Practical Optimal Transport for Machine Learning

Nicolas Courty[1], Laetitia Chapel[1] and Rémi Flamary[2]
[1] IRISA, Université Bretagne Sud, France
[2] CMAP – Ecole Polytechnique

Ecole de recherche (part. 2) Rennes – Juin 2022
Divergences statistiques et géométriques pour l'apprentissage machine

Overview

Planning of the course:

- (Morning) 1h30 of introductory course to *computational optimal transport* (Nicolas)
- (Afternoon) 1h30 of introduction to Unbalanced OT and Transport between metric spaces (Laetitia)
- (Afternoon) 1h30 ofpractical sessions in Python (POT)
- (Tomorrow Morning) 1h30 of of applications of OT to Machine Learning (Rémi)

Table of content

Optimal transport : introduction

What is optimal transport ?

The natural geometry of probability measures


Monge


Kantorovich


Koopmans

Nobel '75


Dantzig


Brenier


Otto


McCann


Villani

Fields '10

The origins of optimal transport

666 · MÉMOIRES DE L'ACADÉMIE ROYALE

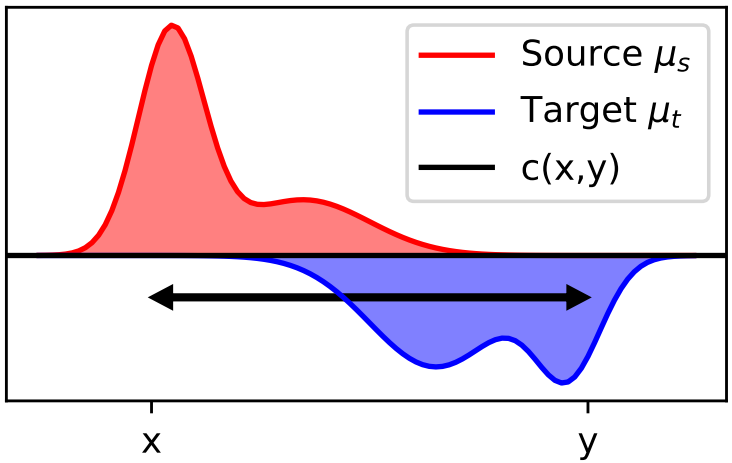
MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.

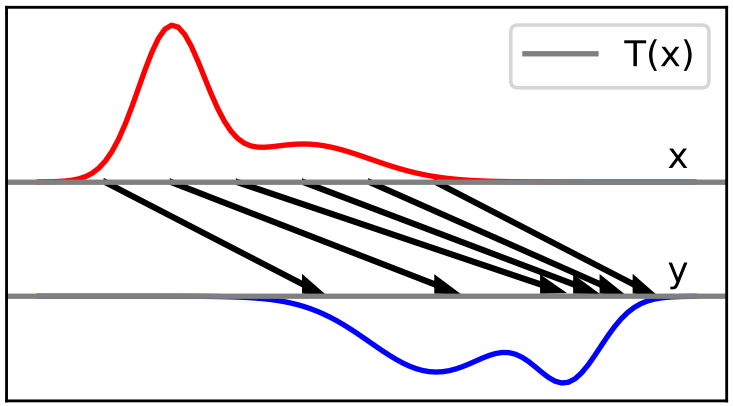
Par M. MONGE.

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).

The origins of optimal transport

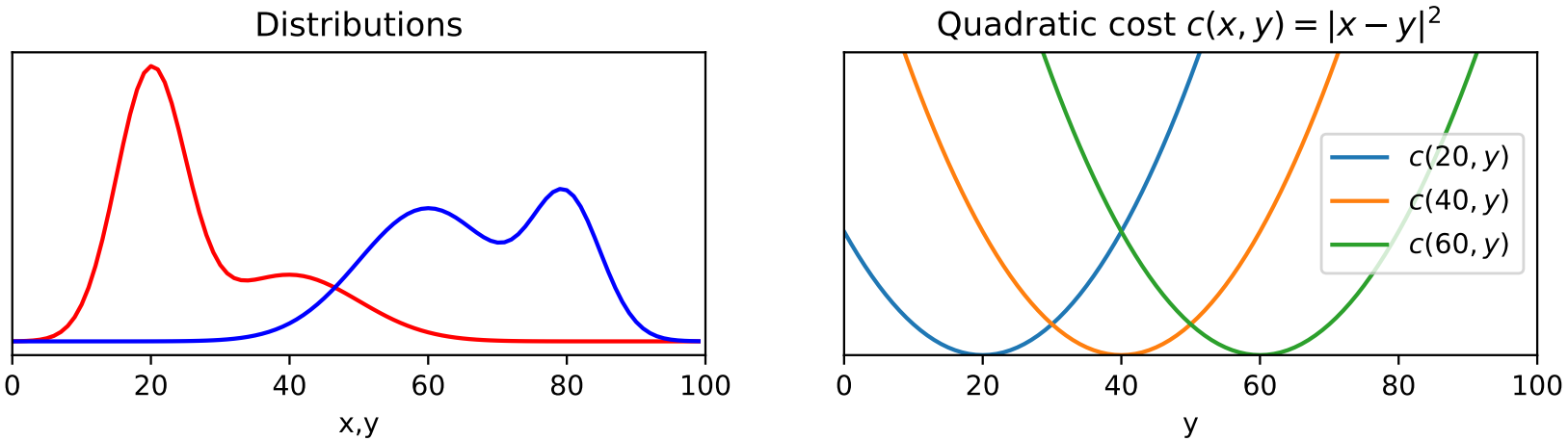




Problem [Monge, 1781]

- 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).

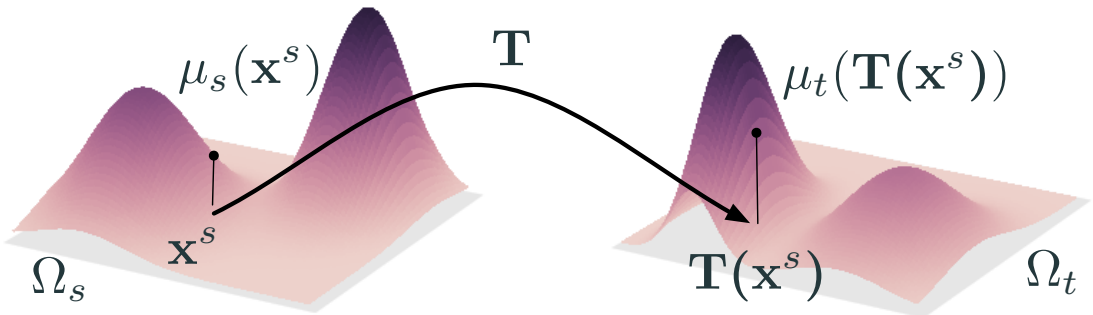
Optimal transport (Monge formulation)



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

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

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 $\mu_t = \eta(x)dx$

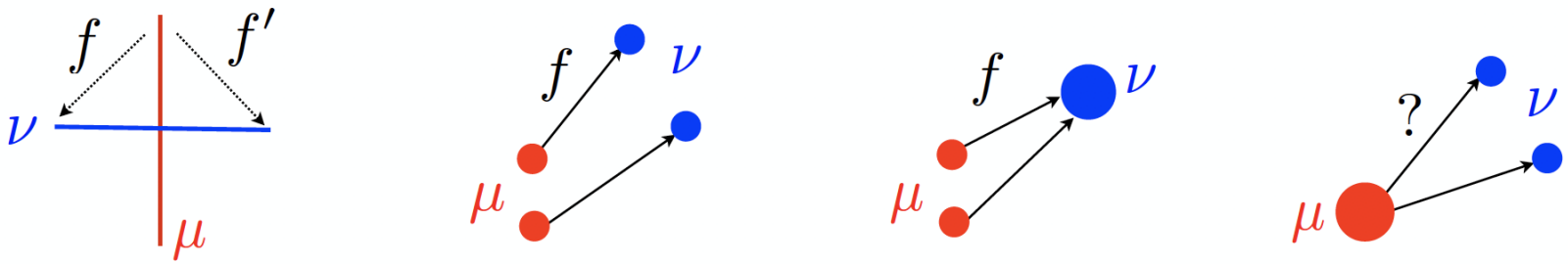
$$T\#\mu_s = \mu_t \equiv \rho(T(x))|\mathsf{det}(\partial T(x))| = \eta(x)$$

- a.k.a. change of variable formula

Non-existence / Non-uniqueness

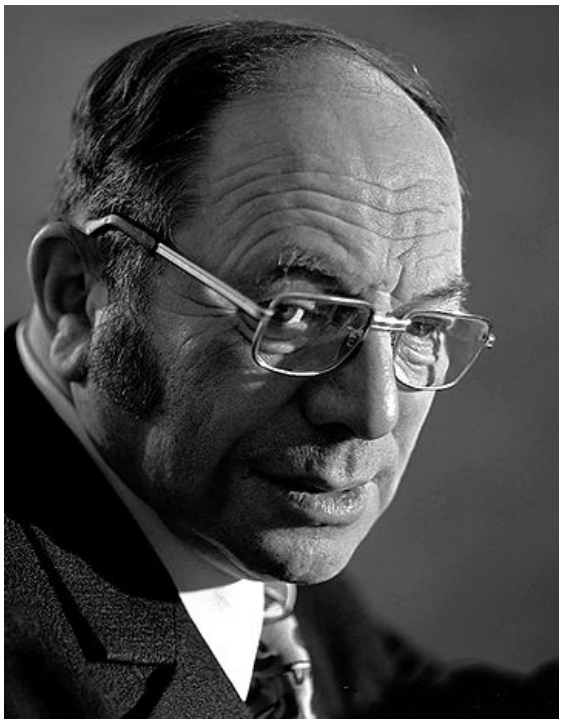
Solving for this push-forward operator is a non-convex optimization problem,

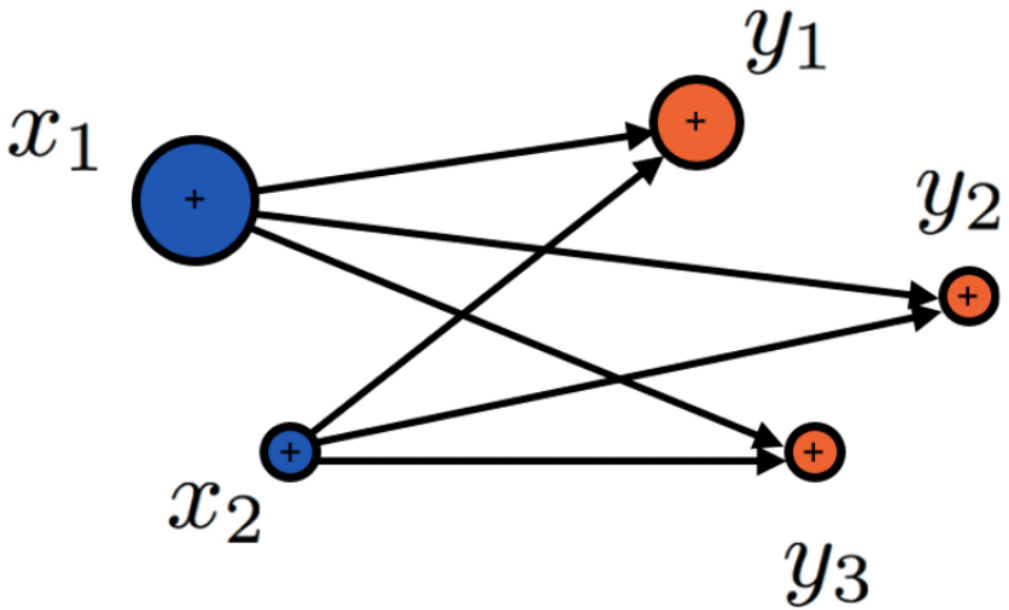
- for which existence is not guaranteed,
- nor unicity



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).

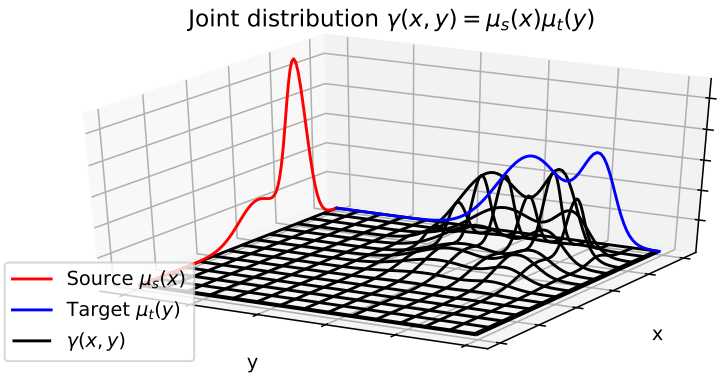
Kantorovich relaxation

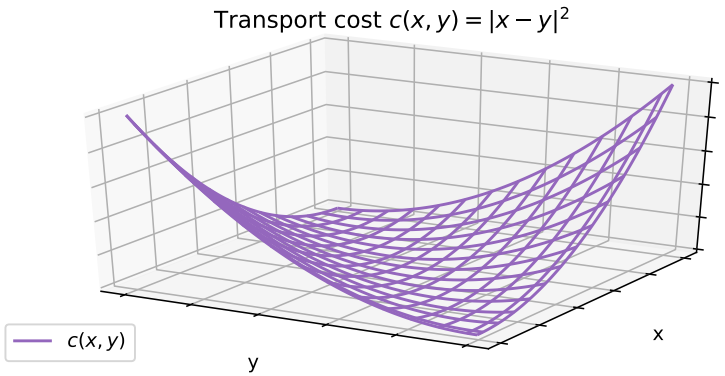




- Leonid Kantorovich (1912–1986), Economy nobelist in 1975, proposed a different formulation of the problem
- with applications mainly for ressource allocation problems

Optimal transport (Kantorovich formulation)





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

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

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

- $\boldsymbol{\gamma}$ 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

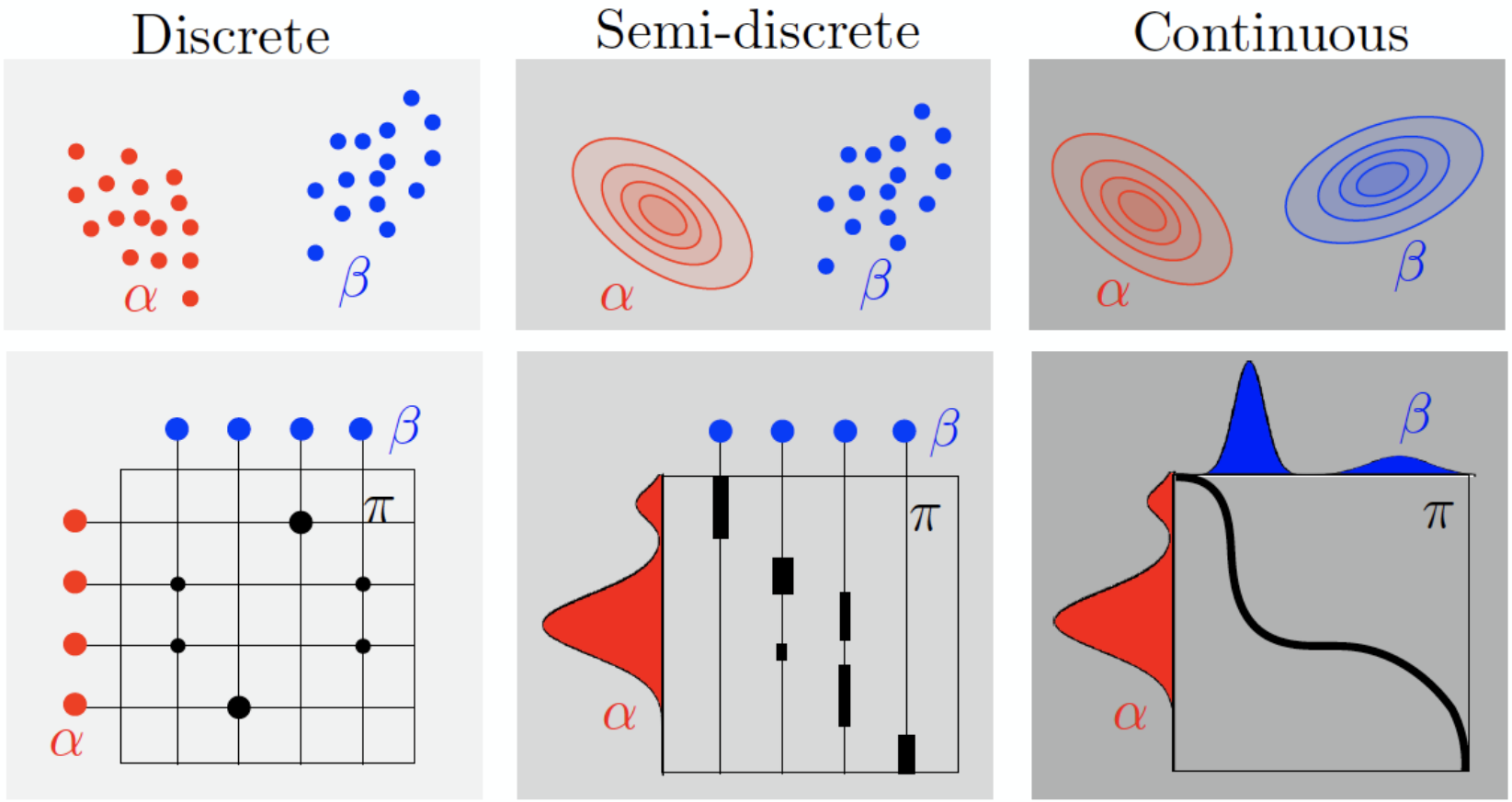


Image from Gabriel Peyré

Couplings

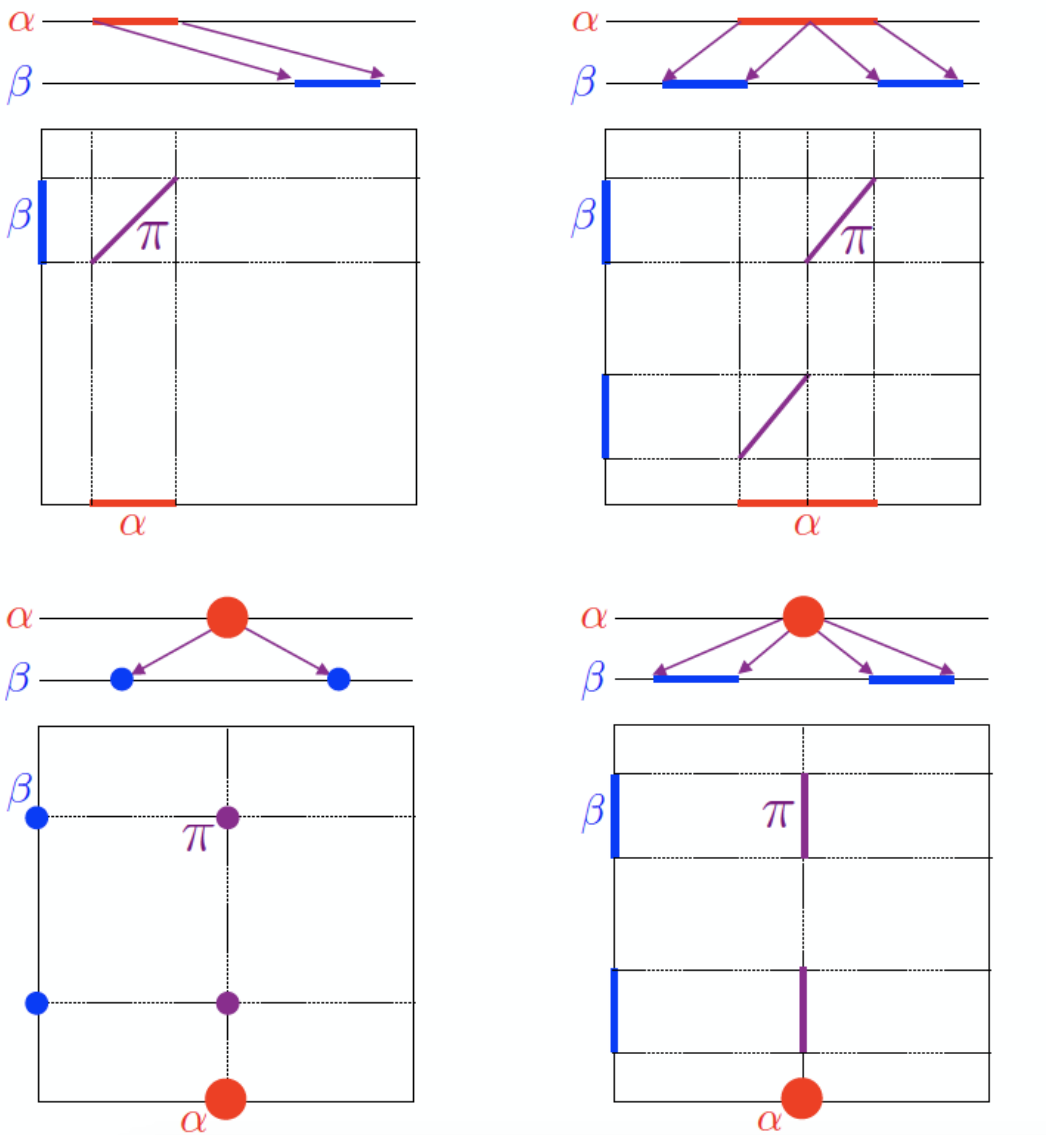


Image from Gabriel Peyré

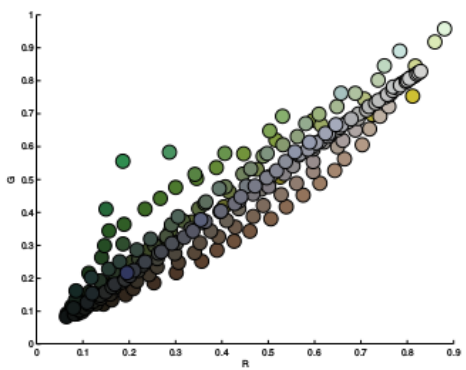
Histogram matching in images : color grading

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

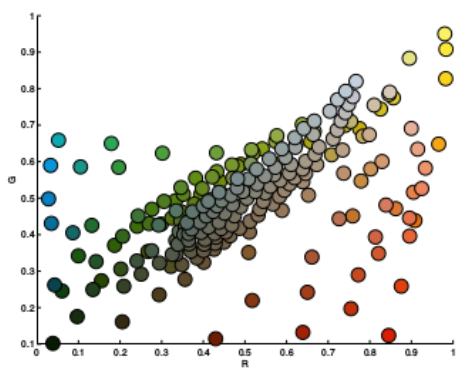
X^0

Y^0

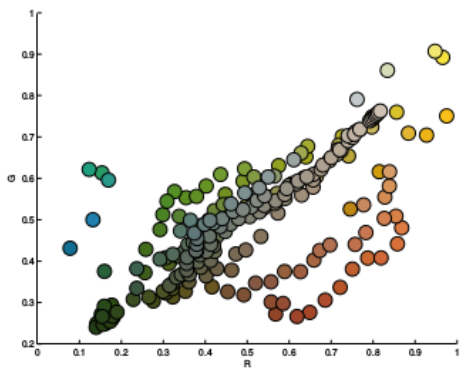
$\tilde{X}^0$



μ_{X^0}



μ_{Y^0}



$\mu_{\tilde{X}^0}$

Histogram matching in images : color grading

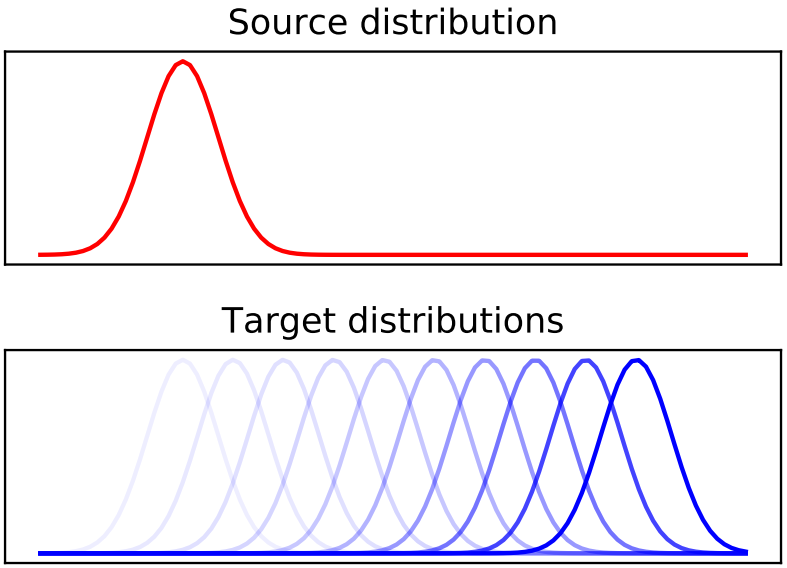
Image colorization [Ferradans et al., 2014]

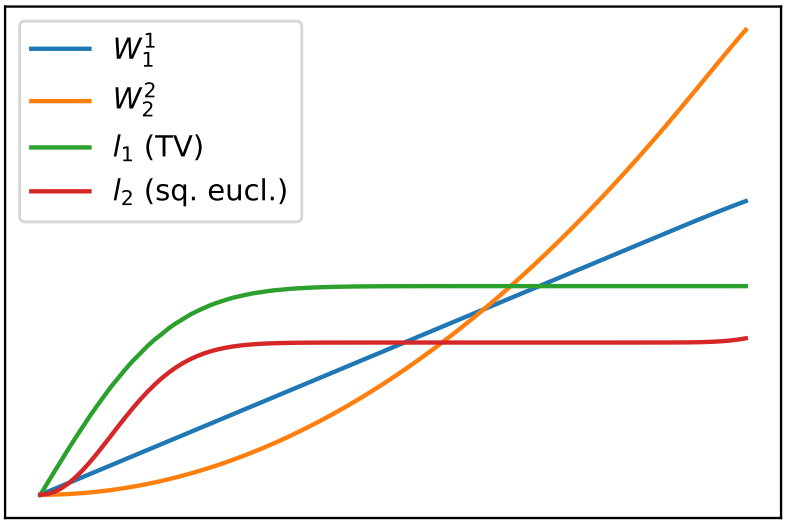
Matching words embedding

- 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

Outline

Wasserstein distance





Wasserstein distance

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

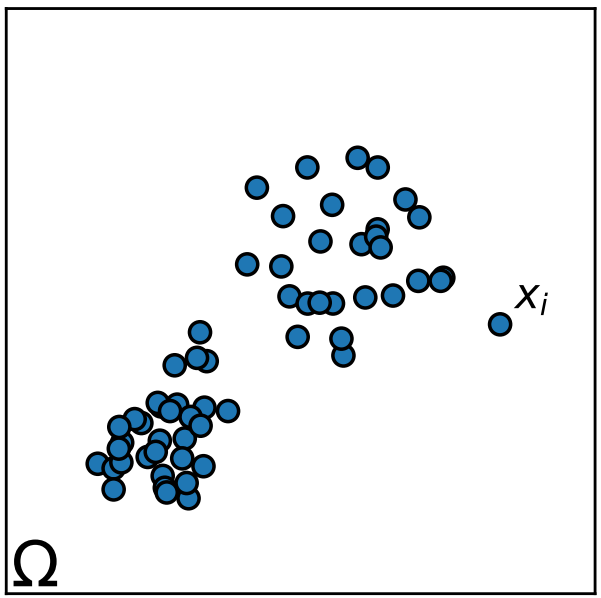
where $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^p$

- 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

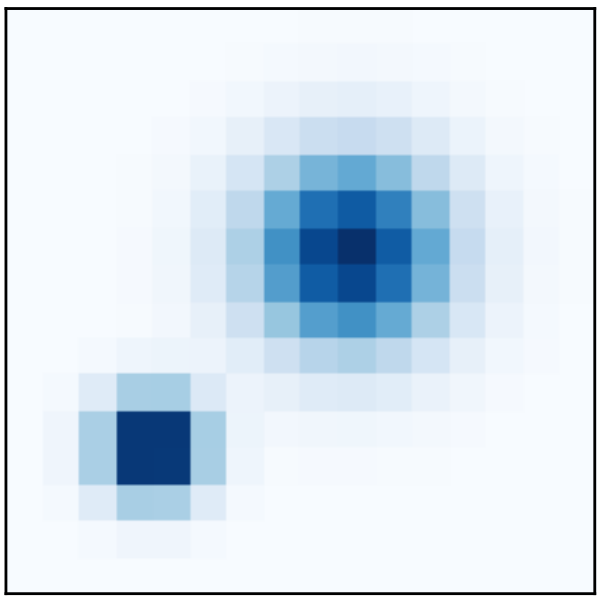
Discrete measure: $\mu = \sum_{i=1}^{n} \mu_i \delta_{\mathbf{x}_i}, \quad \mathbf{x}_i \in \Omega, \quad \sum_{i=1}^{n} \mu_i = 1$

Lagrangian (point clouds)



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

Eulerian (histograms)



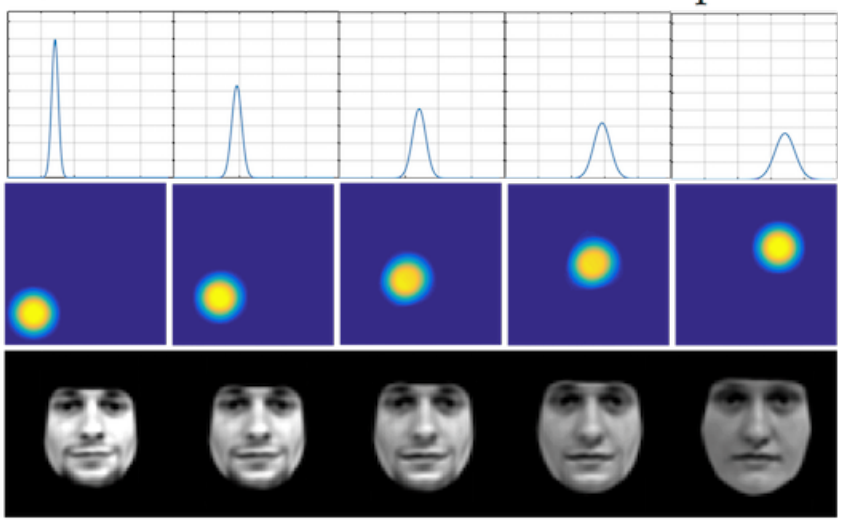
- Fixed positions $\mathbf{x}_i$ e.g. grid
- Convex polytope Σ_n (simplex): $\left\{(\mu_i)_i \geq 0; \sum_i \mu_i = 1\right\}$

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

Geodesic in the 2-Wasserstein space

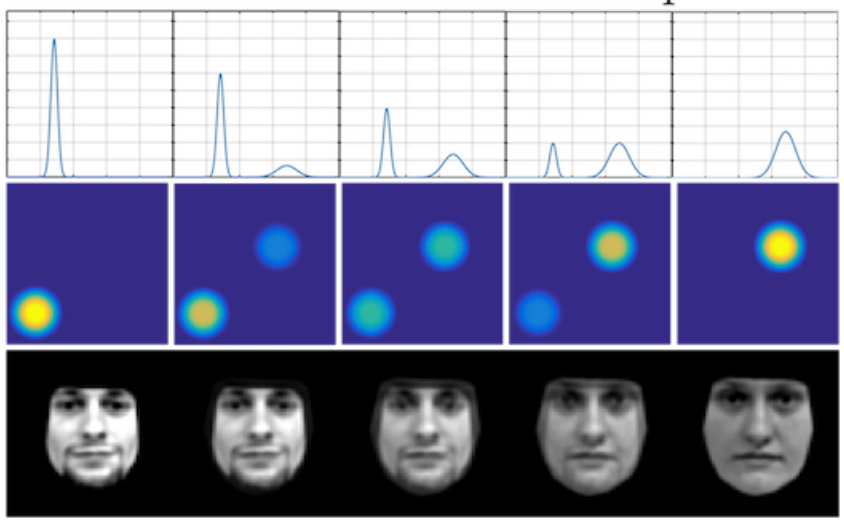


$t = 0 \quad t = 0.25 \quad t = 0.5 \quad t = 0.75 \quad t = 1$

$$\rho^*(.,t) = ((1-t)id + tf^*)_{\#}\mu$$

$$d\rho^*(x,t) = I^*(x,t)dx$$

Geodesic in the Euclidean space

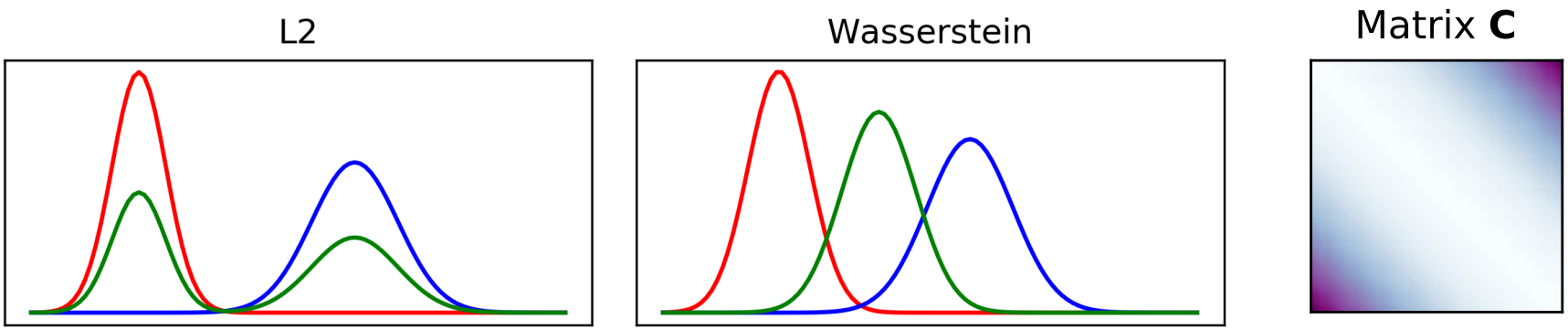


$t = 0 \quad t = 0.25 \quad t = 0.5 \quad t = 0.75 \quad t = 1$

$$I(x,t) = (1-t)I_0(x) + tI_1(x)$$

Illustration by S. Kolhouri

Wasserstein barycenter

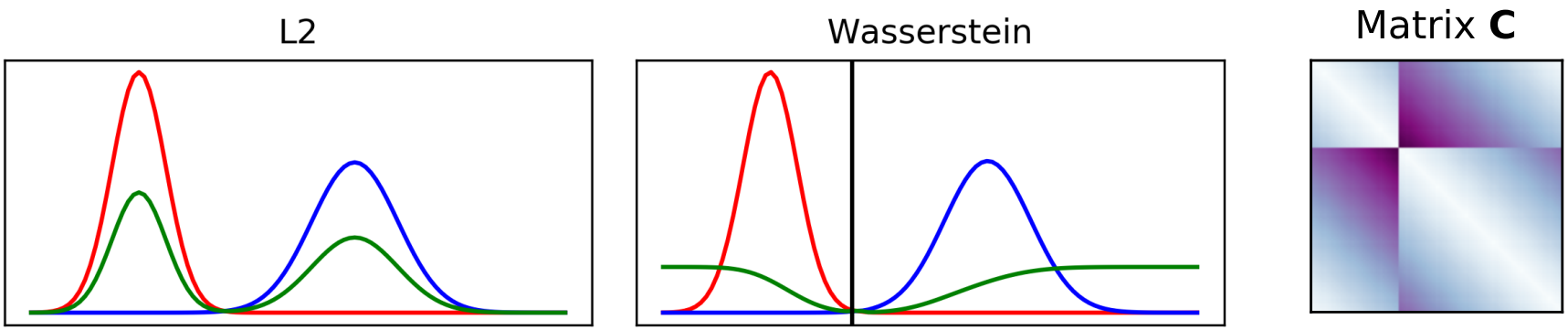


Barycenters [Agueh and Carlier, 2011]

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

- $\lambda_i > 0$ 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 $0 \leq t \leq 1$ [McCann, 1997].
- Regularized barycenters using Bregman projections [Benamou et al., 2015].
- The cost and regularization impacts the interpolation trajectory.

Wasserstein barycenter



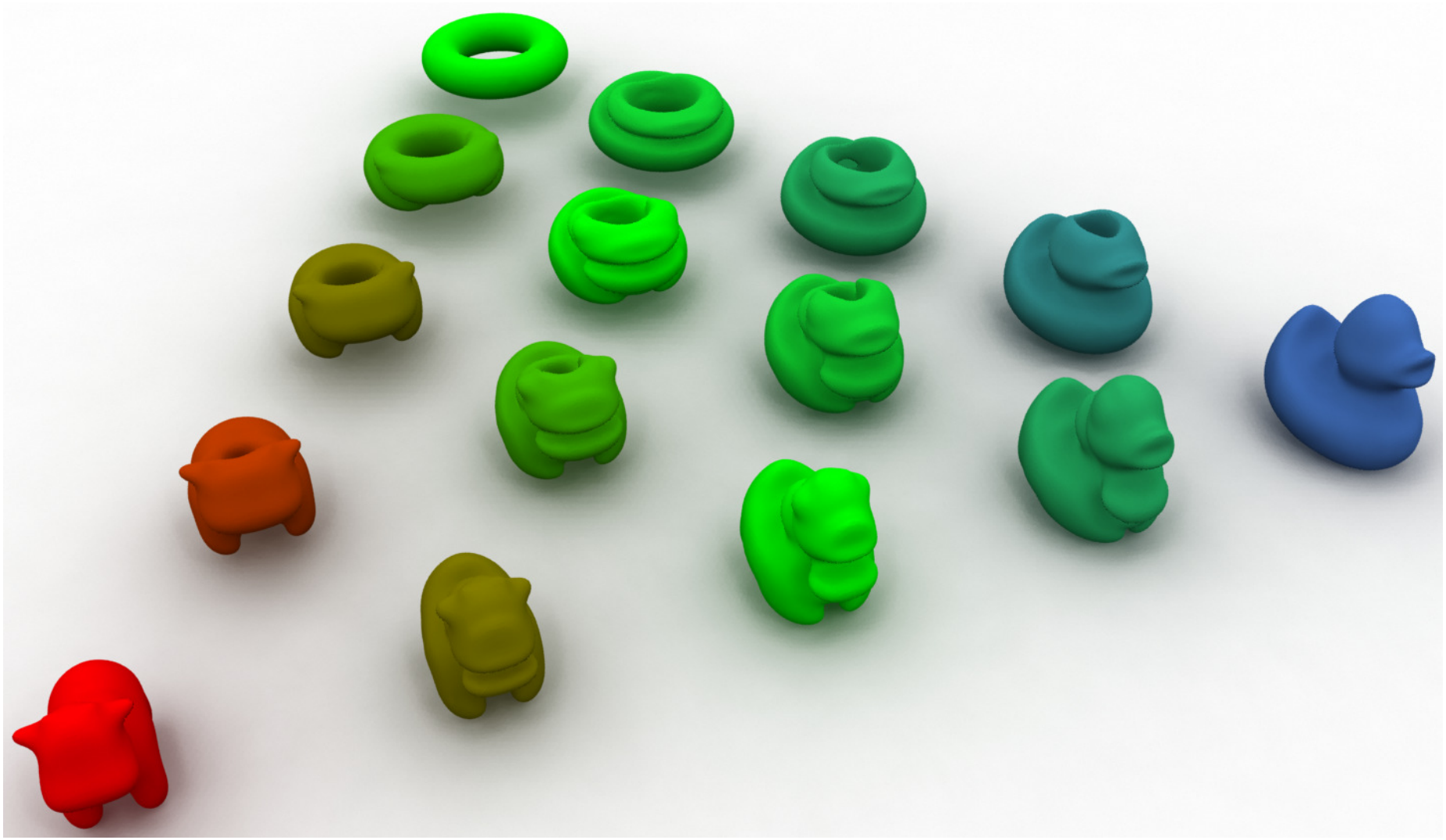
Barycenters [Agueh and Carlier, 2011]

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

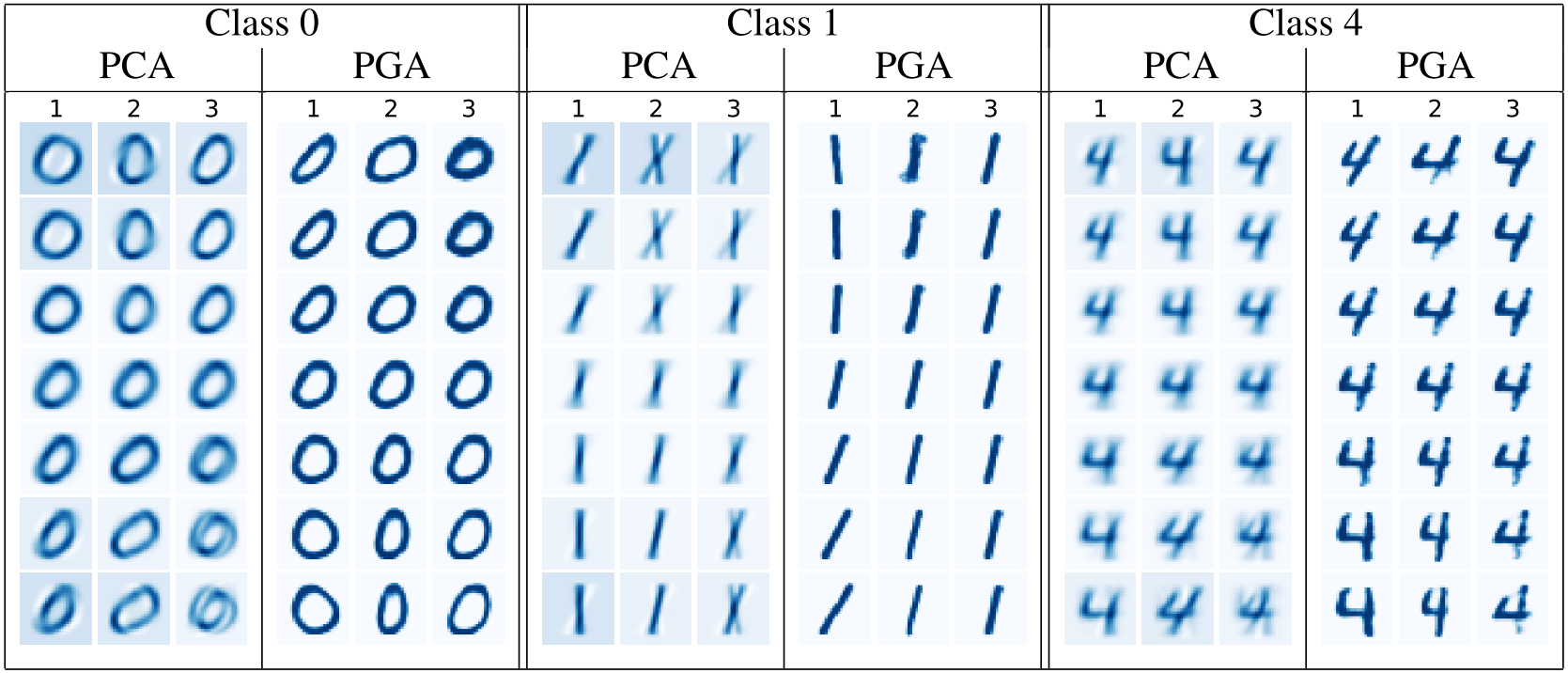
- $\lambda_i > 0$ and $\sum_i^n \lambda_i = 1$.
- Uniform barycenter has $\lambda_i = \frac{1}{n}, \forall i$.
- Interpolation with $n{=}2$ and $\boldsymbol{\lambda} = [1-t, t]$ with $0 \leq t \leq 1$ [McCann, 1997].
- Regularized barycenters using Bregman projections [Benamou et al., 2015].
- The cost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

3D Wasserstein barycenter

Shape interpolation [Solomon et al., 2015]



Principal Geodesics Analysis



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].

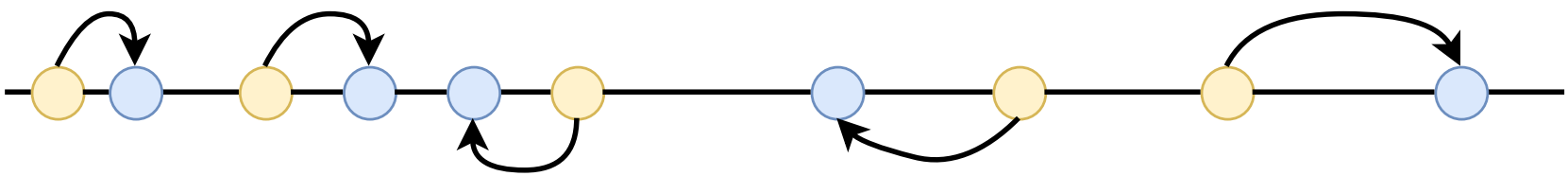
Outline

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.

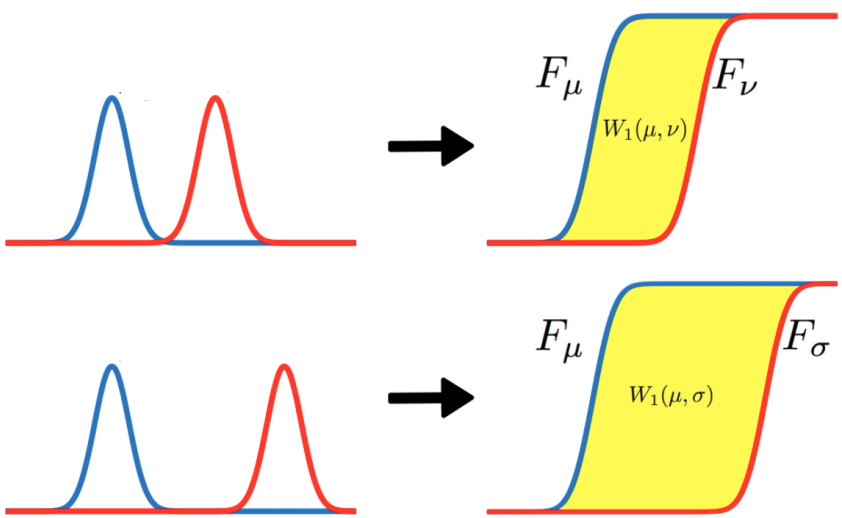
- It is defined such that $F_\mu(t) = \mu(-\infty, t]$.

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-Wassertein on $\mathbb{R}^d$

Wasserstein on $\mathbb{R}$:

$$\forall p \geq 1, \forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}),\ W_p^p(\mu, \nu) = \int_0^1 |F_\mu^{-1}(u) - F_\nu^{-1}(u)|^p \ \mathrm{d}u \tag{4}$$

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.

Sliced-Wasserstein [Rabin et al., 2011b]

Let $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$,

$$SW_p^p(\mu, \nu) = \int_{S^{d-1}} W_p^p(P_\#^\theta \mu, P_\#^\theta \nu) \ \mathrm{d}\lambda(\theta), \tag{5}$$

where $P^\theta(x) = \langle x, \theta \rangle$, λ uniform measure on S^{d-1}.

Properties:

- Distance
- Topogically equivalent to the Wasserstein distance
- Monte-Carlo approximation in $O(Ln \log n)$

This has been extended to a more genral setting. 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]

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

Special cases: Wasserstein on the Circle

Let $\mu, \nu \in \mathcal{P}(S^1)$ where $S^1 = \mathbb{R}/\mathbb{Z}$.

- Parametrize S^1 by $[0, 1[$
- $\forall x, y \in [0, 1[,\ d_{S^1}(x, y) = \min(|x - y|, 1 - |x - y|)$
- For a cost function $c(x, y) = h(d_{S^1}(x, y))$ with $h : \mathbb{R} \to \mathbb{R}^+$ increasing and convex
- $\forall \mu, \nu \in \mathcal{P}(S^1)$, [Rabin et al., 2011a]

$$W_c(\mu, \nu) = \inf_{\alpha \in \mathbb{R}} \int_0^1 h(|F_\mu^{-1}(t) - (F_\nu - \alpha)^{-1}(t)|)\ \mathrm{d}t. \tag{6}$$

- To find α: binary search [Delon et al., 2010]

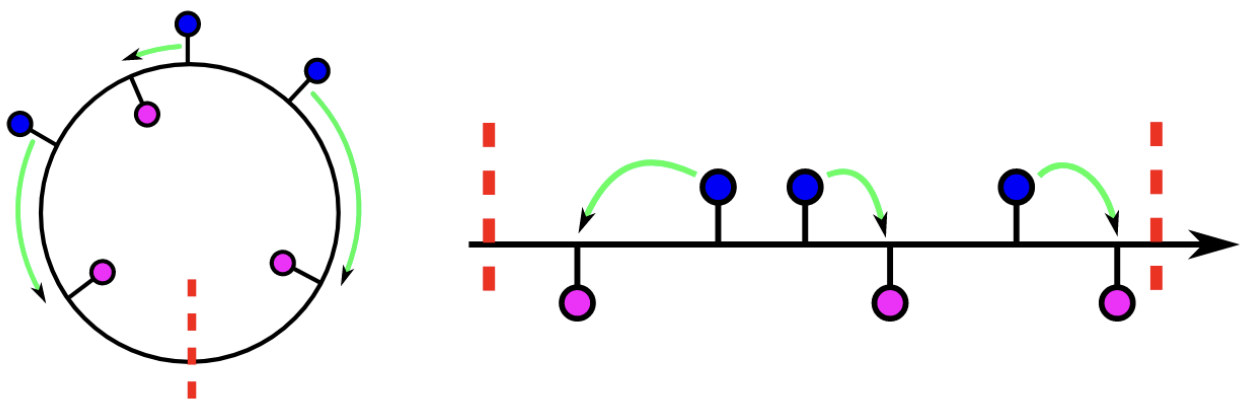


Image from [Rabin et al., 2011a]

Particular Cases

- For $h = \mathrm{Id}$, [Hundrieser et al., 2021]

$$W_1(\mu, \nu) = \int_0^1 |F_\mu(t) - F_\nu(t) - \mathrm{LevMed}(F_\mu - F_\nu)| \, \mathrm{d}t, \tag{7}$$

where

$$\mathrm{LevMed}(f) = \inf \left\{ t \in \mathbb{R}, \ \beta(\{x \in [0,1[, \ f(x) \leq t\}) \geq \frac{1}{2} \right\}. \tag{8}$$

- For $h(x) = x^2$ and $\nu = \mathrm{Unif}(S^1)$,[Bonet et al., 2022]

$$W_2^2(\mu, \nu) = \int_0^1 |F_\mu^{-1}(t) - t - \hat{\alpha}|^2 \, \mathrm{d}t \quad \text{with} \quad \hat{\alpha} = \int x \, \mathrm{d}\mu(x) - \frac{1}{2}. \tag{9}$$

In particular, if $x_1 < \cdots < x_n$ and $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$, then

$$W_2^2(\mu_n, \nu) = \frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i \right)^2 + \frac{1}{n^2} \sum_{i=1}^n (n + 1 - 2i) x_i + \frac{1}{12}. \tag{10}$$

Sliced-Wasserstein on the Sphere

The slicing strategy can be extended to manifolds ! In this case slices are geodesics of the considered manifold. Example on the Sphere [Bonet et al., 2022]:

- Great circle: Intersection between 2-plane and S^{d-1}
- Parametrize 2-plane by the Stiefel manifold

$$\mathbb{V}_{d,2} = \{U \in \mathbb{R}^{d\times 2},\ U^T U = I_2\}$$

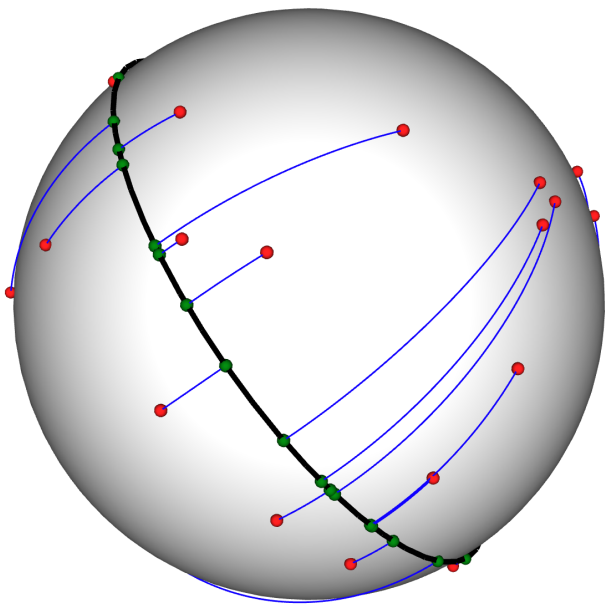
- Projection on great circle C: For a.e. $x \in S^{d-1}$,

$$P^C(x) = \operatorname*{argmin}_{y\in C}\ d_{S^{d-1}}(x, y),$$

where $d_{S^{d-1}}(x, y) = \arccos(\langle x, y\rangle)$.

- For $U \in \mathbb{V}_{d,2}$, $C = \operatorname{span}(UU^T) \cap S^{d-1}$,

$$P^U(x) = U^T \operatorname*{argmin}_{y\in C} d_{S^{d-1}}(x, y)$$
$$= \frac{U^T x}{\|U^T x\|_2}.$$



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

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

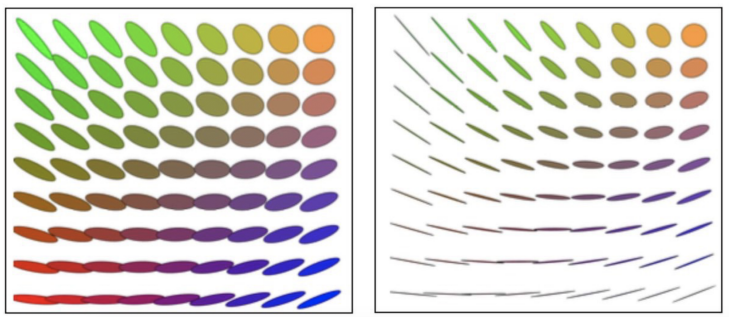
where $\mathbb{B}(,)$ is the so-called Bures metric:

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

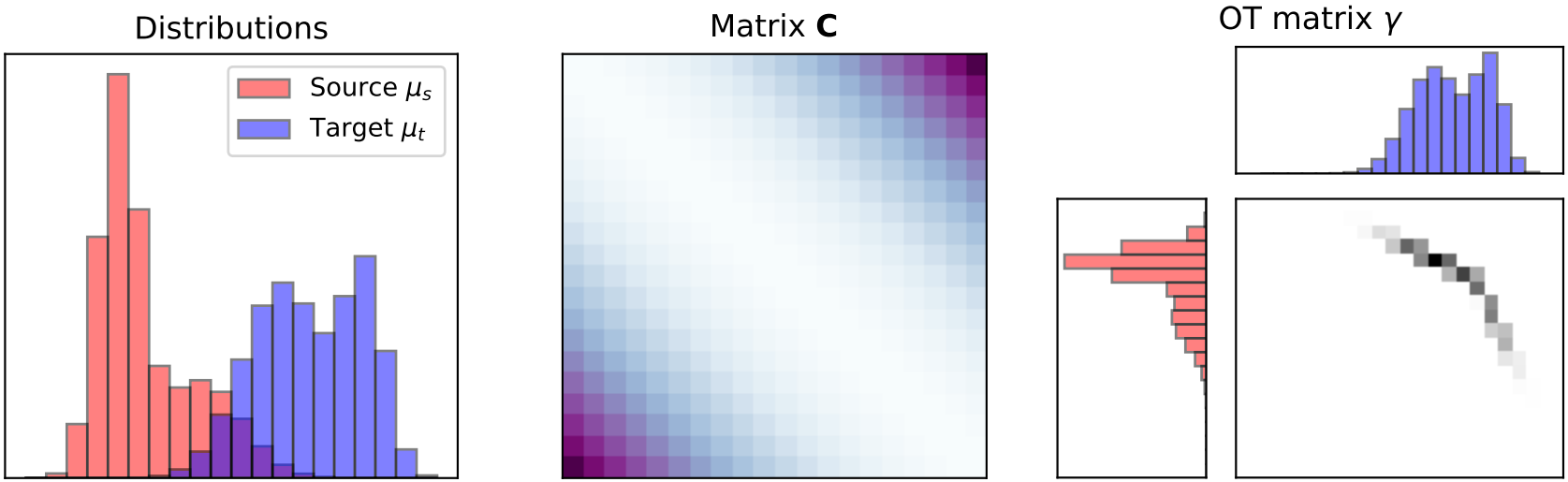
The optimal map T is given by

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

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

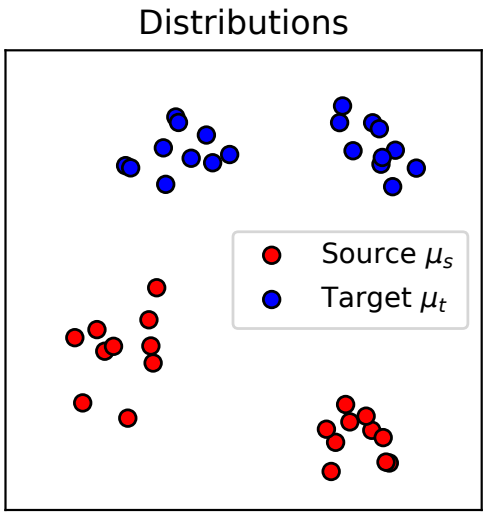
$$\gamma_0 = \underset{\gamma \in \mathcal{P}}{\operatorname{argmin}} \quad \left\{ \langle \gamma, \mathbf{C} \rangle_F = \sum_{i,j} \gamma_{i,j} c_{i,j} \right\}$$

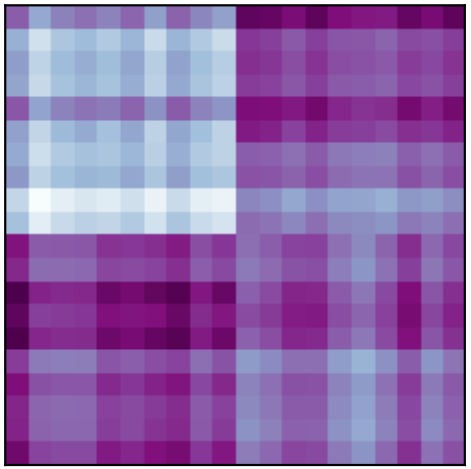
where $\mathbf{C}$ is a cost matrix with $c_{i,j} = c(\mathbf{x}_i^s, \mathbf{x}_j^t)$ and the marginals constraints are

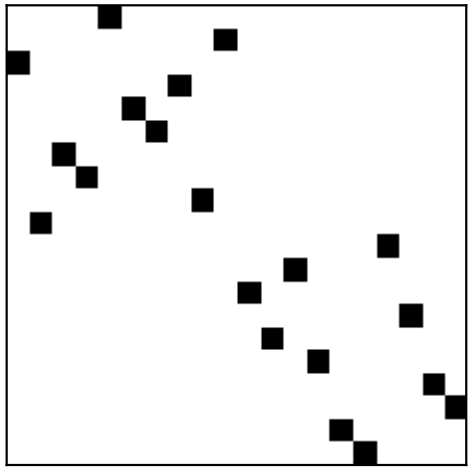
$$\mathcal{P} = \left\{ \gamma \in (\mathbb{R}^+)^{\mathbf{n_s} \times \mathbf{n_t}} \,|\, \gamma \mathbf{1}_{\mathbf{n_t}} = \mu_{\mathbf{s}}, \gamma^{\mathbf{T}} \mathbf{1}_{\mathbf{n_s}} = \mu_{\mathbf{t}} \right\}$$

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

Optimal transport with discrete distributions







OT Linear Program

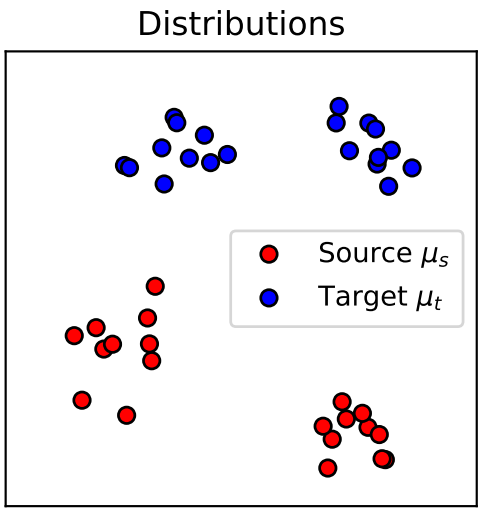
$$\gamma_0 = \underset{\gamma \in \mathcal{P}}{\operatorname{argmin}} \quad \left\{ \langle \gamma, \mathbf{C} \rangle_F = \sum_{i,j} \gamma_{i,j} c_{i,j} \right\}$$

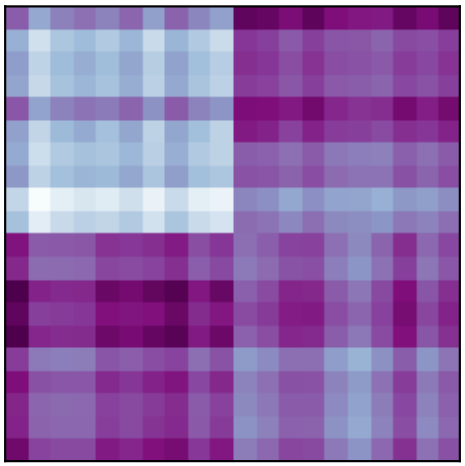
where $\mathbf{C}$ is a cost matrix with $c_{i,j} = c(\mathbf{x}_i^s, \mathbf{x}_j^t)$ and the marginals constraints are

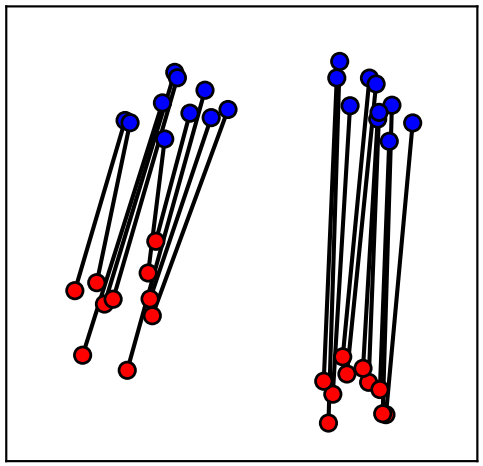
$$\mathcal{P} = \left\{ \boldsymbol{\gamma} \in (\mathbb{R}^+)^{\mathbf{n_s} \times \mathbf{n_t}} \,|\, \boldsymbol{\gamma} \mathbf{1}_{\mathbf{n_t}} = \boldsymbol{\mu_s}, \boldsymbol{\gamma}^{\mathbf{T}} \mathbf{1}_{\mathbf{n_s}} = \boldsymbol{\mu_t} \right\}$$

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

Optimal transport with discrete distributions







OT Linear Program

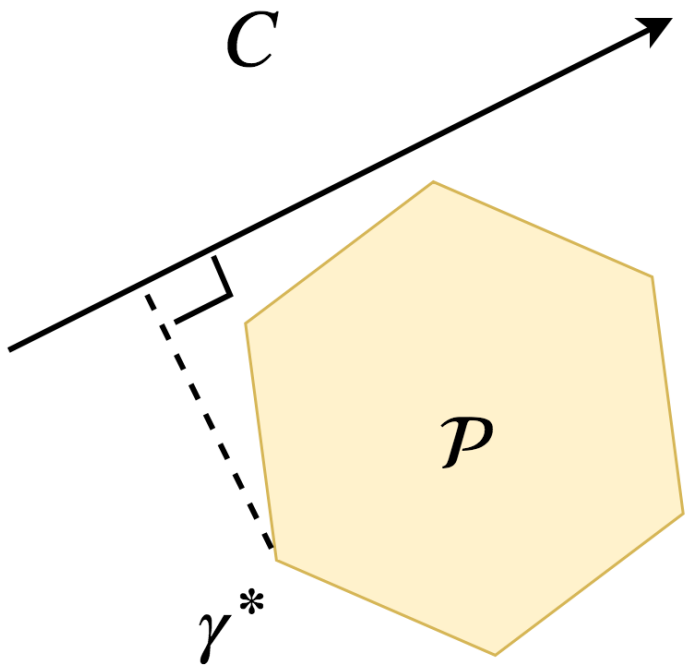
$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma} \in \mathcal{P}}{\operatorname{argmin}} \quad \left\{ \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F = \sum_{i,j} \gamma_{i,j} c_{i,j} \right\}$$

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



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

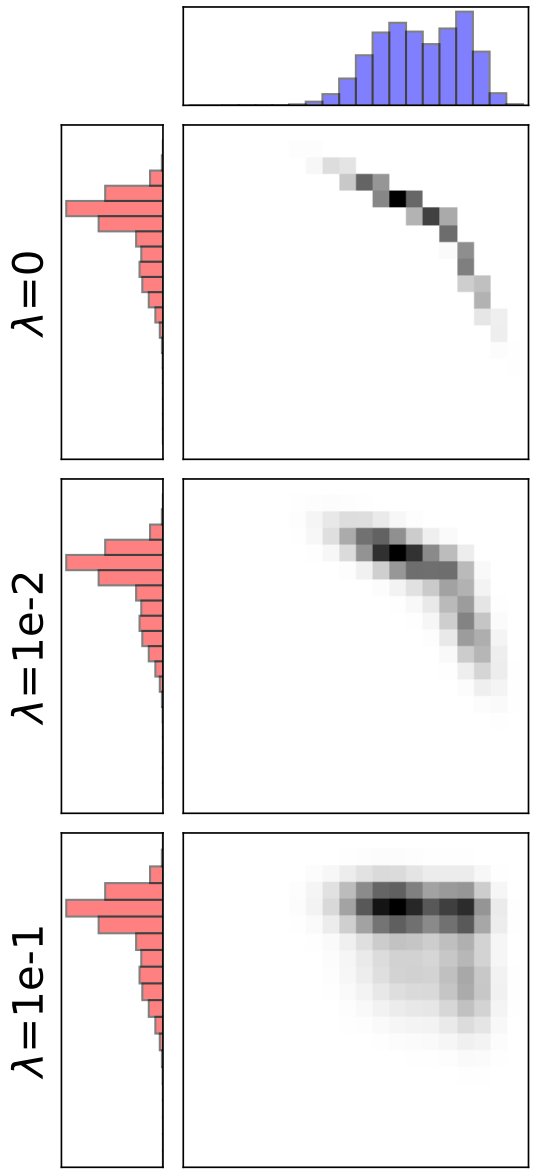
$$\gamma_0^\lambda = \underset{\gamma \in \mathcal{P}}{\operatorname{argmin}} \quad \langle \gamma, \mathbf{C} \rangle_F + \lambda \Omega(\gamma), \qquad (11)$$

Regularization term $\Omega(\gamma)$

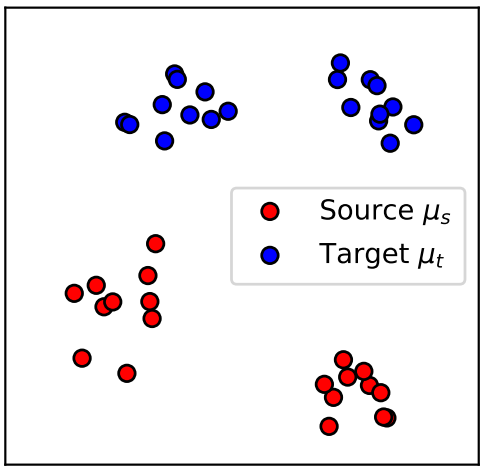
- Entropic regularization [Cuturi, 2013].
- Group Lasso [Courty et al., 2016].
- KL, Itakura Saito, β-divergences, [Dessein et al., 2016].

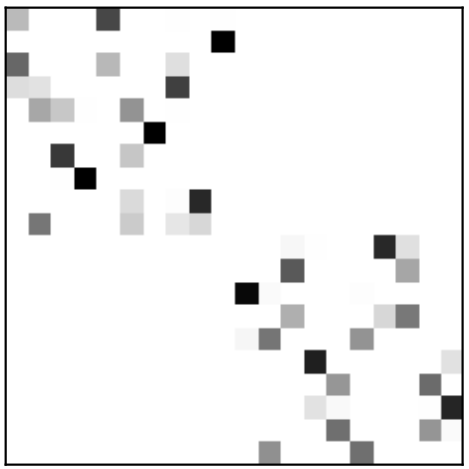
Why regularize?

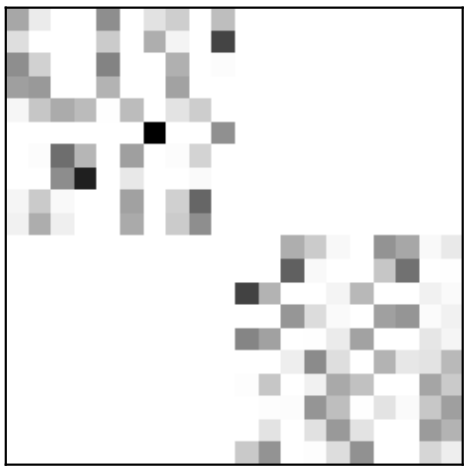
- Smooth the "distance" estimation:
$$W_\lambda(\mu_s, \mu_t) = \langle \gamma_0^\lambda, \mathbf{C} \rangle_F$$
- Encode prior knowledge on the data.
- Better posed problem (convex, stability).
- Fast algorithms to solve the OT problem.



Entropic regularized optimal transport





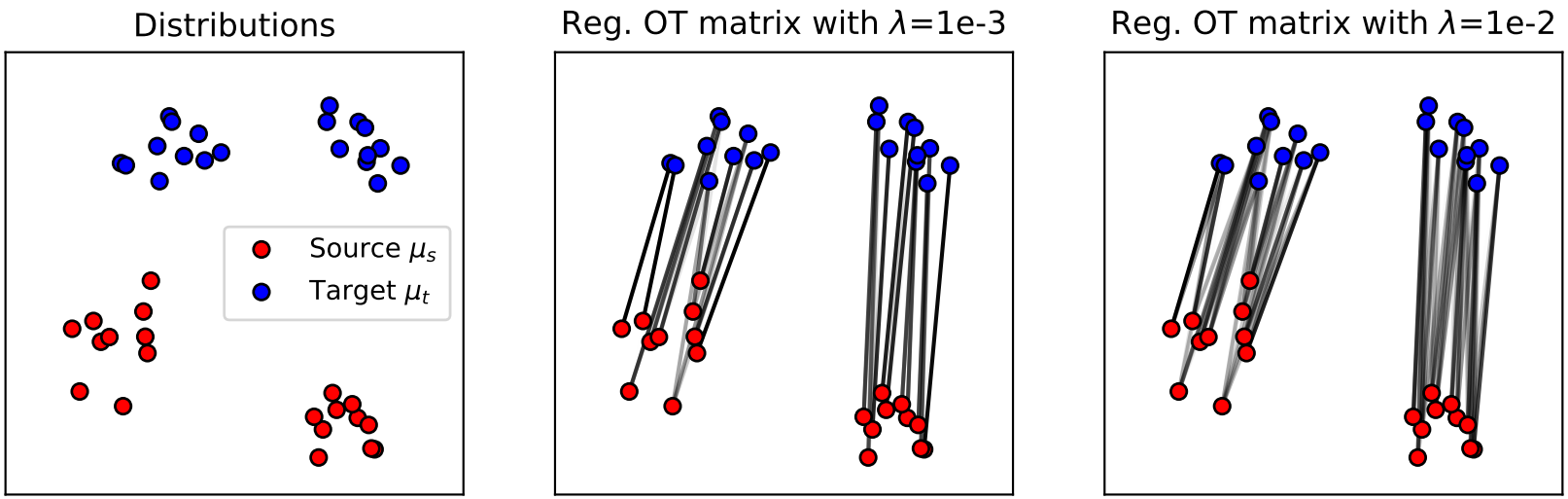


Entropic regularization [Cuturi, 2013]

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

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

Entropic regularized optimal transport



Entropic regularization [Cuturi, 2013]

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

- Regularization with the negative entropy of γ.

Resolving the entropy regularized problem

Entropy-regularized transport

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

$$\boldsymbol{\gamma}_0^\lambda = \text{diag}(\mathbf{u}) \exp(-\mathbf{C}/\lambda) \text{diag}(\mathbf{v})$$

Why ? Consider the Lagrangian of the optimization problem:

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

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

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

Sinkhorn-Knopp algorithm

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 $1, \dots, n_{it}$ **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 $\mathbf{T} = \text{diag}(\mathbf{u}^{(n_{it})})\mathbf{K}\text{diag}(\mathbf{v}^{(n_{it})})$

- 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]

Sinkhorn as Bregman projections

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

$$\boldsymbol{\gamma}^\star = \underset{\boldsymbol{\gamma} \in \mathcal{P}}{\operatorname{argmin}} \, \mathrm{KL}(\boldsymbol{\gamma}, \zeta), \tag{12}$$

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:

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

- 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:

$$\max_{\phi,\psi} \left\{ \int \phi d\mu_s + \int \psi d\mu_t \;\; | \;\; \phi(x) + \psi(y) \leq c(x,y) \right\} \tag{13}$$

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

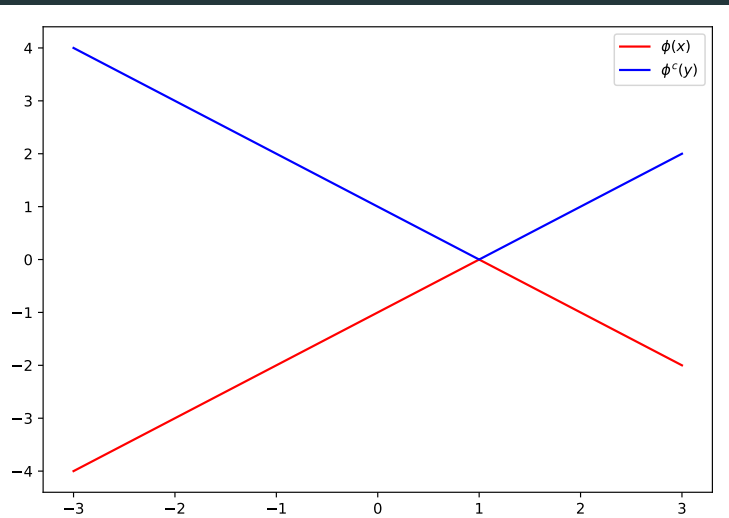
Introducing the *c-transform* (or *c-conjugate*) H^c which is in spirit close to a Legendre transform:

$$\phi^c \stackrel{\text{def}}{=} H^c(\phi) = \inf_x c(x,y) - \phi(x) \tag{14}$$

then the following problem is equivalent:

$$\max_{\phi} \left\{ \int \phi d\mu_s + \int \phi^c d\mu_t \;\; | \;\; \phi(x) + \phi^c(y) \leq c(x,y) \right\} \tag{15}$$

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

$$\sup_{\phi \in \text{Lip}^1} \int \phi d(\mu_s - \mu_t) = \sup_{\phi \in \text{Lip}^1} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mu_s} [\phi(x)] - \mathop{\mathbb{E}}_{\mathbf{y} \sim \mu_t} [\phi(y)] \tag{16}$$

- 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 $\Phi(x)$

$$T(x) = x - \nabla\phi(x) = \nabla(\frac{x^2}{2} - \phi(x)) = \nabla(\Phi(x)) \tag{17}$$

Brenier's Theorem

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, μ absolutely continuous with respect to the Lebesgue measure. Then, the optimal coupling γ^* is unique and of the form $\gamma^* = (Id, \nabla\Phi)_{\#}\mu$ with Φ a convex function.

- can be optimized for instance with Input Convex Neural Networks (ICNN) [Amos et al., 2017] to model the convex functions

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

$$W(\mu_s, \mu_t) = \max_{\alpha \in \mathbb{R}^{n^s}, \beta \in \mathbb{R}^{n^t}, \alpha_i + \beta_j \leq c(\mathbf{X^s_i}, \mathbf{X^t_j})} \alpha^T \mathbf{a} + \beta^T \mathbf{b} \tag{18}$$

i.e. find a scalar values per sample

Regularized case

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

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

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

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

$$\max_{\alpha, \beta} \alpha^T \mu_s + \beta^T \mu_t - \frac{1}{\lambda} \exp(\frac{\alpha}{\lambda})^T \mathbf{K} \exp(\frac{\beta}{\lambda}) \tag{19}$$

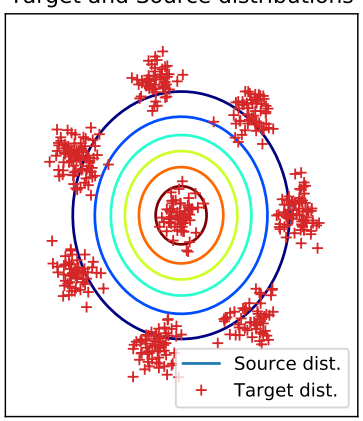
with $\mathbf{K} = \exp(-\frac{\mathbf{C}}{\lambda})$.

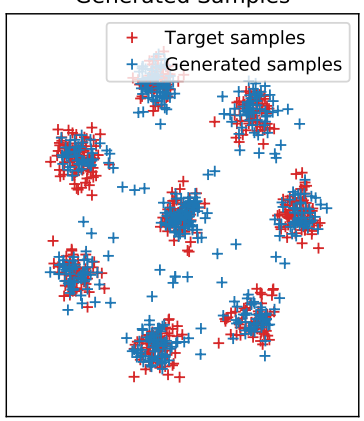
Remark: The Sinkhorn algorithm is a gradient ascent on the dual variables !

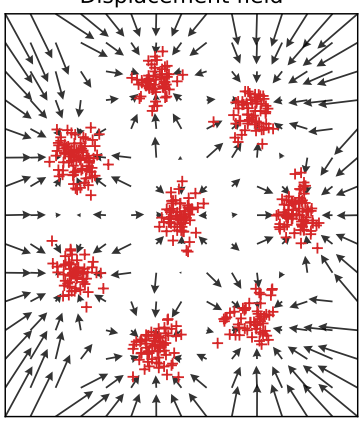
Regularized case

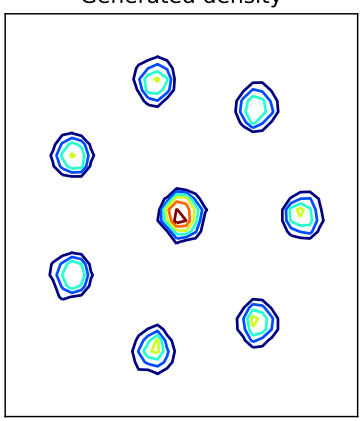
With this unconstrained problem, incremental gradients techniques (SGD, SAG) can be used to solve the problem !

- [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.









2 ways of minimizing the Wasserstein distance

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 $\mathbf{X}$

Case 1: fixed support

Recalling the form of the dual

$$W(\mu, \mu_t) = \max_{\alpha \in \mathbb{R}^{n^s}, \beta \in \mathbb{R}^{n^t}, \alpha_i + \beta_j \leq c(\mathbf{x}, \mathbf{x_j^t})} \alpha^T \mathbf{m} + \beta^T \mathbf{b} \tag{20}$$

- $W(\mu, \mu_t)$ is convex wrt. $\mathbf{m}$
- $\partial_{\mathbf{m}} W(\mu, \mu_t) = \alpha^*$
- **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

$$W_2^2(\mu, \mu_t) = \min_{\boldsymbol{\gamma} \in \mathcal{P}} \quad < \boldsymbol{\gamma}, \mathbf{1_{n^s}} \mathbf{1_{n^t}^T} \mathbf{X}^2 + \mathbf{X^t}^{2T} \mathbf{1_{n^t}} \mathbf{1_{n^s}} - 2\mathbf{X}\mathbf{X^t} > \tag{21}$$

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

General case: autodifferentiation

Automatic differentiation to the rescue !

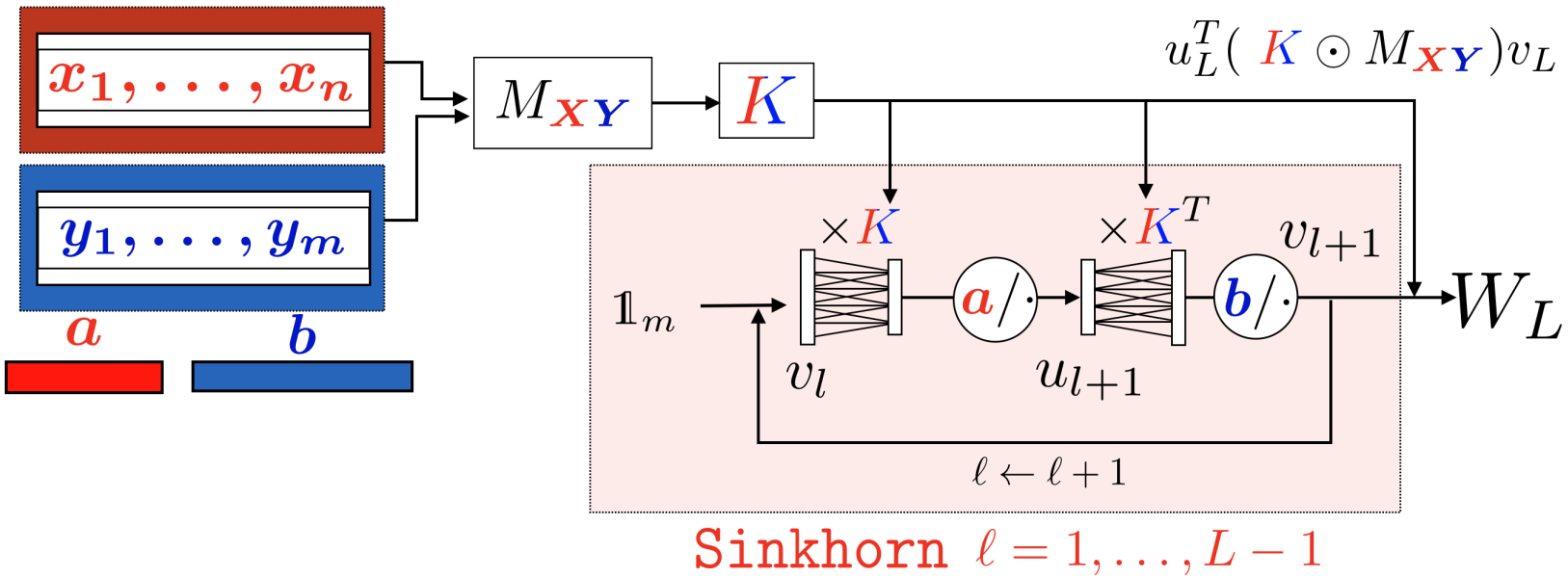


Image from Marco Cuturi

But also consider using the Enveloppe theorem, *i.e.* take the gradient in the optimal solution of the primal problem.

Contents

Outline

Gradient flows

Point of View:

- Define Wasserstein gradient flows
- Analogies with gradient flows in Euclidean space
- For more abstract views, see [Santambrogio, 2017, Ambrosio et al., 2008]
- Talk from Anna Korba https://mathtube.org/lecture/video/wasserstein-gradient-flows-machine-learning

Gradient Flows on $\mathbb{R}^p$

Let $X = \mathbb{R}^p$, d a distance (*e.g.* $d(x, y) = \|x - y\|_2$), $F : X \to \mathbb{R}$.

Goal:

$$\min_x \; F(x)$$

Gradient Flows on $\mathbb{R}^p$

Let $X = \mathbb{R}^p$, d a distance (*e.g.* $d(x, y) = \|x - y\|_2$), $F : X \to \mathbb{R}$.

Goal:

$$\min_x \ F(x)$$

Definition (Gradient Flow on $\mathbb{R}^p$)
A gradient flow is a curve $x : [0, T] \to X$ which decreases as much as possible along the functional F.

i.e. If F is differentiable, x follows the Cauchy problem

$$\begin{cases} \frac{\mathrm{d}x}{\mathrm{d}t}(t) = -\nabla F(x(t)) \\ x(0) = x_0 \end{cases}$$

Gradient Flows on $\mathbb{R}^p$

If F is differentiable, x follows the Cauchy problem

$$\begin{cases} \frac{\mathrm{d}x}{\mathrm{d}t}(t) = -\nabla F(x(t)) \\ x(0) = x_0 \end{cases}$$

Solving the ODE in practice:

- Explicit Euler scheme ($x_k = x(k\tau)$):

$$x_{k+1} = x_k - \tau \nabla F(x_k)$$

- Implicit Euler scheme:

$$\begin{aligned} x_{k+1} = x_k - \tau \nabla F(x_{k+1}) &\iff 0 = \frac{x_{k+1} - x_k}{\tau} + \nabla F(x_{k+1}) \\ &\iff x_{k+1} \in \operatorname*{argmin}_{x \in X} \frac{\|x - x_k\|_2^2}{2\tau} + F(x) \\ &\iff x_{k+1} = \operatorname{prox}_{\tau F}(x_k) \end{aligned}$$

- Any ODE solver (Runge-Kutta...)

Other characterization

See [Santambrogio, 2017, Ambrosio et al., 2008]

- Energy Dissipation Equality (EDE):

$$\begin{aligned} &\frac{\mathrm{d}x}{\mathrm{d}t}(t) = -\nabla F(x(t)) \iff \forall 0 \leq s < t \leq 1, \\ &F(x(s)) - F(x(t)) = \int_s^t \left(\frac{1}{2}|x'(u)|^2 + \frac{1}{2}|\nabla F(x(u))|^2 \right) \mathrm{d}u, \end{aligned} \tag{22}$$

  where (speed)

$$|x'(t)| = \lim_{h \to 0} \frac{d\big(x(t+h), x(t)\big)}{h}$$

  and (descending slope)

$$|\nabla F|(x) = \limsup_{y \to x} \frac{|F(y) - F(x)|}{d(x,y)}.$$

- Evolution Variational Inequality (EVI): For F λ-geodesically convex,

$$x'(t) \in \partial F(x(t)) \iff \forall y \in X, \ \frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}|x(t) - y|^2 \leq F(y) - F(x(t)) - \frac{\lambda}{2}|x(t) - y|^2 \tag{23}$$

JKO Scheme

Let $\mathcal{P}_2(\mathbb{R}^p) = \{\mu \in \mathcal{P}(\mathbb{R}^p),\ \int \|x\|^2 \,\mathrm{d}\mu(x) < +\infty\}$.

Define Gradient Flows in $(\mathcal{P}_2(\mathbb{R}^p), W_2)$ via the JKO Scheme [Jordan et al., 1998]: Let $F : \mathcal{P}_2(\mathbb{R}^p) \to \mathbb{R}$, $\mu_0 \in \mathcal{P}_2(\mathbb{R}^p)$,

$$\forall k \geq 0,\ \mu_{k+1}^\tau \in \underset{\mu \in \mathcal{P}_2(\mathbb{R}^p)}{\operatorname{argmin}} \ \frac{W_2^2(\mu, \mu_k^\tau)}{2\tau} + F(\mu) = \mathrm{JKO}_{\tau F}(\mu_k^\tau) \tag{24}$$

Define a piecewise constant interpolation μ^τ, *i.e.*

$$\forall t \in [k\tau, (k+1)\tau[,\ \mu_t^\tau = \mu_k^\tau$$

Wasserstein gradient flows: $t \mapsto \mu_t$ such that, for $\mu^\tau \xrightarrow[\tau \to 0]{} \mu$.

(also called (generalized) minimizing movement [Bonnotte, 2013, Liutkus et al., 2019])

Gradient Flow in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$:

Iterated Minimization scheme (JKO Scheme) [Jordan et al., 1998]:

$$\mu_{k+1}^{\tau} \in \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \frac{1}{2\tau} W_2^2(\mu, \mu_k^{\tau}) + F(\mu)$$

Examples

- $F(\mu) = \int \rho(x) \log \rho(x) \, \mathrm{d}x + \int V(x)\rho(x)\mathrm{d}x$ if $\mathrm{d}\mu = \rho \mathrm{dLeb}$
  Solution in the limit $\tau \to 0$ to the PDE: (Fokker-Planck)

$$\partial_t \rho_t = \operatorname{div}(\rho_t \nabla V) + \Delta \rho_t$$

- $F(\mu) = \frac{1}{2} SW_2^2(\mu, \nu) + \lambda \mathcal{H}(\mu)$ [Bonnotte, 2013, Liutkus et al., 2019]
- $F(\mu) = \frac{1}{2} MMD^2(\mu, \nu)$ [Arbel et al., 2019]
- $F(\mu) = \frac{1}{2} \mathrm{KSD}^2(\mu, \nu)$ [Korba et al., 2021]

- If an associated SDE is known, simulate from it
  [Liu et al., 2021, Liutkus et al., 2019, Arbel et al., 2019, Korba et al., 2021]

Examples

Let $F(\mu) = \int V(x)\rho(x)\mathrm{d}x + \int \log(\rho(x))\rho(x)\mathrm{d}x$,

Gradient Flow solution of:

$$\partial_t \rho_t = \mathrm{div}(\rho_t \nabla V) + \Delta \rho_t$$

Associated SDE (Langevin Equation):

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\ \mathrm{d}W_t$$

First Algorithms of Resolution of WGFs

If SDE is unknow, need specific numerical methods to solve for Wasserstein Gradient Flows by the JKO Scheme:

- By approximating W_2:
  - By the entropic regularized OT problem + Dykstra's algorithm + $\mu = \sum_{i=1}^n \rho_i \delta_{x_i}$, $(x_i)_i$ grid [Peyré, 2015, Carlier et al., 2017]

$$\forall k,\ \mu_{k+1}^\tau \in \underset{\mu \in \mathcal{P}_2(\mathbb{R}^p)}{\operatorname{argmin}} \ \frac{W_\epsilon^2(\mu, \mu_k^\tau)}{2\tau} + F(\mu) \tag{25}$$

  - By the dual formulation of the entropic OT problem [Caluya and Halder, 2019, Frogner and Poggio, 2020]
  - By using SW_2 [Bonet et al., 2021] + Neural networks g^θ and implicit modeling, *i.e.* $\mu = g_{\#}^\theta p_Z$

$$\forall k,\ \mu_{k+1}^\tau \in \underset{\mu \in \mathcal{P}_2(\mathbb{R}^p)}{\operatorname{argmin}} \ \frac{SW_2^2(\mu, \mu_k^\tau)}{2\tau} + F(\mu) \tag{26}$$

- By using the dynamic formulation of the transport + grid discretization [Laborde, 2016, Carrillo et al., 2021]
- JKO-ICNN [Alvarez-Melis et al., 2021, Mokrov et al., 2021, Bunne et al., 2021]

Theorem (Brenier's Theorem)
Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, μ absolutely continuous with respect to the Lebesgue measure. Then, the optimal coupling γ^* is unique and of the form $\gamma^* = (Id, \nabla\varphi)_{\#}\mu$ with φ is a convex function.

- Reformulate the problem as:

$$u_{k+1}^{\tau} \in \underset{u \in \text{cvx}}{\operatorname{argmin}} \ \frac{1}{2\tau} \int \|\nabla u(x) - x\|_2^2 \ \rho_k^{\tau}(x) \mathrm{d}x + F\big((\nabla u)_{\#}\rho_k^{\tau}\big)$$

- Implicitly define $\rho_{k+1}^{\tau} = (\nabla u_{k+1}^{\tau})_{\#}\rho_k^{\tau}$
- Use Input Convex Neural Networks (ICNN) [Amos et al., 2017] to model the convex functions:

$$\theta_{k+1}^{\tau} \in \underset{\theta \in \{\theta, u_\theta \in \text{ICNN}\}}{\operatorname{argmin}} \ \frac{1}{2\tau} \int \|\nabla_x u_\theta(x) - x\|_2^2 \ \rho_k^{\tau}(x) \mathrm{d}x + F\big((\nabla_x u_\theta)_{\#}\rho_k^{\tau}\big)$$

- Backpropagate through gradient

Resolution of Wasserstein Gradient Flows by the JKO Scheme:

- By approximating W_2
- By using the dynamic formulation of the transport + grid discretization [Laborde, 2016, Carrillo et al., 2021]
- JKOICNN: Modeling the Monge map with ICNNs [Alvarez-Melis et al., 2021, Mokrov et al., 2021, Bunne et al., 2021]

$$\theta_{k+1}^{\tau} \in \underset{\tau \in \{\theta, u_\theta \in \mathrm{ICNN}\}}{\operatorname{argmin}} \frac{1}{2\tau} \int \|\nabla u_\theta(x) - x\|_2^2 \, \mathrm{d}\mu_k^{\tau}(x) + F\big((\nabla_x u_\theta)_{\#}\mu_k^{\tau}\big) \tag{27}$$

- Modeling directly the Monge map

$$T_{k+1}^{\tau} \in \underset{T}{\operatorname{argmin}} \frac{1}{2\tau} \int \|T(x) - x\|_2^2 \, \mathrm{d}\mu_k^{\tau}(x) + F(T_{\#}\mu_k^{\tau}) \tag{28}$$

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.

Alvarez-Melis, D., Schiff, Y., and Mroueh, Y. (2021).

Optimizing functionals on the space of probabilities with input convex neural networks.

Ambrosio, L., Gigli, N., and Savaré, G. (2008).

Gradient flows: in metric spaces and in the space of probability measures.

Springer Science & Business Media.

Amos, B., Xu, L., and Kolter, J. Z. (2017).

Input convex neural networks.

In *International Conference on Machine Learning*, pages 146–155. PMLR.

References ii

Arbel, M., Korba, A., Salim, A., and Gretton, A. (2019).

Maximum mean discrepancy gradient flow.

arXiv preprint arXiv:1906.04370.

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.

Bonet, C., Berg, P., Courty, N., Septier, F., Drumetz, L., and Pham, M.-T. (2022).

Spherical sliced-wasserstein.

References iii


Bonet, C., Courty, N., Septier, F., and Drumetz, L. (2021).
Sliced-wasserstein gradient flows.


Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015).
Sliced and radon Wasserstein barycenters of measures.
Journal of Mathematical Imaging and Vision, 51:22–45.


Bonnotte, N. (2013).
Unidimensional and evolution methods for optimal transportation.
PhD thesis, Paris 11.


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


Bunne, C., Meng-Papaxanthos, L., Krause, A., and Cuturi, M. (2021).
Jkonet: Proximal optimal transport modeling of population dynamics.


Caluya, K. F. and Halder, A. (2019).
Proximal recursion for solving the fokker-planck equation.
In *2019 American Control Conference (ACC)*, pages 4098–4103. IEEE.


Carlier, G., Duval, V., Peyré, G., and Schmitzer, B. (2017).
Convergence of entropic schemes for optimal transport and gradient flows.
SIAM Journal on Mathematical Analysis, 49(2):1385–1418.


Carrillo, J. A., Craig, K., Wang, L., and Wei, C. (2021).
Primal dual methods for wasserstein gradient flows.
Foundations of Computational Mathematics, pages 1–55.


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.


Delon, J., Salomon, J., and Sobolevski, A. (2010).
Fast transport optimization for monge costs on the circle.
SIAM Journal on Applied Mathematics, 70(7):2239–2258.


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).


Frogner, C. and Poggio, T. (2020).

Approximate inference with wasserstein gradient flows.

In *International Conference on Artificial Intelligence and Statistics*, pages 2581–2590. PMLR.


Genevay, A., Cuturi, M., Peyré, G., and Bach, F. (2016).

Stochastic optimization for large-scale optimal transport.

In *NIPS*, pages 3432–3440.


Hundrieser, S., Klatt, M., and Munk, A. (2021).

The statistics of circular optimal transport.

arXiv preprint arXiv:2103.15426.


Jordan, R., Kinderlehrer, D., and Otto, F. (1998).
The variational formulation of the fokker–planck equation.
SIAM journal on mathematical analysis, 29(1):1–17.


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


Korba, A., Aubin-Frankowski, P.-C., Majewski, S., and Ablin, P. (2021).
Kernel stein discrepancy descent.
arXiv preprint arXiv:2105.09994.


Laborde, M. (2016).
Interacting particles systems, Wasserstein gradient flow approach.
PhD thesis, PSL Research University.

Liu, S., Sun, H., and Zha, H. (2021).

Approximating the optimal transport plan via particle-evolving method.

Liutkus, A., Simsekli, U., Majewski, S., Durmus, A., and Stöter, F.-R. (2019).

Sliced-wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions.

In *International Conference on Machine Learning*, pages 4104–4113. PMLR.

McCann, R. J. (1997).

A convexity principle for interacting gases.

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

Mokrov, P., Korotin, A., Li, L., Genevay, A., Solomon, J., and Burnaev, E. (2021).

Large-scale wasserstein gradient flows.


Monge, G. (1781).

Mémoire sur la théorie des déblais et des remblais.

De l'Imprimerie Royale.


Peyré, G. (2015).

Entropic approximation of wasserstein gradient flows.

SIAM Journal on Imaging Sciences, 8(4):2323–2351.


Rabin, J., Delon, J., and Gousseau, Y. (2011a).

Transportation distances on the circle.

Journal of Mathematical Imaging and Vision, 41(1):147–167.


Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2011b).

Wasserstein barycenter and its application to texture mixing.

In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer.


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.


Santambrogio, F. (2017).

{**Euclidean, metric, and Wasserstein**} **gradient flows: an overview.**

Bulletin of Mathematical Sciences, 7(1):87–154.


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.