)+\psi(y) = c(x,y)$ for γ almost every (x,y)
- γ is concentrated on a "c-cyclically monotone" set

Generalizing Convex Analysis Tools (I)

Definition (Cyclical monotonicity)

$\Gamma \subset \mathcal{X} \times \mathcal{Y}$ is c-cyclical monotone iff for all $(x_i, y_i)_{i=1}^n \in \Gamma^n$

$$\sum_{i=1}^{n} c(x_i, y_i) \leq \sum_{i=1}^{n} c(x_i, y_{\sigma(i)}) \text{ for all permutation } \sigma.$$

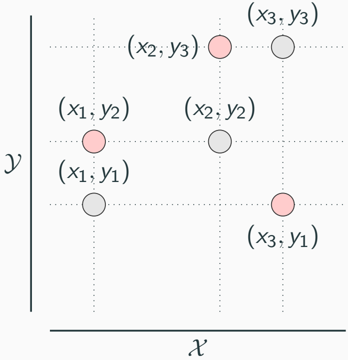


Generalizing Convex Analysis Tools (I)

Definition (Cyclical monotonicity)

$\Gamma \subset \mathcal{X} \times \mathcal{Y}$ is c-cyclical monotone iff for all $(x_i, y_i)_{i=1}^n \in \Gamma^n$

$$\sum_{i=1}^{n} c(x_i, y_i) \leq \sum_{i=1}^{n} c(x_i, y_{\sigma(i)}) \text{ for all permutation } \sigma.$$



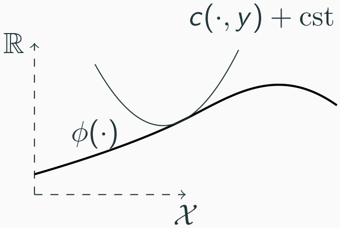
Generalizing Convex Analysis Tools (II)

Definition (c-conjugacy)

For $\mathcal{X} = \mathcal{Y}$ and $c : \mathcal{X}^2 \to \mathbb{R}$ symmetric :

$$\phi^c(y) := \inf_{x \in \mathcal{X}} c(x, y) - \phi(x)$$

A function ϕ is *c-concave* iff there exists ψ such that $\phi = \psi^c$.



Generalizing Convex Analysis Tools (II)

Definition (c-conjugacy)

For $\mathcal{X} = \mathcal{Y}$ and $c : \mathcal{X}^2 \to \mathbb{R}$ symmetric :

$$\phi^c(y) := \inf_{x \in \mathcal{X}} c(x, y) - \phi(x)$$

A function ϕ is *c-concave* iff there exists ψ such that $\phi = \psi^c$.

- on $\mathbb{R}^d$, for $c(x, y) = x \cdot y$: ψ c-concave $\Leftrightarrow$ ψ concave;
- for all ϕ, $\phi^{ccc} = \phi^c$;
- consequence :

$$C(\mu, \nu) = \max_{\phi \text{ } c\text{-concave}} \left\{ \int_{\mathcal{X}} \phi(x) d\mu(x) + \int_{\mathcal{Y}} \phi^c(y) d\nu(y) \right\} \qquad \text{(Dual)}$$

Special Cases

- real line ($\mathcal{X} = \mathcal{Y} = \mathbb{R}$)
- distance cost ($c = \text{dist}$)
- quadratic cost ($c = \| \cdot - \cdot \|^2$)

Real Line

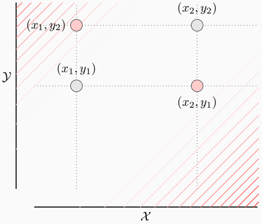
Theorem (Monotone Rearrangement)

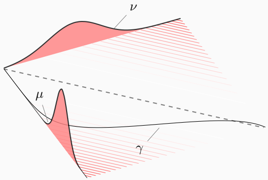
If $\mu, \nu \in \mathcal{P}(\mathbb{R})$ and $c(x, y) = h(y - x)$ with h strictly convex:

- unique optimal transport plan γ^*
- denoting $F^{[-1]}$ the quantile functions:

$$C(\mu, \nu) = \int_0^1 h(F_\mu^{[-1]}(s) - F_\nu^{[-1]}(s))ds$$

"Proof". Here, c-cyclically monotone $\Leftrightarrow$ increasing graph. □



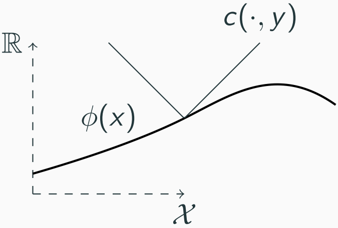


Distance Cost

If $\mathcal{X} = \mathcal{Y}$ and $c(x, y) = \text{dist}(x, y)$

- ϕ c-concave $\Leftrightarrow$ ϕ 1-Lipschitz
- $\phi^c(y) = \inf_x d(x, y) - \phi(x) = -\phi(y)$
- consequence :

$$C(\mu, \nu) = \max_{\phi \text{ 1-Lipschitz}} \left\{ \int_{\mathcal{X}} \phi(x) d(\mu - \nu)(x) \right\} \tag{Dual}$$



Quadratic Cost

Reformulation

- $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ with finite moments of order 2
- cost $c(x, y) := \frac{1}{2}\|y - x\|^2$
- note that $c(x, y) = (\|x\|^2 + \|y\|^2)/2 - x \cdot y$, thus solve:

$$\max_{\gamma \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} \langle x, y \rangle d\gamma(x, y) : \gamma \in \Pi(\mu, \nu) \right\} \qquad \text{(Primal)}$$

Theorem (Brenier '87)

(i) At optimality, $\operatorname{spt} \gamma \subset \partial \phi$, *where* $\phi : \mathbb{R}^n \to \mathbb{R}$ *convexe.*
(ii) If μ *has a density,* $T = \nabla \phi$ *is the unique optimal map.*

"Proof". (i) $\phi(x) + \phi^*(y) = x \cdot y$, γ-a.e (ii) $\nabla \phi$ defined $\mathcal{L}$-a.e.

Transport of Covariance

Whenever the dual potential ϕ is quadratic: transport of covariance

Theorem (Affine transport map)

Let $c(x, y) = \frac{1}{2}\|y - x\|^2$ on $\mathbb{R}^d$ and let $A, B \in S_+^d$. It holds

$$\min_{\substack{\mathrm{cov}(\mu)=A \\ \mathrm{cov}(\nu)=B}} C(\mu, \nu) = \mathrm{dist}_b(A, B)^2.$$

- $\mathrm{dist}_b(A, B)^2 = \mathrm{tr}\, A + \mathrm{tr}\, B - 2\,\mathrm{tr}(A^{\frac{1}{2}} B A^{\frac{1}{2}})^{\frac{1}{2}}$ Bures metric on S_+^d
- Transport map $T = A^{-1} \# B$ ($\cdot\#\cdot$ geometric mean).

[Refs]:

Bhatia, Jain, Lim (2017). *On the Bures-Wasserstein distance between positive definite matrices*

Wasserstein distance

Definition

Let $\text{dist} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a metric. The Wasserstein distance is

$$W_2(\mu,\nu) := \left\{ \min_{\gamma \in \mathcal{M}_+(\mathcal{X}^2)} \int_{X^2} \text{dist}(x,y)^2 d\gamma(x,y) : \gamma \in \Pi(\mu,\nu) \right\}^{\frac{1}{2}}$$

- W_2 metrizes weak convergence + 2-nd order moments
- if $(\mathcal{X}, \text{dist})$ is a geodesic space, so is $(\mathcal{P}(\mathcal{X}), W_2)$
- similar definition for W_p with $p \geq 1$

Constant speed geodesic for W_2 on $\mathcal{P}(\mathbb{R})$

$$((1-t)\mathrm{Id} + tT)_{\#}\mu$$

Summing up

First Properties

- rich duality with concepts from convex analysis
- rich structure in specific cases

Properties of the distance W_2 on $\mathbb{R}^d$

- optimal plans supported on $\partial\phi$ with $\phi : \mathbb{R}^d \to \mathbb{R}$ convex
- the space $(\mathcal{P}(\mathbb{R}^d), W_2)$ is a complete geodesic space
- some explicit cases (real line, linear maps)

Outline

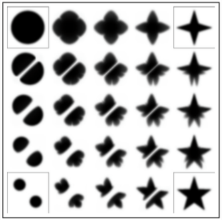
Histogram & shapes processing

Color transfer

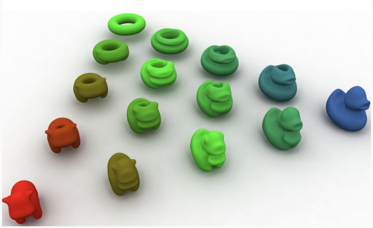
 +  =  or 

color | target | OT | unbalanced OT

Barycenters



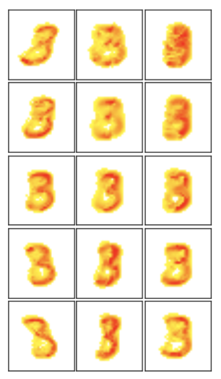
(Benamou et al. '15)



(Solomon et al. '15)

Histograms and shape processing

- compute barycenter $\bar{\mu}$ of a family $(\mu_k)_k$
- transport maps from $\bar{\mu}$ gives a Hilbertian parameterization
- apply your favorite data analysis method!



Three PCs from the MNIST dataset (Seguy and Cuturi, 2015)

[Refs]:
Seguy, Cuturi (2015). *Principal Geodesic Analysis for Probability Measures [...].*
Wang, Slepcev, Basu, Ozolek, Rohde (2012). *A linear optimal transportation framework.*

Loss for regression:

Learn predictor $f_\theta : \mathcal{X} \to \mathcal{Y} := \mathcal{P}(\{1, \dots, k\})$

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{(X,Y)} \left[W_2^2(f_\theta(X), Y) \right].$$

(a) **Flickr user tags**: zoo, run, mark; **our proposals**: running, summer, fun; **baseline proposals**: running, country, lake.

(b) **Flickr user tags**: travel, architecture, tourism; **our proposals**: sky, roof, building; **baseline proposals**: art, sky, beach.

(c) **Flickr user tags**: spring, race, training; **our proposals**: road, bike, trail; **baseline proposals**: dog, surf, bike.

Predict probability over tags from an image (Frogner et al. 2015)

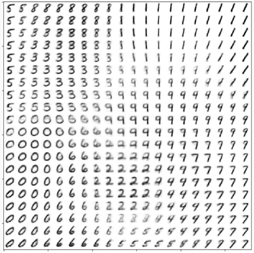
[Refs]:
Frogner, Zhang, Mobahi, Araya, Poggio (2015). *Learning with a Wasserstein loss.*

Machine learning

Loss for density fitting:
Given $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$,
learn map $f_\theta : \mathcal{X} \to \mathcal{Y}$

$$\min_{\theta \in \mathbb{R}^d} W_2^2((f_\theta)_\# \mu, \nu)$$

$\Rightarrow$ more in part II.



Generating figure from MNIST (Genevay et al. 2018)

[Refs]:
Genevay, Peyré, Cuturi (2017). *Learning Generative Models with Sinkhorn Divergences.*

Often needed

- differentiability properties
- unbalanced optimal transport

Outline

Unbalanced OT

Optimal Transport has an intrinsic constraint:

$$\mu(\mathcal{X}) = \nu(\mathcal{Y})$$

What if $\mu(\mathcal{X}) \neq \nu(\mathcal{Y})$?

Unbalanced OT

Optimal Transport has an intrinsic constraint:

$$\mu(\mathcal{X}) = \nu(\mathcal{Y})$$

What if $\mu(\mathcal{X}) \neq \nu(\mathcal{Y})$?

Unbalanced Optimal Transport

- often comes up in applications
- normalization is generally a poor choice
- are there approaches that stand out?

Unbalanced OT

Optimal Transport has an intrinsic constraint:

$$\mu(\mathcal{X}) = \nu(\mathcal{Y})$$

What if $\mu(\mathcal{X}) \neq \nu(\mathcal{Y})$?

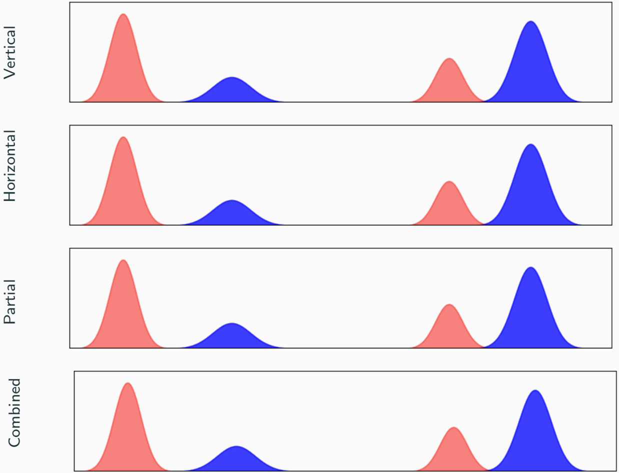
Unbalanced Optimal Transport

- often comes up in applications
- normalization is generally a poor choice
- are there approaches that stand out?

Strategy

- preserve key properties of optimal transport
- combine *horizontal* (transport) and *vertical* (linear) geometries

Vertical/Horizontal



Optimal Partial Transport

Setting: $\mu \in \mathcal{M}_+(\mathcal{X})$ and $\nu \in \mathcal{M}_+(\mathcal{Y})$ nonnegative measures.

Variational Problem

Choose $0 < m \leq \min\{\mu(\mathcal{X}), \nu(\mathcal{Y})\}$ and solve

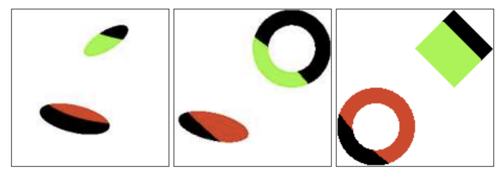
$$\min_{\gamma \in \mathcal{M}_+(\mathbb{R}^{2d})} \int c(x,y) d\gamma(x,y)$$

$$\text{subject to} \quad \pi^x_\# \gamma \leq \mu$$

$$\pi^y_\# \gamma \leq \nu$$

$$\gamma(\mathcal{X} \times \mathcal{Y}) = m$$

- old & simple modification of the original problem
- "equivalent" formulations: dynamic, entropy-transport
- alternatively, add a sink/source reachable at a certain cost



Optimal partial transport in 2D (Benamou et al. 2015)

[Refs]:
Benamou, Carlier, Cuturi, Nenna, Peyré (2015) *Iterative Bregman Projections for Regularized Transportation Problems*

Wasserstein Fisher-Rao a.k.a. Hellinger-Kantorovich

Setting: $\mu \in \mathcal{M}_+(\mathcal{X})$ and $\nu \in \mathcal{M}_+(\mathcal{Y})$ nonnegative measures.

Definition

The natural generalization of W_2 to this setting is

$$\widehat{W_2^2}(\mu, \nu) = \min_{\gamma \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})} \mathrm{KL}(\pi^x_\# \gamma | \mu) + \mathrm{KL}(\pi^y_\# \gamma | \nu) + \int c_\ell(x, y) d\gamma(x, y)$$

where $c_\ell(x, y) = -\log \cos^2(\min\{\mathrm{dist}(x, y), \pi/2\})$.

where KL is the Kullback-Leibler divergence, defined if $\mu_1 \ll \mu_2$ as

$$\mathrm{KL}(\mu_1 | \mu_2) = \int \log\left(\frac{d\mu_1}{d\mu_2}\right) \mathrm{d}\mu_1 - \mu_1(\mathcal{X}) + \mu_2(\mathcal{X})$$

Wasserstein Fisher-Rao a.k.a. Hellinger-Kantorovich

Setting: $\mu \in \mathcal{M}_+(\mathcal{X})$ and $\nu \in \mathcal{M}_+(\mathcal{Y})$ nonnegative measures.

Definition

The natural generalization of W_2 to this setting is

$$\widehat{W}_2^2(\mu,\nu) = \min_{\gamma \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \mathrm{KL}(\pi^x_\# \gamma | \mu) + \mathrm{KL}(\pi^y_\# \gamma | \nu) + \int c_\ell(x,y) d\gamma(x,y)$$

where $c_\ell(x,y) = -\log\cos^2(\min\{\mathrm{dist}(x,y), \pi/2\})$.

Main properties

- geodesic space, Riemannian-like structure
- growth and displacement intertwined
- various explicit formulations: lifted problem, dynamic problem

[Refs]:
Liero, Mielke, Savaré (2015). *Optimal Entropy-Transport Problems and a New Hellinger–Kantorovich Distance [...]*
Kondratyev, Monsaingeon, Vorotnikov (2015). *A New Optimal Transport Distance on the Space of [...] Measures.*
Chizat, Peyré, Schmitzer, Vialard (2015). *An Interpolating Distance between Optimal Transport and Fisher-Rao.*

Outline

Vertical Perturbations

Reminder

Optimal transport between $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ with cost c:

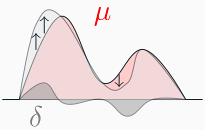
$$C(\mu, \nu) = \sup_{(\varphi,\psi) \text{ admissible}} \int_{\mathbb{R}^d} \varphi \, d\mu + \int_{\mathbb{R}^d} \psi \, d\nu$$

Vertical Perturbations

Reminder

Optimal transport between $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ with cost c:

$$C(\mu, \nu) = \sup_{(\varphi,\psi) \text{ admissible}} \int_{\mathbb{R}^d} \varphi \, d\mu + \int_{\mathbb{R}^d} \psi \, d\nu$$



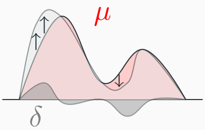
Perturbed marginal: $\mu + \epsilon\delta$

Vertical Perturbations

Reminder

Optimal transport between $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ with cost c:

$$C(\mu, \nu) = \sup_{(\varphi,\psi) \text{ admissible}} \int_{\mathbb{R}^d} \varphi \, d\mu + \int_{\mathbb{R}^d} \psi \, d\nu$$



Perturbed marginal: $\mu + \epsilon\delta$

Vertical (linear) derivative

Let δ a signed measure with $\int \delta = 0$. If optimal φ unique,

$$\frac{d}{d\epsilon} C(\mu + \epsilon\delta, \nu)|_{\epsilon=0} = \int_{\mathbb{R}^d} \varphi \, d\delta$$

Horizontal Perturbations

Reminder

Optimal transport between $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ with cost c:

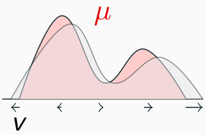
$$C(\mu, \nu) = \inf_{\gamma \text{ admissible}} \int_{(\mathbb{R}^d)^2} c(x, y)\, d\gamma(x, y)$$

Horizontal Perturbations

Reminder

Optimal transport between $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ with cost c:

$$C(\mu, \nu) = \inf_{\gamma \text{ admissible}} \int_{(\mathbb{R}^d)^2} c(x, y)\, d\gamma(x, y)$$



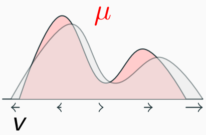
Perturbed cost: $c(x + \epsilon v(x), y) \approx c(x, y) + \epsilon \nabla_x c(x, y) \cdot v(x)$

Horizontal Perturbations

Reminder

Optimal transport between $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ with cost c:

$$C(\mu,\nu) = \inf_{\gamma \text{ admissible}} \int_{(\mathbb{R}^d)^2} c(x,y)\, d\gamma(x,y)$$



Perturbed cost: $c(x+\epsilon v(x), y) \approx c(x,y) + \epsilon \nabla_x c(x,y) \cdot v(x)$

Horizontal (Wasserstein) perturbation

Let $v : \mathbb{R}^d \to \mathbb{R}^d$ a curl free map. If optimal γ unique,

$$\frac{d}{d\epsilon} C((\mathrm{id} + \epsilon v)_{\#}\mu, \nu)|_{\epsilon=0} = \int_{(\mathbb{R}^d)^2} \nabla_x c(x,y) \cdot v(x) d\gamma(x,y).$$

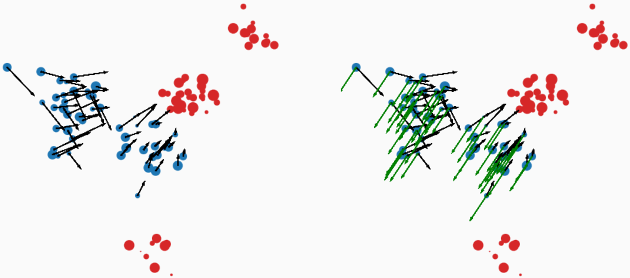
Special case of W_2

Setting: quadratic cost on $\mathbb{R}^d$, $v : \mathbb{R}^d \to \mathbb{R}^d$ a curl free map.

Differentiability of W_2

If unique optimal transport plan γ, then

$$\frac{d}{d\epsilon}\frac{1}{2}W_2^2((\mathrm{id}+\epsilon v)_{\#}\mu,\nu)|_{\epsilon=0} = \int_{(\mathbb{R}^d)^2}(x-y)\cdot v(x)d\gamma(x,y)$$



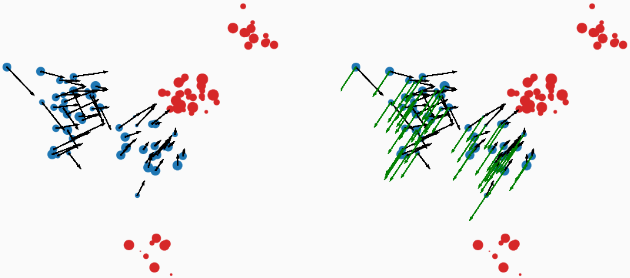
Special case of W_2

Setting: quadratic cost on $\mathbb{R}^d$, $v : \mathbb{R}^d \to \mathbb{R}^d$ a curl free map.

Differentiability of W_2

If unique optimal transport plan γ, then

$$\frac{d}{d\epsilon} \frac{1}{2} W_2^2((\mathrm{id} + \epsilon v)_{\#}\mu, \nu)|_{\epsilon=0} = \int_{(\mathbb{R}^d)^2} (x - y) \cdot v(x) d\gamma(x, y)$$



Next section: regularized W_2, always differentiable.

Outline

Main Theoretical Facts

A Glimpse of Applications

Unbalanced Optimal Transport

Differentiability

Computation and Approximation

Density Fitting

Losses between Probability Measures

Discrete Optimal Transport

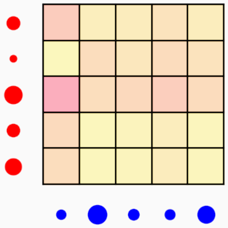
Discrete Setting

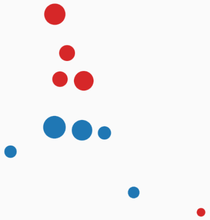
- Discrete measures $\mu = \sum_{i=1}^{n} p_i \delta_{x_i}$, $\nu = \sum_{j=1}^{n} q_j \delta_{y_i}$.
- Cost matrix $C_{i,j} = c(x_i, y_j)$

Linear Program

$$\min_{\gamma \in \mathcal{S}(p,q)} \sum_{i,j} C_{i,j} \gamma_{i,j}$$

where $\mathcal{S}(p,q) = \{\gamma \in \mathbb{R}_+^{n \times m} \; ; \; p_i = \sum_j \gamma_{i,j} \text{ and } q_j = \sum_i \gamma_{i,j}\}$.





Discrete Optimal Transport

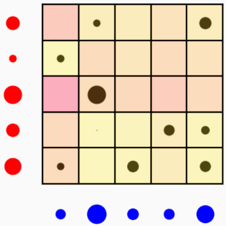
Discrete Setting

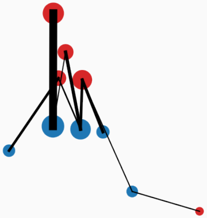
- Discrete measures $\mu = \sum_{i=1}^{n} p_i \delta_{x_i}$, $\nu = \sum_{j=1}^{n} q_j \delta_{y_i}$.
- Cost matrix $C_{i,j} = c(x_i, y_j)$

Linear Program

$$\min_{\gamma \in \mathcal{S}(p,q)} \sum_{i,j} C_{i,j} \gamma_{i,j}$$

where $\mathcal{S}(p,q) = \{\gamma \in \mathbb{R}_+^{n \times m} \; ; \; p_i = \sum_j \gamma_{i,j} \text{ and } q_j = \sum_i \gamma_{i,j}\}$.



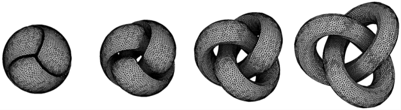


Exact solvers

Algorithm	Setting	Complexity
Network simplex	–	$\tilde{O}(n^3)$
Hungarian	bistochastic	$O(n^3)$
Auction	$C_{i,j}$ integers	$O(n^3)$

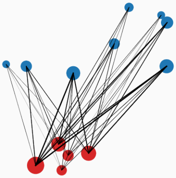
Efficient methods in $\mathbb{R}^2$ or $\mathbb{R}^3$

- semi-discrete solver based on Laguerre cells
- minimizing Benamou-Brenier functional (finite elements)
- resolution of Monge-Ampère equation (finite elements)

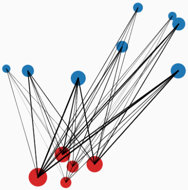


(Levy, 15')

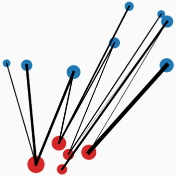
Approximate Solver



Product coupling

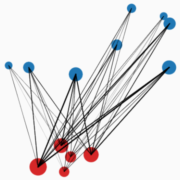


$0 < \beta^{-1} < \infty$

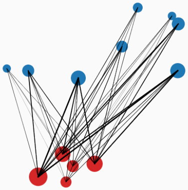


Optimal coupling

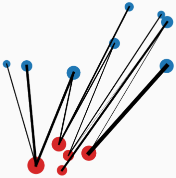
Approximate Solver



Product coupling



$0 < \beta^{-1} < \infty$

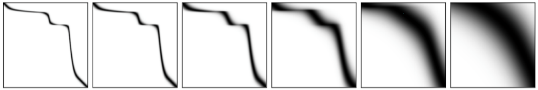


Optimal coupling

Entropic regularization (Cuturi '13)

$$\min_{\gamma \in \mathcal{S}(p,q)} \sum_{i,j} C_{i,j}\gamma_{i,j} + \beta^{-1}\,\mathrm{KL}(\gamma, \mu \otimes \nu)$$

where $\mathrm{KL}(a, b) = \sum_i a_i(\log(a_i/b_i) - 1)$.



Optimal transport plan as β decreases (Nenna et al. '15)

Sinkhorn's algorithm

Proposition (Optimality Condition)

Define the kernel $K_{i,j} = \exp(-\beta \cdot C_{i,j})$. There exists $a, b \in \mathbb{R}_+^n$ such that at optimality:

$$\gamma^*_{i,j} = a_i K_{i,j} b_j$$

Sinkhorn's algorithm

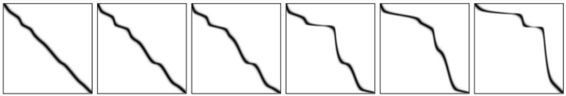
Proposition (Optimality Condition)

Define the kernel $K_{i,j} = \exp(-\beta \cdot C_{i,j})$. There exists $a, b \in \mathbb{R}^n_+$ such that at optimality:

$$\gamma^*_{i,j} = a_i K_{i,j} b_j$$

Sinkhorn's Algorithm

1. initialize $b = (1, \dots, 1)$ and repeat until convergence
   1.1 $a \leftarrow p \oslash (Kb)$ **[rescale rows]**
   1.2 $b \leftarrow q \oslash (K^T a)$ **[rescale columns]**
2. return $\gamma^*_{i,j} = a_i K_{i,j} b_j$.



Evolution of $(a_i K_{i,j} b_j)_{i,j}$, in (Benamou et al. 2015)

Complexity Results

One iteration

- matrix/vector product in $O(n^2)$ (sometimes better)
- highly parallelizable on GPUs

Solving entropy-regularized OT

- Linear convergence of a, b in *Hilbert metric*
- ϵ-accurate solution in $O(n^2 \log(1/\epsilon))$
- stochastic algorithms (see later), accelerations

Complexity Results

One iteration

- matrix/vector product in $O(n^2)$ (sometimes better)
- highly parallelizable on GPUs

Solving entropy-regularized OT

- Linear convergence of a, b in *Hilbert metric*
- ϵ-accurate solution in $O(n^2 \log(1/\epsilon))$
- stochastic algorithms (see later), accelerations

Solving OT

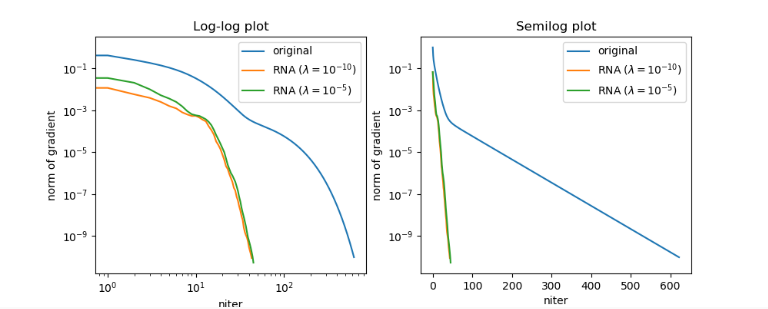
- Sinkhorn's algorithm allows to build an ϵ-accurate feasible transport plan in $\tilde{O}(n^2/\epsilon^2)$
- best bound in $\tilde{O}(n^2/\epsilon)$ (active research)

[Refs (see ref therein)]:
Lin, Ho, Jordan (2019). *On Efficient Optimal Transport [...]*
Dvurechensky, Gasnikov, Kroshnin (2018). *Computational Optimal Transport [...]*
Blanchet, Jambulapati, Kent, Sidford (2018). *Towards Optimal Running Times for Optimal Transport*

Fast In Practice

Overrelaxation and Nonlinear Acceleration

- average/extrapolate the iterates, possibly adaptively
- typical fixed-point algorithms accelerations
- preserves the iteration complexity and parallelizable

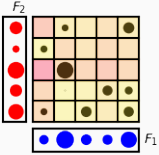


[Refs]:
Scieur, D'Aspremont, Bach (2016). *Regularized Nonlinear Acceleration*
Thibault, Chizat, Dossal, Papadakis (2017). *Overrelaxed Sinkhorn Algorithm for Regularized Optimal Transport*

Generalization

Solving barycenters, unbalanced OT, inverse problems...

$$\min \sum C_{i,j}\gamma_{i,j} + F_1(\gamma \cdot 1_n) + F_2(\gamma^T \cdot 1_n) + \beta^{-1}\,\mathrm{KL}(\gamma, \mu \otimes \nu)$$



Scaling iterates (alternate maximization on the dual)

1. initialize $b = 1_n$ and repeat until convergence
   1.1 $a \leftarrow \mathrm{prox}_{F_1}(Kb) \oslash (Kb)$ **[descent on rows]**
   1.2 $b \leftarrow \mathrm{prox}_{F_2}(K^T a) \oslash (K^T a)$ **[descent on columns]**
2. return $\gamma^*_{i,j} = a_i K_{i,j} b_j$.

$$\mathrm{prox}_F(\bar{s}) := \arg\min_s \{F(s) + \epsilon\,\mathrm{KL}(s|\bar{s})\}$$

[Refs]:
Chizat, Peyré, Schmitzer, Vialard (2016). *Scaling algorithms for unbalanced optimal transport problems*

Outline

Density Fitting

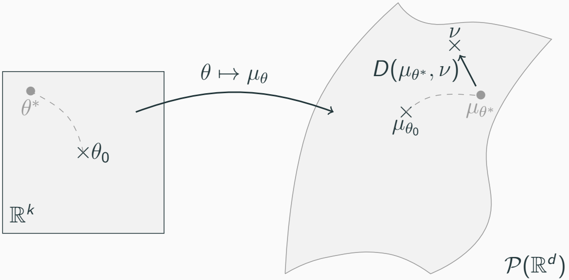
Ingredients

- a parametric family $\theta \in \mathbb{R}^k \to \mu_\theta \in \mathcal{P}(\mathbb{R}^d)$
- a target $\nu \in \mathcal{P}(\mathbb{R}^d)$

General problem

Chose a loss $D : \mathcal{P}(\mathbb{R}^d)^2 \to [0, \infty]$ and solve

$$\min_{\theta \in \mathbb{R}^k} D(\mu_\theta, \nu).$$



Examples (I)

Statistical inference

- μ_θ is an exponential family
- ν is known through samples $\hat{\nu} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$

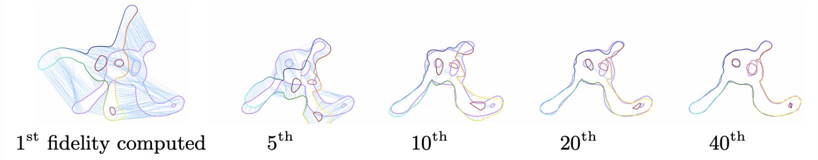
Choosing $D = \mathrm{KL}$ gives the maximum likelihood estimator:

$$\min_{\theta\in\mathbb{R}^k} \mathrm{KL}(\nu|\mu_\theta) \rightsquigarrow \min_{\theta\in\mathbb{R}^k} \mathbb{E}_{x\sim\nu}\left[-\log\left(\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mathcal{L}}(x)\right)\right]$$
$$\rightsquigarrow \max_{\theta\in\mathbb{R}^k} \frac{1}{n}\sum_{i=1}^n \log\left(\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mathcal{L}}(x_i)\right)$$

Examples (II)

Shapes matching

- μ_θ is $(f_\theta)_\# \mu$ where f_θ is a smooth deformation of $\mathbb{R}^d$ and μ a reference shape
- ν is a target shape
- goal : find a smooth deformation f_{θ^*} from μ to ν



(Feydy et al. '17)

[Refs]:

Feydy, Charlier, Vialard, Peyré (2017). *Optimal Transport for Diffeomorphic Registration*

Examples (III)

Generative modeling

- μ_θ is $(f_\theta)_\# \mu$ where f_θ is a neural network and μ is a simple distribution (Gaussian) on a low dimensional space
- ν is a target distribution observed through samples
- goal : generate new samples from ν using $f_\theta(X)$, $X \sim \mu$

Random bedrooms (Arjovsky et al. '14)

[Refs]:
Arjovsky, Chintala, Bottou (2014). *Wasserstein GAN*
Genevay, Peyré, Cuturi (2017). *Learning Generative Models with Sinkhorn Divergences*

Properties Needed

Gradient-based minimization

Choose step-size η, start from $\theta^{(0)}$ and (ideally) define

$$\theta^{(k+1)} = \theta^{(k)} - \eta \nabla_\theta [D(\mu_{\theta^{(k)}}, \nu)].$$

Requires

- differentiability
- low computational complexity
- low sample complexity
- to incorporate geometry

Outline

Classes of losses

- φ-divergence (includes KL, Hellinger, TV,...)
- integral probability metrics (includes MMD, W_1)
- Sinkhorn divergences
- Wasserstein loss

φ-divergences

Definition

Let $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ be a convex function with $\varphi(1) = 0$ and superlinear (to simplify):

$$D_\varphi(\mu, \nu) = \begin{cases} \int_{\mathbb{R}^d} \varphi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right) \mathrm{d}\nu(x) & \text{if } \mu \ll \nu \\ +\infty & \text{otherwise} \end{cases}$$

- pointwise comparison of the density (no geometry)
- recovers KL when $\varphi(s) = s \log(s)$
- computational cost $O(n)$ (on a discrete space)
- estimation: depends on the class of density considered

Integral Probability Metrics

Definition

Let $\mathcal{F}$ a subset of functions $\mathbb{R}^d \to \mathbb{R}$ that contains 0 and define

$$D_{\mathcal{F}}(\mu,\nu) = \sup_{f\in\mathcal{F}} \int_{\mathbb{R}^d} f(x)\mathrm{d}(\mu-\nu)(x)$$

It $\mathcal{F}$ is the set of 1-Lipschitz functions then $D_{\mathcal{F}} = W_1$.

Maximum Mean Discrepancy

Let $\mathcal{F}$ be the 1-ball of a RKHS $\mathcal{H}$ with kernel k, then

$$D_{\mathcal{F}}(\mu,\nu) = \|\mu-\nu\|_k^2 \quad \text{where} \quad \|\mu\|_k^2 := \iint k(x,y)\mathrm{d}\mu(x)\otimes\mathrm{d}\mu(y)$$

- computational cost $O(n^2)$
- sample complexity : accuracy in $O(1/n)$

[Refs]:
Sriperumbudur et al.(2012). *On the Empirical Estimation of Integral Probability Metrics.*

Optimal Transport

We know the definition...

$$C(\mu, \nu) = \min_{\gamma \in \Pi(\mu,\nu)} \int c \mathrm{d}\gamma$$

- "a lot" of geometry
- computational cost: $O(n^3)$ or $O(n^2/\epsilon^2)$

Sample Complexity

- $|\mathbb{E}[W_2^2(\hat{\mu}_n, \hat{\nu}_n) - W_2^2(\mu, \nu)|] = O(n^{-2/d})$ for $d > 4$
- there exists better estimators if the density is assumed smooth

[Refs]:
Weed, Bach (2017). *Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance*
Weed, Berthet (2019). *Estimation of smooth densities in Wasserstein distance.*

Sinkhorn divergence

$$C_\beta(\mu,\nu) = \min_{\gamma\in\Pi(\mu,\nu)} \int c\mathrm{d}\gamma + \beta^{-1}\,\mathrm{KL}(\gamma|\mu\otimes\nu)$$

Definition

$$D_\beta(\mu,\nu)^2 := 2C_\beta(\mu,\nu) - C_\beta(\mu,\mu) - C_\beta(\nu,\nu)$$

Properties

- converges to $C(\mu,\nu)$ as $\beta\to\infty$
- converges to $\|\mu-\nu\|^2_{-c}$ as $\beta\to 0$
- it is positive definite if $e^{-\beta c}$ is a positive definite kernel

[Refs]:
Feydy, Séjourné, Vialard, Amari, Trouvé, Peyré (2018). *Interpolating between Optimal Transport and MMD using Sinkhorn Divergences*
Ramdas, Trillos, Cuturi, (2017). *On Wasserstein two-sample testing and related families of nonparametric tests.*

Sinkhorn divergence (II)

Proposition (sample complexity)

$$\mathbb{E}[|D_\beta(\mu,\nu) - D_\beta(\hat{\mu}_n, \hat{\nu}_n)|] = O(1/\sqrt{n})$$

Computational Properties

- computation through Sinkhorn algorithm in $O(n^2 \log(1/\epsilon))$
- or, with stochastic algorithms
  $\rightsquigarrow$ SGD achieves the $O(1/\sqrt{n})$ rate

$\rightsquigarrow$ the "constants" deteriorate as $\beta \to \infty$.

[Refs]:
Mena, Weed (2019). *Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem.*
Genevay, Chizat, Bach, Cuturi, Peyré (2018). *Sample Complexity of Sinkhorn divergences.*
Genevay, Cuturi, Peyré, Bach (2016). *Stochastic Optimization for Large-scale Optimal Transport*

Comparison

Loss D	computational compl.	sample compl.	geometry
φ-divergence	$-$	$-$	- -
MMD	$O(n^2)$	$O(n^{-1})$	-
Sinkhorn div.	$\tilde{O}(n^2 \log 1/\epsilon)$	$O(n^{-1/2})$	$+$
Wasserstein	$\tilde{O}(n^3)$ or $\tilde{O}(n^2/\epsilon^2)$	$O(n^{-2/d})$	$++$

- (disclaimer) these quantities are not exactly comparable
- ideally, deal with computational and statistical aspects jointly
- for density fitting, study ideally the complexity of the whole scheme

Part 1: qualitative overview

- **classical theory**
- **selection of properties and variants**

Part 2: Algorithms and Approximations

- **computational aspects**
- **entropic regularization**
- **statistical aspects**

[Some reference textbooks:]
- Peyré, Cuturi (2018). Computational Optimal Transport
- Santambrogio (2015). Optimal Transport for Applied Mathematicians
- Villani (2008). Optimal Transport, Old and New