Optimal Transport for Machine Learning

Part 1: Introduction to Computational Optimal Transport

Rémi Flamary, Nicolas Courty

July 4 2022

Hi Paris Summer School, 2022, Paris

Overview of OTML part of the course

Part 1 : Introduction to optimal transport

- Optimal transport problem
- Wasserstein distance and geometry
- Computational aspects and regularized OT
- Optimal Transport extensions

Part 2 : Learning with optimal transport

- Learning to map with OT
- Learning from histograms
- Learning from empirical distributions

Table of content (Part 1)

Optimal transport

What is optimal transport ?

The natural geometry of probability measures


Monge


Kantorovich


Koopmans

Nobel '75


Dantzig


Brenier


Otto


McCann


Villani

Fields '10


Figalli

Fields '18

The origins of optimal transport

666 · MÉMOIRES DE L'ACADÉMIE ROYALE

MÉMOIRE

SUR LA

THÉORIE DES DÉBLAIS

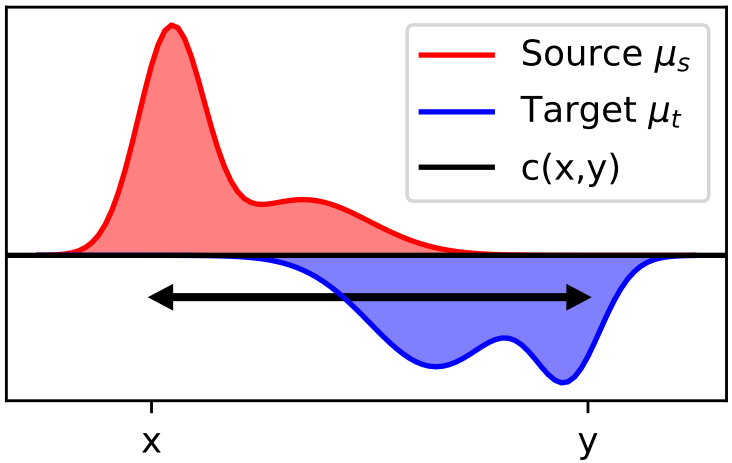
ET DES REMBLAIS.

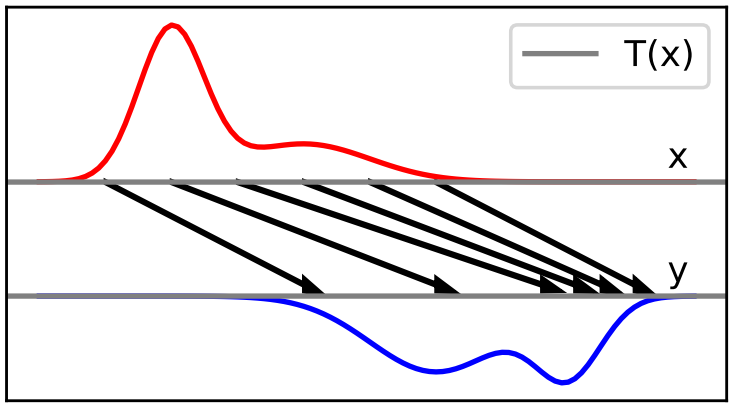
Par M. MONGE.

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

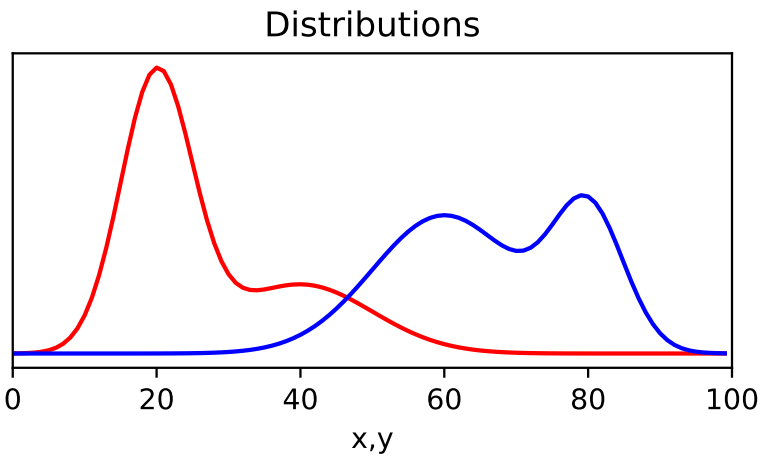


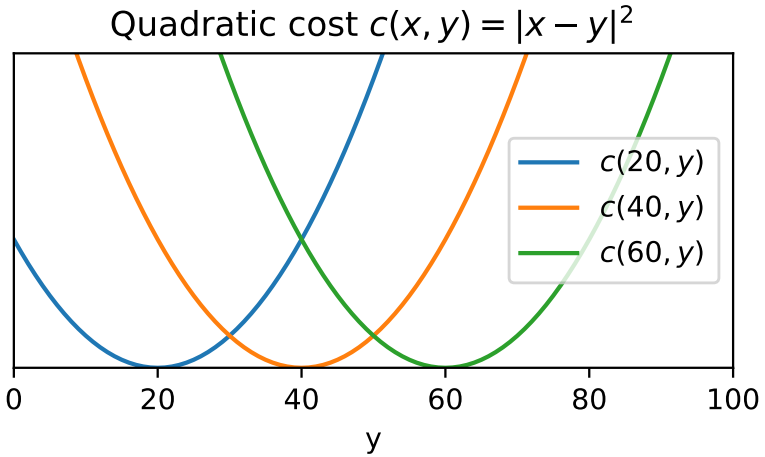


Problem [Monge, 1781]

- 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)



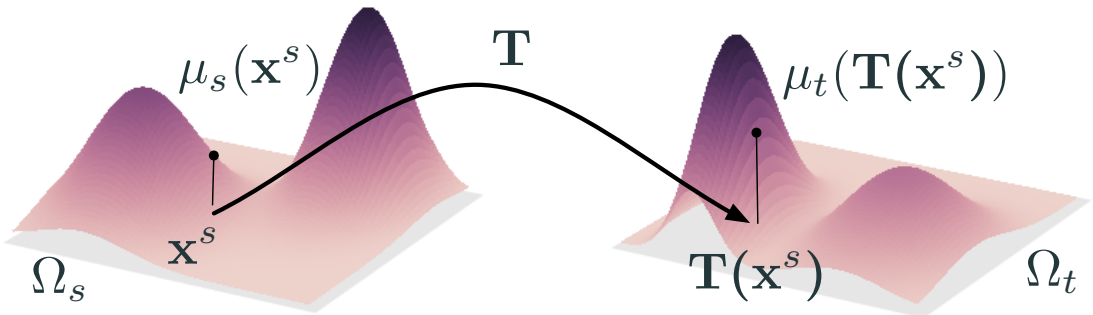


- Probability measures μ_s and μ_t on and a cost function $c : \Omega_s \times \Omega_t \to \mathbb{R}^+$.
- The Monge formulation [Monge, 1781] aim at finding a mapping $T : \Omega_s \to \Omega_t$

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

- Non convex problem because of the constraint $T\#\mu_s = \mu_t$.

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



Pushforward operator $T\#$

- Transfers measures from one space Ω_s to another space Ω_t

$$\mu_t(A) = \mu_s(T^{-1}(A)), \quad \forall \text{ Borel subset } A \in \Omega_s$$

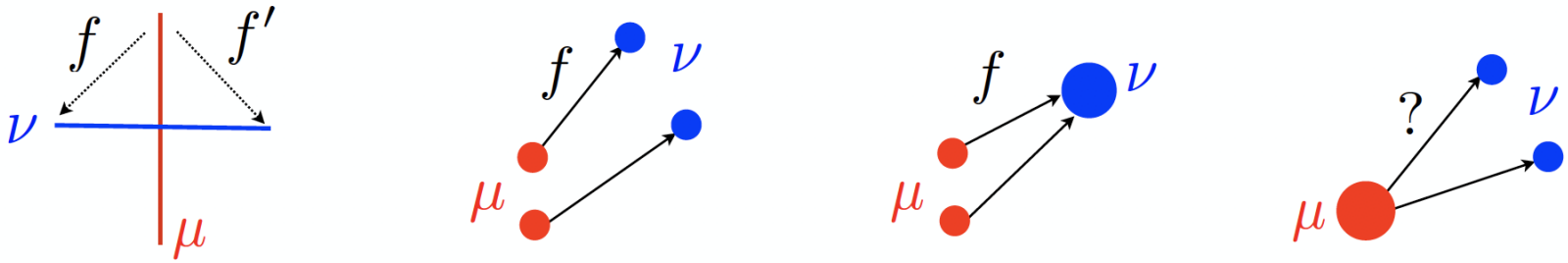
- For smooth measures $\mu_s = \rho(x)dx$ and $\mu_t = \eta(x)dx$

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

  a.k.a. change of variable formula.

- For a discrete distribution $\mu = \sum_i a_i \delta_{\mathbf{x}_i}$ then $T\#\mu = \sum_i a_i \delta_{T(\mathbf{x}_i)}$.

Properties of mapping T

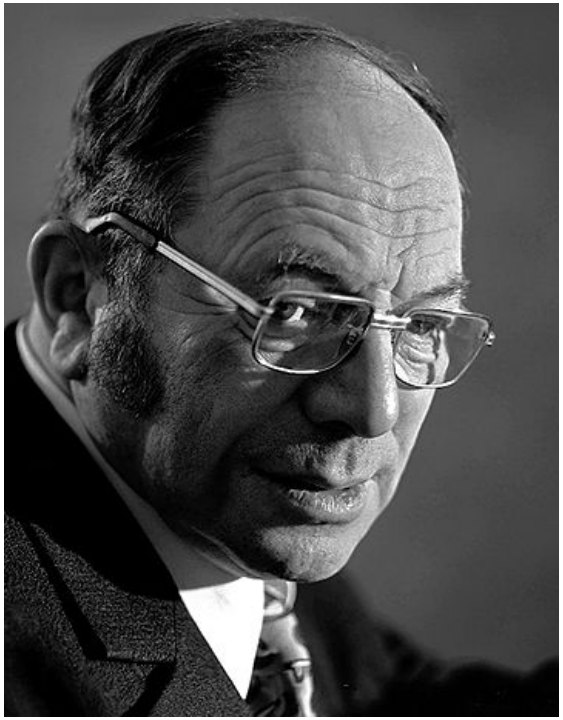


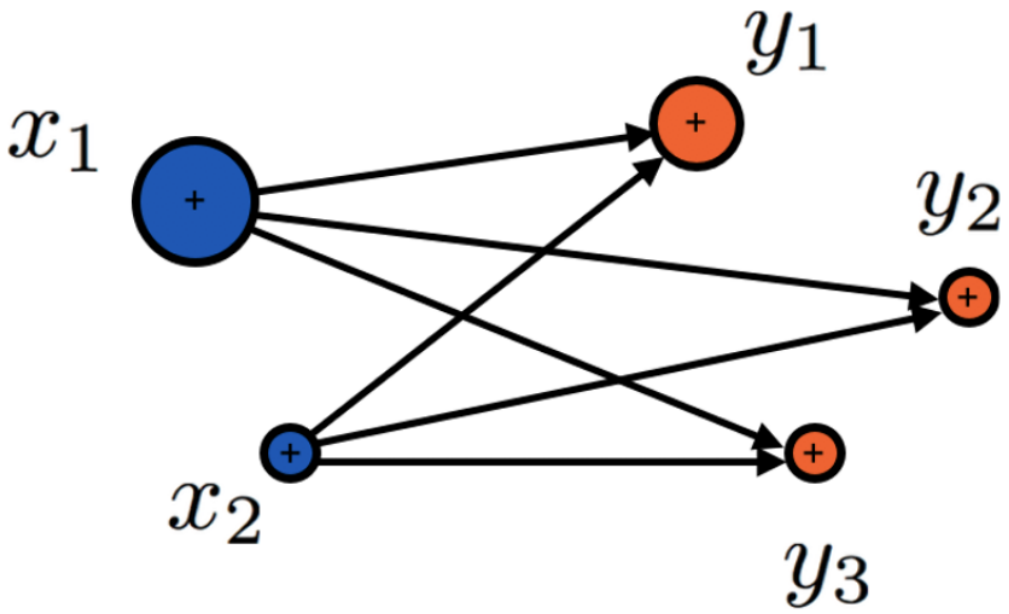
Non-existence / Non-uniqueness

- $T\#\mu_s = \mu_t$ is a non-convex constraint.
- Existence of T is not guaranteed.
- Unicity of T is not guaranteed.
- [Brenier, 1991] proved existence and unicity of the Monge map for $c(x, y) = \|x - y\|^2$ and distributions with densities (i.e. continuous).

Image from Gabriel Peyré

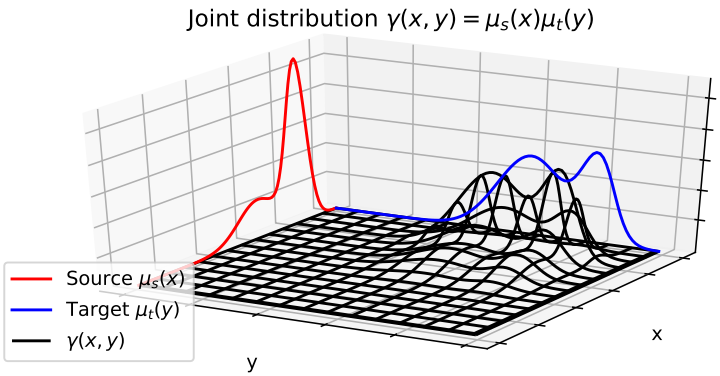
Kantorovich relaxation

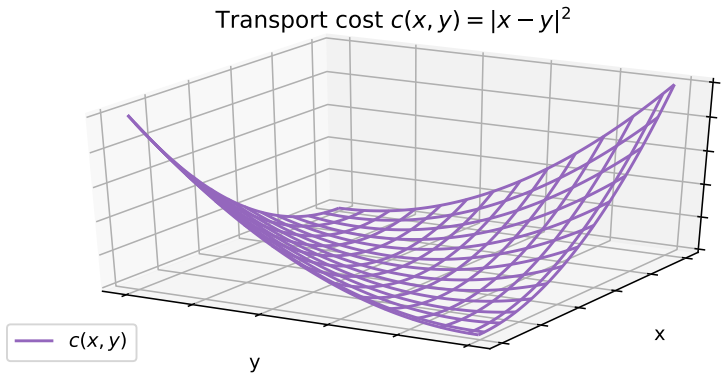




- Leonid Kantorovich (1912–1986), Economy nobelist in 1975
- Focus on where the mass goes, allow splitting [Kantorovich, 1942].
- Applications mainly for resource allocation problems

Optimal transport (Kantorovich formulation)





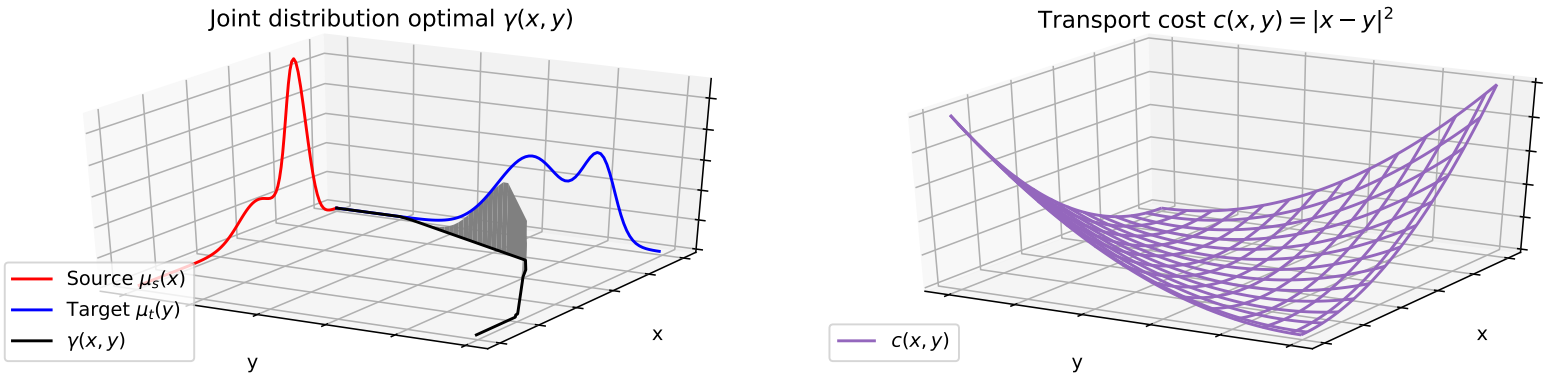
- The Kantorovich formulation [Kantorovich, 1942] seeks for a probabilistic coupling $T \in \mathcal{P}(\Omega_s \times \Omega_t)$ between Ω_s and Ω_t:

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

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

- T is a joint probability measure with marginals μ_s and μ_t.
- Linear Program that always has a solution ($\mu_s \otimes \mu_t \in \mathcal{P}$).

Optimal transport (Kantorovich formulation)



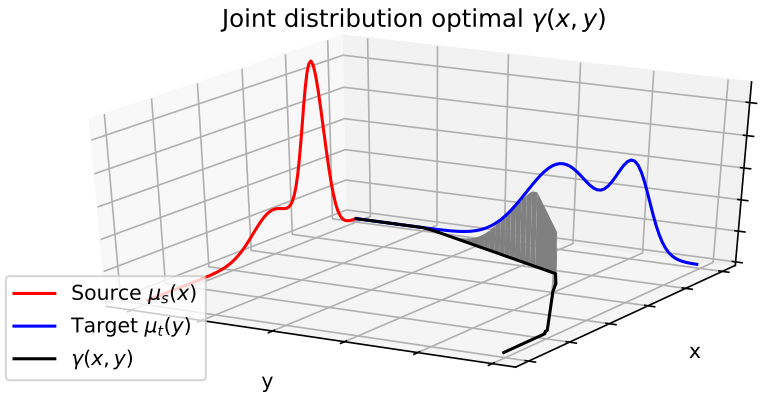
- The Kantorovich formulation [Kantorovich, 1942] seeks for a probabilistic coupling $T \in \mathcal{P}(\Omega_s \times \Omega_t)$ between Ω_s and Ω_t:

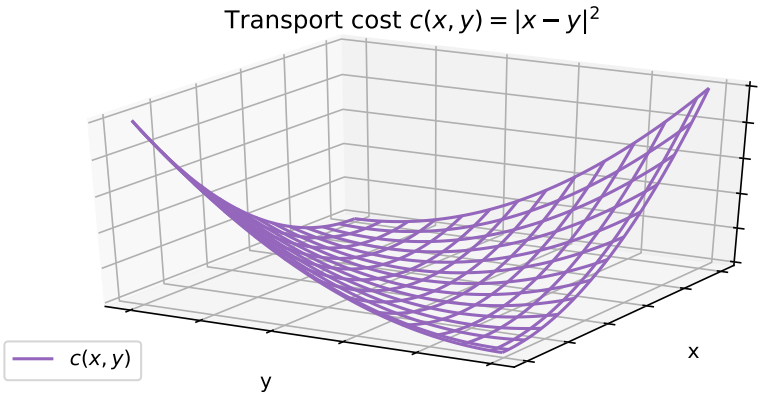
$$\boldsymbol{\gamma}_0 = \underset{T}{\operatorname{argmin}} \int_{\Omega_s \times \Omega_t} c(\mathbf{x}, \mathbf{y}) T(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}, \tag{2}$$

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

- T is a joint probability measure with marginals μ_s and μ_t.
- Linear Program that always has a solution ($\mu_s \otimes \mu_t \in \mathcal{P}$).

Optimal transport (Kantorovich dual formulation)



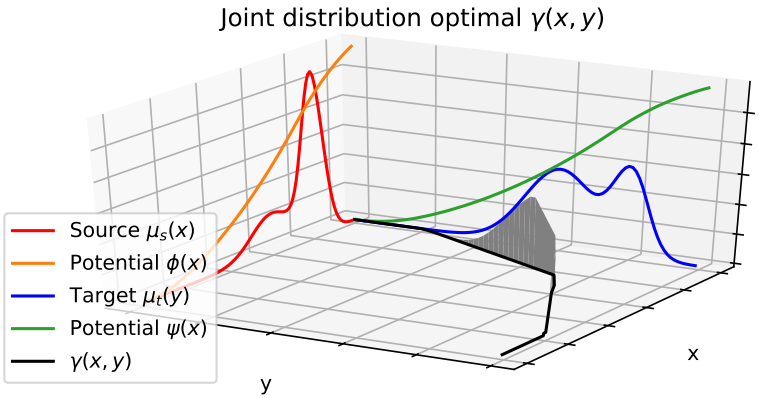


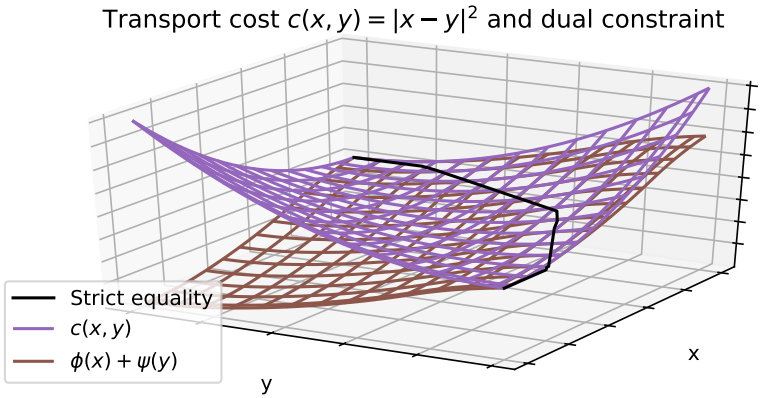
Dual formulation of the OT linear program

$$\max_{\phi,\psi} \quad \left\{ \int \phi d\mu_s + \int \psi d\mu_t \;\middle|\; \phi(\mathbf{x}) + \psi(\mathbf{y}) \leq c(\mathbf{x}, \mathbf{y}) \right\} \tag{3}$$

- ϕ and ψ are scalar function also known as Kantorovich potentials.
- Equivalent problem by the Rockafellar-Fenchel theorem.
- Objective value separable wrt μ_s and μ_t.
- Primal-dual relation : the support of T is where $\phi(\mathbf{x}) + \psi(\mathbf{y}) = c(\mathbf{x}, \mathbf{y})$

Optimal transport (Kantorovich dual formulation)





Dual formulation of the OT linear program

$$\max_{\phi,\psi} \quad \left\{ \int \phi d\mu_s + \int \psi d\mu_t \;\middle|\; \phi(\mathbf{x}) + \psi(\mathbf{y}) \leq c(\mathbf{x}, \mathbf{y}) \right\} \tag{3}$$

- ϕ and ψ are scalar function also known as Kantorovich potentials.
- Equivalent problem by the Rockafellar-Fenchel theorem.
- Objective value separable wrt μ_s and μ_t.
- Primal-dual relation : the support of T is where $\phi(\mathbf{x}) + \psi(\mathbf{y}) = c(\mathbf{x}, \mathbf{y})$

Optimal transport (Kantorovich dual formulation)

The linear dual constraint suggest that there exits an optimal ψ for a given ϕ.

c-transform (or c-conjugate)

$$\phi^c(\mathbf{y}) \stackrel{\text{def}}{=} H^c(\phi) = \inf_{\mathbf{x}} \quad c(\mathbf{x}, \mathbf{y}) - \phi(\mathbf{x}) \tag{4}$$

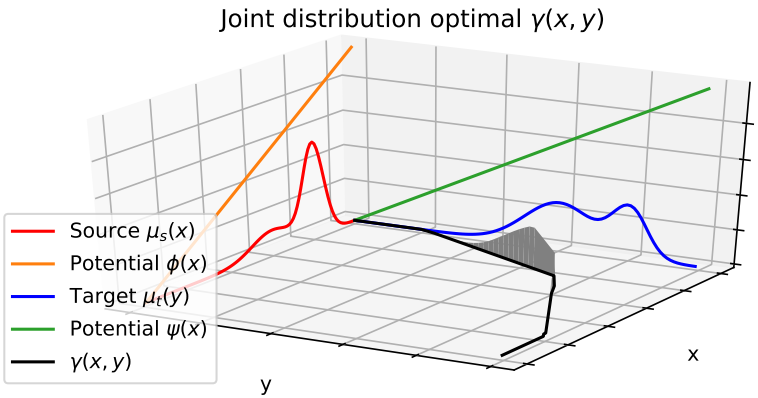
Similar a Legendre transform (equal when $c(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$).

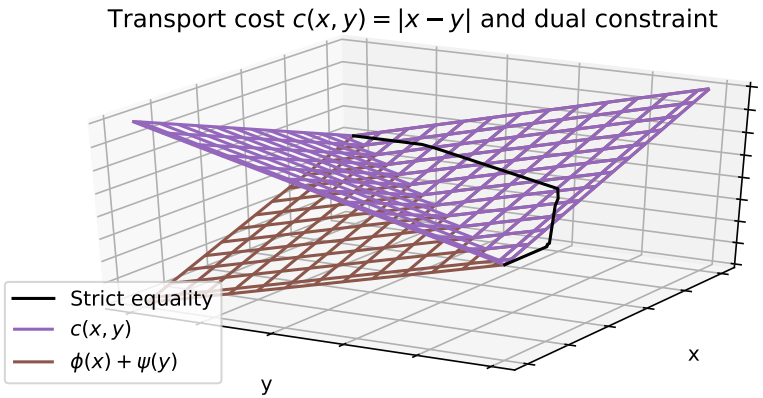
Semi-dual formulation

$$\max_{\phi} \quad \left\{ \int \phi d\mu_s + \int \phi^c d\mu_t \right\} \tag{5}$$

- Depends only on one dual potential through the c-transform.
- Nice reformulation when H^c is easy to compute of close form.
- Special case when $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$.

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





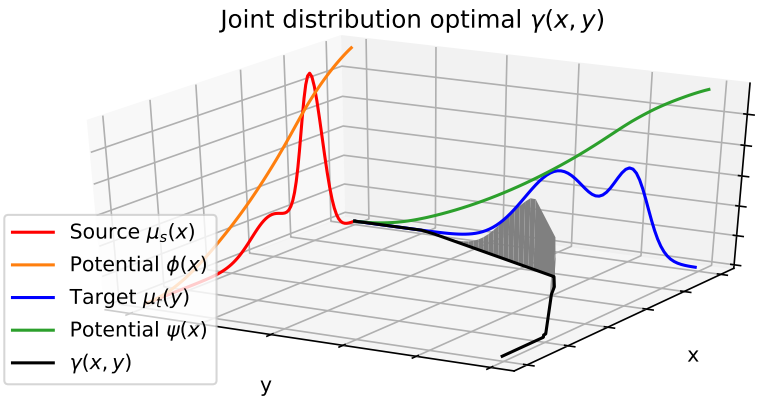
Case $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$

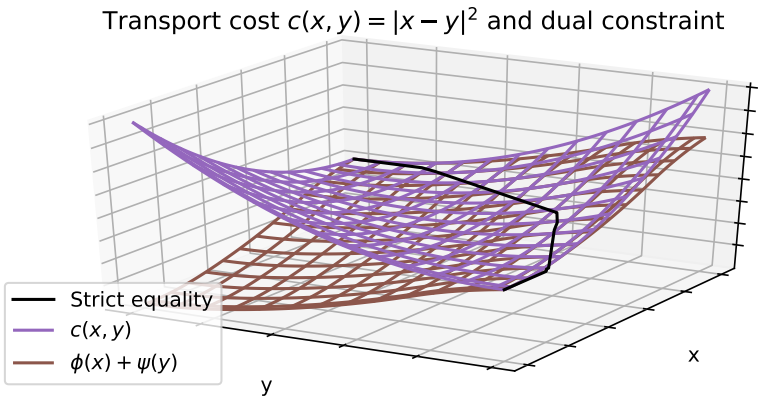
- Existence of a solution but not unique.
- For any $\phi \in \text{Lip}^1$ (set of 1-Lipschitz functions), we have $\phi^c(x) = -\phi(x)$.
- The dual OT problem can be reformulated as

$$\sup_{\phi \in \text{Lip}^1} \int \phi d(\mu_s - \mu_t) = \sup_{\phi \in \text{Lip}^1} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mu_s} [\phi(x)] - \mathop{\mathbb{E}}_{\mathbf{y} \sim \mu_t} [\phi(y)] \tag{6}$$

- Also known as **Kantorovich-Rubinstein duality**
- Formulation used for Wasserstein GAN (more details in next part).

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





Case $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2/2$

- When μ_s and μ_t are continuous, $T(x)$ the OT mapping exists and is unique.
- More remarkably, it is a gradient of a convex functions $\Phi(x)$

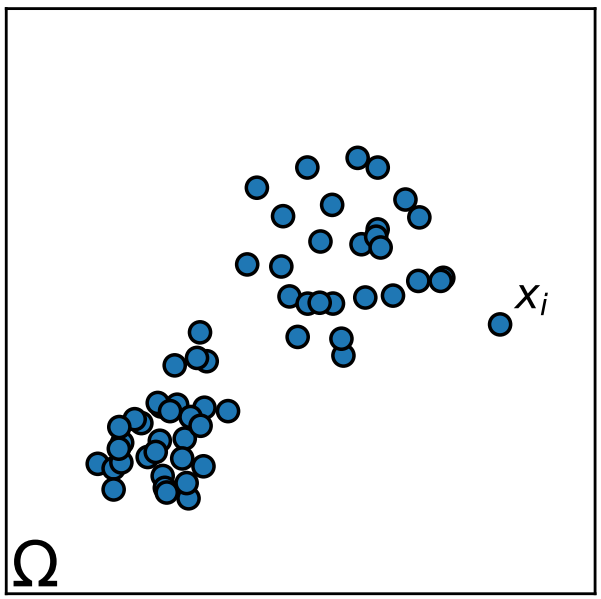
$$T(x) = x - \nabla\phi(x) = \nabla\left(\frac{\|x\|^2}{2} - \phi(x)\right) = \nabla(\Phi(x)) \tag{7}$$

- This is also known as **Brenier's Theorem** [Brenier, 1991].

Discrete distributions: Empirical vs Histogram

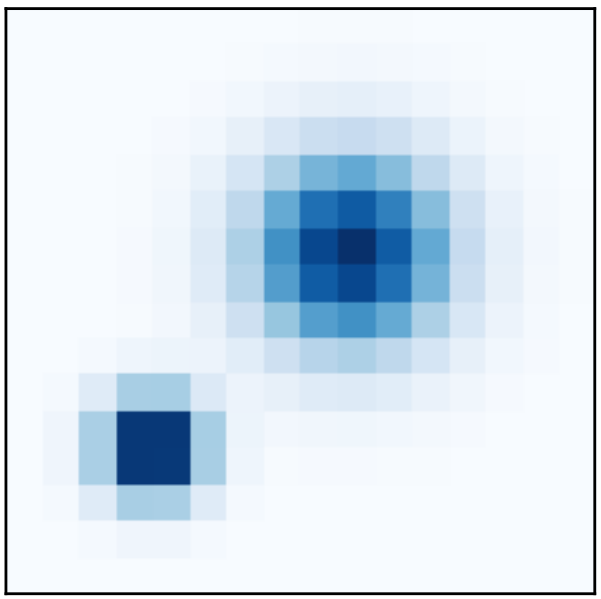
Discrete measure: $\mu = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}, \quad \mathbf{x}_i \in \Omega, \quad \sum_{i=1}^{n} a_i = 1$

Lagrangian (point clouds)



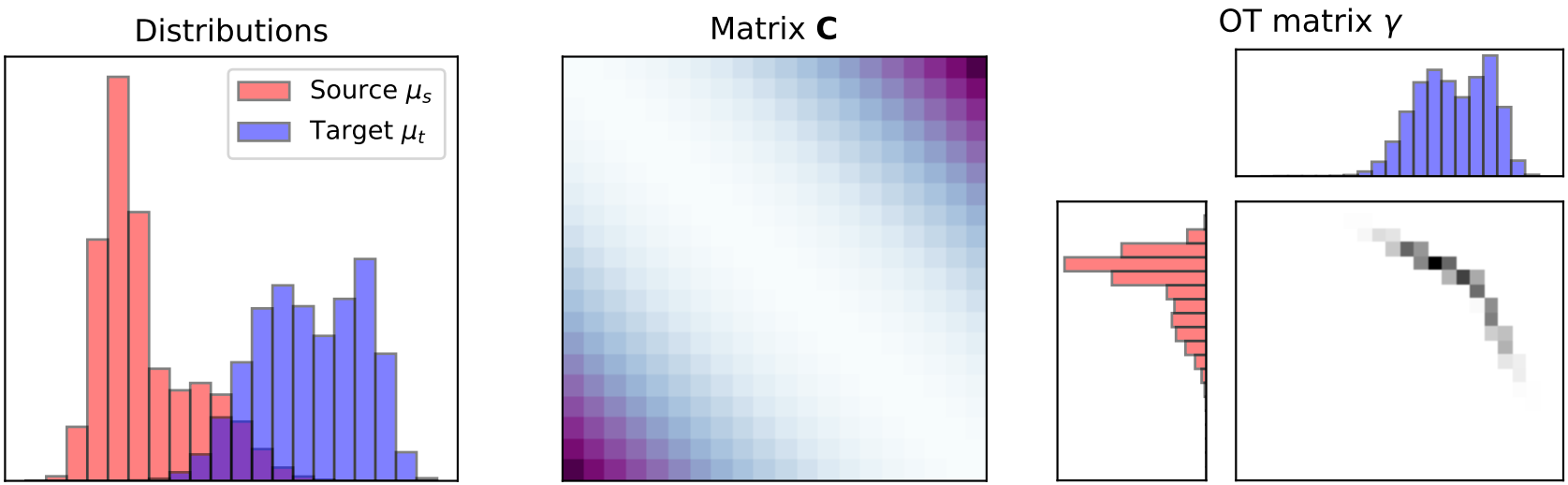
- Constant weight: $a_i = \frac{1}{n}$
- Quotient space: Ω^n, Σ_n

Eulerian (histograms)



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

Optimal transport with discrete distributions



OT Linear Program

When $\mu_s = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i^s}$ and $\mu_t = \sum_{i=1}^{n} b_i \delta_{\mathbf{x}_i^t}$

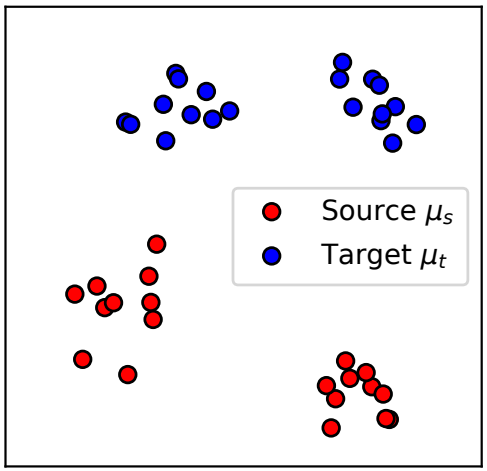
$$\mathbf{T}_0 = \underset{\mathbf{T} \in \Pi(\mu_s, \mu_t)}{\operatorname{argmin}} \left\{ \langle \mathbf{T}, \mathbf{C} \rangle_F = \sum_{i,j} T_{i,j} c_{i,j} \right\}$$

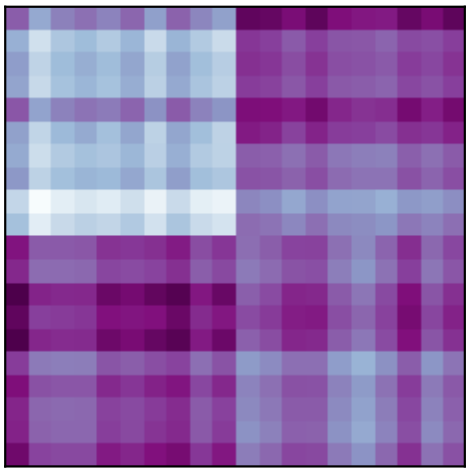
where $\mathbf{C}$ is a cost matrix with $c_{i,j} = c(\mathbf{x}_i^s, \mathbf{x}_j^t)$ and the marginals constraints are

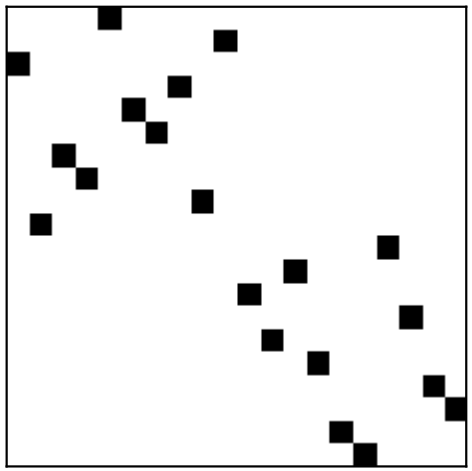
$$\Pi(\mu_s, \mu_t) = \left\{ \mathbf{T} \in (\mathbb{R}^+)^{n_s \times n_t} \,|\, \mathbf{T}\mathbf{1}_{n_t} = \mathbf{a}, \mathbf{T}^T \mathbf{1}_{n_s} = \mathbf{b} \right\}$$

Linear program with $n_s n_t$ variables and $n_s + n_t$ constraints. Demo

Optimal transport with discrete distributions







OT Linear Program

When $\mu_s = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i^s}$ and $\mu_t = \sum_{i=1}^{n} b_i \delta_{\mathbf{x}_i^t}$

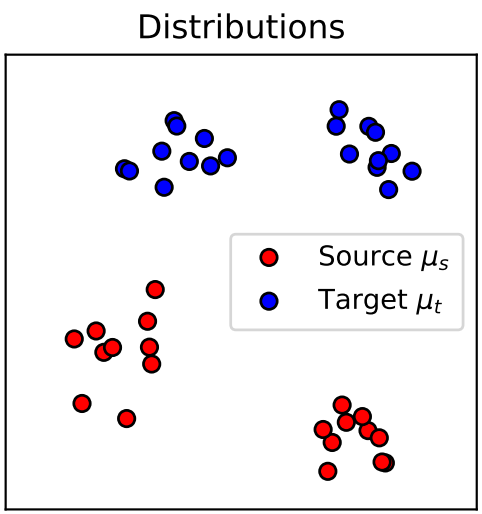
$$\mathbf{T}_0 = \underset{\mathbf{T} \in \Pi(\mu_s, \mu_t)}{\operatorname{argmin}} \left\{ \langle \mathbf{T}, \mathbf{C} \rangle_F = \sum_{i,j} T_{i,j} c_{i,j} \right\}$$

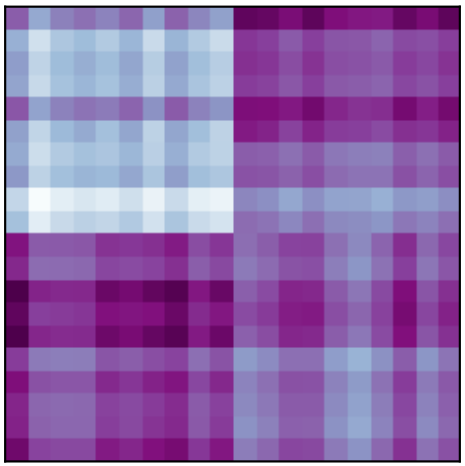
where $\mathbf{C}$ is a cost matrix with $c_{i,j} = c(\mathbf{x}_i^s, \mathbf{x}_j^t)$ and the marginals constraints are

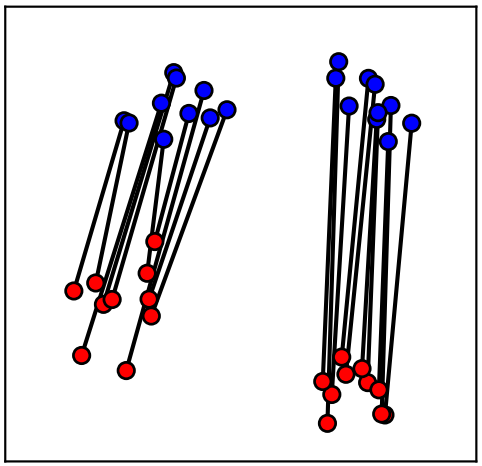
$$\Pi(\mu_s, \mu_t) = \left\{ \mathbf{T} \in (\mathbb{R}^+)^{n_s \times n_t} \,|\, \mathbf{T}\mathbf{1}_{n_t} = \mathbf{a}, \mathbf{T}^T \mathbf{1}_{n_s} = \mathbf{b} \right\}$$

Linear program with $n_s n_t$ variables and $n_s + n_t$ constraints. Demo

Optimal transport with discrete distributions







OT Linear Program

When $\mu_s = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i^s}$ and $\mu_t = \sum_{i=1}^{n} b_i \delta_{\mathbf{x}_i^t}$

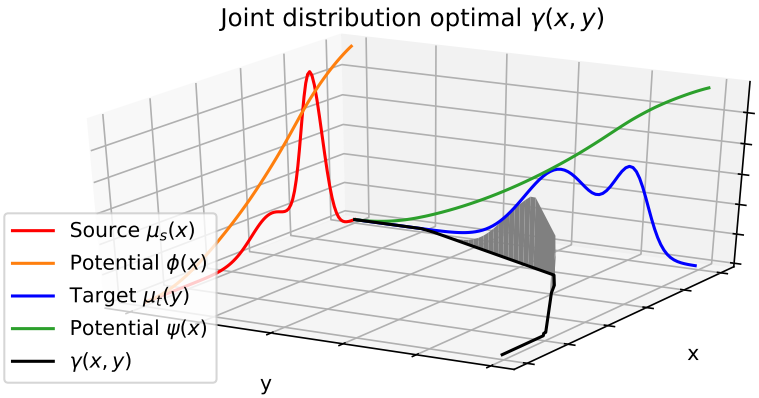
$$\mathbf{T}_0 = \underset{\mathbf{T} \in \Pi(\mu_s, \mu_t)}{\operatorname{argmin}} \left\{ \langle \mathbf{T}, \mathbf{C} \rangle_F = \sum_{i,j} T_{i,j} c_{i,j} \right\}$$

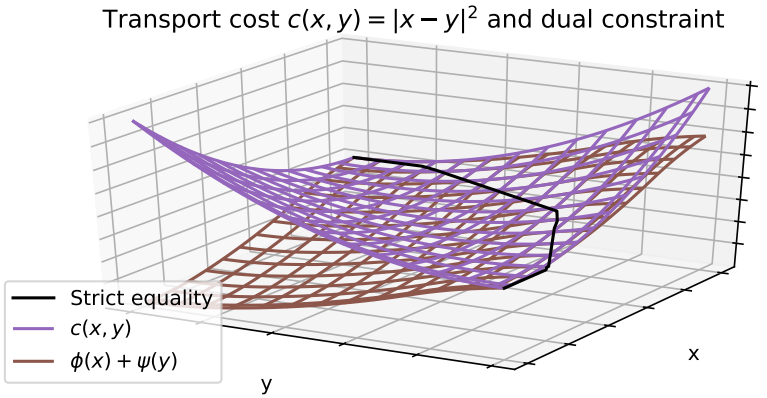
where $\mathbf{C}$ is a cost matrix with $c_{i,j} = c(\mathbf{x}_i^s, \mathbf{x}_j^t)$ and the marginals constraints are

$$\Pi(\mu_s, \mu_t) = \left\{ \mathbf{T} \in (\mathbb{R}^+)^{n_s \times n_t} \,|\, \mathbf{T}\mathbf{1}_{n_t} = \mathbf{a}, \mathbf{T}^T \mathbf{1}_{n_s} = \mathbf{b} \right\}$$

Linear program with $n_s n_t$ variables and $n_s + n_t$ constraints. Demo

OT Dual for discrete distributions





Discrete OT dual formulation

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^{n^s}, \boldsymbol{\beta} \in \mathbb{R}^{n^t}} \quad \boldsymbol{\alpha}^T \mathbf{a} + \boldsymbol{\beta}^T \mathbf{b} \tag{8}$$

$$\text{s.t.} \quad \alpha_i + \beta_j \leq c_{i,j} \quad \forall i, j \tag{9}$$

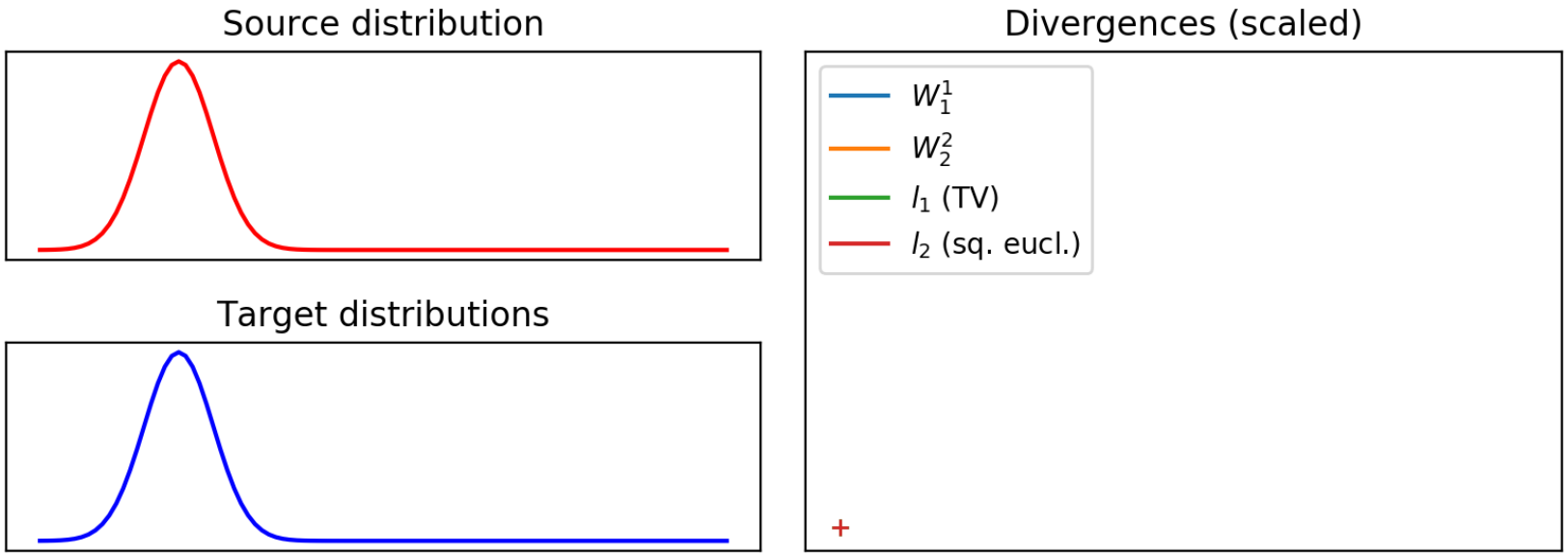
- With $\mu_s = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i^s}$ and $\mu_t = \sum_{i=1}^{n} b_i \delta_{\mathbf{x}_i^t}$
- Linear program with $n_s + n_t$ variables and $n_s n_t$ constraints.
- Solved with Network Flow solver of complexity $O(n^3 \log(n))$ with $n = \max(n_s, n_t)$.

Matching words embedding

Word mover's distance [Kusner et al., 2015]

- Words 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.
- Small value of the objective means similar documents.
- OT matrix provide interpretability (word correspondance).

Wasserstein distance



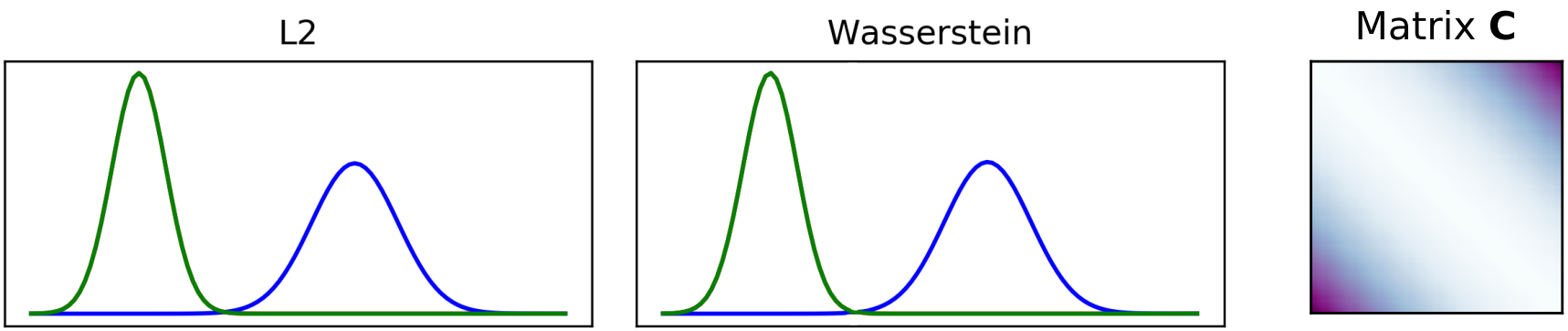
Wasserstein distance

$$W_p^p(\mu_s, \mu_t) = \min_{T \in \mathcal{P}} \int_{\Omega_s \times \Omega_t} \|\mathbf{x} - \mathbf{y}\|^p T(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y}) \sim T} [\|\mathbf{x} - \mathbf{y}\|^p] \qquad (10)$$

In this case we have $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^p$

- 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).

Wasserstein barycenter

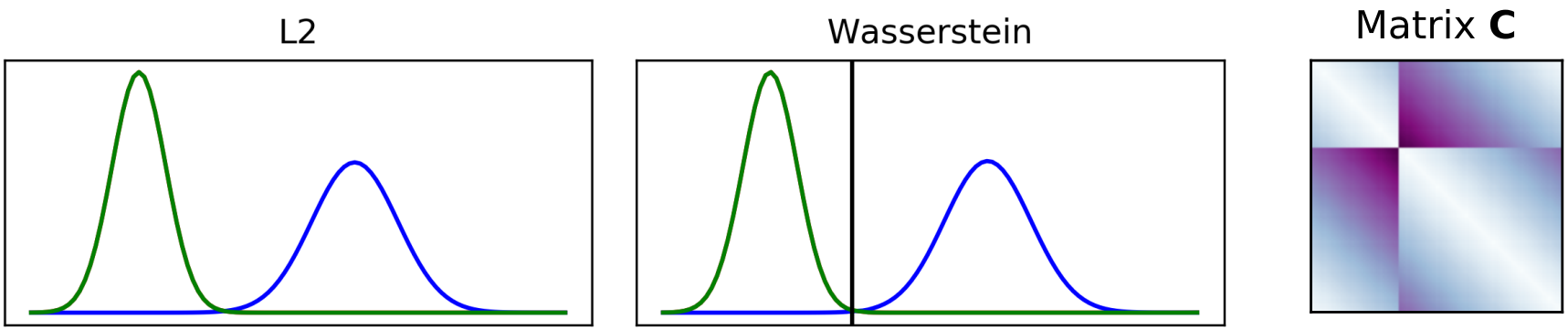


Barycenters [Agueh and Carlier, 2011]

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

- $\lambda_i > 0$ 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

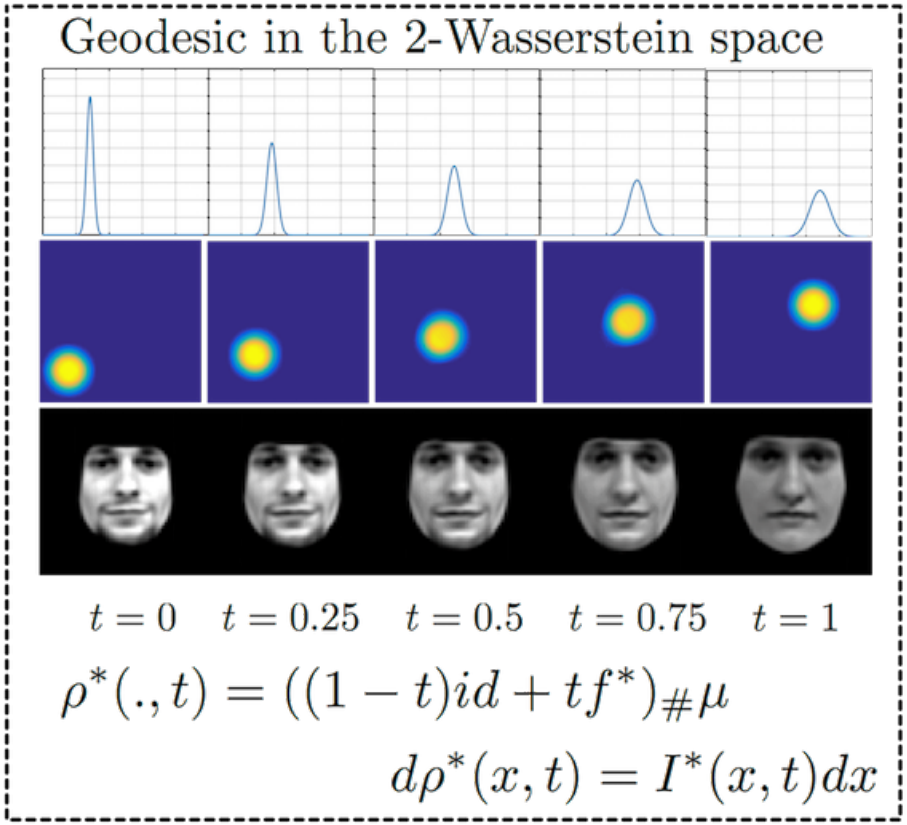


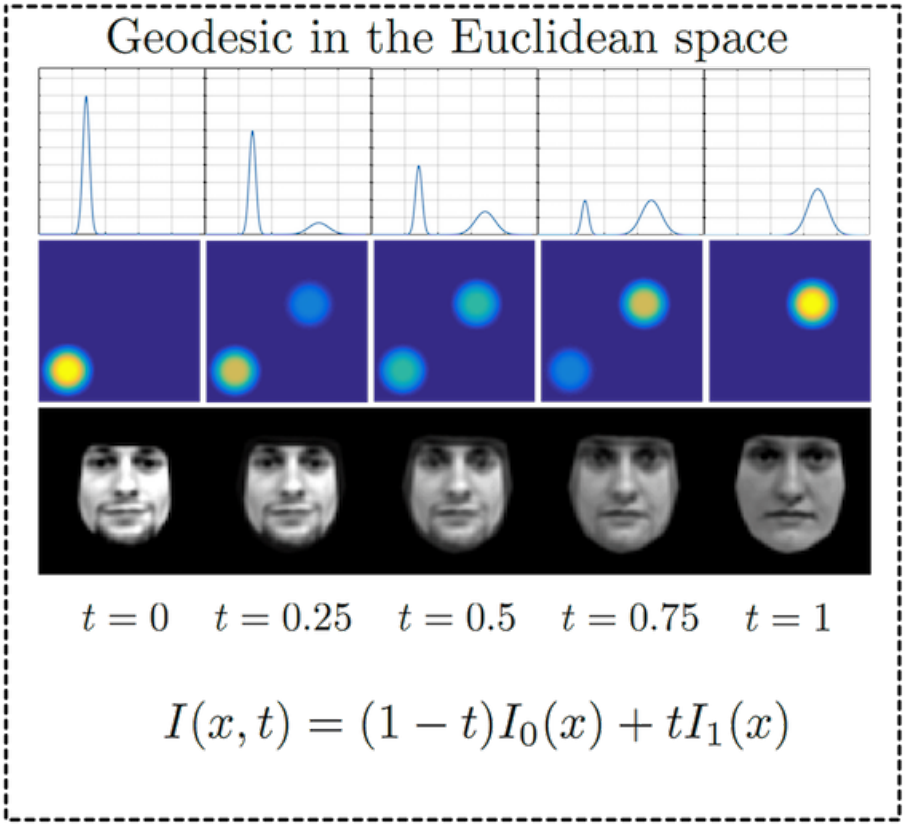
Barycenters [Agueh and Carlier, 2011]

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

- $\lambda_i > 0$ 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.

Wasserstein space

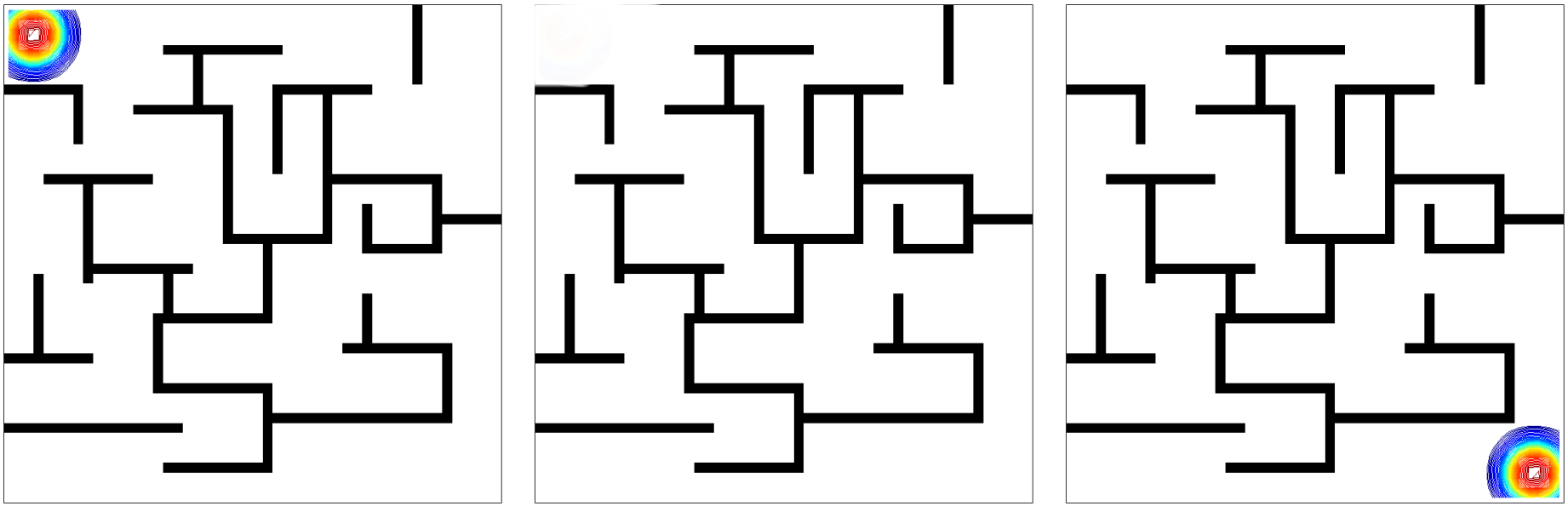




- The space of probability distribution equipped with the Wasserstein metric $(\mathcal{P}_p(X), W_2^2(X))$ defines a geodesic space with a Riemannian structure [Santambrogio, 2014].
- Geodesics are shortest curves on $\mathcal{P}_p(X)$ that link two distributions

Illustration from [Kolouri et al., 2017] and maze example from [Papadakis et al., 2014]

Wasserstein space

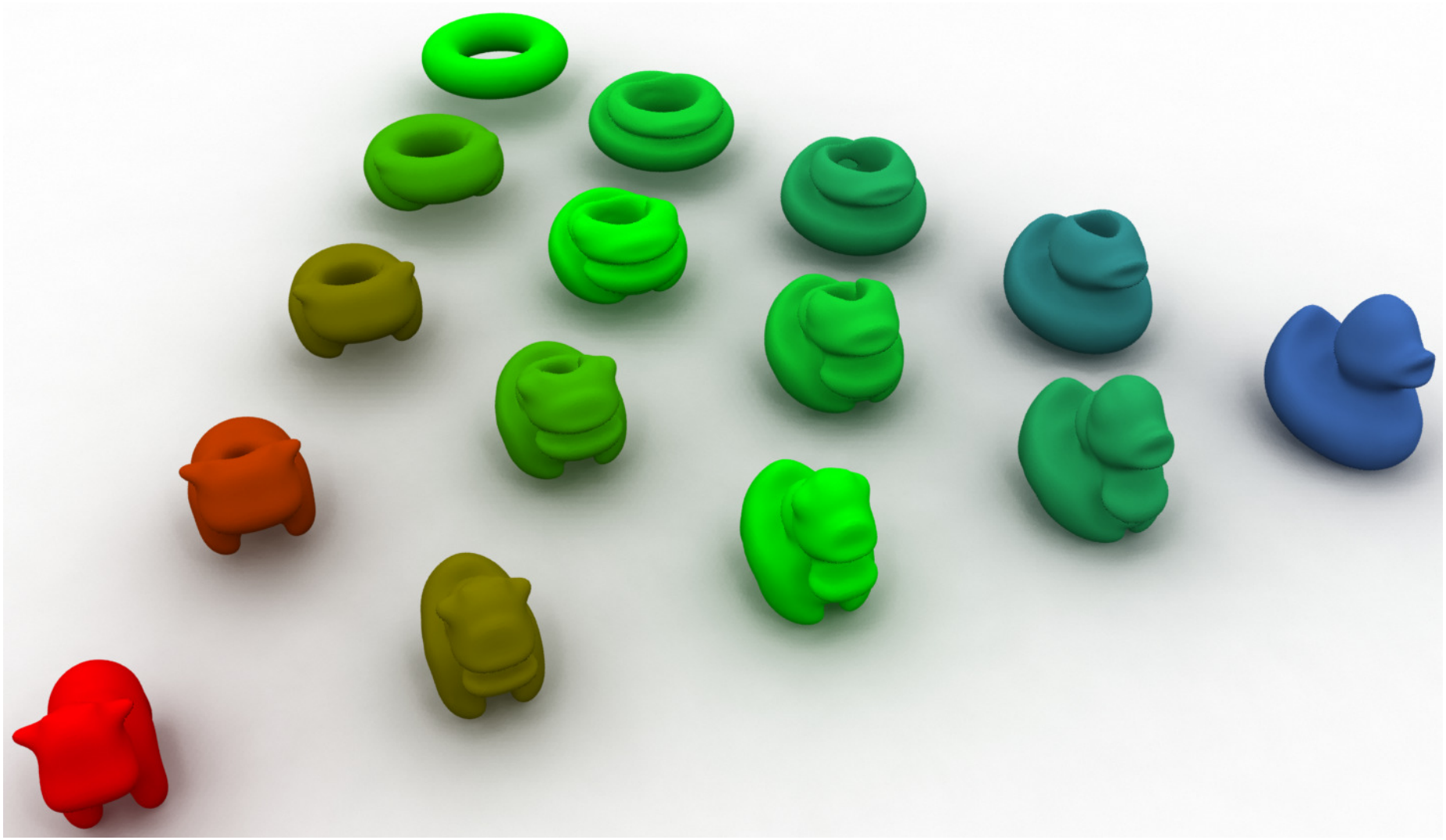


- The space of probability distribution equipped with the Wasserstein metric $(\mathcal{P}_p(X), W_2^2(X))$ defines a geodesic space with a Riemannian structure [Santambrogio, 2014].
- Geodesics are shortest curves on $\mathcal{P}_p(X)$ that link two distributions
- Cost between two pixels is the shortest path in the maze (Riemannian metric).

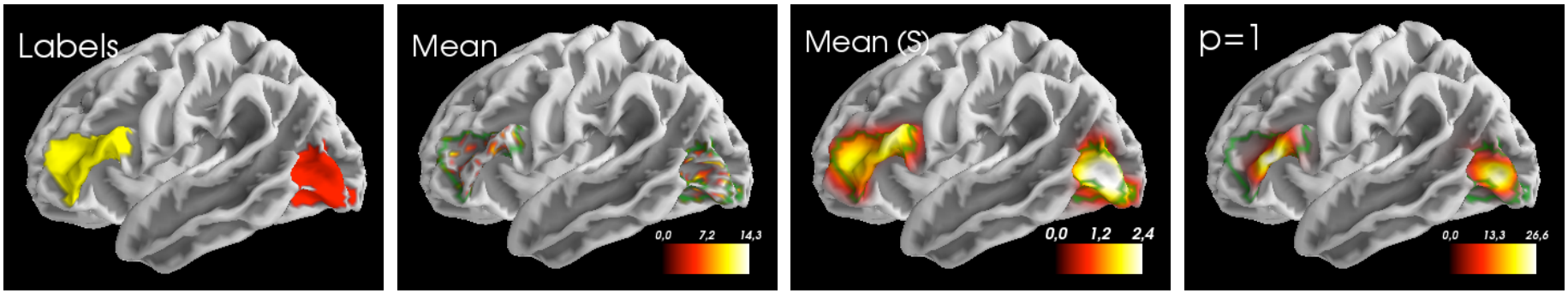
Illustration from [Kolouri et al., 2017] and maze example from [Papadakis et al., 2014]

3D Wasserstein barycenter

Shape interpolation [Solomon et al., 2015]



Wasserstein averaging of fMRI

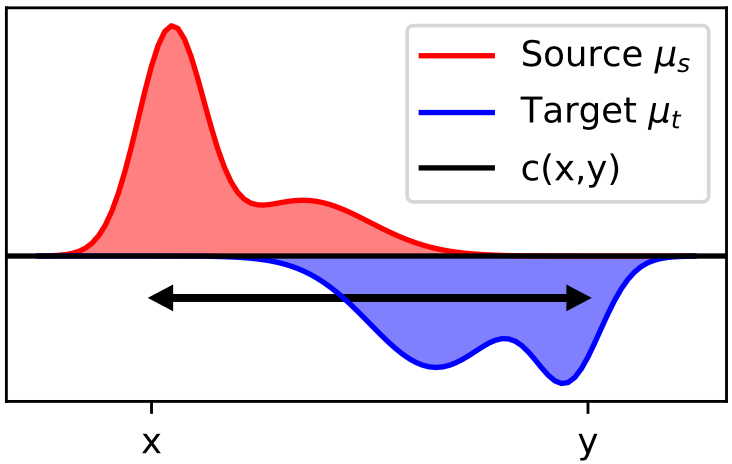


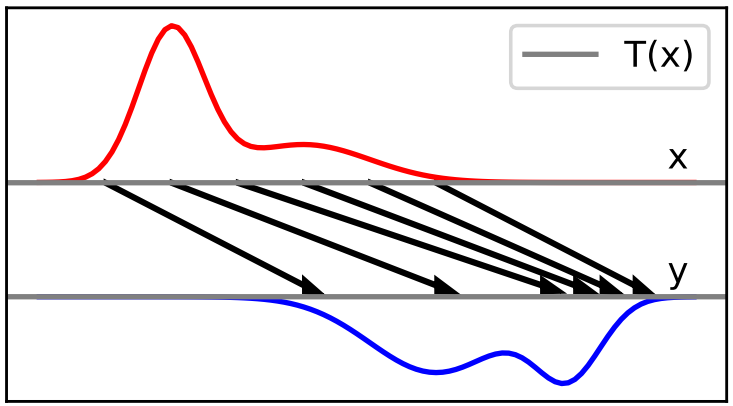
OT averaging of neurological data [Gramfort et al., 2015]

- Average fMRI activation maps on voxels or cortical surface (natural metric).
- Classical average across subjects and gaussian blur loose information.
- OT averaging recover central activation areas with better precision.
- Can encode both geometrical (3D position) or anatomical connectivity information.
- Extension using OT-Lp seems more robust to noise [Wang et al., 2018].

Outline

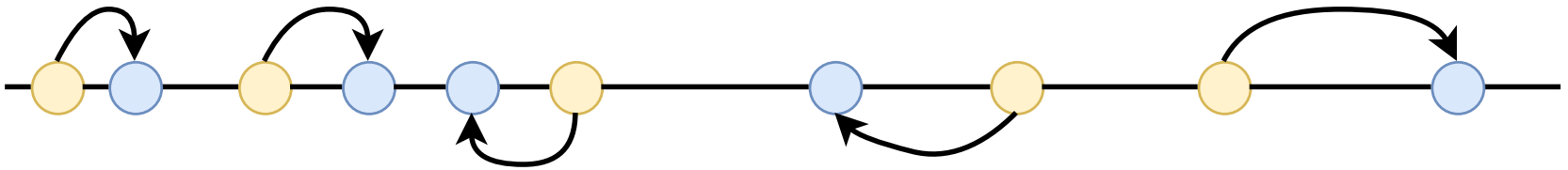
Special case: OT in 1D





- When $c(x, y)$ is a strictly convex and increasing function of $|x - y|$.
- If $x_1 < x_2$ and $y_1 < y_2$, we have $c(x_1, y_1) + c(x_2, y_2) < c(x_1, y_2) + c(x_2, y_1)$
- The OT plan respects the ordering of the elements.
- Solution is given by the monotone rearrangement of μ_1 onto μ_2.
- Simple algorithm for discrete distribution by sorting $O(N \log N)$.

Special case: OT in 1D



- When $c(x, y)$ is a strictly convex and increasing function of $|x - y|$.
- If $x_1 < x_2$ and $y_1 < y_2$, we have $c(x_1, y_1) + c(x_2, y_2) < c(x_1, y_2) + c(x_2, y_1)$
- The OT plan respects the ordering of the elements.
- Solution is given by the monotone rearrangement of μ_1 onto μ_2.
- Simple algorithm for discrete distribution by sorting $O(N \log N)$.

Special case: OT in 1D

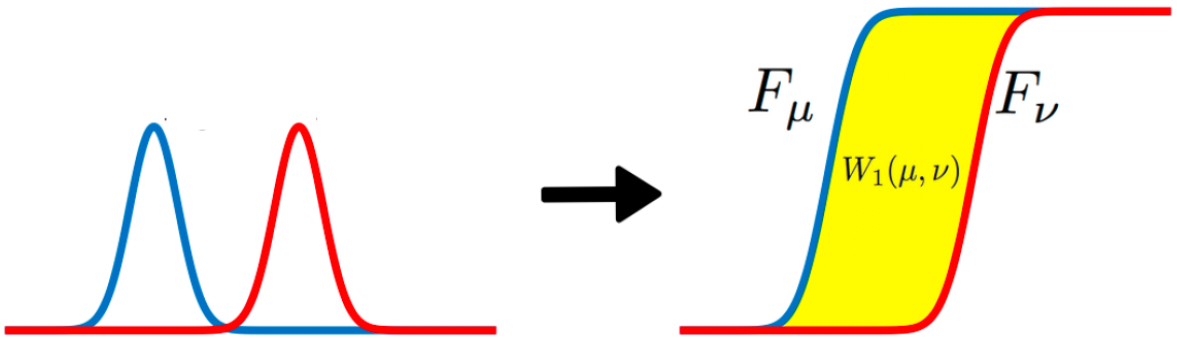


Illustration with cumulative distributions

- F_μ cumulative distribution function of μ : $F_\mu(t) = \mu(-\infty, t]$.
- $F_\mu^{-1}(q),\ q \in [0,1]$ is the quantile function: $F_\mu^{-1}(q) = \inf\{x \in \mathbb{R} : F_\mu(x) \geq q\}$.
- The value of the W_1 Wasserstein distance

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

- Very fast $O(n\log(n))$ computation on discrete distributions.

Special case: OT in 1D

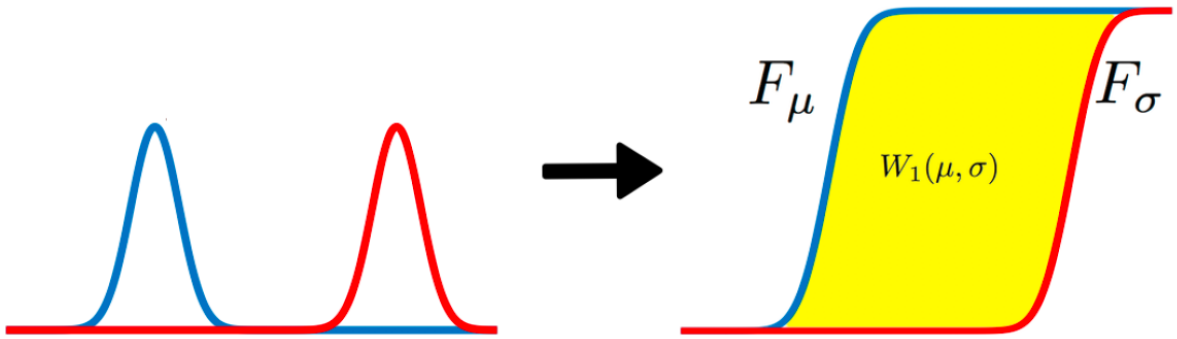


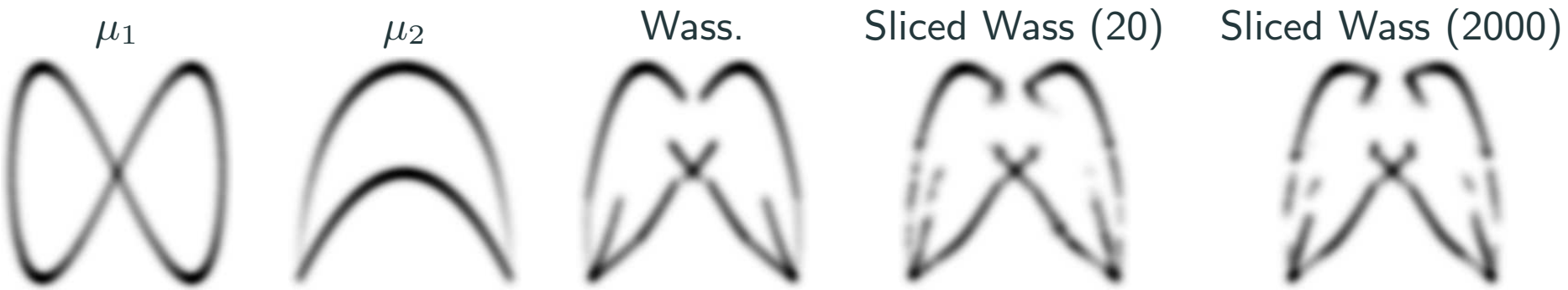
Illustration with cumulative distributions

- F_μ cumulative distribution function of μ : $F_\mu(t) = \mu(-\infty, t]$.
- $F_\mu^{-1}(q),\ q \in [0,1]$ is the quantile function: $F_\mu^{-1}(q) = \inf\{x \in \mathbb{R} : F_\mu(x) \geq q\}$.
- The value of the W_1 Wasserstein distance

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

- Very fast $O(n \log(n))$ computation on discrete distributions.

Sliced Radon Wasserstein



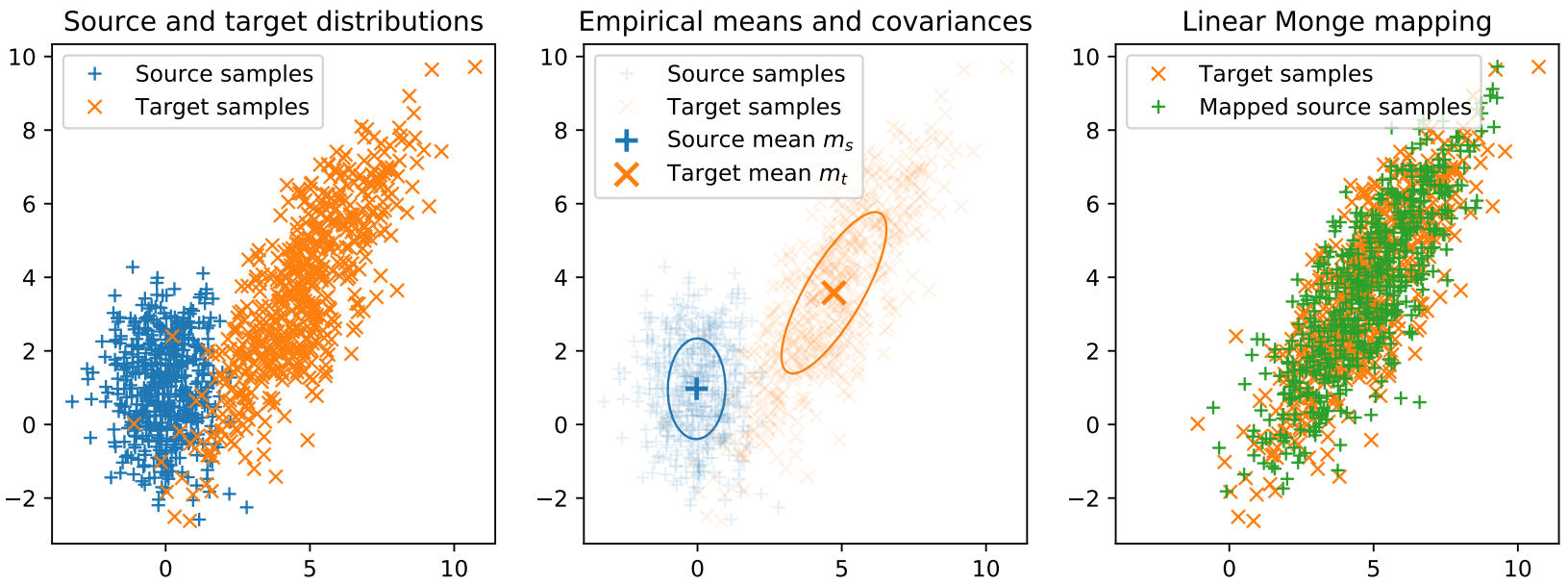
p-sliced Wasserstein distance (pSW) [Bonneel et al., 2015]

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

where $\mathcal{R}$ is the Radon transform $\mathcal{R}(\mu, \theta) = \int_{\mathbb{S}^{d-1}} \mu(\mathbf{x}) \delta(t - \theta^\top \mathbf{x}) d\mathbf{x} \; \forall \theta \in \mathbb{S}^{d-1}$

- Can be approximated by discrete sampling of the directions θ.
- Fast 1D wasserstein on 1D projections when $d > 1$, fast distance estimation and barycenter computation.
- p-sliced Wasserstein distance used for kernel learning between distributions [Kolouri et al., 2016].

Special case: OT between Gaussians (1)



Wasserstein between Gaussian distributions

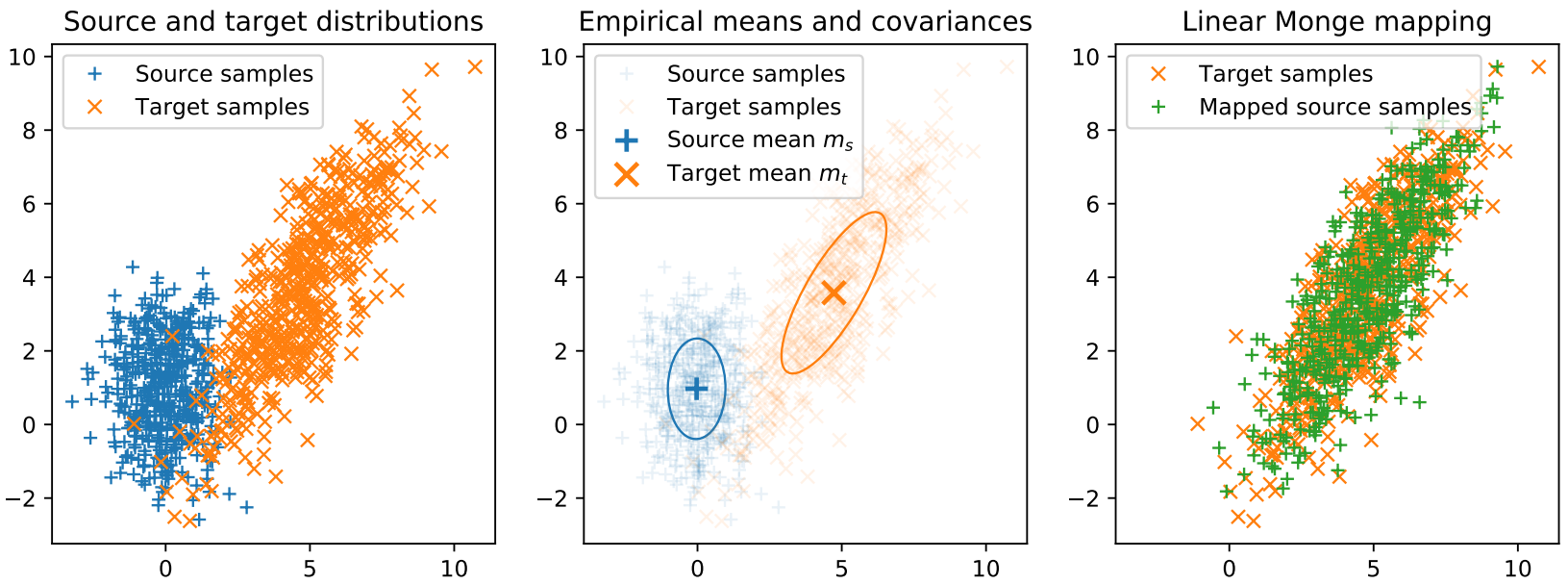
- $\mu_s \sim \mathcal{N}(\mathbf{m}_1, \Sigma_1)$ and $\mu_t \sim \mathcal{N}(\mathbf{m}_2, \Sigma_2)$
- Wasserstein distance with $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$ reduces to:

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

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

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

Special case: OT between Gaussians (2)



OT mapping between Gaussian distributions

- $\mu_s \sim \mathcal{N}(\mathbf{m}_1, \Sigma_1)$ and $\mu_t \sim \mathcal{N}(\mathbf{m}_2, \Sigma_2)$
- The optimal map T for $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$ is given by

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

with

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

Regularized optimal transport

$$\mathbf{T}_0^\lambda = \underset{\mathbf{T}\in\Pi(\mu_s,\mu_t)}{\operatorname{argmin}} \quad \langle \mathbf{T}, \mathbf{C}\rangle_F + \lambda\Omega(\mathbf{T}), \qquad (11)$$

Regularization term $\Omega(\mathbf{T})$

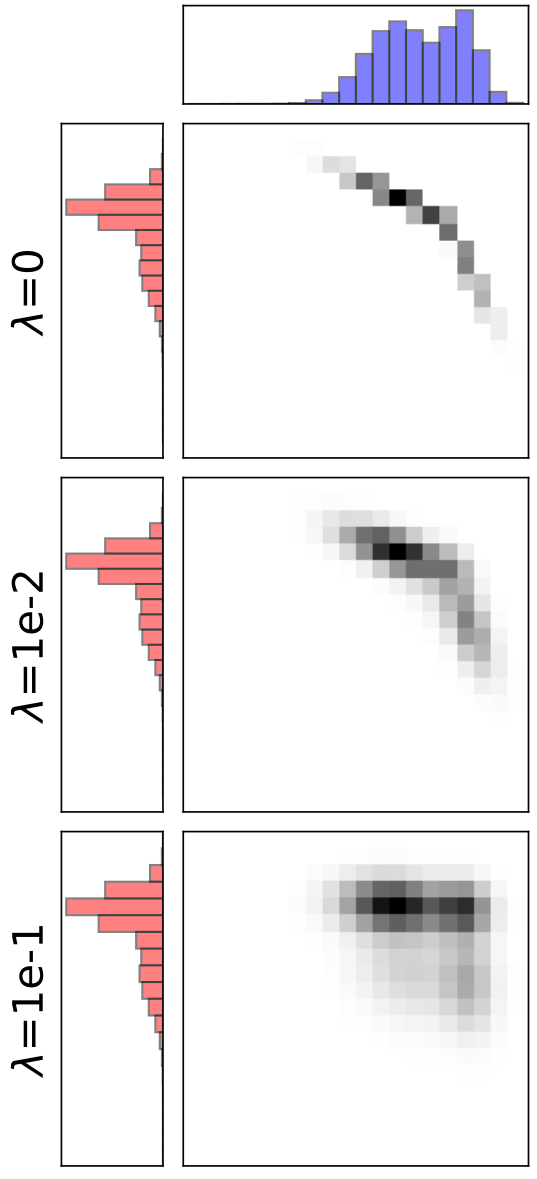
- Entropic regularization [Cuturi, 2013].
- Group Lasso [Courty et al., 2016a].
- KL, Itakura Saito, β-divergences, [Dessein et al., 2016].

Why regularize?

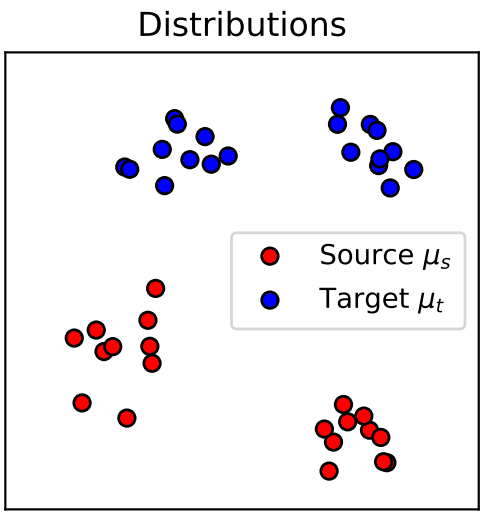
- Smooth the "distance" estimation:

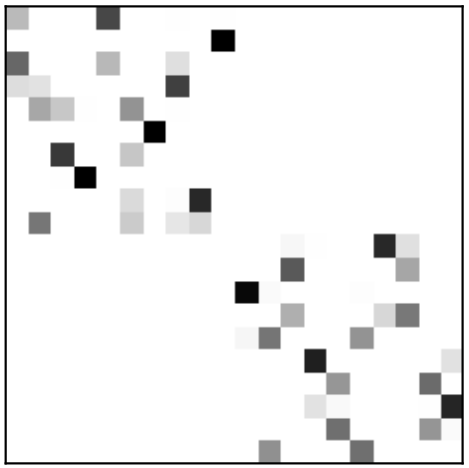
$$W_\lambda(\mu_s, \mu_t) = \left\langle \mathbf{T}_0^\lambda, \mathbf{C}\right\rangle_F$$

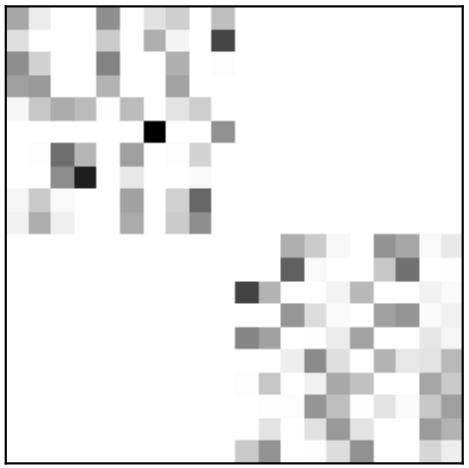
- Encode prior knowledge on the data.
- Better posed problem (convex, stability).
- Fast algorithms to solve the OT problem.



Entropic regularized optimal transport





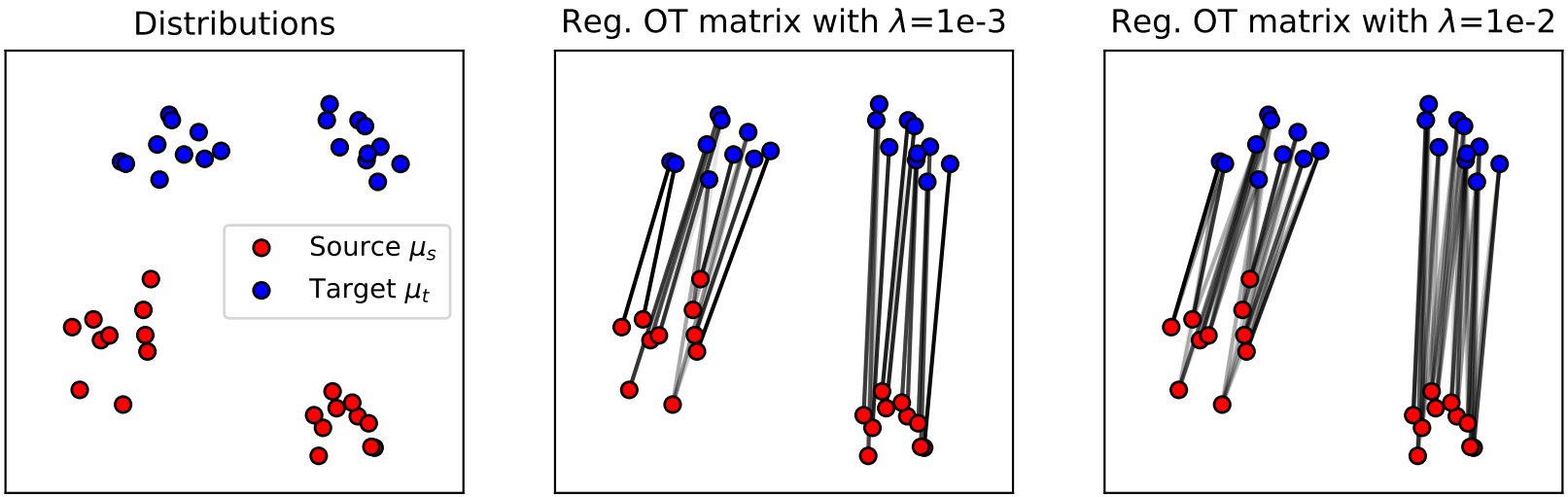


Entropic regularization [Cuturi, 2013]

$$\mathbf{T}_0^\lambda = \underset{\mathbf{T}\in\Pi(\mu_s,\mu_t)}{\operatorname{argmin}} \quad \langle \mathbf{T}, \mathbf{C} \rangle_F + \lambda \sum_{i,j} T_{i,j}(\log T_{i,j} - 1)$$

- Regularization with the negative entropy of T.
- Looses sparsity, gains stability.
- Strictly convex optimization problem.
- Loss and OT matrix are differentiable.

Entropic regularized optimal transport



Entropic regularization [Cuturi, 2013]

$$\mathbf{T}_0^\lambda = \underset{\mathbf{T}\in\Pi(\mu_s,\mu_t)}{\operatorname{argmin}} \quad \langle \mathbf{T}, \mathbf{C} \rangle_F + \lambda \sum_{i,j} T_{i,j}(\log T_{i,j} - 1)$$

- Regularization with the negative entropy of T.
- Looses sparsity, gains stability.
- Strictly convex optimization problem.
- Loss and OT matrix are differentiable.

Solving the entropy regularized problem

Lagrangian of the optimization problem

$$\mathcal{L}(\mathbf{T}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{ij} T_{ij} C_{ij} + \lambda T_{ij} (\log T_{ij} - 1) + \boldsymbol{\alpha}^{\mathbf{T}} (\mathbf{T}\mathbf{1}_{n_t} - \mathbf{a}) + \boldsymbol{\beta}^{\mathbf{T}} (\mathbf{T}^T \mathbf{1}_{n_s} - \mathbf{b})$$

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

Entropy-regularized transport

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

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

- Through the **Sinkhorn theorem** $\text{diag}(\mathbf{u})$ and $\text{diag}(\mathbf{v})$ exist and are unique.
- Relation with dual variables: $u_i = \exp(\alpha_i/\lambda), \quad v_j = \exp(\beta_j/\lambda)$.
- Can be solved by the **Sinkhorn-Knopp** algorithm.

Sinkhorn-Knopp algorithm

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

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

Dual formulation of entropic OT

Primal formulation of entropic OT

$$\min_{\mathbf{T}\in\Pi(\mu_s,\mu_t)} \quad \langle \mathbf{T}, \mathbf{C} \rangle_F + \lambda \sum_{i,j} T_{i,j}(\log T_{i,j} - 1)$$

Dual formulation of entropic OT

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad \boldsymbol{\alpha}^T \mathbf{a} + \boldsymbol{\beta}^T \mathbf{b} - \frac{1}{\lambda} \exp\left(\frac{\boldsymbol{\alpha}}{\lambda}\right)^T \mathbf{K} \exp\left(\frac{\boldsymbol{\beta}}{\lambda}\right) \qquad \text{with } \mathbf{K} = \exp\left(-\frac{\mathbf{C}}{\lambda}\right) \quad (12)$$

- Sinkhorn algorithm is a gradient ascent on the dual variables.
- Dual problem is unconstrained: stochastic gradient descent (SGD) [Genevay et al., 2016, Seguy et al., 2017] or L-BFGS [Blondel et al., 2017].
- Semi-dual : closed form for $\boldsymbol{\beta}$ for a fixed $\boldsymbol{\alpha}$ (`sumlogexp`) leads to fast SAG algorithm [Genevay et al., 2016].

Solving entropic OT with Bregman Projections

Kullback Leibler (KL) divergence

$$\mathrm{KL}(\mathbf{T}, \rho) = \sum_{ij} T_{ij} \log \frac{T_{ij}}{\rho_{ij}} =< \mathbf{T}, \log \frac{\mathbf{T}}{\rho} >_F,$$

where $\mathbf{T}$ anf ρ are discrete distributions with the same support.

OT as a Bregman projection [Benamou et al., 2015]

$\mathbf{T}^\star$ is the solution of the following Bregman projection

$$\mathbf{T}^\star = \underset{\mathbf{T} \in \Pi(\mu_s, \mu_t)}{\operatorname{argmin}} \ \mathrm{KL}(\mathbf{T}, \mathbf{K}), \quad \text{where } \mathbf{K} = \exp\left(-\frac{C}{\lambda}\right) \tag{13}$$

- Sinkhorn is an iterative projection scheme, with alternative projections on marginal constraints.
- Generalizes to Barycenter computation [Benamou et al., 2015].
- Also generalizes to other regularization but less efficient (Dykstra's Projection algorithm [Dessein et al., 2016]).

Sinkhorn divergence

Sinkhorn loss

$$W_\lambda(\mu_s, \mu_t) = \min_{\mathbf{T} \in \Pi(\mu_s, \mu_t)} \quad \langle \mathbf{T}, \mathbf{C} \rangle_F + \lambda \sum_{i,j} T_{i,j} \log T_{i,j}$$

- Entropic term has smoothing effect.
- Not a divergence ($W_\lambda(\mu, \mu) > 0$ for $\lambda > 0$).

OT loss (aka Sharp Sinkhorn [Luise et al., 2018])

$$OT_\lambda(\mu_s, \mu_t) = \quad \left\langle \mathbf{T}_0^\lambda, \mathbf{C} \right\rangle_F$$

- $\mathbf{T}_0^\lambda$ is the solution of entropic OT above.
- Not a divergence ($OT_\lambda(\mu, \mu) > 0$ for $\lambda > 0$).

Sinkhorn divergence [Genevay et al., 2017]

$$SD_\lambda(\mu_s, \mu_t) = W_\lambda(\mu_s, \mu_t) - \frac{1}{2} W_\lambda(\mu_s, \mu_s) - \frac{1}{2} W_\lambda(\mu_t, \mu_t)$$

- True divergence ($SD_\lambda(\mu, \mu) = 0$).
- Better statistical properties as Wasserstein distance [Genevay et al., 2018].

Regularized OT (general case)

$$\boldsymbol{\gamma}_0^{\lambda} = \underset{T \in \Pi(\mu_s, \mu_t)}{\operatorname{argmin}} \quad \langle T, \mathbf{C} \rangle_F + \lambda \Omega(T),$$

- **Group lasso [Courty et al., 2016b]**

$$\Omega(\mathbf{T}) = \sum_g \sqrt{\sum_{i,j \in \mathcal{G}_g} T_{i,j}^2}$$

  Promotes group sparsity (also submodular reg. [Alvarez-Melis et al., 2017])

- **Frobenius norm [Blondel et al., 2017]**

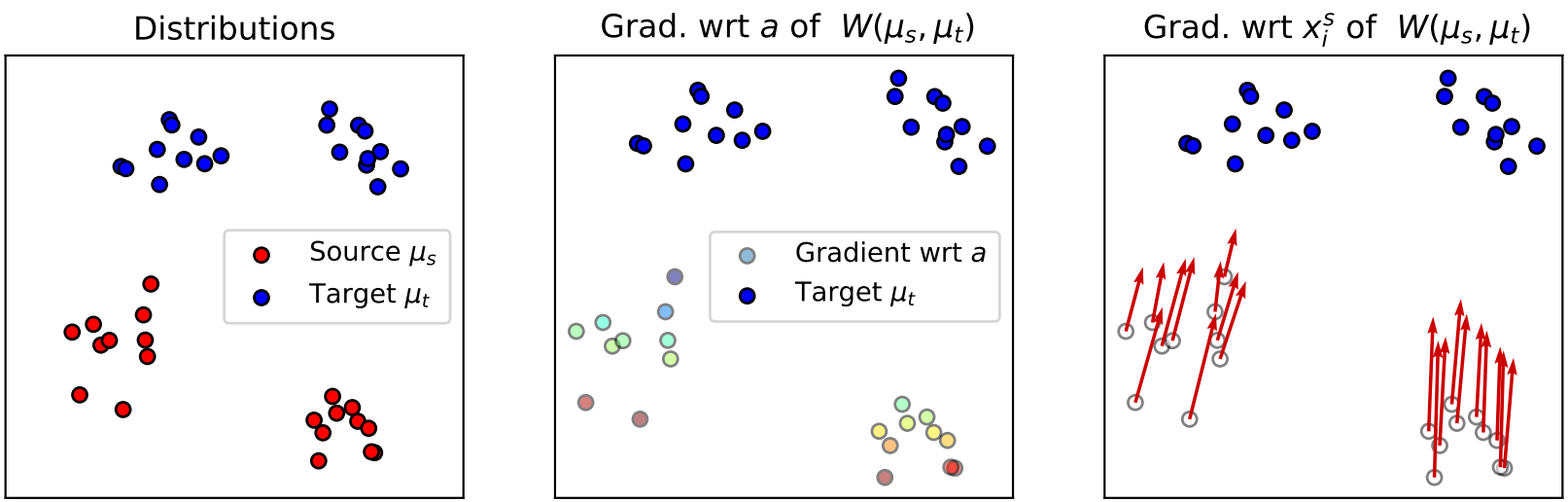
$$\Omega(T) = \sum_{i,j} T_{i,j}^2$$

  Strongly convex regularization that keeps some sparsity in the solution.

- **[Dessein et al., 2016]: KL, Itakura Saito, β-divergences.**

  Solved with Alternative optimization techniques when projection is efficient.

Minimizing the Wasserstein distance



Minimizing the Wasserstein distance

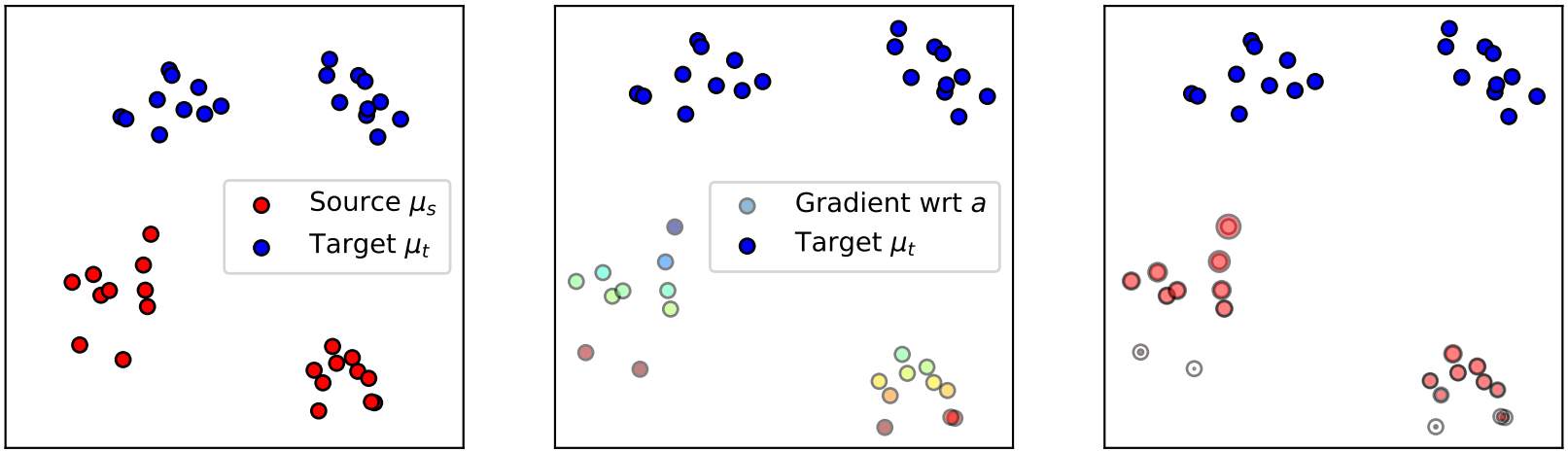
Let $\mu_s = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i^s}$. We seek the minimal Wasserstein estimator:

$$\min_{\mu_s} \quad W(\mu_s, \mu_t)$$

In practice for a discrete distribution μ_s there are two ways of doing this:

- **Case 1:** For a fixed support $\mathbf{X}_s = \{\mathbf{x}_i^s\}$ find the optimal weights $\mathbf{a}$ (Eulerian).
- **Case 2:** For fixed weights $\mathbf{a}$ find the optimal support $\mathbf{X}_s = \{\mathbf{x}_i^s\}$ (Lagrangian).

Case 1: fixed support

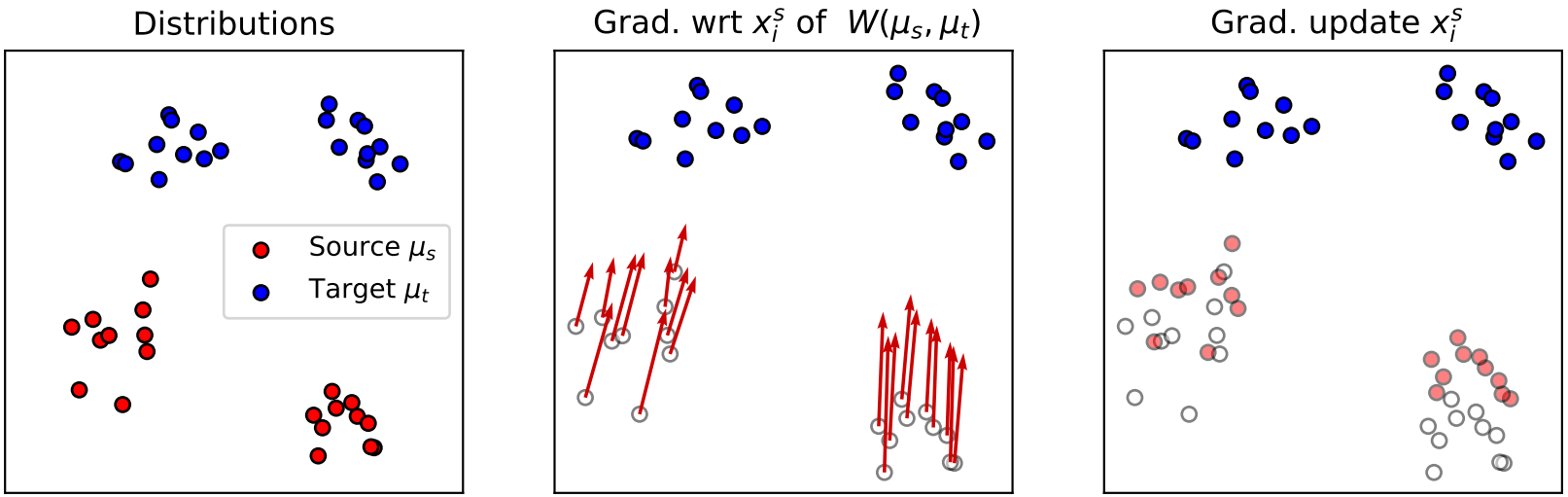


Gradient with respect to weigths $\mathbf{a}$

$$W(\mu_s, \mu_t) = \max_{\boldsymbol{\alpha}\in\mathbb{R}^{n^s}, \boldsymbol{\beta}\in\mathbb{R}^{n^t}, \alpha_i+\beta_j\leq c(\mathbf{x_i^s}, \mathbf{x_j^t})} \boldsymbol{\alpha}^T \mathbf{a} + \boldsymbol{\beta}^T \mathbf{b} \tag{14}$$

- $W(\mu_s, \mu_t)$ is convex wrt. $\mathbf{a}$
- Dual solution $\boldsymbol{\alpha}^*$ is a sub-gradient : $\boldsymbol{\alpha}^* \in \partial_{\mathbf{a}} W(\mu_s, \mu_t)$
- **Entropy regularized**: $W(\mu_s, \mu_t)$ is smooth, convex and $\nabla_{\mathbf{a}} W_\lambda(\mu_s, \mu_t) = \lambda \log \mathbf{u}$.
- **OT loss**: $\nabla_{\mathbf{a}} OT_\lambda(\mu_s, \mu_t)$ computed using the implicit function theorem [Luise et al., 2018].

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

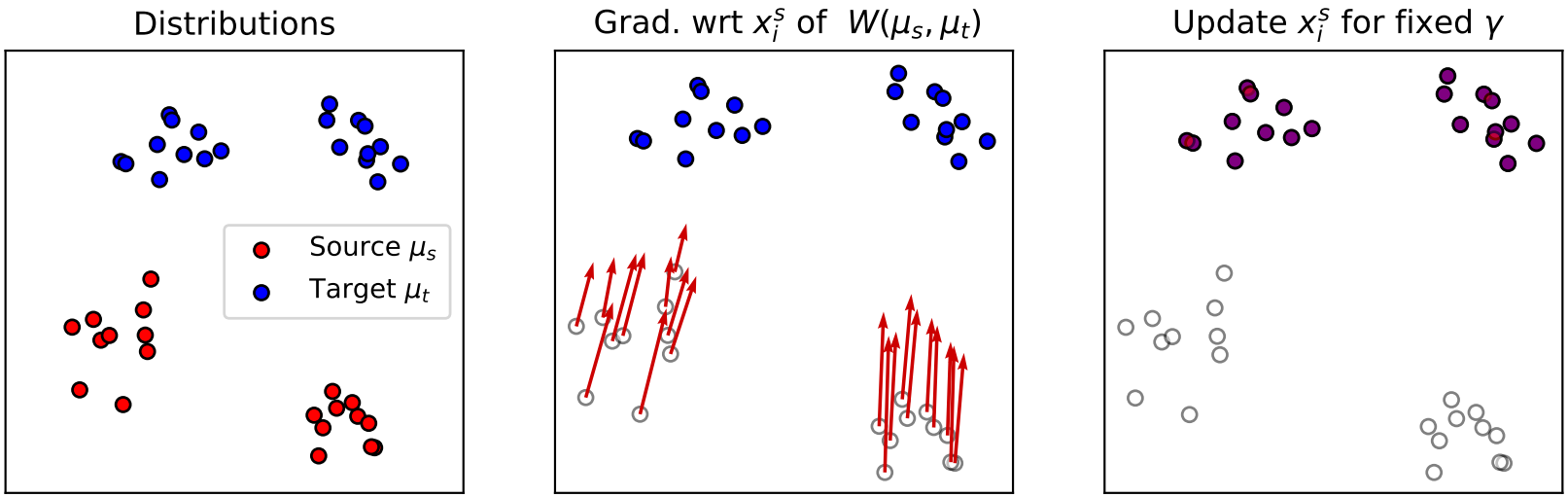


Gradient and update respect to weigths $\mathbf{X}_s = \{\mathbf{x}_i^s\}$ for $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$

$$W_2^2(\mu_s, \mu_t) = \min_{\mathbf{T} \in \Pi(\mu_s, \mu_t)} \sum_{i,j} T_{i,j} \|\mathbf{x}_i^s - \mathbf{x}_j^t\|^2 \tag{15}$$

- Gradient: $\nabla_{\mathbf{x}_i^s} W_2^2(\mu_s, \mu_t) = 2\mathbf{x}_i^s - 2\frac{1}{a_i} \sum_j T_{i,j} \mathbf{x}_j^t$
- $W_2^2(\mu_s, \mu_t)$ decreases if $\mathbf{X_s} \leftarrow \text{diag}(\mathbf{a}^{-1})\mathbf{T}^* \mathbf{X_t}$
- Expression above called barycentric interpolation [Ferradans et al., 2014].

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



Gradient and update respect to weigths $\mathbf{X}_s = \{\mathbf{x}_i^s\}$ for $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$

$$W_2^2(\mu_s, \mu_t) = \min_{\mathbf{T} \in \Pi(\mu_s, \mu_t)} \sum_{i,j} T_{i,j} \|\mathbf{x}_i^s - \mathbf{x}_j^t\|^2 \tag{15}$$

- Gradient: $\nabla_{\mathbf{x}_i^s} W_2^2(\mu_s, \mu_t) = 2\mathbf{x}_i^s - 2\frac{1}{a_i} \sum_j T_{i,j} \mathbf{x}_j^t$
- $W_2^2(\mu_s, \mu_t)$ decreases if $\mathbf{X_s} \leftarrow \text{diag}(\mathbf{a}^{-1})\mathbf{T}^* \mathbf{X_t}$
- Expression above called barycentric interpolation [Ferradans et al., 2014].

General case for entropic OT: autodifferentiation

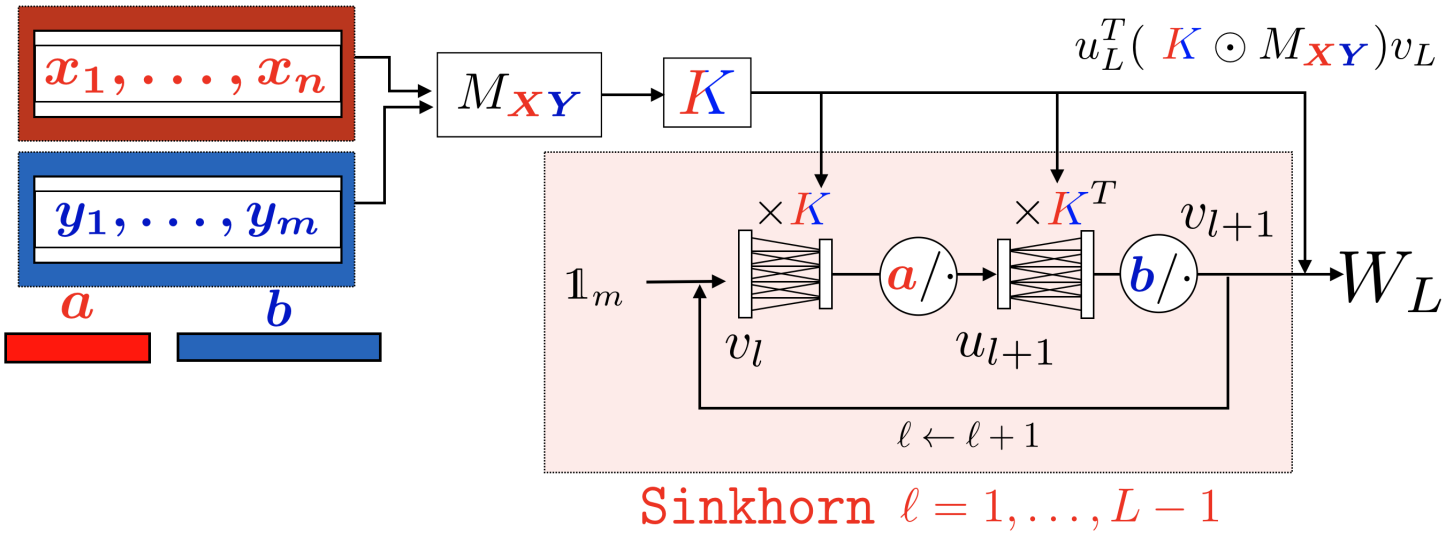


Image from Marco Cuturi

Sinkhorn Autodiff [Genevay et al., 2017]

- Computing gradients through implicit function theorem can be costly [Luise et al., 2018].
- Each iteration of the SInkhorn algorithm is differentiable.
- Modern neural network toolboxes can perform autodiff (Pytorch, Tensorflow).
- Fast but needs log-stabilization for numerical stability.
- At convergence, closed form solution of the gradients exist (no need to autodiff).

Outline

Extensions of Optimal Transport

Relaxation and extensions

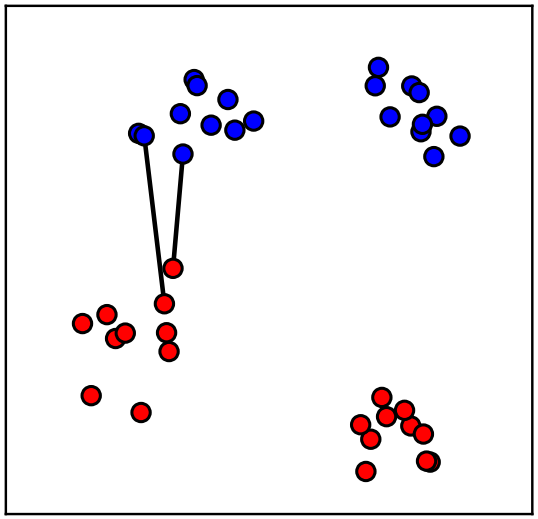
- OT is a powerful formulation for several ML applications.
- But as illustrated by entropic regularization, one can also change the optimization problem to get a better/more representative problem.
- Several extensions and variants of OT has been studied by mathematicians and ML practitioners.

Extensions of Optimal Transport

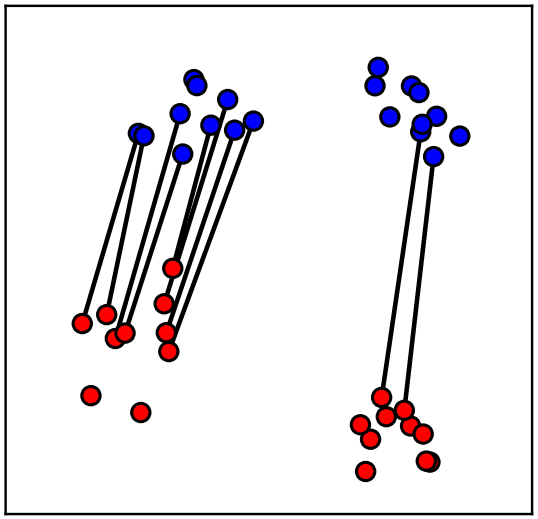
- **Partial OT**, only a portion of the mass is required to be transported.
- **Unbalanced OT**, can transport between distributions with different total mass.
- **Multi-marginal OT**, searches for a transport between more than two distributions.
- **Gromov-Wasserstein OT**, searches for a transport across metric spaces.

Partial Optimal Transport

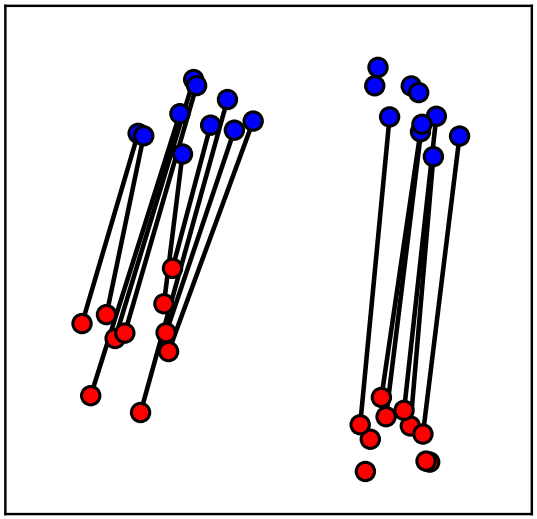
Partial OT with $m = 0.1$



Partial OT with $m = 0.5$



Partial OT with $m = 0.8$



Partial OT [Caffarelli and McCann, 2010, Figalli, 2010]

$$\min_{\mathbf{T} \in \Pi^m(\mu_s, \mu_t)} \left\{ \langle \mathbf{T}, \mathbf{C} \rangle_F = \sum_{i,j} T_{i,j} c_{i,j} \right\}$$

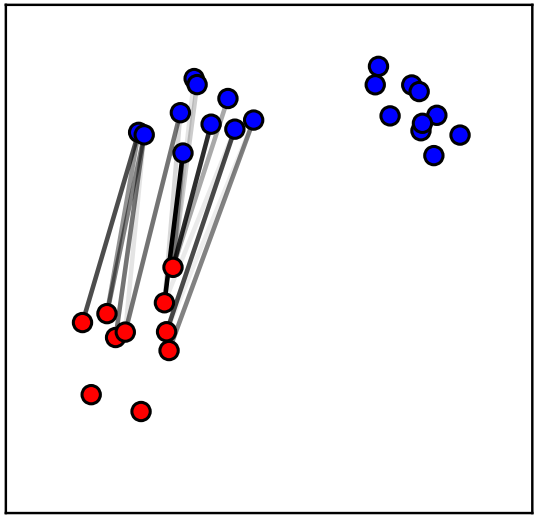
where $\mathbf{C}$ is a cost matrix with $c_{i,j} = c(\mathbf{x}_i^s, \mathbf{x}_j^t)$ and the marginals constraints are

$$\Pi^m(\mu_s, \mu_t) = \left\{ \mathbf{T} \in (\mathbb{R}^+)^{n_s \times n_t} \,|\, \mathbf{T}\mathbf{1}_{n_t} \leq \mathbf{a}, \mathbf{T}^T \mathbf{1}_{n_s} \leq \mathbf{b}, \mathbf{1}_{n_s}^T \mathbf{T} \mathbf{1}_{n_t} = m \right\}$$

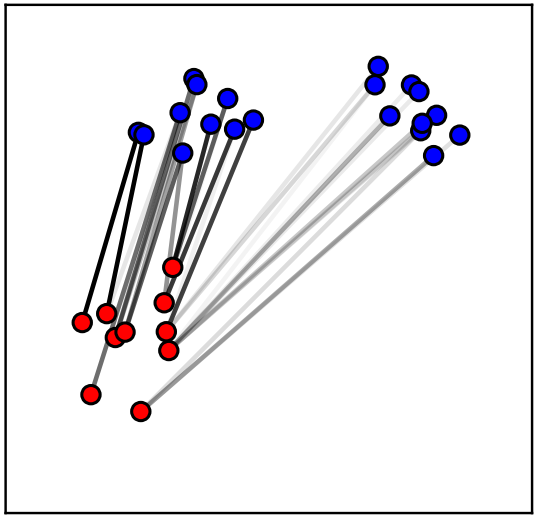
- The equality constraint is on the total transported mass that must be equal to m.
- Allows distributions with different total mass when $m \leq \min(\mathbf{1}_{n_s}^T \mathbf{a}, \mathbf{1}_{n_t}^T \mathbf{b})$

Unbalanced Optimal Transport

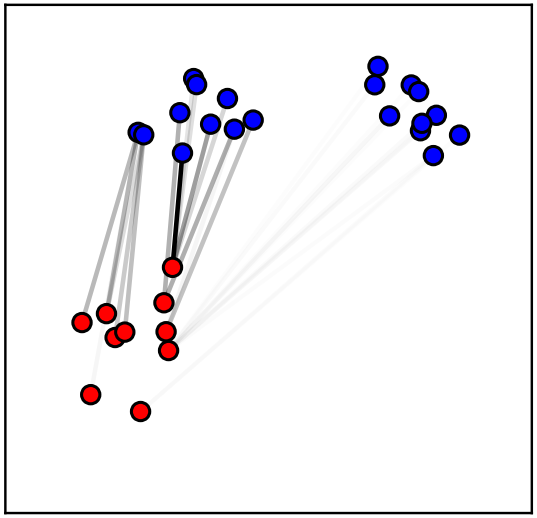
L2 UOT with $\lambda^u = 30$



L2 UOT with $\lambda^u = 50$



KL UOT with $\lambda^u = 1$

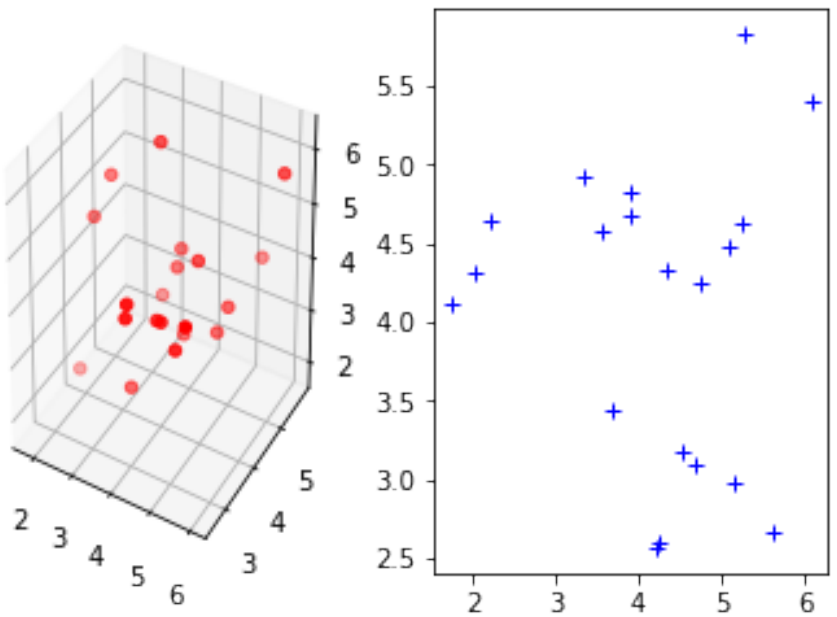


Unbalanced Optimal transport (UOT) [Benamou, 2003]

$$\min_{\mathbf{T} \geq 0} \quad \langle \mathbf{T}, \mathbf{C} \rangle_F + \lambda^u D_\varphi(\mathbf{T}\mathbf{1}_m, \mathbf{a}) + \lambda^u D_\varphi(\mathbf{T}^\top \mathbf{1}_n, \mathbf{b}) \tag{16}$$

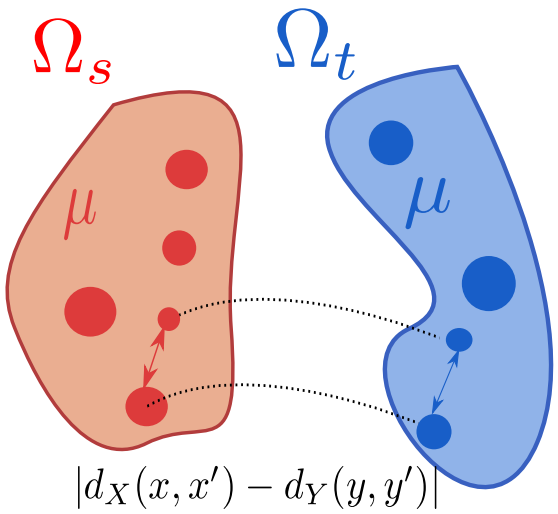
- D_φ is a a Bregman divergence penalizing the violation of the marginal constraints.
- Only a portion of the total mass is transported, total mass can be unbalanced between source and target due to constraint relaxation.
- Closed form exists between Gaussians [Janati et al., 2020, Janati, 2021].
- Can be reformulated and solved like a penalized regression problem [Chapel et al., 2021].

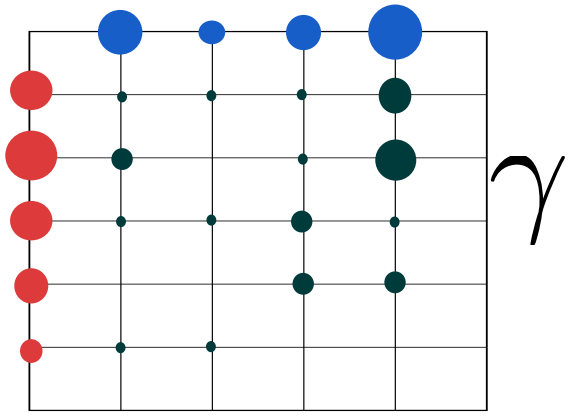
Can you transport between different spaces ?



- Ω_s : source space, Ω_t : target space.
- Both domains/spaces do not share the same variables.
- There is no $c(\mathbf{x}, \mathbf{y})$ between the two domains.
- They are related (observe similar objects) but not registered.
- Example: multi-modality with observations on different objects.

Gromov-Wasserstein divergence





Inspired from Gabriel Peyré

GW for discrete distributions [Memoli, 2011]

$$\mathcal{GW}_p(\mu_s, \mu_t) = \left(\min_{T \in \Pi(\mu_s, \mu_t)} \sum_{i,j,k,l} |D_{i,k} - D'_{j,l}|^p T_{i,j}\, T_{k,l} \right)^{\frac{1}{p}}$$

with $\mu_s = \sum_i a_i \delta_{\mathbf{x}_i^s}$ and $\mu_t = \sum_j b_j \delta_{x_j^t}$ and $D_{i,k} = \|\mathbf{x}_i^s - \mathbf{x}_k^s\|, D'_{j,l} = \|\mathbf{x}_j^t - \mathbf{x}_l^t\|$

- Distance between metric measured spaces : across different spaces.
- Search for an OT plan that preserve the pairwise relationships between samples.
- Invariant to isometry in either spaces (e.g. rotations and translation).

Entropic Gromov-Wasserstein

Optimization Problem [Peyré et al., 2016]

$$\mathcal{GW}^p_{p,\epsilon}(\mu_s, \mu_t) = \min_{\mathbf{T}\in\Pi(\mu_s,\mu_t)} \sum_{i,j,k,l} |D_{i,k} - D'_{j,l}|^p T_{i,j}\, T_{k,l} + \epsilon \sum_{i,j} T_{i,j} \log T_{i,j} \qquad (17)$$

with $\mu_s = \sum_i a_i \delta_{\mathbf{x}^s_i}$ and $\mu_t = \sum_j b_j \delta_{x^t_j}$ and $D_{i,k} = \|\mathbf{x}^s_i - \mathbf{x}^s_k\|, D'_{j,l} = \|\mathbf{x}^t_j - \mathbf{x}^t_l\|$

- Smoothing the original GW with a convex and smooth entropic term.

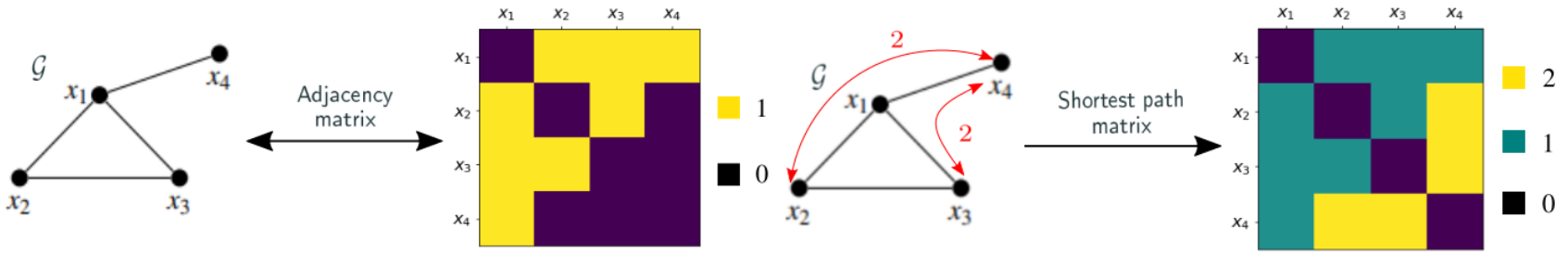
Solving the entropic $\mathcal{GW}$ [Peyré et al., 2016]

- Problem (17) can be solved using a KL mirror descent.
- This is equivalent to solving at each iteration t

$$\mathbf{T}^{(t+1)} = \min_{T\in\mathcal{P}} \quad \left\langle \mathbf{T}, \mathbf{G}^{(t)} \right\rangle_F + \epsilon \sum_{i,j} T_{i,j} \log T_{i,j}$$

  Where $G^{(t)}_{i,j} = 2\sum_{k,l} |D_{i,k} - D'_{j,l}|^p T^{(t)}_{k,l}$ is the gradient of the GW loss at previous point $\mathbf{T}^{(k)}$.
- Problem above can be solved using a Sinkhorn-Knopp algorithm of entropic OT.
- Very fast approximation exist for low rank distances [Scetbon et al., 2021].

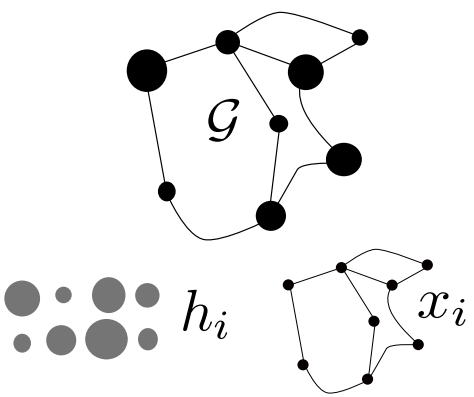
Gromov-Wasserstein between graphs



Modeling the graph structure with a pairwise matrix D

- An undirected graph $\mathcal{G} := (\mathbf{V}, \mathbf{E})$ is defined by $\mathbf{V} = \{\mathbf{x_i}\}_{\mathbf{i} \in [\mathbf{N}]}$ set of the $\mathbf{N}$ nodes and $\mathbf{E} = \{(\mathbf{x_i}, \mathbf{x_j}) | \mathbf{x_i} \leftrightarrow \mathbf{x_j}\}$ set of edges.
- Structure represented as a symmetric matrix $\boldsymbol{D}$ of relations between the nodes.
- Possible choices : **Adjacency matrix** (used in this study), Laplacian matrix, Shortest path matrix.

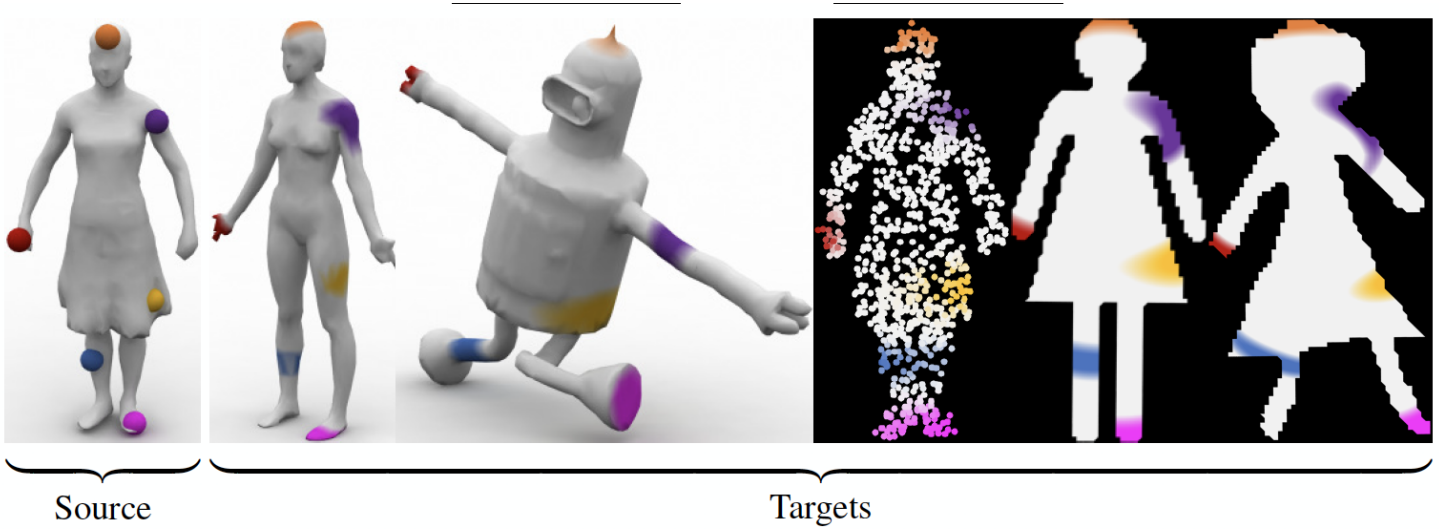
Graph as a distribution (D, h)



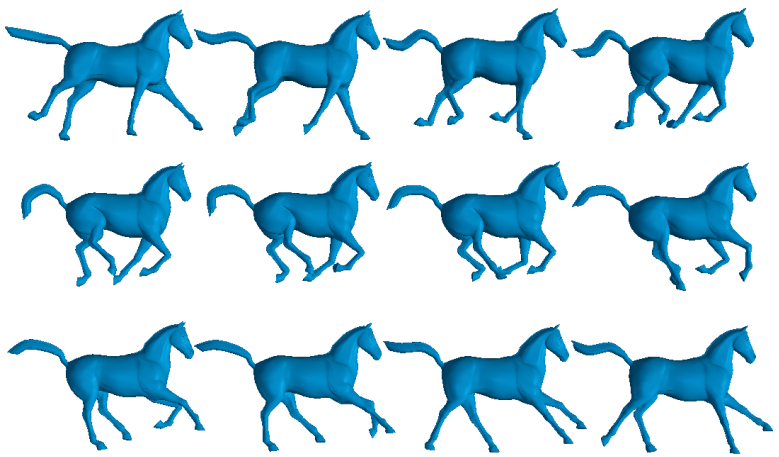
- Graph represented as $\mu_X = \sum_i h_i \delta_{x_i}$.
- The positions x_i are implicit and represented as the pairwise matrix $\boldsymbol{D}$.
- h_i are the masses on the nodes of the graphs (uniform by default).

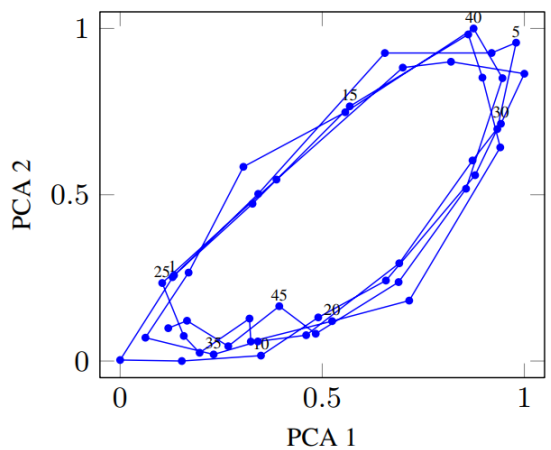
Applications of GW [Solomon et al., 2016]

Shape matching between 3D and 2D surfaces

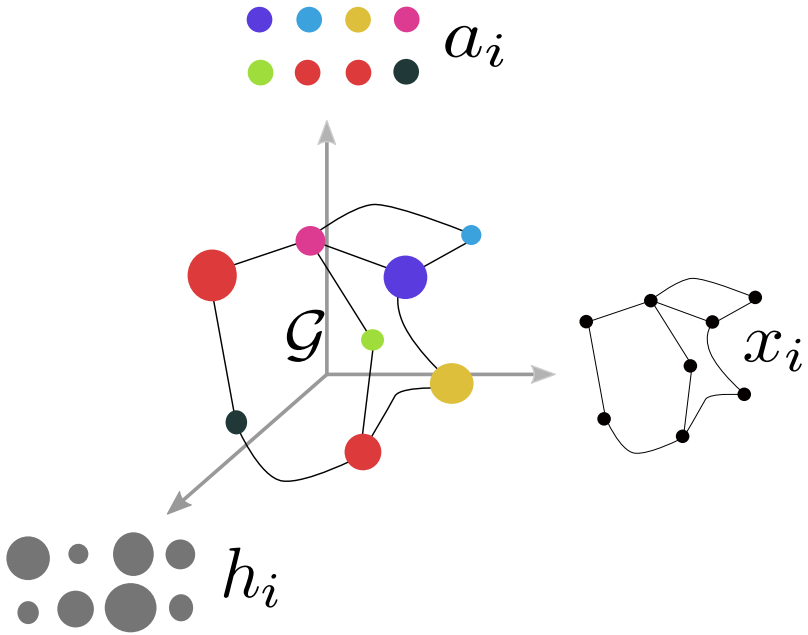


Multidimensional scaling (MDS) of shape collection





Labeled graphs as distributions



$$\left.\vphantom{\sum}\right\} \mu = \sum_i h_i \delta_{(x_i, a_i)}$$

$$\left.\vphantom{\sum}\right\} \mu_A = \sum_i h_i \delta_{a_i}$$

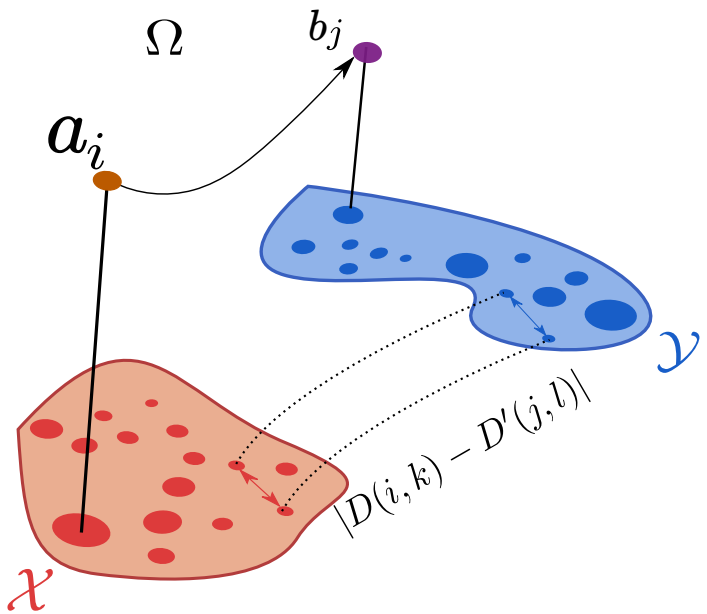
$$\left.\vphantom{\sum}\right\} \mu_X = \sum_i h_i \delta_{x_i}$$

Graph data representation

$$\mu = \sum_{i=1}^{n} h_i \delta_{(x_i a_i)}$$

- Nodes are weighted by their mass h_i.
- But no common metric between the structure points x_i of two different graphs.
- Features values a_i can be compared through the common metric

Fused Gromov-Wasserstein distance



Fused Gromov Wasserstein distance

$\mu_s = \sum_{i=1}^{n} h_i \delta_{x_i, a_i}$ and $\mu_t = \sum_{j=1}^{m} g_j \delta_{y_j, b_j}$

$$\mathcal{FGW}_{p,q,\alpha}(D, D', \mu_s, \mu_t) = \left(\min_{\mathbf{T} \in \Pi(\mu_s, \mu_t)} \sum_{i,j,k,l} \left((1-\alpha) C_{i,j}^q + \alpha |D_{i,k} - D'_{j,l}|^q\right)^p T_{i,j}\, T_{k,l} \right)^{\frac{1}{p}}$$

with $D_{i,k} = \|x_i - x_k\|$ and $D'_{j,l} = \|y_i - y_l\|$ and $C_{i,j} = \|a_i - b_j\|$

- Parameters $q > 1$, $\forall p \geq 1$.
- $\alpha \in [0, 1]$ is a trade off parameter between structure and features.

FGW Properties (1)

$$\mathcal{FGW}^p_{p,q,\alpha}(D, D', \mu_s, \mu_t) = \min_{\mathbf{T} \in \Pi(\mu_s, \mu_t)} \sum_{i,j,k,l} \left((1-\alpha) C^q_{i,j} + \alpha |D_{i,k} - D'_{j,l}|^q\right)^p T_{i,j}\, T_{k,l}$$

Metric properties [Vayer et al., 2020]

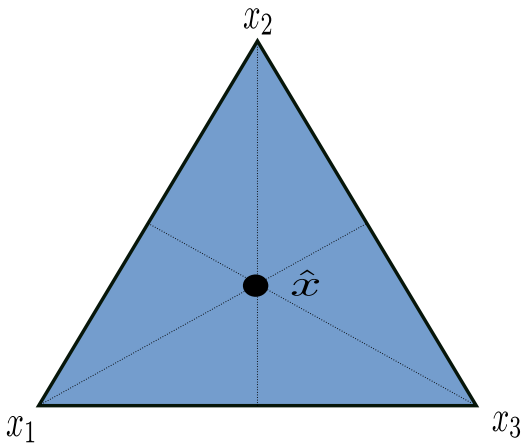
- $\mathcal{FGW}$ defines a metric over structured data with **measure and features preserving isometries** as invariants.
- $\mathcal{FGW}$ is a metric for $q = 1$ a semi metric for $q > 1$, $\forall p \geq 1$.
- The distance is nul *iff* :
  - There exists a Monge map $T\#\mu_s = \mu_t$.
  - Structures are equivalent through this Monge map (isometry).
  - Features are equal through this Monge map.

Other properties for continuous distributions

- Interpolation between $\mathcal{W}$ ($\alpha = 0$) and $\mathcal{GW}$ ($\alpha = 1$) distances.
- Geodesic properties (constant speed, unicity).

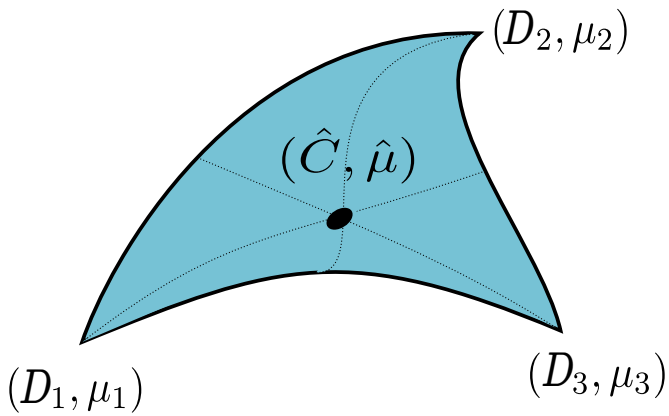
FGW barycenter

Euclidean barycenter



$$\min_x \sum_k \lambda_k \|x - x_k\|^2$$

FGW barycenter

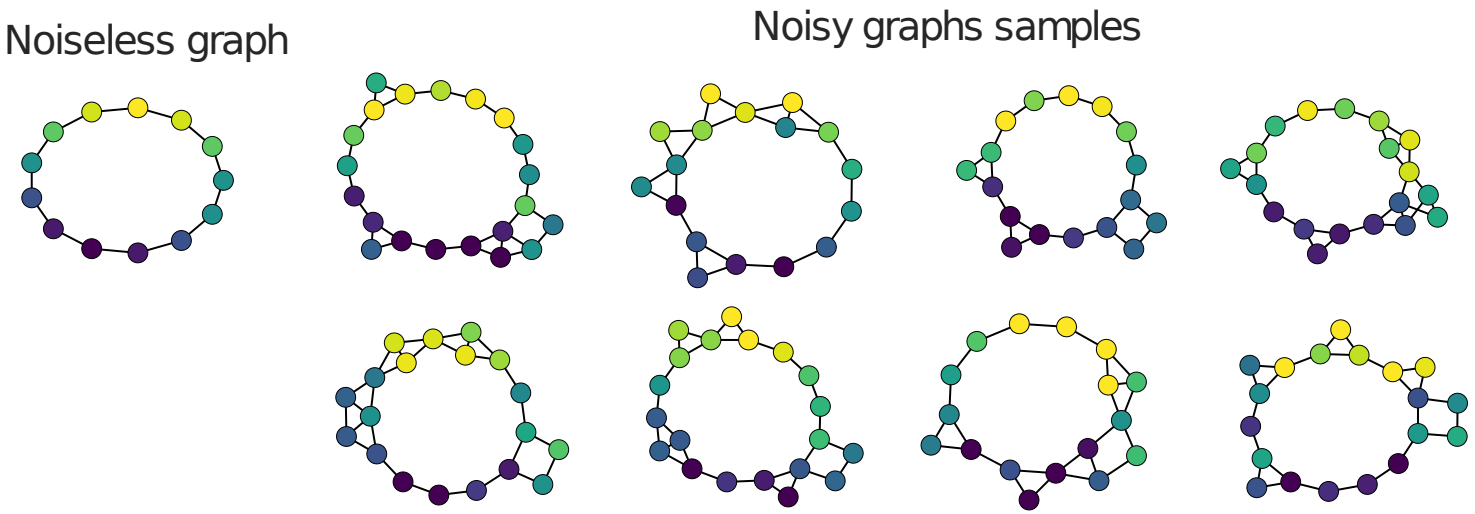


$$\min_{D \in \mathbb{R}^{n \times n}, \mu} \sum_i \lambda_i \mathcal{FGW}(D_i, D, \mu_i, \mu)$$

FGW barycenter $p = 1, q = 2$

- Estimate FGW barycenter using Frechet means (similar to [Peyré et al., 2016]).
- Barycenter optimization solved via block coordinate descent (on $T, D, \{a_i\}_i$).
- Can chose to fix the structure (D) or the features $\{a_i\}_i$ in the barycenter.
- a_{ii}, and D updates are weighted averages using T.

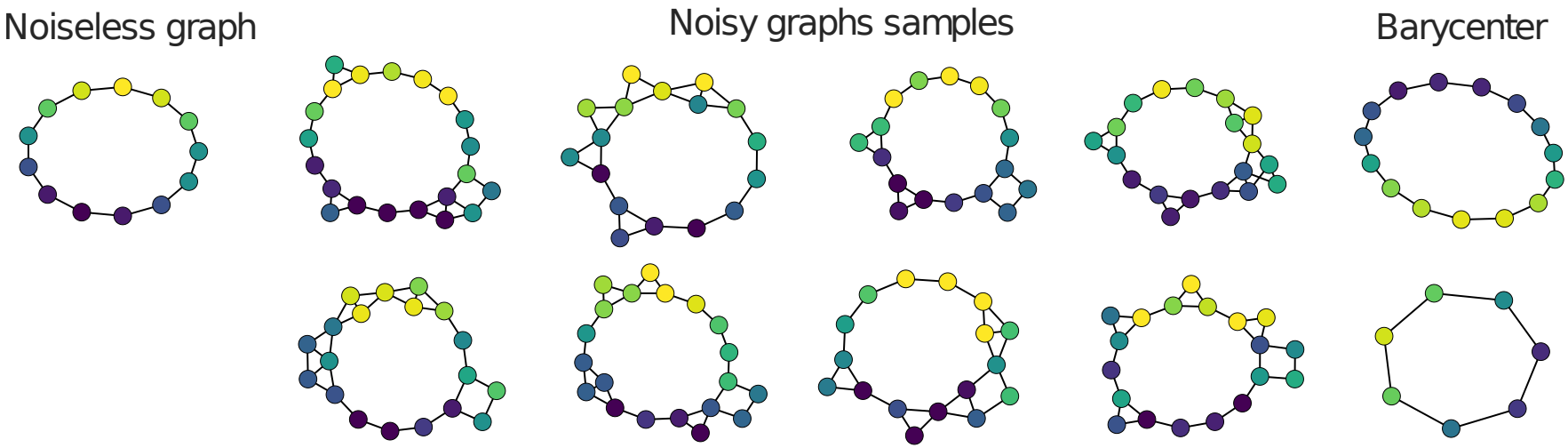
FGW barycenter on labeled graphs



Barycenter of noisy graphs

- We select a clean graph, change the number of nodes and add label noise and random connections.
- We compute the barycenter on $n = 15$ and $n = 7$ nodes.
- Barycenter graph is obtained through thresholding of the D matrix.

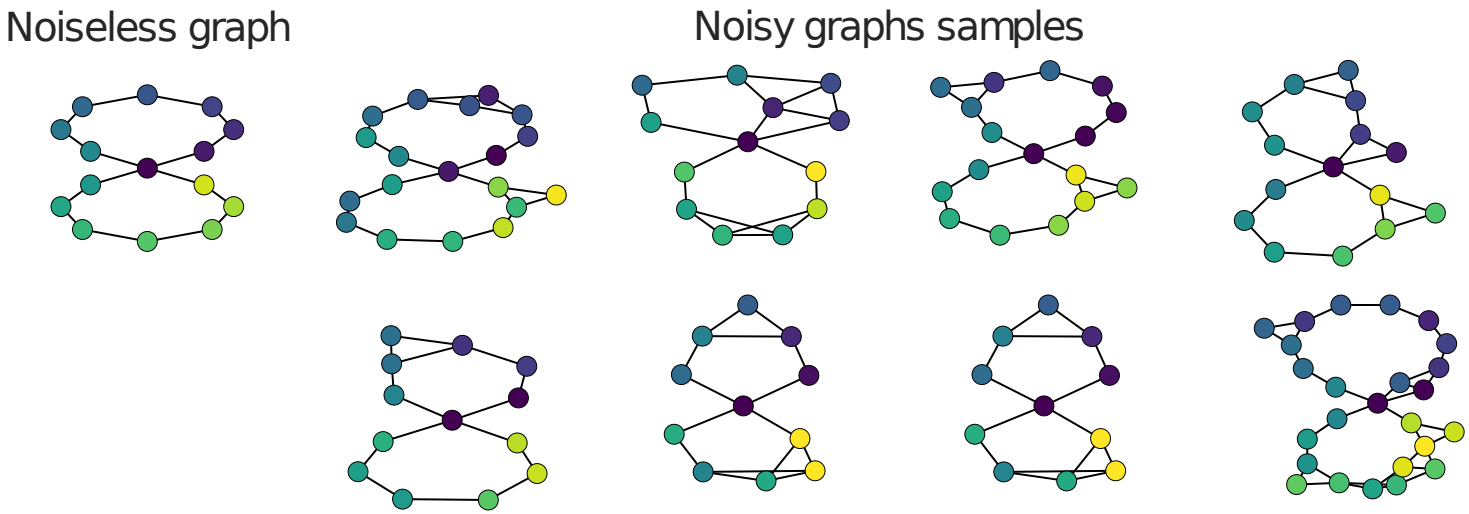
FGW barycenter on labeled graphs



Barycenter of noisy graphs

- We select a clean graph, change the number of nodes and add label noise and random connections.
- We compute the barycenter on $n = 15$ and $n = 7$ nodes.
- Barycenter graph is obtained through thresholding of the D matrix.

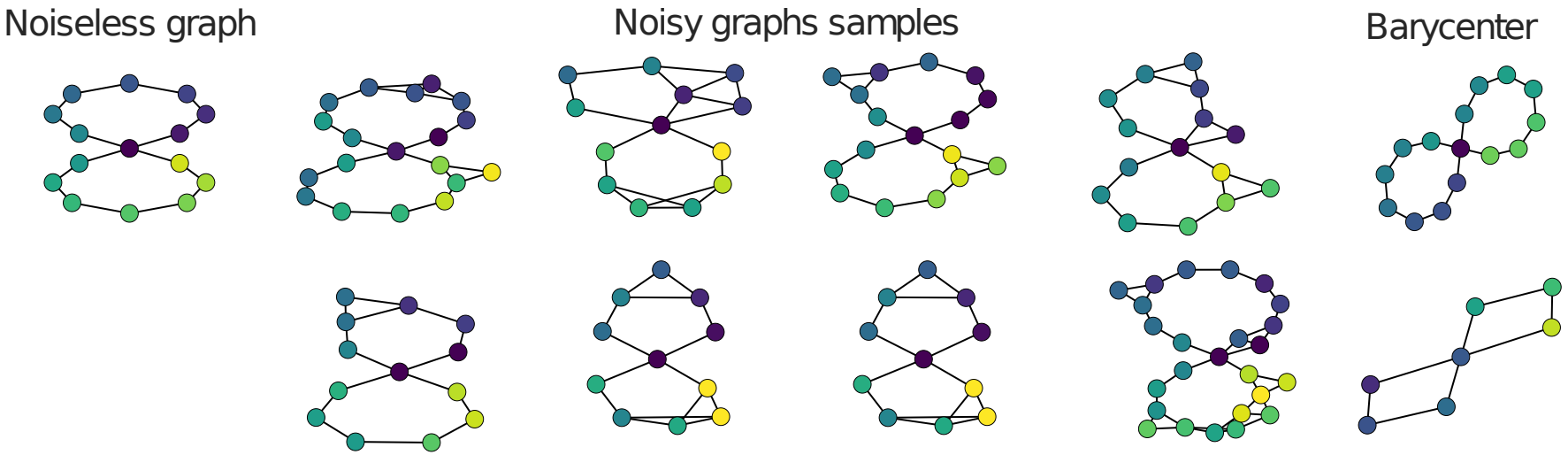
FGW barycenter on labeled graphs



Barycenter of noisy graphs

- We select a clean graph, change the number of nodes and add label noise and random connections.
- We compute the barycenter on $n = 15$ and $n = 7$ nodes.
- Barycenter graph is obtained through thresholding of the D matrix.

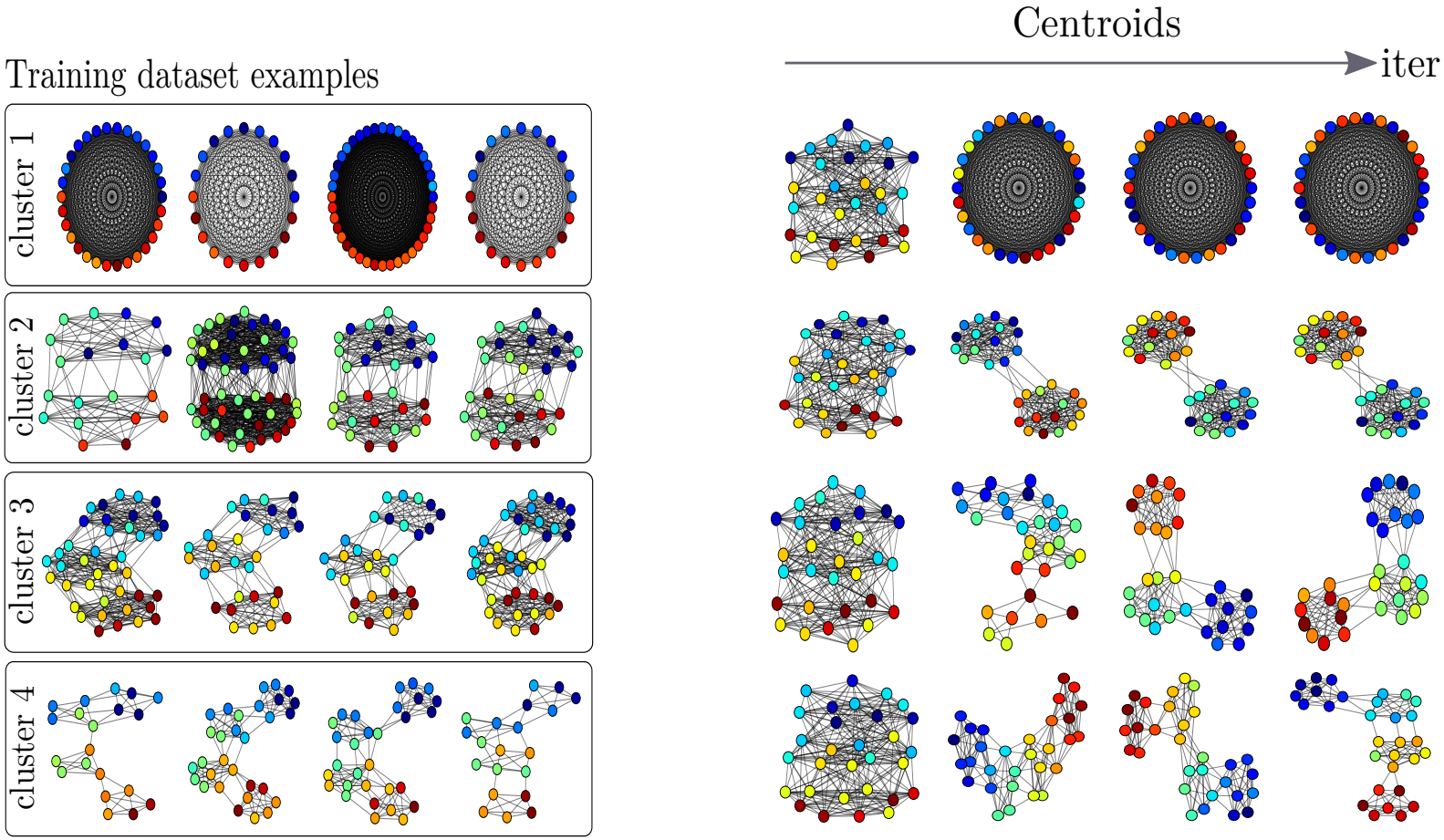
FGW barycenter on labeled graphs



Barycenter of noisy graphs

- We select a clean graph, change the number of nodes and add label noise and random connections.
- We compute the barycenter on $n = 15$ and $n = 7$ nodes.
- Barycenter graph is obtained through thresholding of the D matrix.

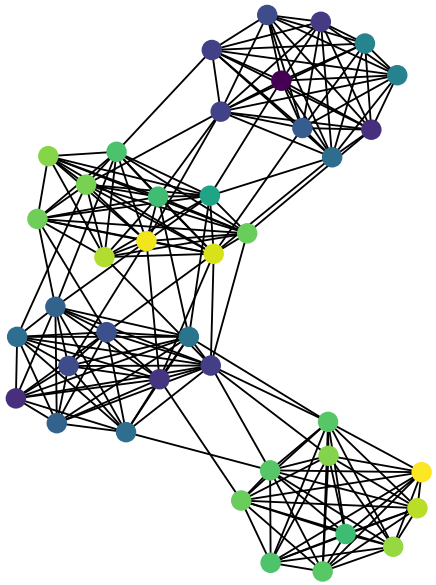
FGW for graphs based clustering



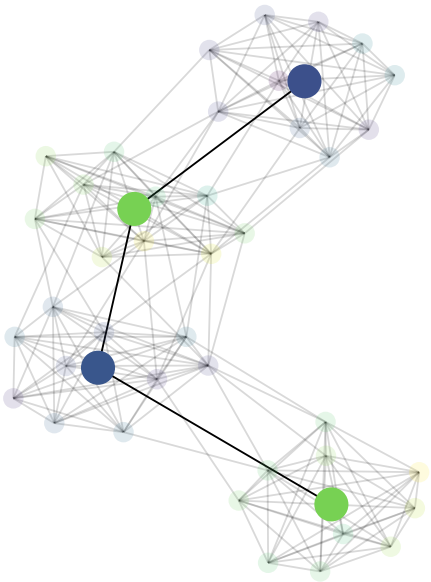
- Clustering of multiple real-valued graphs. Dataset composed of 40 graphs (10 graphs $\times$ 4 types of communities)
- k-means clustering using the FGW barycenter

FGW for community clustering

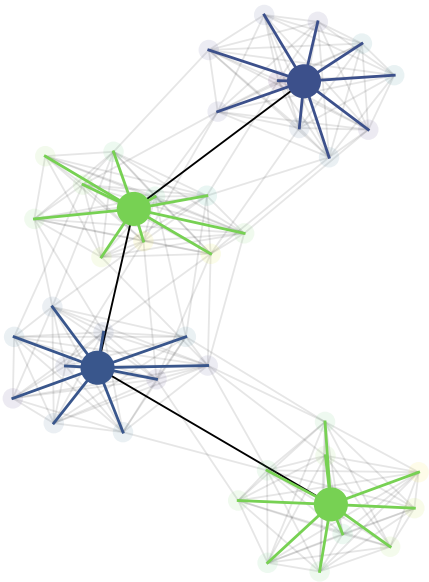
Graph with communities



Approximate Graph



Clustering with transport matrix



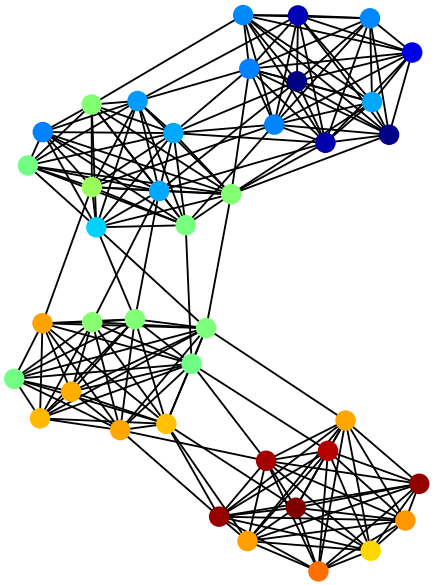
Graph approximation and comunity clustering

$$\min_{D,\mu} \quad \mathcal{FGW}(D, D_0, \mu, \mu_0)$$

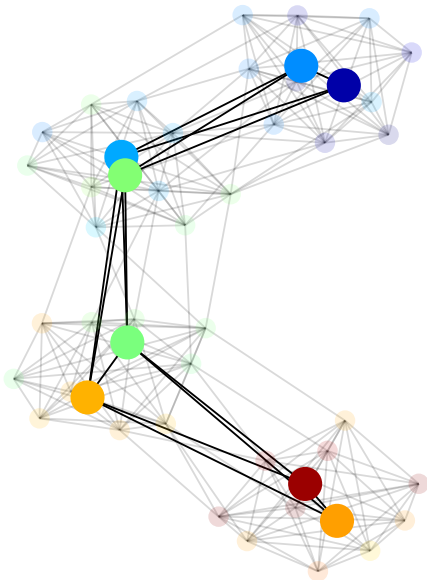
- Approximate the graph (D_0, μ_0) with a small number of nodes.
- OT matrix give the clustering affectation.
- Works for signals and multiple modes in the clusters.

FGW for community clustering

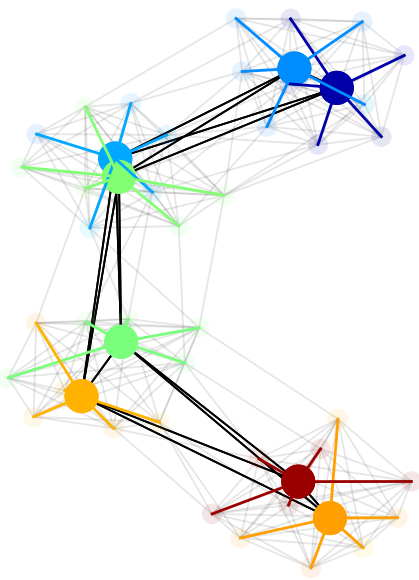
Graph with bimodal communities



Approximate Graph



Clustering with transport matrix



Graph approximation and comunity clustering

$$\min_{D,\mu} \quad \mathcal{FGW}(D, D_0, \mu, \mu_0)$$

- Approximate the graph (D_0, μ_0) with a small number of nodes.
- OT matrix give the clustering affectation.
- Works for signals and multiple modes in the clusters.

Summary for Part 1

Optimal transport

- Theoretically grounded ways of comparing probability distributions.
- Non-parametric comparison (between empirical distributions).
- Ground metric encode the geometry of the space (barycenters, geodesic).
- Two aspects: mapping (Monge) vs coupling (Kantorovitch).
- Several variants exists depending on the application.

Optimization

- Solving OT is a linear program.
- Regularization (entropic) leads to faster algorithms.
- Minimization of Wasserstein distance can be done.
- Reference for computational OT : [Peyré et al., 2019]

Next step: how to use it in machine learning applications ?

References i

[Agueh and Carlier, 2011] Agueh, M. and Carlier, G. (2011).

Barycenters in the wasserstein space.

SIAM Journal on Mathematical Analysis, 43(2):904–924.

[Alvarez-Melis et al., 2017] Alvarez-Melis, D., Jaakkola, T. S., and Jegelka, S. (2017).

Structured optimal transport.

arXiv preprint arXiv:1712.06199.

[Benamou, 2003] Benamou, J.-D. (2003).

Numerical resolution of an "unbalanced" mass transport problem.

ESAIM: Mathematical Modelling and Numerical Analysis, 37(5):851–868.

[Benamou et al., 2015] Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. (2015).

Iterative Bregman projections for regularized transportation problems.

SISC.

References ii

[Blondel et al., 2017] Blondel, M., Seguy, V., and Rolet, A. (2017).

Smooth and sparse optimal transport.

arXiv preprint arXiv:1710.06276.

[Bonneel et al., 2015] 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.

[Brenier, 1991] Brenier, Y. (1991).

Polar factorization and monotone rearrangement of vector-valued functions.

Communications on pure and applied mathematics, 44(4):375–417.

[Caffarelli and McCann, 2010] Caffarelli, L. A. and McCann, R. J. (2010).

Free boundaries in optimal transport and monge-ampere obstacle problems.

Annals of mathematics, pages 673–730.

[Chapel et al., 2021] Chapel, L., Flamary, R., Wu, H., Févotte, C., and Gasso, G. (2021).

Unbalanced optimal transport through non-negative penalized linear regression.

In *Neural Information Processing Systems (NeurIPS)*.

References iii

[Courty et al., 2016a] Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016a).

Optimal transport for domain adaptation.

IEEE Transactions on Pattern Analysis and Machine Intelligence.

[Courty et al., 2016b] Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016b).

Optimal transport for domain adaptation.

Pattern Analysis and Machine Intelligence, IEEE Transactions on.

[Cuturi, 2013] Cuturi, M. (2013).

Sinkhorn distances: Lightspeed computation of optimal transportation.

In *Neural Information Processing Systems (NIPS)*, pages 2292–2300.

[Dessein et al., 2016] Dessein, A., Papadakis, N., and Rouas, J.-L. (2016).

Regularized optimal transport and the rot mover's distance.

arXiv preprint arXiv:1610.06447.

[Ferradans et al., 2014] Ferradans, S., Papadakis, N., Peyré, G., and Aujol, J.-F. (2014).

Regularized discrete optimal transport.

SIAM Journal on Imaging Sciences, 7(3).

References iv

[Figalli, 2010] Figalli, A. (2010).

The optimal partial transport problem.

Archive for rational mechanics and analysis, 195(2):533–560.

[Genevay et al., 2018] Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2018).

Sample complexity of sinkhorn divergences.

arXiv preprint arXiv:1810.02733.

[Genevay et al., 2016] Genevay, A., Cuturi, M., Peyré, G., and Bach, F. (2016).

Stochastic optimization for large-scale optimal transport.

In *NIPS*, pages 3432–3440.

[Genevay et al., 2017] Genevay, A., Peyré, G., and Cuturi, M. (2017).

Sinkhorn-autodiff: Tractable wasserstein learning of generative models.

arXiv preprint arXiv:1706.00292.

References v

[Gramfort et al., 2015] Gramfort, A., Peyré, G., and Cuturi, M. (2015).

Fast optimal transport averaging of neuroimaging data.

In *International Conference on Information Processing in Medical Imaging*, pages 261–272. Springer.

[Janati, 2021] Janati, H. (2021).

Advances in Optimal transport and applications to neuroscience.

PhD thesis, Institut Polytechnique de Paris.

[Janati et al., 2020] Janati, H., Muzellec, B., Peyré, G., and Cuturi, M. (2020).

Entropic optimal transport between unbalanced gaussian measures has a closed form.

Advances in Neural Information Processing Systems, 33.

[Kantorovich, 1942] Kantorovich, L. (1942).

On the translocation of masses.

C.R. (Doklady) Acad. Sci. URSS (N.S.), 37:199–201.

References vi

[Kolouri et al., 2017] Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. (2017).

Optimal mass transport: Signal processing and machine-learning applications.

IEEE signal processing magazine, 34(4):43–59.

[Kolouri et al., 2016] Kolouri, S., Zou, Y., and Rohde, G. K. (2016).

Sliced wasserstein kernels for probability distributions.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267.

[Kusner et al., 2015] Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015).

From word embeddings to document distances.

In *International Conference on Machine Learning*, pages 957–966.

[Luise et al., 2018] Luise, G., Rudi, A., Pontil, M., and Ciliberto, C. (2018).

Differential properties of sinkhorn approximation for learning with wasserstein distance.

In *Advances in Neural Information Processing Systems*, pages 5864–5874.

References vii

[McCann, 1997] McCann, R. J. (1997).

A convexity principle for interacting gases.

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

[Memoli, 2011] Memoli, F. (2011).

Gromov wasserstein distances and the metric approach to object matching.

Foundations of Computational Mathematics, pages 1–71.

[Monge, 1781] Monge, G. (1781).

Mémoire sur la théorie des déblais et des remblais.

De l'Imprimerie Royale.

[Papadakis et al., 2014] Papadakis, N., Peyré, G., and Oudet, E. (2014).

Optimal Transport with Proximal Splitting.

SIAM Journal on Imaging Sciences, 7(1):212–238.

[Peyré et al., 2019] Peyré, G., Cuturi, M., et al. (2019).

Computational optimal transport: With applications to data science.

Foundations and Trends® in Machine Learning, 11(5-6):355–607.

References viii

[Peyré et al., 2016] Peyré, G., Cuturi, M., and Solomon, J. (2016).

Gromov-wasserstein averaging of kernel and distance matrices.

In *ICML*, pages 2664–2672.

[Rubner et al., 2000] 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, 2014] Santambrogio, F. (2014).

Introduction to optimal transport theory.

Notes.

[Scetbon et al., 2021] Scetbon, M., Peyré, G., and Cuturi, M. (2021).

Linear-time gromov wasserstein distances using low rank couplings and costs.

arXiv preprint arXiv:2106.01128.

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

Large-scale optimal transport and mapping estimation.

References ix

[Solomon et al., 2015] 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.

[Solomon et al., 2016] Solomon, J., Peyré, G., Kim, V. G., and Sra, S. (2016).

Entropic metric alignment for correspondence problems.

ACM Transactions on Graphics (TOG), 35(4):72.

[Vayer et al., 2020] Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. (2020).

Fused gromov-wasserstein distance for structured objects.

Algorithms, 13 (9):212.

[Wang et al., 2018] Wang, Q., Redko, I., and Takerkart, S. (2018).

Population averaging of neuroimaging data using l p distance-based optimal transport.

In *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pages 1–4. IEEE.