Optimal transport for machine learning

Nicolas Courty[1], Rémi Flamary[2]
[1] IRISA, University of Bretagne Sud, France
[2] OCA, Lagrange, Université Nice - Sophia-Antipolis

Statlearn 2018

Planning of the day:

- (Morning) 3h of introductory course to optimal transport and related applications to machine learning
  - 1h20: introduction to computational optimal transport (nicolas)
  - small break
  - 1h20: applications to machine learning problems (rémi)
- (Afternoon) 3h of practical sessions in Python

Table of content

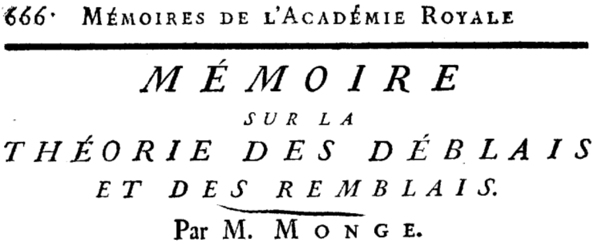
Optimal transport : introduction

What is optimal transport ?

The natural geometry of probability measures

Monge Kantorovich Koopmans Dantzig Brenier Otto McCann Villani

Nobel '75 Fields '10



666· MÉMOIRES DE L'ACADÉMIE ROYALE

MÉMOIRE

SUR LA

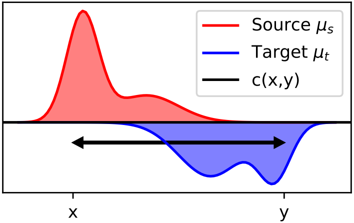
THÉORIE DES DÉBLAIS

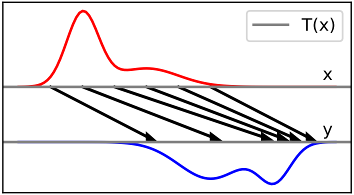
ET DES REMBLAIS.

Par M. MONGE.

Problem [Monge, 1781]

- How to move dirt from one place (déblais) to another (remblais) while minimizing the effort ?
- Find a mapping T between the two distributions of mass (transport).
- Optimize with respect to a displacement cost $c(x, y)$ (optimal).

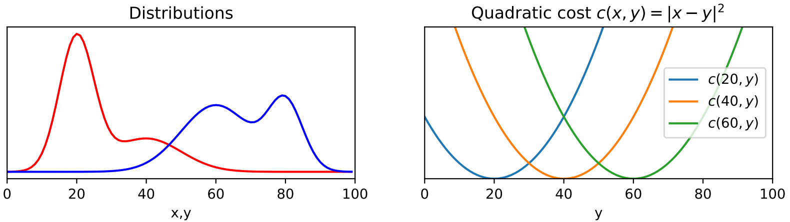




Problem [Monge, 1781]

- How to move dirt from one place (déblais) to another (remblais) while minimizing the effort ?
- Find a mapping T between the two distributions of mass (transport).
- Optimize with respect to a displacement cost $c(x, y)$ (optimal).

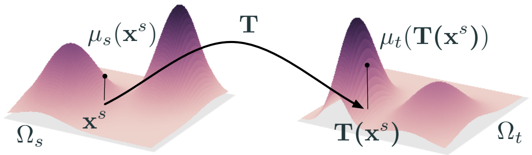
Optimal transport (Monge formulation)



- Probability measures μ_s and μ_t on and a cost function $c : \Omega_s \times \Omega_t \rightarrow \mathbb{R}^+$.
- The Monge formulation [Monge, 1781] aim at finding a mapping $T : \Omega_s \rightarrow \Omega_t$

$$\inf_{T\#\mu_s=\mu_t} \quad \int_{\Omega_s} c(\mathbf{x}, T(\mathbf{x}))\mu_s(\mathbf{x})d\mathbf{x} \tag{1}$$

What is $T\#\mu_s = \mu_t$?



- $T\#$ is the so called push forward operator
- it transfers measures from one space Ω_s to another space Ω_t
- it is equivalent to:

$$\mu_t(A) = \mu_s(T^{-1}(A))$$

$$\int_{\Omega_t} g(y) d\mu_t(y) = \int_{\Omega_s} g(T(x)) d\mu_s(x)$$

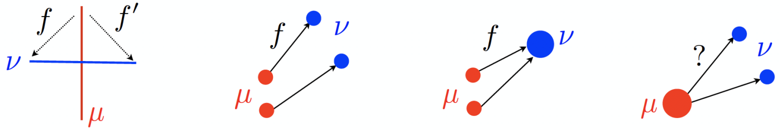
- for smooth measures $\mu_s = \rho(x)dx$ and $\mu_t = \eta(x)dx$

$$T\#\mu_s = \mu_t \equiv \rho(T(x))|\text{det}(\partial T(x))| = \eta(x)$$

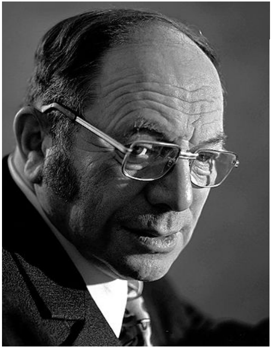
- a.k.a. change of variable formula

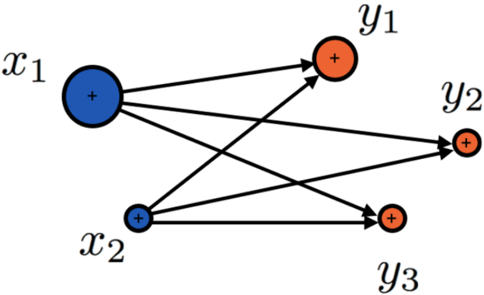
Solving for this push-forward operator is a non-convex optimization problem,

- for which existence is not guaranteed,
- nor unicity



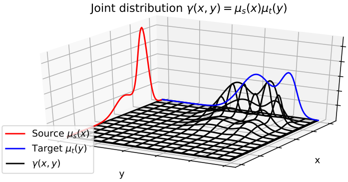
Note: [Brenier, 1991] proved existence and unicity of the Monge map for $c(x, y) = \|x - y\|^2$ and distributions with densities (i.e. continuous).

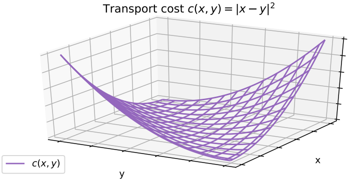




- Leonid Kantorovich (1912–1986), Economy nobelist in 1975, proposed a different formulation of the problem
- with applications mainly for ressource allocation problems

Optimal transport (Kantorovich formulation)





- The Kantorovich formulation [Kantorovich, 1942] seeks for a probabilistic coupling $\gamma \in \mathcal{P}(\Omega_s \times \Omega_t)$ between Ω_s and Ω_t:

$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \int_{\Omega_s \times \Omega_t} c(\mathbf{x}, \mathbf{y}) \boldsymbol{\gamma}(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}, \tag{2}$$

$$\text{s.t.} \quad \boldsymbol{\gamma} \in \mathcal{P} = \left\{ \boldsymbol{\gamma} \geq \mathbf{0}, \int_{\boldsymbol{\Omega_t}} \boldsymbol{\gamma}(\mathbf{x}, \mathbf{y}) \mathbf{dy} = \boldsymbol{\mu_s}, \int_{\boldsymbol{\Omega_s}} \boldsymbol{\gamma}(\mathbf{x}, \mathbf{y}) \mathbf{dx} = \boldsymbol{\mu_t} \right\}$$

- γ is a joint probability measure with marginals μ_s and μ_t.
- Linear Program that always have a solution.

The 3 ways of optimal transport

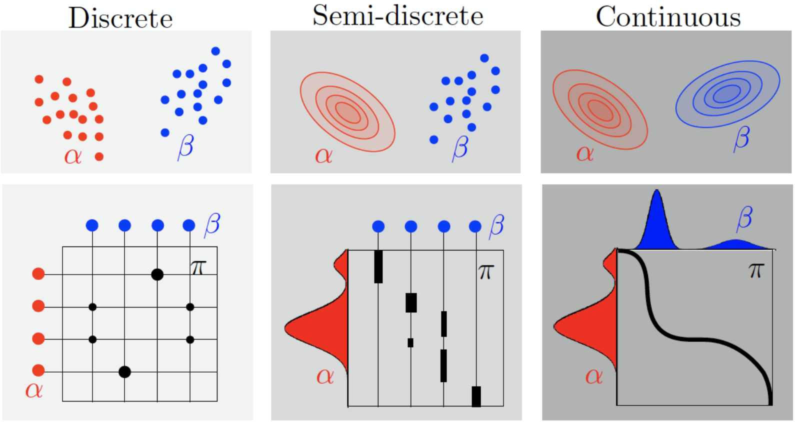


Image from Gabriel Peyré

Couplings

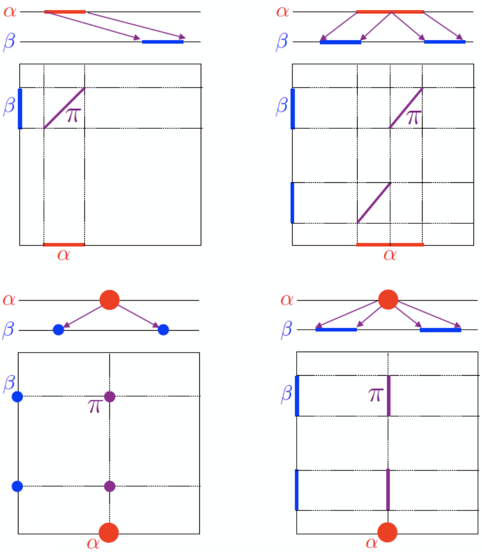
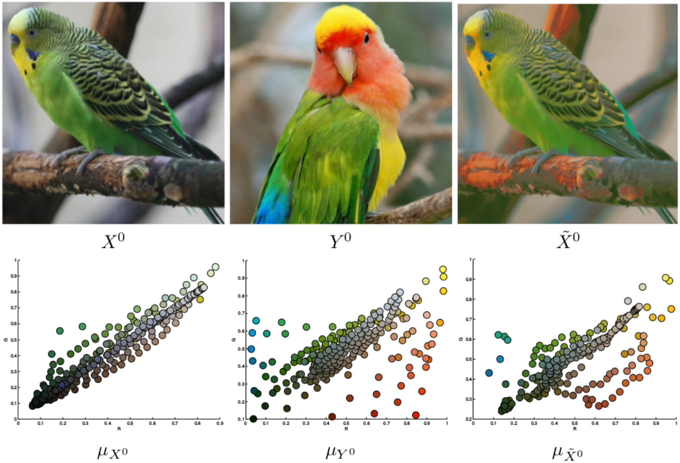


Image from Gabriel Peyré

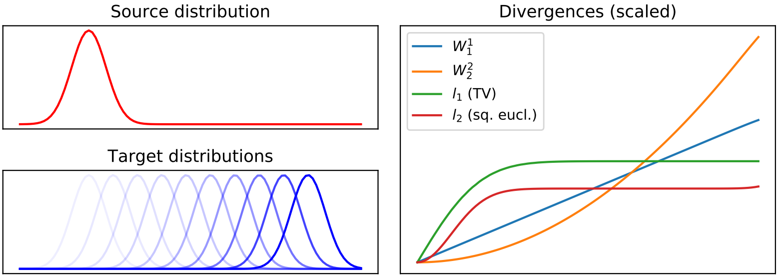
Pixels as empirical distribution [Ferradans et al., 2014]



Histogram matching in images : color grading

Image colorization [Ferradans et al., 2014]

- Words are embedded in a high-dimensional space with neural networks
- Matching two documents is an OT problem, with the cost being the l_2 distance in the embedded space



Wasserstein distance

$$W_p^p(\mu_s, \mu_t) = \min_{\boldsymbol{\gamma}\in\mathcal{P}} \int_{\Omega_s\times\Omega_t} c(\mathbf{x},\mathbf{y})\boldsymbol{\gamma}(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y} = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\boldsymbol{\gamma}}[c(\mathbf{x},\mathbf{y})] \tag{3}$$

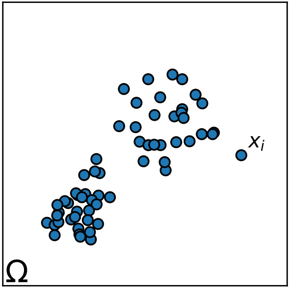
where $c(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^p$

- A.K.A. Earth Mover's Distance (W_1^1) [Rubner et al., 2000].
- Do not need the distribution to have overlapping support.
- Works for continuous and discrete distributions (histograms, empirical).

Discrete distributions: Empirical vs Histogram

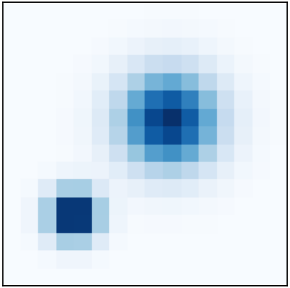
Discrete measure: $$\mu = \sum_{i=1}^{n} \mu_i \delta_{\mathbf{x}_i}, \quad \mathbf{x}_i \in \Omega, \quad \sum_{i=1}^{n} \mu_i = 1$$

Lagrangian (point clouds)



- Constant weight: $\mu_i = \frac{1}{n}$
- Quotient space: Ω^n, Σ_n

Eulerian (histograms)



- Fixed positions $\mathbf{x}_i$ e.g. grid
- Convex polytope Σ_n (simplex): $\left\{(\mu_i)_i \geq 0; \sum_i \mu_i = 1\right\}$

Wasserstein space

The space of probability distribution equipped with the Wasserstein metric ($\mathcal{P}_p(X)$, $W_2^2(X)$) defines a geodesic space with a Riemannian structure [Santambrogio, 2014].

- Geodesics are shortest curves on $\mathcal{P}_p(X)$ that link two distributions

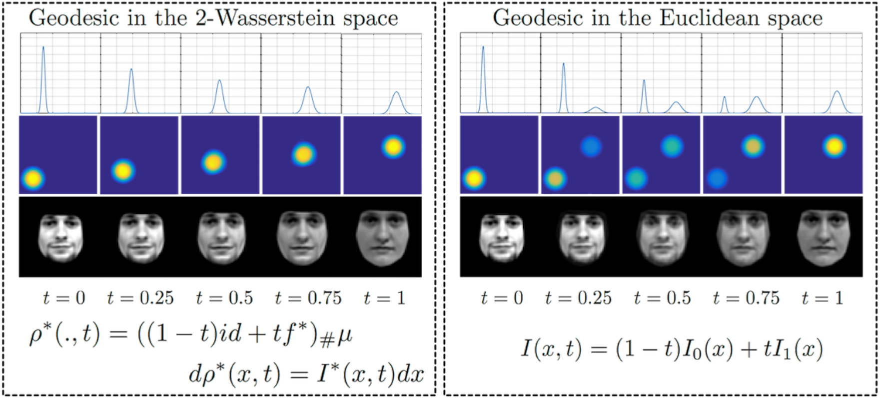
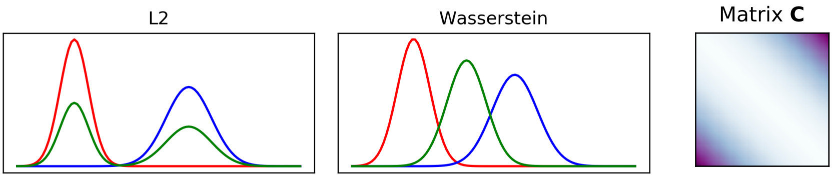


Illustration by S. Kolhouri

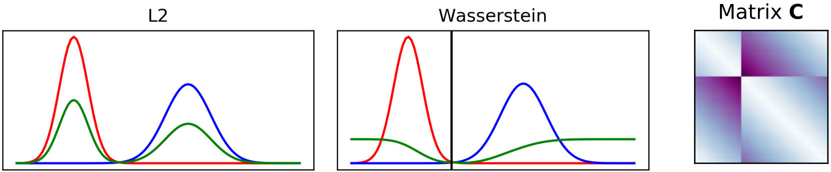
Wasserstein barycenter



Barycenters [Agueh and Carlier, 2011]

$$\bar{\mu} = \arg\min_{\mu} \quad \sum_{i}^{n} \lambda_i W_p^p(\mu^i, \mu)$$

- $\lambda_i > 0$ and $\sum_i^n \lambda_i = 1$.
- Uniform barycenter has $\lambda_i = \frac{1}{n}, \forall i$.
- Interpolation with n=2 and $\boldsymbol{\lambda} = [1-t, t]$ with $0 \le t \le 1$ [McCann, 1997].
- Regularized barycenters using Bregman projections [Benamou et al., 2015].
- The cost and regularization impacts the interpolation trajectory.

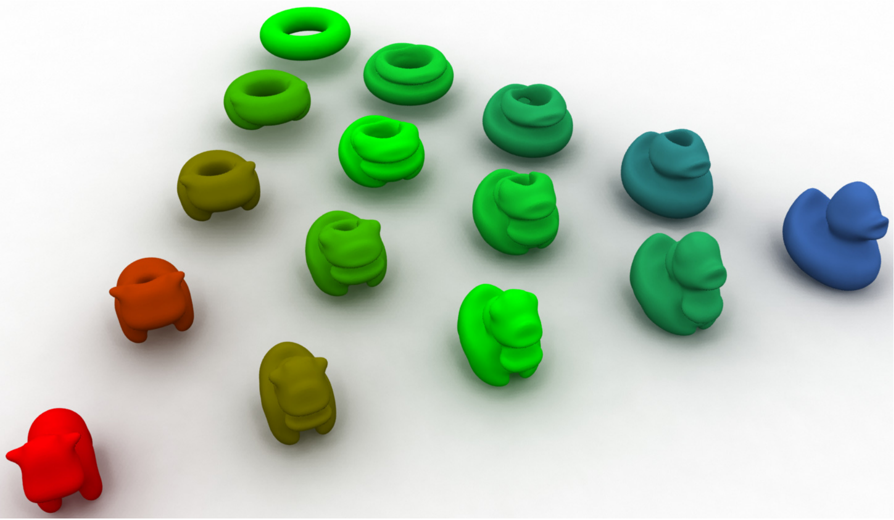


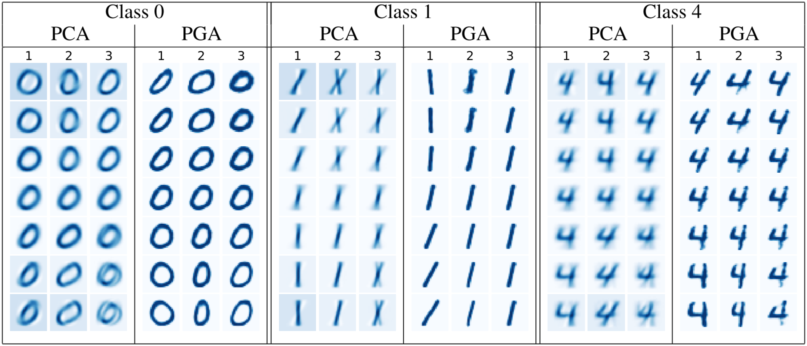
Barycenters [Agueh and Carlier, 2011]

$$\bar{\mu} = \arg\min_{\mu} \quad \sum_{i}^{n} \lambda_i W_p^p(\mu^i, \mu)$$

- $\lambda_i > 0$ and $\sum_i^n \lambda_i = 1$.
- Uniform barycenter has $\lambda_i = \frac{1}{n}, \forall i$.
- Interpolation with n=2 and $\boldsymbol{\lambda} = [1-t, t]$ with $0 \leq t \leq 1$ [McCann, 1997].
- Regularized barycenters using Bregman projections [Benamou et al., 2015].
- The cost and regularization impacts the interpolation trajectory.

Shape interpolation [Solomon et al., 2015]





Geodesic PCA in the Wasserstein space [Bigot et al., 2017]

- Generalization of Principal Component Analysis to the Wassertsein manifold.
- Regularized OT [Seguy and Cuturi, 2015].
- Approximation using Wasserstein embedding [Courty et al., 2017].
- Also note recent Wasserstein Dictionary Learning approaches [Schmitz et al., 2017].

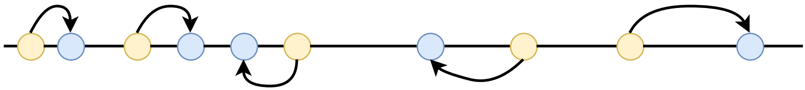
Outline

Special case: 1D distribution

We consider the case where $c(x, y)$ is a strictly convex and increasing function of $|x - y|$.

- if $x_1 < x_2$ and $y_1 < y_2$, it is easy to check that $c(x_1, y_1) + c(x_2, y_2) < c(x_1, y_2) + c(x_2, y_1)$
- As such, any optimal transport plan respects the ordering of the elements, and the solution is given by the monotone rearrangement of μ_1 onto μ_2

This gives very simple algorithm to compute the transport in $O(N \log N)$, by sorting both x_i and y_i and summing the absolute values of differences.



Special case: 1D distribution

Consider the cumulative distribution functions F_μ associated to the μ distribution.

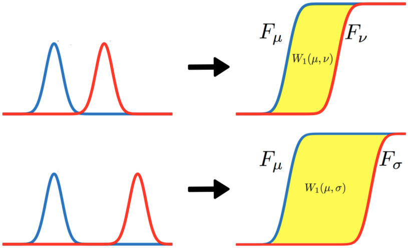
- It is defined such that $F_\mu(t) = \mu(-\infty, t]$.

We will note $F_\mu^{-1}(q),\ q \in [0,1]$ the corresponding generalized inverse distribution (or quantile function)

- defined as $F_\mu^{-1}(q) = \inf\{x \in \mathbb{R} : F_\mu(x) \geq q\}$.

Then,

$$W_1(\mu_s, \mu_t) = \int_0^1 c(F_{\mu_s}^{-1}(q), F_{\mu_t}^{-1}(q))dq$$



Sliced Radon Wasserstein

This property gives a method for computing Wasserstein in higher dimensions ($n > 1$).

The principle is simple. Slice the distribution along lines, project the measures onto it and compute $1D$ Wasserstein along those projections. More formally, consider the Radon transform $\mathcal{R}$:

$$\mathcal{R}(\mu, \theta) = \int_{\mathbb{S}^{d-1}} \mu(\mathbf{x})\delta(t - \theta.\mathbf{x})dx$$

where $t \in \mathbb{R}$ parametrizes the support and $\forall\theta \in \mathbb{S}^{d-1}$ (unit sphere in $\mathbb{R}^d$). Then, the p-sliced Wasserstein distance is given by:

p-sliced Wasserstein distance pSW [Bonneel et al., 2015]

$$pSW_p^p(\mu_s, \mu_t) = \int_{\mathbb{S}^{d-1}} W_p^p(\mathcal{R}(\mu_s, \theta), \mathcal{R}(\mu_t, \theta))d\theta$$

works well in $2D$, impractical in larger dimensions.

Special case: transport between Gaussians

In the case where $\mu_s \sim \mathcal{N}(\mathbf{m}_1, \Sigma_1)$ and $\mu_t \sim \mathcal{N}(\mathbf{m}_2, \Sigma_2)$ the Wasserstein distance with $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$ reduces to:

W_2^2 between Gaussians

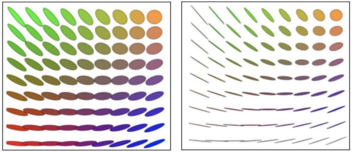
$$W_2^2(\mu_s, \mu_t) = ||\mathbf{m}_1 - \mathbf{m}_2||_2^2 + \mathcal{B}(\Sigma_1, \Sigma_2)^2$$

where $\mathbb{B}(,)$ is the so-called Bures metric:

$$\mathcal{B}(\Sigma_1, \Sigma_2)^2 = \text{trace}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}).$$

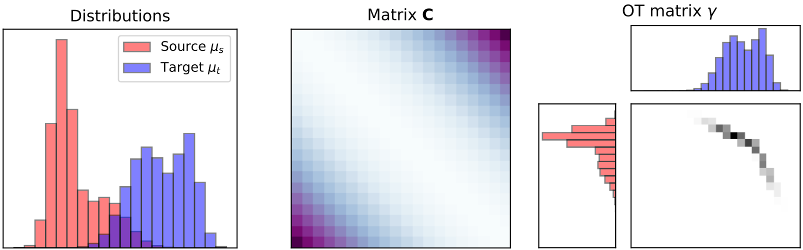
The optimal map T is given by

$$T(\mathbf{x}) = \mathbf{m}_2 + A(\mathbf{x} - \mathbf{m}_1)$$



with $A = \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}$

Optimal transport with discrete distributions



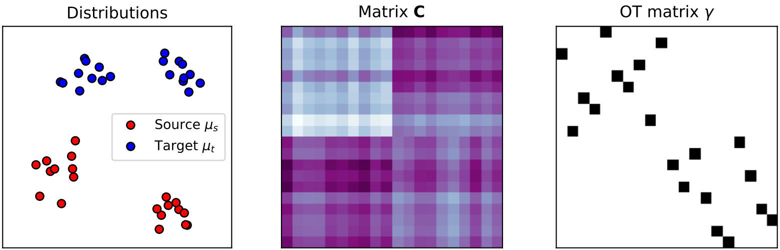
OT Linear Program

$$\gamma_0 = \underset{\gamma \in \mathcal{P}}{\operatorname{argmin}} \quad \left\{ \langle \gamma, \mathbf{C} \rangle_F = \sum_{i,j} \gamma_{i,j} c_{i,j} \right\}$$

where $\mathbf{C}$ is a cost matrix with $c_{i,j} = c(\mathbf{x}_i^s, \mathbf{x}_j^t)$ and the marginals constraints are

$$\mathcal{P} = \left\{ \gamma \in (\mathbb{R}^+)^{\mathbf{n_s} \times \mathbf{n_t}} | \, \gamma \mathbf{1}_{\mathbf{n_t}} = \mu_{\mathbf{s}}, \gamma^{\mathbf{T}} \mathbf{1}_{\mathbf{n_s}} = \mu_{\mathbf{t}} \right\}$$

Solved with Network Flow solver of complexity $O(n^3 \log(n))$.



OT Linear Program

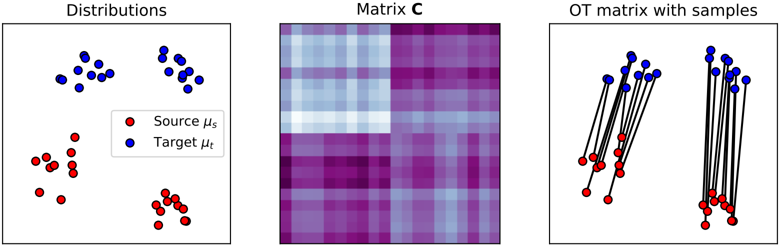
$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma} \in \mathcal{P}}{\operatorname{argmin}} \quad \left\{ \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F = \sum_{i,j} \gamma_{i,j} c_{i,j} \right\}$$

where $\mathbf{C}$ is a cost matrix with $c_{i,j} = c(\mathbf{x}_i^s, \mathbf{x}_j^t)$ and the marginals constraints are

$$\mathcal{P} = \left\{ \boldsymbol{\gamma} \in (\mathbb{R}^+)^{\mathbf{n_s} \times \mathbf{n_t}} \,|\, \boldsymbol{\gamma} \mathbf{1}_{\mathbf{n_t}} = \boldsymbol{\mu_s}, \boldsymbol{\gamma}^{\mathbf{T}} \mathbf{1}_{\mathbf{n_s}} = \boldsymbol{\mu_t} \right\}$$

Solved with Network Flow solver of complexity $O(n^3 \log(n))$.

Optimal transport with discrete distributions



OT Linear Program

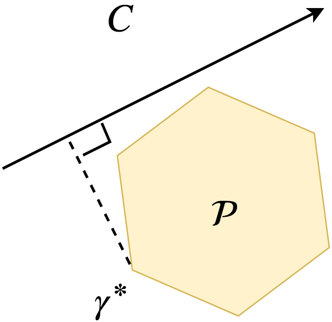
$$\gamma_0 = \underset{\gamma \in \mathcal{P}}{\operatorname{argmin}} \quad \left\{ \langle \gamma, \mathbf{C} \rangle_F = \sum_{i,j} \gamma_{i,j} c_{i,j} \right\}$$

where $\mathbf{C}$ is a cost matrix with $c_{i,j} = c(\mathbf{x}_i^s, \mathbf{x}_j^t)$ and the marginals constraints are

$$\mathcal{P} = \left\{ \gamma \in (\mathbb{R}^+)^{\mathbf{n_s} \times \mathbf{n_t}} | \, \gamma \mathbf{1}_{\mathbf{n_t}} = \mu_{\mathbf{s}}, \gamma^{\mathbf{T}} \mathbf{1}_{\mathbf{n_s}} = \mu_{\mathbf{t}} \right\}$$

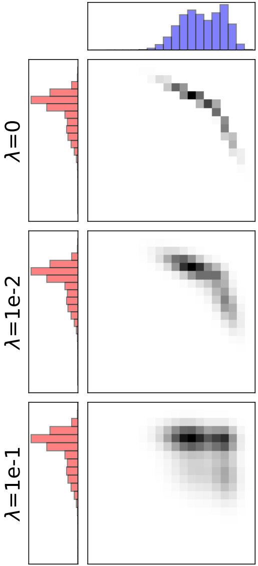
Solved with Network Flow solver of complexity $O(n^3 \log(n))$.

Optimal transport with discrete distributions



- $\mathcal{P}$ is the Birkhoff polytope
- No unique solution in some cases, numerical instabilities
- Not differentiable !

$$\gamma_0^\lambda = \underset{\gamma \in \mathcal{P}}{\operatorname{argmin}} \quad \langle \gamma, \mathbf{C} \rangle_F + \lambda \Omega(\gamma), \qquad (4)$$



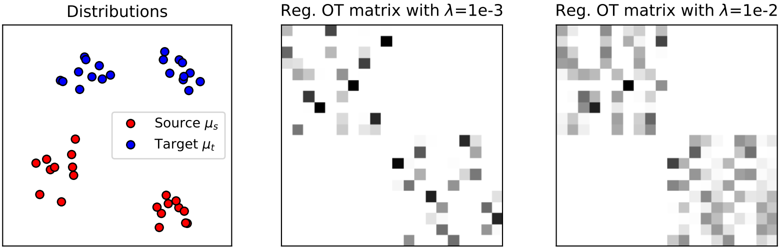
Regularization term $\Omega(\gamma)$

- Entropic regularization [Cuturi, 2013].
- Group Lasso [Courty et al., 2016].
- KL, Itakura Saito, β-divergences, [Dessein et al., 2016].

Why regularize?

- Smooth the "distance" estimation:
  $$W_\lambda(\mu_s, \mu_t) = \langle \gamma_0^\lambda, \mathbf{C} \rangle_F$$
- Encode prior knowledge on the data.
- Better posed problem (convex, stability).
- Fast algorithms to solve the OT problem.

Entropic regularized optimal transport

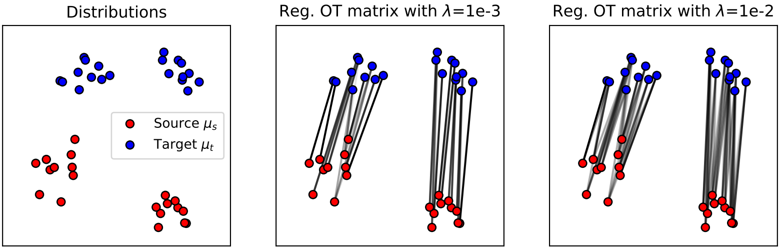


Entropic regularization [Cuturi, 2013]

$$\Omega(\boldsymbol{\gamma}) = \sum_{i,j} \boldsymbol{\gamma}(i,j)(\log \boldsymbol{\gamma}(i,j) - 1)$$

- Regularization with the negative entropy of $\boldsymbol{\gamma}$.

Entropic regularized optimal transport



Entropic regularization [Cuturi, 2013]

$$\Omega(\boldsymbol{\gamma}) = \sum_{i,j} \boldsymbol{\gamma}(i,j)(\log \boldsymbol{\gamma}(i,j) - 1)$$

- Regularization with the negative entropy of $\boldsymbol{\gamma}$.

Resolving the entropy regularized problem

Entropy-regularized transport

The solution of entropy regularized optimal transport problem is of the form

$$\gamma_0^\lambda = \text{diag}(\mathbf{u}) \exp(-\mathbf{C}/\lambda) \text{diag}(\mathbf{v})$$

Why ? Consider the Lagrangian of the optimization problem:

$$\mathcal{L}(\boldsymbol{\gamma}, \alpha, \beta) = \sum_{ij} \boldsymbol{\gamma}_{ij} \mathbf{C}_{ij} + \lambda \boldsymbol{\gamma}_{ij} (\log \boldsymbol{\gamma}_{ij} - 1) + \alpha^{\mathbf{T}} (\boldsymbol{\gamma} \mathbf{1}_{n_t} - \mu_s) + \beta^{\mathbf{T}} (\boldsymbol{\gamma}^T \mathbf{1}_{n_s} - \mu_t)$$

$$\begin{aligned}
\partial \mathcal{L}(\boldsymbol{\gamma}, \alpha, \beta) / \partial \boldsymbol{\gamma}_{ij} \quad &= \quad \mathbf{C}_{ij} + \lambda \log \boldsymbol{\gamma}_{ij} + \alpha_i + \beta_j \\
\partial \mathcal{L}(\boldsymbol{\gamma}, \alpha, \beta) / \partial \boldsymbol{\gamma}_{ij} = 0 \quad &\Longrightarrow \quad \boldsymbol{\gamma}_{ij} = \exp(\frac{\alpha_i}{\lambda}) \exp(-\frac{\mathbf{C}_{ij}}{\lambda}) \exp(\frac{\beta_j}{\lambda})
\end{aligned}$$

- Through the **Sinkhorn theorem** diag$(\mathbf{u})$ and diag$(\mathbf{v})$ exist and are unique.
- Can be solved by the **Sinkhorn-Knopp** algorithm (implementation in parallel, GPU).

The Sinkhorn-Knopp algorithm performs alternatively a scaling along the rows and columns of $\mathbf{K} = \exp(-\frac{\mathbf{C}}{\lambda})$ to match the desired marginals.

Algorithm 1 Sinkhorn-Knopp Algorithm (SK).

Require: $\mathbf{a}, \mathbf{b}, \mathbf{C}, \lambda$
  $\mathbf{u}^{(0)} = \mathbf{1}, \mathbf{K} = \exp(-\mathbf{C}/\lambda)$
  for i in $1, \ldots, n_{it}$ **do**
    $\mathbf{v}^{(i)} = \mathbf{b} \oslash \mathbf{K}^\top \mathbf{u}^{(i-1)}$ // Update right scaling
    $\mathbf{u}^{(i)} = \mathbf{a} \oslash \mathbf{K}\mathbf{v}^{(i)}$ // Update left scaling
  end for
  return $\mathcal{T} = \text{diag}(\mathbf{u}^{(n_{it})})\mathbf{K}\text{diag}(\mathbf{v}^{(n_{it})})$

- Complexity $O(kn^2)$, where k iterations are required to reach convergence
- Fast implementation in parallel, GPU friendly
- Convolutive/Heat structure for K [Solomon et al., 2015]

Recalling that the Kullback Leibler (KL) divergence between two distribution is

$$\mathrm{KL}(\boldsymbol{\gamma}, \rho) = \sum_{ij} \boldsymbol{\gamma}_{ij} \log \frac{\boldsymbol{\gamma}_{ij}}{\rho_{ij}} =< \boldsymbol{\gamma}, \log \frac{\boldsymbol{\gamma}}{\rho} >_F,$$

Benamou *et al.* [Benamou et al., 2015] showed that solving for the OT problem is actually a Bregman projection

OT as a Bregman projection

$\boldsymbol{\gamma}^\star$ is the solution of the following Bregman projection

$$\boldsymbol{\gamma}^\star = \operatorname*{argmin}_{\boldsymbol{\gamma} \in \mathcal{P}} \mathrm{KL}(\boldsymbol{\gamma}, \zeta), \tag{5}$$

where $\zeta = \exp(-\frac{C}{\lambda})$.

- Sinkhorn in this case is an iterative projection scheme, with alternative projections on marginal constraints.
- Generalizes well for barycenters computation

Dual formulation of optimal transport

- Yet, solving for γ is impractical to intractable when dealing with high-dimensional distributions
- especially if one is interested in computing the gradients of the Wasserstein distance
- Other solving strategies should be taken into consideration
- Recalling that any LP problem can be turnt into its dual form:

$$\begin{array}{rrcl|rrcl}
\multicolumn{4}{c|}{\textbf{primal form}:} & \multicolumn{4}{c}{\textbf{dual form}:} \\
\text{minimize} & z & = & \mathbf{c}^T\mathbf{x}, & \text{maximize} & \tilde{z} & = & \mathbf{b}^T\mathbf{y}, \\
\text{so that} & \mathbf{A}\mathbf{x} & = & \mathbf{b} & \text{so that} & \mathbf{A}^T\mathbf{y} & \leq & \mathbf{c} \\
\text{and} & \mathbf{x} & \geq & \mathbf{0} & & & &
\end{array}$$

- Weak duality: $\tilde{z}$ is a lower bound of z, Strong duality $\tilde{z} = z$
- Strong duality is usually achieved via Farkas Theorem

Duality: general case with continuous distributions

We now introduce two functions scalar functions ϕ and ψ (also known as Kantorovich potentials) that will act as our dual variables. Then, we consider the optimal problem is equivalent (by the Rockafellar-Fenchel theorem) to:

$$\max_{\phi,\psi} \left\{ \int \phi d\mu_s + \int \psi d\mu_t \;\mid\; \phi(x) + \psi(y) \leq c(x,y) \right\} \tag{6}$$

Note that the marginal constraint has been turned into an equality constraint on ϕ and ψ

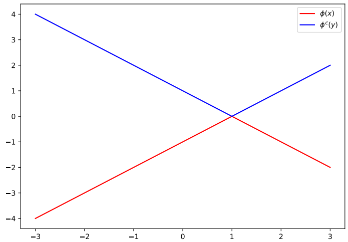
Introducing the *c-transform* (or *c-conjugate*) H^c which is in spirit close to a Legendre transform:

$$\phi^c \overset{\text{def}}{=} H^c(\phi) = \inf_x c(x,y) - \phi(x) \tag{7}$$

then the following problem is equivalent:

$$\max_{\phi} \left\{ \int \phi d\mu_s + \int \phi^c d\mu_t \;\mid\; \phi(x) + \phi^c(y) \leq c(x,y) \right\} \tag{8}$$

Case $c(x, y) = |x - y|$ (a.k.a W_1^1)



Whenever $c(x, y) = |x - y|$, then:

- existence of a solution but not unique
- For any $\phi \in \text{Lip}^1$ (set of 1-Lipschitz functions), we have $\phi^c(x) = -\phi(x)$

The optimal transport problem then amounts to find $\phi \in \text{Lip}^1$ as

$$\sup_{\phi \in \text{Lip}^1} \int \phi d(\mu_s - \mu_t) = \sup_{\phi \in \text{Lip}^1} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mu_s} [\phi(x)] - \mathop{\mathbb{E}}_{\mathbf{y} \sim \mu_t} [\phi(y)] \tag{9}$$

- also known as **Kantorovich-Rubinstein duality**
- ϕ can be learnt as a neural network constrained to the set Lip^1, see next section on GAN

Case $c(x, y) = |x - y|^2/2$ (a.k.a W_2^2)

Whenever the cost is quadratic, $c(x, y) = |x - y|^2/2$, then:

- $T(x)$ the transport mapping exists and is unique
- More remarkably, it is a gradient of a convex functions $\Phi(x)$

$$T(x) = x - \nabla\phi(x) = \nabla(\frac{x^2}{2} - \phi(x)) = \nabla(\Phi(x)) \tag{10}$$

- This is also known as **Brenier's Theorem**

Dual: empirical version

In the case when we have access to discrete distributions, μ_s (resp. μ_t) is characterized by a set of locations $\mathbf{X^s}$ and masses $\mathbf{a} \in \mathbb{R}^{n^s}$ (resp. $\mathbf{X^t}$ and $\mathbf{b} \in \mathbb{R}^{n^t}$)

Discrete dual version of OT

$$W(\mu_s, \mu_t) = \max_{\alpha \in \mathbb{R}^{n^s}, \beta \in \mathbb{R}^{n^t}, \alpha_i + \beta_j \leq c(\mathbf{X^s_i}, \mathbf{X^t_j})} \alpha^T \mathbf{a} + \beta^T \mathbf{b} \tag{11}$$

i.e. find a scalar values per sample

Adding regularization to the original problem turns the dual computation to an unconstrained problem !

In the case of entropy regularization, *i.e.*

$W_\lambda(\mu_s, \mu_t) = \min_{\boldsymbol{\gamma} \in \mathcal{P}} \quad \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F + \lambda \Omega(\boldsymbol{\gamma})$ with $\Omega(\boldsymbol{\gamma}) = \sum_{i,j} \boldsymbol{\gamma}(i,j) \log \boldsymbol{\gamma}(i,j)$,

the dual now reads (in a discrete settings, measures are collections of Diracs):

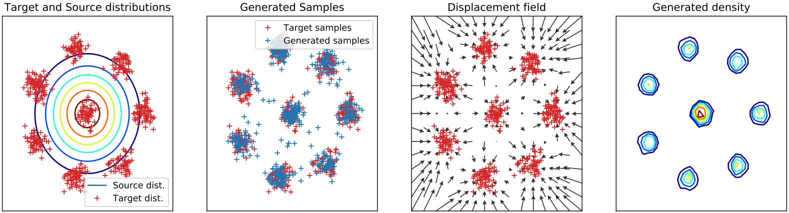
$$\max_{\alpha,\beta} \alpha^T \mu_s + \beta^T \mu_t - \frac{1}{\lambda} \exp(\frac{\alpha}{\lambda})^T \mathbf{K} \exp(\frac{\beta}{\lambda}) \tag{12}$$

with $\mathbf{K} = \exp(-\frac{\mathbf{C}}{\lambda})$.

Remark: The Sinkhorn algorithm is a gradient ascent on the dual variables !

With this unconstrained problem, incremental gradients techniques (SGD, SAG) can be used to solve the problem !

- [Genevay et al., 2016] used the semi-dual formulation (one variable is removed by replacing it with its c-transform) int the first stochastic version of Optimal Transport problem
- [Seguy et al., 2017] used the full dual version with entropic and L2 regularizations, together with neural networks to parameterize the problem.



In machine learning applications, one can be interested in finding distributions that minimize the Wasserstein distance wrt. a reference measure. There are two ways of understanding this:

- case 1: **for a fixed support** $\mathbf{X}$, find the corresponding probability masses $\mathbf{m}$
- case 2: **for a fixed vector of probability masses** $\mathbf{m}$, e.g. uniform distribution, find the corresponding support $\mathbf{X}$

Recalling the form of the dual

$$W(\mu, \mu_t) = \max_{\alpha \in \mathbb{R}^{n^s}, \beta \in \mathbb{R}^{n^t}, \alpha_i + \beta_j \leq c(\mathbf{X}, \mathbf{X_j^t})} \alpha^T \mathbf{m} + \beta^T \mathbf{b} \tag{13}$$

- $W(\mu, \mu_t)$ is convex wrt. **m**
- $\partial_{\mathbf{m}} W(\mu, \mu_t) = \alpha^*$
- **Entropy regularized case**: $W_\lambda(\mu, \mu_t)$ is convex and $\nabla_{\mathbf{m}} W_\lambda(\mu, \mu_t) = \lambda \log \mathbf{u}$

Case 2: fixed probability masses $\mathbf{m}$

Recalling the form of the primal problem

$$W_2^2(\mu, \mu_t) = \min_{\boldsymbol{\gamma} \in \mathcal{P}} \quad < \boldsymbol{\gamma}, \mathbf{1}_{\mathbf{n^s}} \mathbf{1}_{\mathbf{n^t}}^{\mathbf{T}} \mathbf{X}^2 + \mathbf{X^t}^{2T} \mathbf{1}_{\mathbf{n^t}} \mathbf{1}_{\mathbf{n^s}} - 2\mathbf{X}\mathbf{X^t} > \tag{14}$$

- $W_2^2(\mu, \mu_t)$ decreases if $\mathbf{X} \leftarrow \mathbf{X^t} \boldsymbol{\gamma}^{*T} \mathrm{diag}(\mathbf{m}^{-1})$
- explicit gradient for the regularized case.
- Barycentric interpolation !
- see Rémi next slides

Automatic differentiation to the rescue !

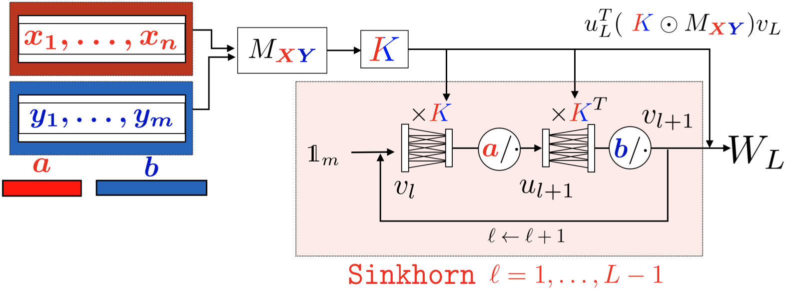
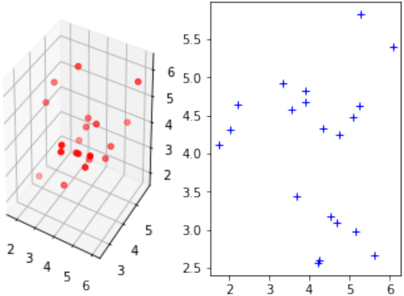


Image from Marco Cuturi

Outline

Taking into account spaces discrepancy



$$\begin{cases} \Omega_s & : \text{Source space} \\ \Omega_t & : \text{Target space} \end{cases} \quad \text{such that} \quad dim(\Omega_s) \neq dim(\Omega_t)$$

$\Rightarrow$ We can't define direct dissimilarities between source and target samples

If Ω_s and Ω_t are two spaces of different dimensions, Mémoli [Mémoli, 2011] proposed the Gromov-Wasserstein Distance between the two measured dissimilarity matrices (C, p) and $(\overline{C}, q)$:

Gromov-Wasserstein distance

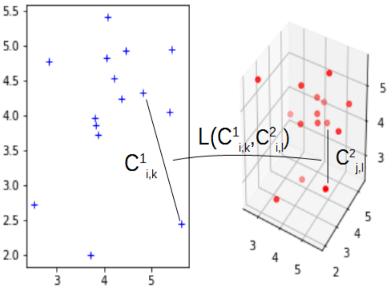
$$GW(C, \overline{C}, \mu_s, \mu_t) = \text{argmin}_{\gamma \in \mathcal{P}} \left(\sum_{i,j,k,l} L(C_{i,k}, \overline{C}_{j,l}) * \boldsymbol{\gamma}_{i,j} * \boldsymbol{\gamma}_{k,l} \right)$$

- This is related to a Quadratic Assignment Problem (QAP), opposed to the linear assignment problem as with the classical OT problem.
- non-convex problem, NP-hard

Gromov-Wasserstein distance

What is $L(C_{i,k}, \overline{C}_{j,l})$?

- Distance/dissimilarity between distances
- Several Choices are possible :
  - $L(a, b) = \frac{1}{2}|a - b|^2$
  - $L(a, b) = \text{KL}(a|b) = a * log(\frac{a}{b}) - a + b$



Computing GW coupling

Peyré and colleagues consider the entropic regularization of this problem [Peyré et al., 2016] :

$$GW(C, \overline{C}, \mu_s, \mu_t) = \underset{\gamma \in \mathcal{P}}{\operatorname{argmin}} \left(\sum_{i,j,k,l} L(C_{i,k}, \overline{C}_{j,l}) * \boldsymbol{\gamma}_{i,j} * \boldsymbol{\gamma}_{k,l} - \gamma H(\boldsymbol{\gamma}) \right)$$

One can easily compute **GW** by using projected gradient descent. With the right parameters, iterations can be simplified in :

Iteration :

$$\boldsymbol{\gamma}^{k+1} \leftarrow \underset{\gamma \in \mathcal{P}}{\operatorname{argmin}} \quad \left\langle \boldsymbol{\gamma}, \mathcal{L}(C, \overline{C}) \otimes \boldsymbol{\gamma}^{k} \right\rangle - \gamma H(\boldsymbol{\gamma})$$

Where $\otimes$ denotes the tensorial product:

$$\mathcal{L}(C, \overline{C}) \otimes \boldsymbol{\gamma} = \left(\sum_{k,l} L(C_{i,k}, \overline{C}_{j,l}) \boldsymbol{\gamma}_{k,l} \right)_{i,j}$$

The projection can be solved by simply applying a Sinkhorn algorithm.

We can show that, if $L(a,b)$ can be written as $f_1(a) + f_2(b) - h_1(a)h_2(b)$,

$$\mathcal{L}(C,\overline{C}) \otimes \boldsymbol{\gamma} = c_{C,\overline{C}} - h_1(C)\boldsymbol{\gamma}h_2(\overline{C})^T$$

with $c_{C,\overline{C}} = f_1(C)p\mathbf{I}_{N_2}^T + \mathbf{I}_{N_1}q^T f_2(\overline{C})^T$ (independant of $\boldsymbol{\gamma}$)

example :

$$L(a,b) = \frac{1}{2}|a-b|^2 \Rightarrow \begin{cases} f_1(a) & = \frac{1}{2}a^2 \\ f_2(b) & = \frac{1}{2}b^2 \\ h_1(a) & = a \\ h_2(b) & = b \end{cases}$$

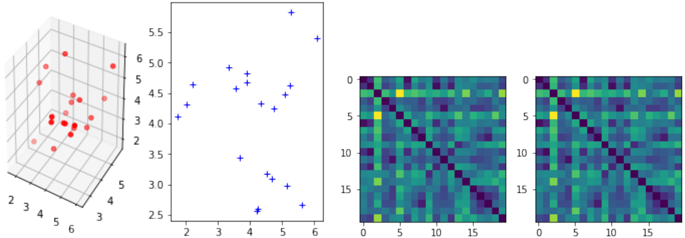


Figure 1: Source and target measures and associated cost matrices C and $\overline{C}$

GW coupling matrix :

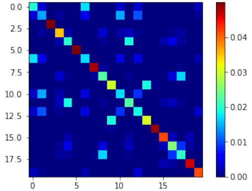


Illustration of applications of GW

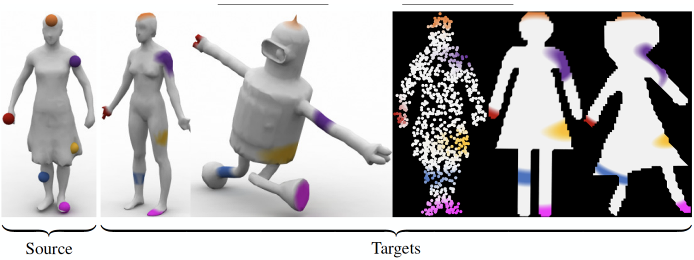


Figure 2: Shape matching between 3D and 2D objects

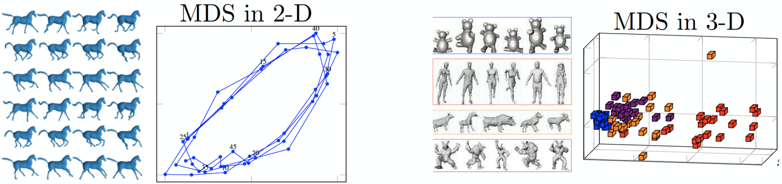
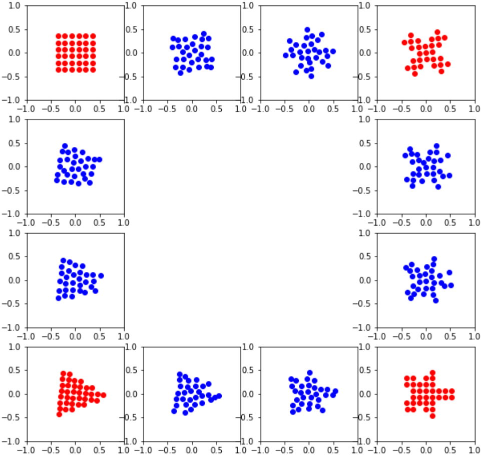


Figure 3: Visualization/classification of shapes datasets

Gromov-Wasserstein barycenters

Since we have defined a distance between two measured similarity matrices, we can compute barycenters between those spaces.

Example : progressive shape interpolation with Gromov-Wasserstein barycenters



Summary

Optimal transport is a well theoretically grounded ways of comparing probability distributions

- that allows to compare empirical distributions in a non-parametric ways
- that leverages on a ground metric in the embedding space
- for which exist several algorithmic solutions

It comes in several flavours:

- Monge problem: find a mapping (transport map)
- Kantorovich problem: find a coupling (transport plan)

References I

Agueh, M. and Carlier, G. (2011).
Barycenters in the wasserstein space.
SIAM Journal on Mathematical Analysis, 43(2):904–924.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. (2015).
Iterative Bregman projections for regularized transportation problems.
SISC.

Bigot, J., Gouet, R., Klein, T., López, A., et al. (2017).
Geodesic pca in the wasserstein space by convex pca.
In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 53, pages 1–26. Institut Henri Poincaré.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015).
Sliced and radon Wasserstein barycenters of measures.
Journal of Mathematical Imaging and Vision, 51.

Brenier, Y. (1991).
Polar factorization and monotone rearrangement of vector-valued functions.
Communications on pure and applied mathematics, 44(4):375–417.

Courty, N., Flamary, R., and Ducoffe, M. (2017).
Learning wasserstein embeddings.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016).
Optimal transport for domain adaptation.
IEEE Transactions on Pattern Analysis and Machine Intelligence.

Cuturi, M. (2013).
Sinkhorn distances: Lightspeed computation of optimal transportation.
In *Neural Information Processing Systems (NIPS)*, pages 2292–2300.

References III

Dessein, A., Papadakis, N., and Rouas, J.-L. (2016).
Regularized optimal transport and the rot mover's distance.
arXiv preprint arXiv:1610.06447.

Ferradans, S., Papadakis, N., Peyré, G., and Aujol, J.-F. (2014).
Regularized discrete optimal transport.
SIAM Journal on Imaging Sciences, 7(3).

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. (2016).
Stochastic optimization for large-scale optimal transport.
In *NIPS*, pages 3432–3440.

Kantorovich, L. (1942).
On the translocation of masses.
C.R. (Doklady) Acad. Sci. URSS (N.S.), 37:199–201.

References IV

McCann, R. J. (1997).

A convexity principle for interacting gases.

Advances in mathematics, 128(1):153–179.

Mémoli, F. (2011).

Gromov-Wasserstein distances and the metric approach to object matching.

Foundations of Computational Mathematics, pages 1–71.

Monge, G. (1781).

Mémoire sur la théorie des déblais et des remblais.

De l'Imprimerie Royale.

Peyré, G., Cuturi, M., and Solomon, J. (2016).

Gromov-Wasserstein Averaging of Kernel and Distance Matrices.

In *ICML 2016*, Proc. 33rd International Conference on Machine Learning, New-York, United States.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000).

The earth mover's distance as a metric for image retrieval.

International journal of computer vision, 40(2):99–121.

Santambrogio, F. (2014).

Introduction to optimal transport theory.

Notes.

Schmitz, M. A., Heitz, M., Bonneel, N., Mboula, F. M. N., Coeurjolly, D., Cuturi, M., Peyré, G., and Starck, J.-L. (2017).

Wasserstein dictionary learning: Optimal transport-based unsupervised non-linear dictionary learning.

arXiv preprint arXiv:1708.01955.

Seguy, V., Bhushan Damodaran, B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. (2017).

Large-scale optimal transport and mapping estimation.

Seguy, V. and Cuturi, M. (2015).
Principal geodesic analysis for probability measures under the optimal transport metric.
In *Advances in Neural Information Processing Systems*, pages 3312–3320.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. (2015).
Convolutional wasserstein distances: Efficient optimal transportation on geometric domains.
ACM Transactions on Graphics (TOG), 34(4):66.