Optimal transport for machine learning

Rémi Flamary

AG GDR ISIS, Sète, 16 Novembre 2017

Collaborators


N. Courty


A. Rakotomamonjy


D. Tuia


A. Habrard

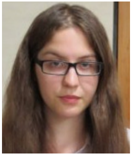

M. Cuturi


M. Perrot


C. Févotte


V. Emiya

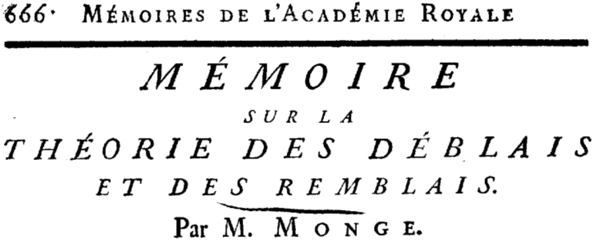

V. Seguy


M. Ducoffe

+ ANR OATMIL project members

Table of content

The origins of optimal transport


666· MÉMOIRES DE L'ACADÉMIE ROYALE

MÉMOIRE

SUR LA

THÉORIE DES DÉBLAIS

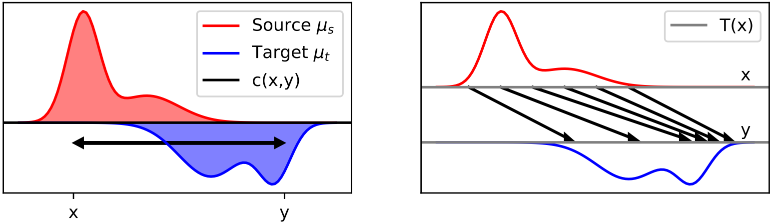
ET DES REMBLAIS.

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).

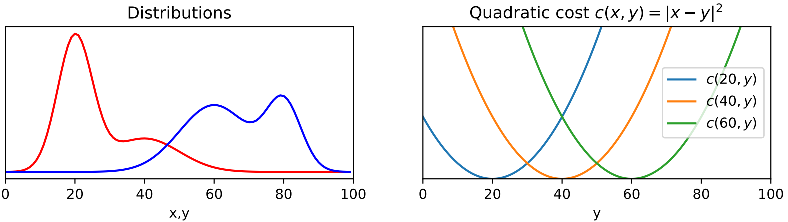
The origins of optimal transport



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)

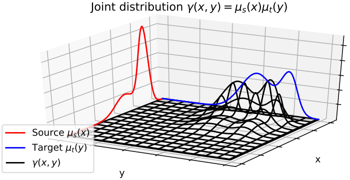


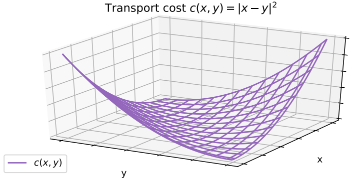
- Probability measures μ_s and μ_t on and a cost function $c : \Omega_s \times \Omega_t \rightarrow \mathbb{R}^+$.
- The Monge formulation [Monge, 1781] aim at finding a mapping $T : \Omega_s \rightarrow \Omega_t$

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

- Non-convex optimization problem, mapping does not exist in the general case.
- [Brenier, 1991] proved existence and unicity of the Monge map for $c(x, y) = \|x - y\|^2$ and distributions with densities.

Optimal transport (Kantorovich formulation)





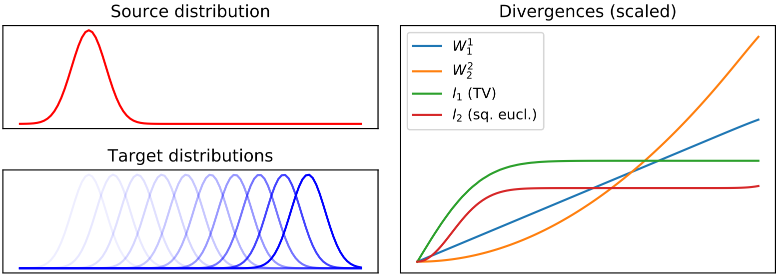
- The Kantorovich formulation [Kantorovich, 1942] seeks for a probabilistic coupling $\boldsymbol{\gamma} \in \mathcal{P}(\Omega_s \times \Omega_t)$ between Ω_s and Ω_t:

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

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

- $\boldsymbol{\gamma}$ is a joint probability measure with marginals μ_s and μ_t.
- Linear Program that always have a solution.

Wasserstein distance



Wasserstein distance

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

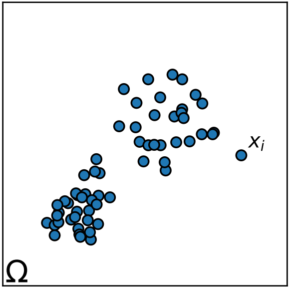
where $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^p$

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

Discrete distributions: Empirical vs Histogram

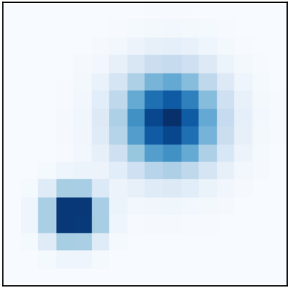
Discrete measure: $$\mu = \sum_{i=1}^{n} \mu_i \delta_{\mathbf{x}_i}, \quad \mathbf{x}_i \in \Omega, \quad \sum_{i=1}^{n} \mu_i = 1$$

Lagrangian (point clouds)



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

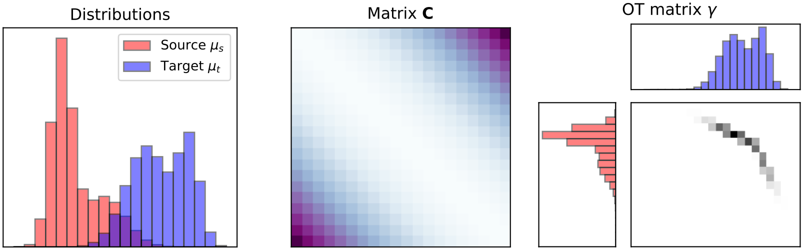
Eulerian (histograms)



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

Slide stolen from Gabriel Peyré

Optimal transport with discrete distributions



OT Linear Program

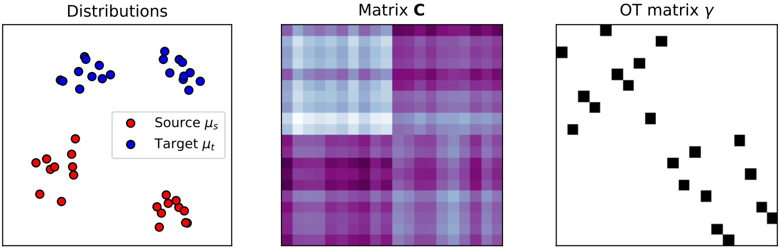
$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma} \in \mathcal{P}}{\arg\min} \quad \left\{ \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F = \sum_{i,j} \gamma_{i,j} c_{i,j} \right\}$$

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 $O(n^3)$.

Optimal transport with discrete distributions



OT Linear Program

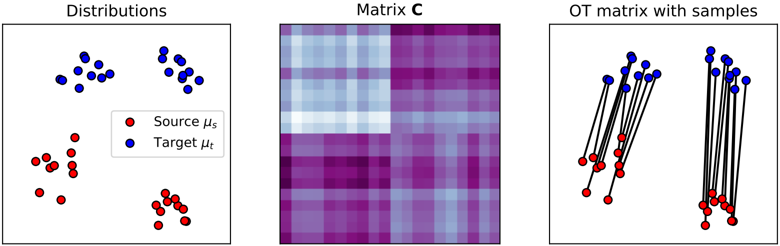
$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma} \in \mathcal{P}}{\arg\min} \quad \left\{ \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F = \sum_{i,j} \gamma_{i,j} c_{i,j} \right\}$$

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

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

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

Optimal transport with discrete distributions



OT Linear Program

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

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 $O(n^3)$.

Regularized optimal transport

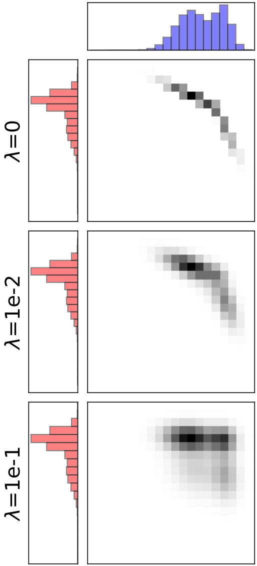
$$\gamma_0^\lambda = \underset{\gamma \in \mathcal{P}}{\arg\min} \quad \langle \gamma, \mathbf{C} \rangle_F + \lambda \Omega(\gamma), \tag{4}$$

Regularization term $\Omega(\gamma)$

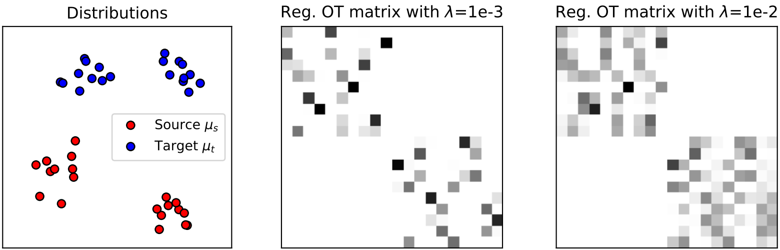
- Entropic regularization [Cuturi, 2013].
- Group Lasso [Courty et al., 2016a].
- KL, Itakura Saito, β-divergences, [Dessein et al., 2016].

Why regularize?

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



Entropic regularized optimal transport

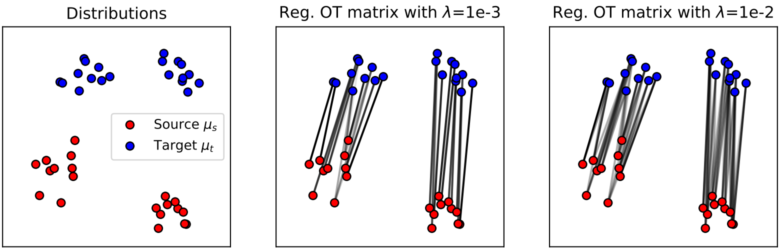


Entropic regularization [Cuturi, 2013]

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

- Regularization with the negative entropy of $\boldsymbol{\gamma}$.
- Solution of the form $\boldsymbol{\gamma}_0^\lambda = \text{diag}(\mathbf{u}) \exp(-\mathbf{C}/\lambda)\text{diag}(\mathbf{v})$.
- **Sinkhorn-Knopp** algorithm (implementation in parallel, GPU).
- Smooth problem in the dual can be solved with BFGS [Cuturi and Peyré, 2016], SGD [Genevay et al., 2016, Seguy et al., 2017].

Entropic regularized optimal transport

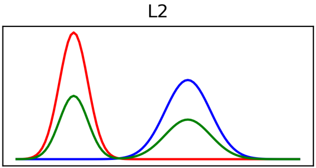


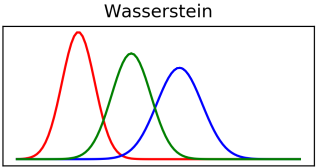
Entropic regularization [Cuturi, 2013]

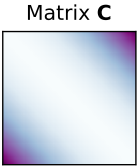
$$\Omega(\boldsymbol{\gamma}) = \sum_{i,j} \boldsymbol{\gamma}(i,j) \log \boldsymbol{\gamma}(i,j)$$

- Regularization with the negative entropy of $\boldsymbol{\gamma}$.
- Solution of the form $\boldsymbol{\gamma}_0^\lambda = \text{diag}(\mathbf{u}) \exp(-\mathbf{C}/\lambda) \text{diag}(\mathbf{v})$.
- **Sinkhorn-Knopp** algorithm (implementation in parallel, GPU).
- Smooth problem in the dual can be solved with BFGS [Cuturi and Peyré, 2016], SGD [Genevay et al., 2016, Seguy et al., 2017].

Wasserstein barycenter





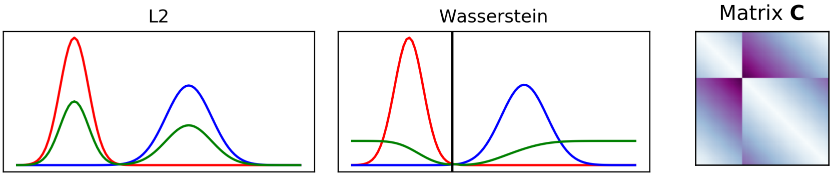


Barycenters [Agueh and Carlier, 2011] and Wasserstein Geodesic

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

- $\lambda_i > 0$ 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



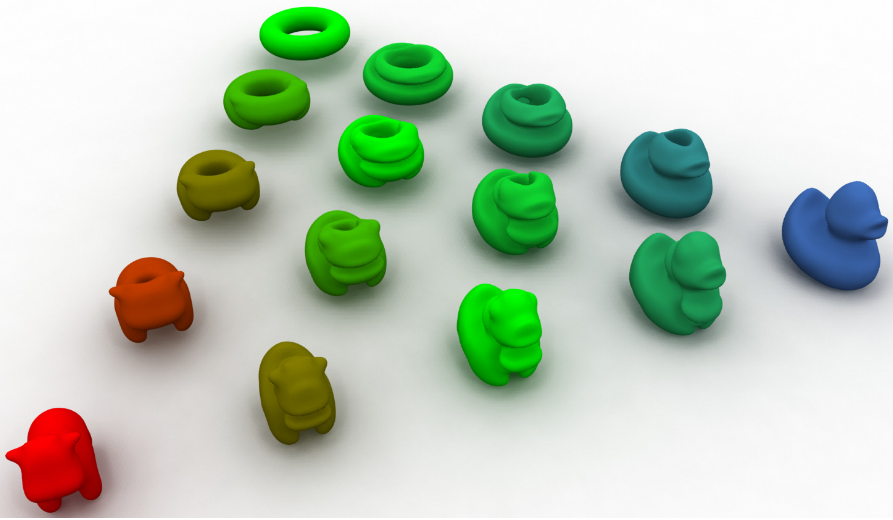
Barycenters [Agueh and Carlier, 2011] and Wasserstein Geodesic

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

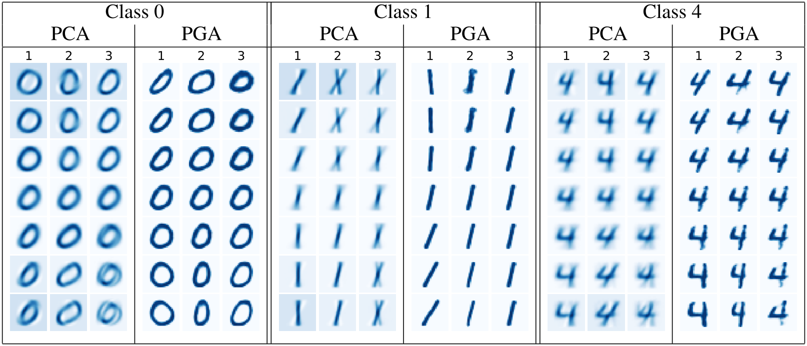
- $\lambda_i > 0$ and $\sum_i^n \lambda_i = 1$.
- Uniform barycenter has $\lambda_i = \frac{1}{n}, \forall i$.
- Interpolation with n=2 and $\boldsymbol{\lambda} = [1 - t, t]$ with $0 \leq t \leq 1$ [McCann, 1997].
- Regularized barycenters using Bregman projections [Benamou et al., 2015].
- The cost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

3D Wasserstein barycenter

Shape interpolation [Solomon et al., 2015]



Principal Geodesics Analysis

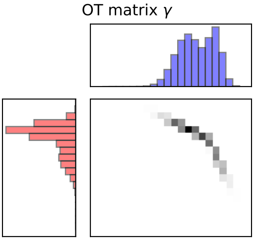


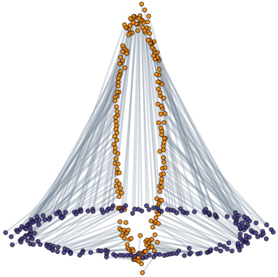
Geodesic PCA in the Wasserstein space [Bigot et al., 2017]

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

Section

Learning with optimal transport





Learning from histograms

- Wasserstein distance.
- Ground metric design.
- Loss for multilabel classifier [Frogner et al., 2015]
- Loss for linear unmixing [Flamary et al., 2016b].

Learning from empirical distributions

- Non parametric divergence between non overlapping distributions.
- Estimate discriminant subspace [Flamary et al., 2016a].
- Objective function for GAN [Arjovsky et al., 2017].

Supervised learning with Wasserstein Loss

Siberian husky

Eskimo dog

Flickr : street, parade, dragon
Prediction : people, protest, parade

Flickr : water, boat, ref ection, sun-shine
Prediction : water, river, lake, summer;

Learning with a Wasserstein Loss [Frogner et al., 2015]

$$\min_{f} \quad \sum_{k=1}^{N} W_1^1(f(\mathbf{x}_i), \mathbf{l}_i)$$

- Empirical loss minimization with Wasserstein loss.
- Multi-label prediction (labels $\mathbf{l}$ seen as histograms, f output softmax).
- Cost between labels can encode semantic similarity between classes.
- Good performances in image tagging.

Linear unmixing with optimal transport

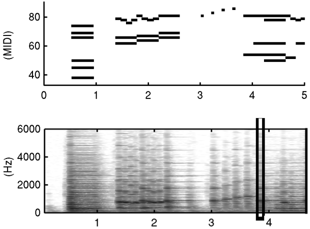
Linear unmixing

$$\min_{\mathbf{h}\in\Delta} \quad W_{\mathbf{C}}(\mathbf{v}, \mathbf{D}\mathbf{h}) \tag{5}$$

- Δ is the probability simplex (positivity, sum to one).
- $\mathbf{v}$ is the observation, $\mathbf{D}$ the dictionary, $\mathbf{h}$ the mixing coefficients.
- Wasserstein as data fitting proposed in [Zen et al., 2014] for matrix factorization.
- Fast algorithm with regularization in [Rolet et al., 2016], non linear unmixing in [Schmitz et al., 2017].

Musical spectral unmixing

- State of the art: KL + designed dictionary.
- Spectra with harmonic structure.
- Variability in the fundamental frequency.
- Variability in the magnitude of the harmonics.



$\Rightarrow$ Optimal spectral transportation [Flamary et al., 2016b].

Linear unmixing with optimal transport

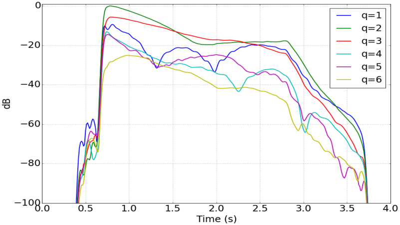
Linear unmixing

$$\min_{\mathbf{h}\in\Delta} \quad W_{\mathbf{C}}(\mathbf{v}, \mathbf{D}\mathbf{h}) \tag{5}$$

- Δ is the probability simplex (positivity, sum to one).
- $\mathbf{v}$ is the observation, $\mathbf{D}$ the dictionary, $\mathbf{h}$ the mixing coefficients.
- Wasserstein as data fitting proposed in [Zen et al., 2014] for matrix factorization.
- Fast algorithm with regularization in [Rolet et al., 2016], non linear unmixing in [Schmitz et al., 2017].

Musical spectral unmixing

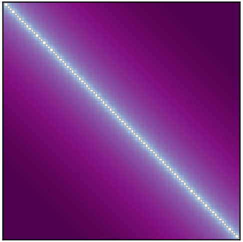
- State of the art: KL + designed dictionary.
- Spectra with harmonic structure.
- Variability in the fundamental frequency.
- Variability in the magnitude of the harmonics.



$\Rightarrow$ Optimal spectral transportation [Flamary et al., 2016b].

Optimal spectral transportation (OST)

Quadratic cost $\mathbf{C}$ (log)



Quadratic cost between frequencies

- Allows small shift in frequencies.
- Very sensitive to harmonics magnitude.

Harmonic invariant cost

$$c_{ij} = \min_{q=1,\ldots,\left\lceil \frac{f_i}{f_j} \right\rceil} (f_i - q f_j)^2 + \epsilon\, \delta_{q \neq 1},$$

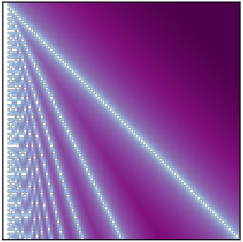
- Allow mass transfer between harmonics.
- $\epsilon > 0$ discriminates between octaves.

Solving the optimization problem

- A good invariant cost allows for extremely simple dictionary elements (diracs on the fundamental frequency).
- We take $\mathbf{D}$ as diracs on the fundamental frequencies of the notes.
- Closed form for solving the OT problem.
- Non-convex Group lasso for sparse estimates and/or entropic regularization.

Optimal spectral transportation (OST)

Harmonic cost **C** (log)



Quadratic cost between frequencies

- Allows small shift in frequencies.
- Very sensitive to harmonics magnitude.

Harmonic invariant cost

$$c_{ij} = \min_{q=1,\dots,\left\lceil \frac{f_i}{f_j} \right\rceil} (f_i - q f_j)^2 + \epsilon\, \delta_{q \neq 1},$$

- Allow mass transfer between harmonics.
- $\epsilon > 0$ discriminates between octaves.

Solving the optimization problem

- A good invariant cost allows for extremely simple dictionary elements (diracs on the fundamental frequency).
- We take $\mathbf{D}$ as diracs on the fundamental frequencies of the notes.
- Closed form for solving the OT problem.
- Non-convex Group lasso for sparse estimates and/or entropic regularization.

OST in action

Simulated data

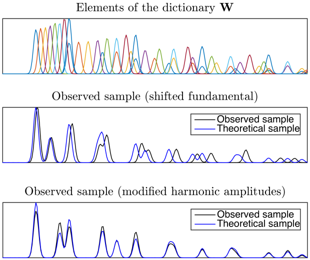
- Robust to shifted fundamental frequency.
- Robust to harmonics magnitude variability.
- Very fast ($\sim$ms per frame).

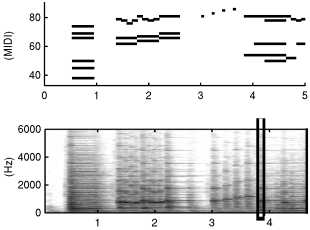
MAPS Dataset [Emiya et al., 2010]

- Several piano sequence from classical music ($m = 60$ notes)
- Comparison with ground truth given as MIDI.
- OST similar of better than KL+Dico while ≥ 70 times quicker.

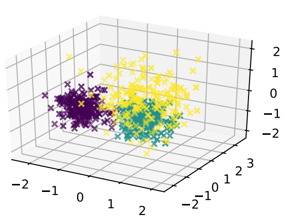
Real time demonstration

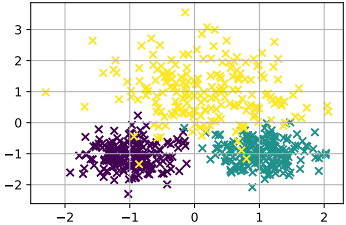
- Python+Pygame implementation.
- Demo url: `https://github.com/rflamary/OST`





Wasserstein Discriminant Analysis (WDA)



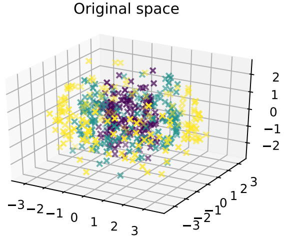


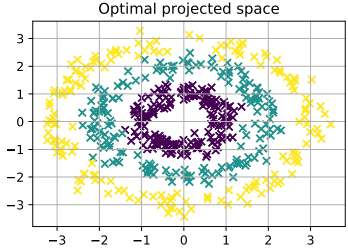
$$\max_{\mathbf{P}\in\Delta} \quad \frac{\sum_{c,c'>c} W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^{c'})}{\sum_c W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^c)} \tag{6}$$

- $\mathbf{X}^c$ are samples from class c.
- $\mathbf{P}$ is an orthogonal projection;

- Converges to Fisher Discriminant when $\lambda \to \infty$.
- Non parametric method that allows nonlinear discrimination.
- Problem solved with gradient ascent in the Stiefel manifold.
- Gradient computed using automatic differentiation of Sinkhorn algorithm.

Wasserstein Discriminant Analysis (WDA)





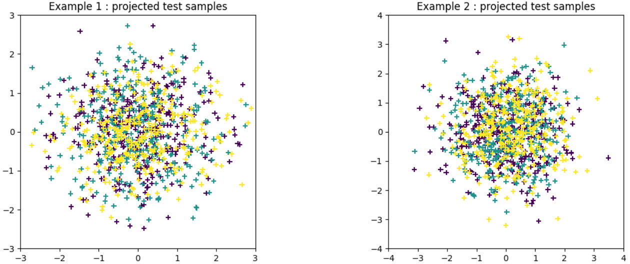
$$\max_{\mathbf{P}\in\Delta} \quad \frac{\sum_{c,c'>c} W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^{c'})}{\sum_c W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^c)} \quad (6)$$

- $\mathbf{X}^c$ are samples from class c.
- $\mathbf{P}$ is an orthogonal projection;

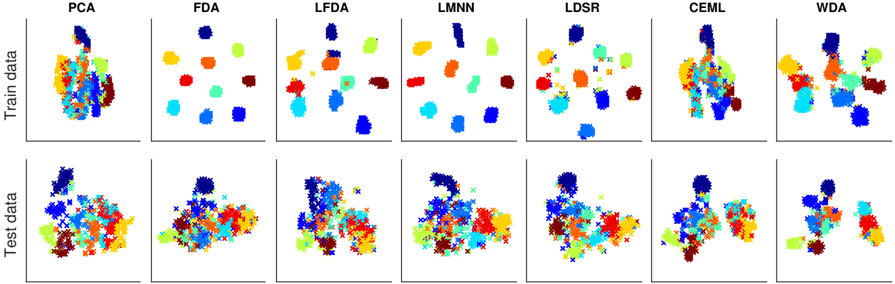
- Converges to Fisher Discriminant when $\lambda \to \infty$.
- Non parametric method that allows nonlinear discrimination.
- Problem solved with gradient ascent in the Stiefel manifold.
- Gradient computed using automatic differentiation of Sinkhorn algorithm.

WDA in action

Simulated datasets : 10→2



MNIST Dataset: 784→10(→2 TSNE)



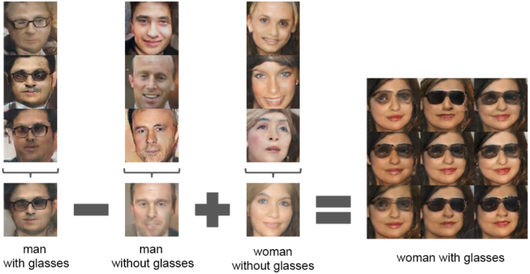
Generative Adversarial Networks (GAN)

Generative Adversarial Networks (GAN) [Goodfellow et al., 2014]

$$\min_G \max_D \quad E_{\mathbf{x}\sim\mu_d}[\log D(\mathbf{x})] + E_{\mathbf{z}\sim\mathcal{N}(0,\mathbf{I})}[\log(1 - D(G(\mathbf{z})))]$$

- Learn a generative model G that outputs realistic samples from data μ_d.
- Learn a classifier D to discriminate between the generated and true samples.
- Make those models compete (Nash equilibrium [Zhao et al., 2016]).
- Generator space has semantic meaning [Radford et al., 2015].
- But extremely hard to train (vanishing gradients).

Generative Adversarial Networks (GAN)

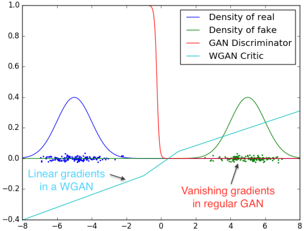


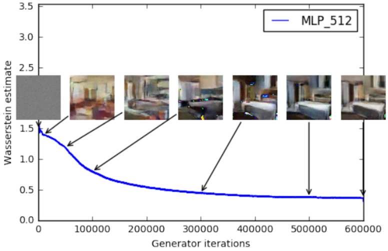
Generative Adversarial Networks (GAN) [Goodfellow et al., 2014]

$$\min_G \max_D \quad E_{\mathbf{x}\sim\mu_d}[\log D(\mathbf{x})] + E_{\mathbf{z}\sim\mathcal{N}(0,\mathbf{I})}[\log(1 - D(G(\mathbf{z})))]$$

- Learn a generative model G that outputs realistic samples from data μ_d.
- Learn a classifier D to discriminate between the generated and true samples.
- Make those models compete (Nash equilibrium [Zhao et al., 2016]).
- Generator space has semantic meaning [Radford et al., 2015].
- But extremely hard to train (vanishing gradients).

Wasserstein Generative Adversarial Networks





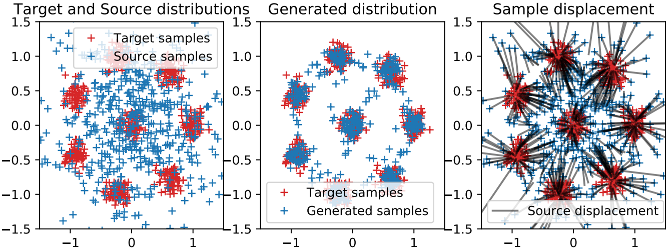
Wasserstein GAN [Arjovsky et al., 2017]

$$\min_{G} \quad W_1^1(G(\mathbf{z}), \mu_d), \quad \text{s.t. } \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}) \tag{7}$$

- Minimize the Wasserstein distance between the data and the generated data.
- Wasserstein approximated in the dual (separable w.r.t. the samples).
- Parametrization of the dual variable D with a neural network.
- Lipschitz constraints in the dual (constrained parameters).
- No vanishing gradients ! Far better convergence in practice.

Section

Mapping with optimal transport



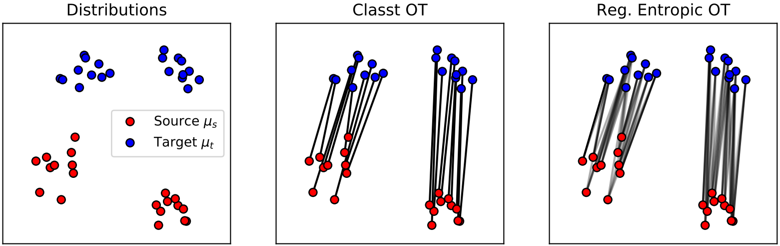
Mapping estimation

- Mapping do not exist in general between empirical distributions.
- Barycentric mapping [Ferradans et al., 2014].
- Smooth mapping estimation [Perrot et al., 2016, Seguy et al., 2017].

Why map ?

- Sensible displacement to align distributions.
- Color adaptation in image [Ferradans et al., 2014].
- Domain adaptation and transfer learning [Courty et al., 2016b].

Transporting the discrete samples

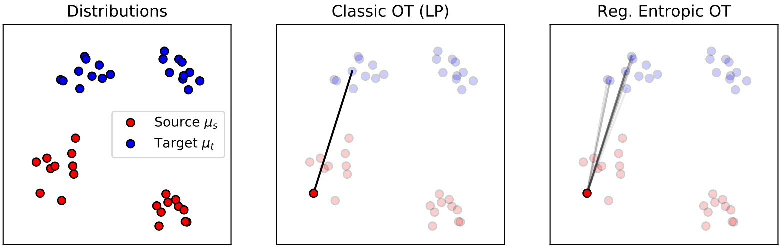


Barycentric mapping [Ferradans et al., 2014]

$$\widehat{T}_{\gamma_0}(\mathbf{x}_i^s) = \arg\min_{\mathbf{x}} \sum_j \gamma_0(i,j) c(\mathbf{x}, \mathbf{x}_j^t). \tag{8}$$

- The mass of each source sample is spread onto the target samples (line of γ_0).
- The mapping is the barycenter of the target samples weighted by γ_0
- Closed form solution for the quadratic loss.
- Limited to the samples in the distribution (no out of sample).

Transporting the discrete samples

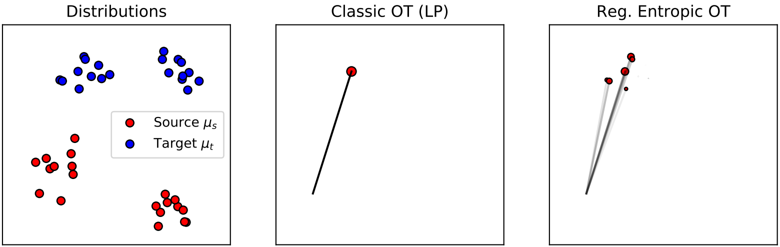


Barycentric mapping [Ferradans et al., 2014]

$$\widehat{T}_{\gamma_0}(\mathbf{x}_i^s) = \arg\min_{\mathbf{x}} \sum_j \gamma_0(i,j) c(\mathbf{x}, \mathbf{x}_j^t). \tag{8}$$

- The mass of each source sample is spread onto the target samples (line of γ_0).
- The mapping is the barycenter of the target samples weighted by γ_0
- Closed form solution for the quadratic loss.
- Limited to the samples in the distribution (no out of sample).

Transporting the discrete samples

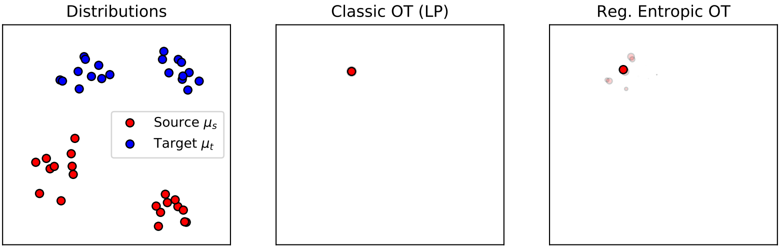


Barycentric mapping [Ferradans et al., 2014]

$$\widehat{T}_{\gamma_0}(\mathbf{x}_i^s) = \arg\min_{\mathbf{x}} \quad \sum_j \gamma_0(i,j) c(\mathbf{x}, \mathbf{x}_j^t). \tag{8}$$

- The mass of each source sample is spread onto the target samples (line of γ_0).
- The mapping is the barycenter of the target samples weighted by γ_0
- Closed form solution for the quadratic loss.
- Limited to the samples in the distribution (no out of sample).

Transporting the discrete samples

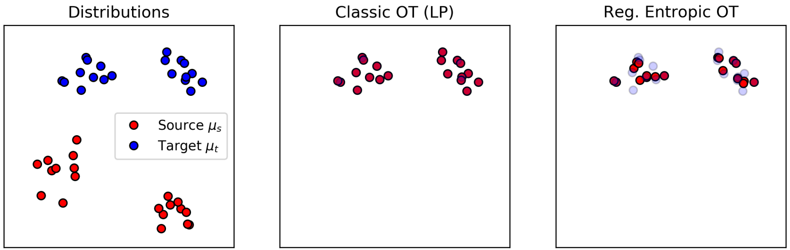


Barycentric mapping [Ferradans et al., 2014]

$$\widehat{T}_{\gamma_0}(\mathbf{x}_i^s) = \arg\min_{\mathbf{x}} \sum_j \gamma_0(i,j) c(\mathbf{x}, \mathbf{x}_j^t). \tag{8}$$

- The mass of each source sample is spread onto the target samples (line of γ_0).
- The mapping is the barycenter of the target samples weighted by γ_0
- Closed form solution for the quadratic loss.
- Limited to the samples in the distribution (no out of sample).

Transporting the discrete samples



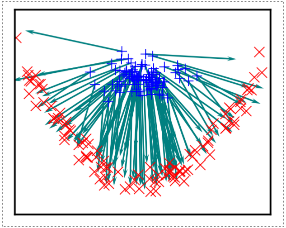
Barycentric mapping [Ferradans et al., 2014]

$$\widehat{T}_{\gamma_0}(\mathbf{x}_i^s) = \arg\min_{\mathbf{x}} \sum_j \gamma_0(i,j)c(\mathbf{x}, \mathbf{x}_j^t). \tag{8}$$

- The mass of each source sample is spread onto the target samples (line of γ_0).
- The mapping is the barycenter of the target samples weighted by γ_0
- Closed form solution for the quadratic loss.
- Limited to the samples in the distribution (no out of sample).

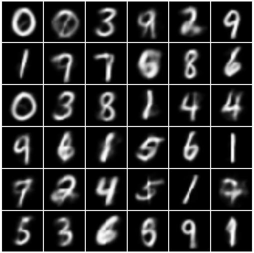
Optimal transport mapping estimation

Joint OT and mapping estimation [Perrot et al., 2016]



- Estimate jointly the OT matrix and a smooth mapping approximating the barycentric mapping.
- The mapping is a regularization for OT.
- Controlled generalization error.
- Linear and kernel mappings limited to small scale datasets.

2-step mapping estimation [Seguy et al., 2017]

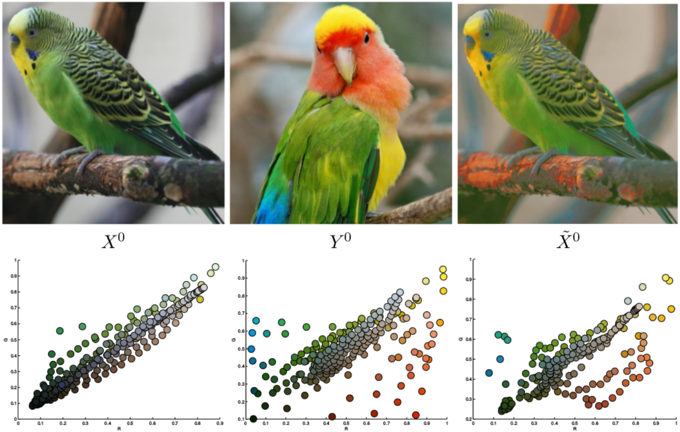


1. Estimate regularized OT in the dual.
2. Estimate a smooth version of the barycentric mapping with a neural network.

- Stochastic Gradient Descent on the OT dual.
- Convergence to the true OT and mapping for small regularization.

Histogram matching in images

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



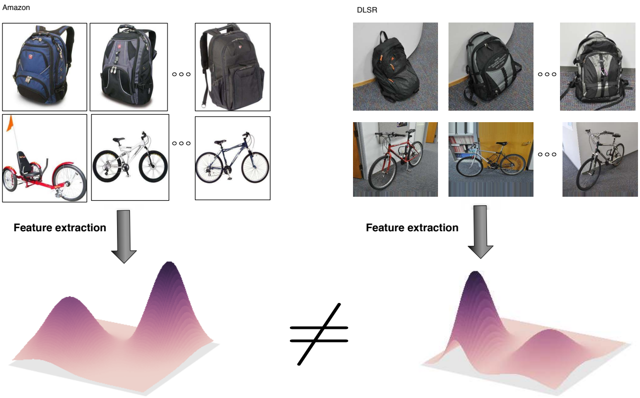
X^0 $\quad$ Y^0 $\quad$ $\tilde{X}^0$

μ_{X^0} $\quad$ μ_{Y^0} $\quad$ $\mu_{\tilde{X}^0}$

Histogram matching in images

Image colorization [Ferradans et al., 2014]

Domain Adaptation problem

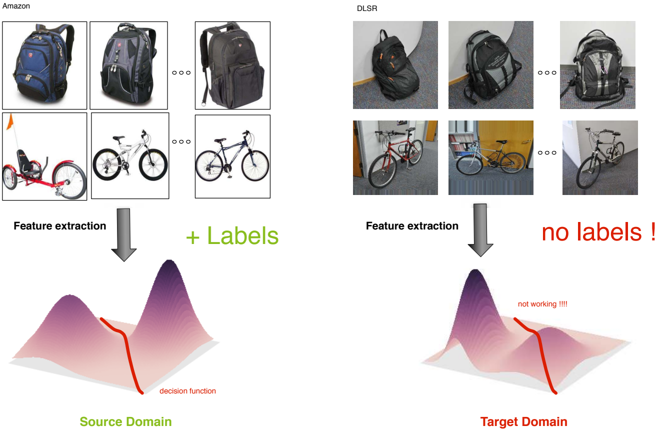


Probability Distribution Functions over the domains

Our context

- Classification problem with data coming from different sources (domains).
- Distributions are different but related.

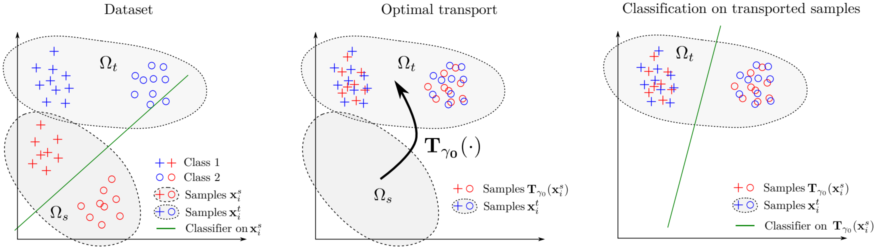
Unsupervised domain adaptation problem



Problems

- Labels only available in the **source domain**, and classification is conducted in the **target domain**.
- Classifier trained on the source domain data performs badly in the target domain

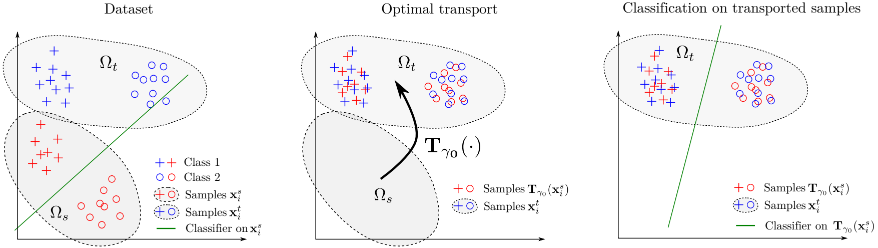
OT for domain adaptation : Step 1



Step 1 : Estimate optimal transport between distributions.

- Choose the ground metric (squared euclidean in our experiments).
- Using regularization allows
  - Large scale and regular OT with entropic regularization [Cuturi, 2013].
  - Class labels in the transport with group lasso [Courty et al., 2016b].
- Efficient optimization based on Bregman projections [Benamou et al., 2015] and
  - Majoration minimization for non-convex group lasso.
  - Generalized Conditionnal gradient for general regularization (cvx. lasso, Laplacian).

OT for domain adaptation : Steps 2 & 3



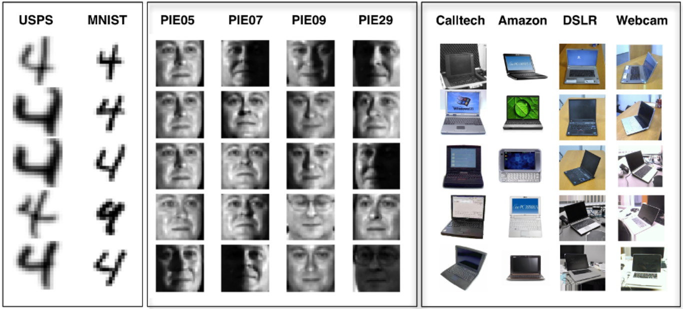
Step 2 : Transport the training samples onto the target distribution.

- The mass of each source sample is spread onto the target samples (line of γ_0).
- Transport using barycentric mapping [Ferradans et al., 2014].
- The mapping can be estimated for out of sample prediction [Perrot et al., 2016, Seguy et al., 2017].

Step 3 : Learn a classifier on the transported training samples

- Transported sample keep their labels.
- Classic ML problem when samples are well transported.

Visual adaptation datasets



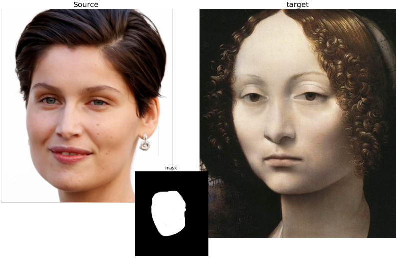
Datasets

- **Digit recognition**, MNIST VS USPS (10 classes, d=256, 2 dom.).
- **Face recognition**, PIE Dataset (68 classes, d=1024, 4 dom.).
- **Object recognition**, Caltech-Office dataset (10 classes, d=800/4096, 4 dom.).

Numerical experiments

- Comparison with state of the art on the 3 datasets.
- OT works very well on digits and object recognition.
- Works well on deep features adaptation and extension to semi-supervised DA.

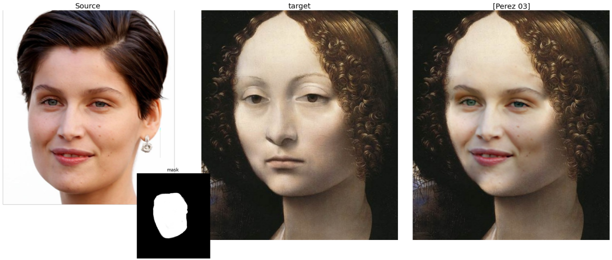
Seamless copy in images



Poisson image editing [Pérez et al., 2003]

- Use the color gradient from the source image.
- Use color border conditions on the target image.
- Solve Poisson equation to reconstruct the new image.

Seamless copy in images



Poisson image editing [Pérez et al., 2003]

- Use the color gradient from the source image.
- Use color border conditions on the target image.
- Solve Poisson equation to reconstruct the new image.

Seamless copy with gradient adaptation [Perrot et al., 2016]

- Transport the gradient from the source to target color gradient distribution.
- Solve the Poisson equation with the mapped source gradients.
- Better respect of the color dynamic and limits false colors.

Seamless copy in images

Poisson image editing [Pérez et al., 2003]

- Use the color gradient from the source image.
- Use color border conditions on the target image.
- Solve Poisson equation to reconstruct the new image.

Seamless copy with gradient adaptation [Perrot et al., 2016]

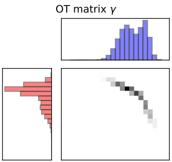
- Transport the gradient from the source to target color gradient distribution.
- Solve the Poisson equation with the mapped source gradients.
- Better respect of the color dynamic and limits false colors.

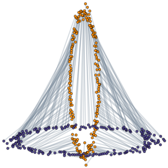
Seamless copy with gradient adaptation

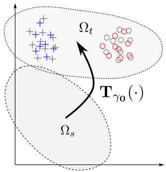
Example and webcam demo: `https://github.com/ncourty/PoissonGradient`

Section

Optimal transport for machine learning







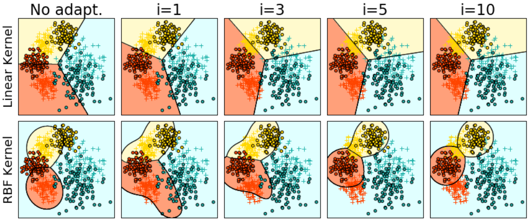
Learning with optimal transport

- Natural divergence for machine learning and estimation.
- Cost encode complex relations in an histogram.
- Regularization is the key (performance, smoothness).
- Recent optimization procedures opened it to medium/large scale datasets.
- Sensible loss between non overlapping distributions.
- Works with both histograms and empirical distributions.

Mapping with optimal transport

- Optimal displacement from one distribution to another.
- Can estimate smooth mapping for out of sample displacement.
- Domain, color and gradient adaptation, transfer learning.

Optimal transport for machine learning



Current and future works

- Joint distribution domain adaptation OT [Courty et al., 2017b].
- Large scale OT and mapping estimation (SGD) [Seguy et al., 2017].
- Approximate Wasserstein embedding for fast data mining [Courty et al., 2017a].

Open questions

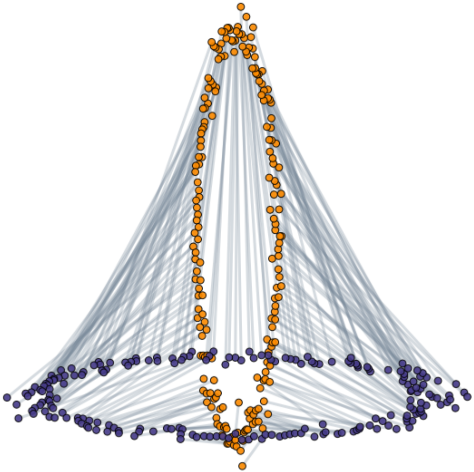
- Generalization bounds for learning with OT.
- Learning the ground metric (supervised, unsupervised).
- Large scale OT and mapping estimation, accelerated stochastic optimization.

Thank you

Python code available on GitHub:
`https://github.com/rflamary/POT`

- OT LP solver, Sinkhorn (stabilized, $\epsilon-$scaling, GPU)
- Domain adaptation with OT.
- Barycenters, Wasserstein unmixing.
- Wasserstein Discriminant Analysis.

Papers available on my website:
`https://remi.flamary.com/`



References I

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

Arjovsky, M., Chintala, S., and Bottou, L. (2017).
Wasserstein gan.
arXiv preprint arXiv:1701.07875.

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.

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

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

References II

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017b).
Joint distribution optimal transportation for domain adaptation.
In *Neural Information Processing Systems (NIPS)*.

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

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

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

Cuturi, M. and Peyré, G. (2016).
A smoothed dual approach for variational wasserstein problems.
SIAM Journal on Imaging Sciences, 9(1):320–343.

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

References III

Emiya, V., Badeau, R., and David, B. (2010).
Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle.
IEEE Transactions on Audio, Speech, and Language Processing, 18(6):1643–1654.

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

Flamary, R., Cuturi, M., Courty, N., and Rakotomamonjy, A. (2016a).
Wasserstein discriminant analysis.
arXiv preprint arXiv:1608.08063.

Flamary, R., Fevotte, C., Courty, N., and Emyia, V. (2016b).
Optimal spectral transportation with application to music transcription.
In *Neural Information Processing Systems (NIPS)*.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. (2015).
Learning with a wasserstein loss.
In *Advances in Neural Information Processing Systems*, pages 2053–2061.

References IV

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014).
Generative adversarial nets.
In *Advances in neural information processing systems*, pages 2672–2680.

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

McCann, R. J. (1997).
A convexity principle for interacting gases.
Advances in mathematics, 128(1):153–179.

Monge, G. (1781).
Mémoire sur la théorie des déblais et des remblais.
De l'Imprimerie Royale.

References V

Pérez, P., Gangnet, M., and Blake, A. (2003).
Poisson image editing.
ACM Trans. on Graphics, 22(3).

Perrot, M., Courty, N., Flamary, R., and Habrard, A. (2016).
Mapping estimation for discrete optimal transport.
In *Neural Information Processing Systems (NIPS)*.

Radford, A., Metz, L., and Chintala, S. (2015).
Unsupervised representation learning with deep convolutional generative adversarial networks.
arXiv preprint arXiv:1511.06434.

Rolet, A., Cuturi, M., and Peyré, G. (2016).
Fast dictionary learning with a smoothed wasserstein loss.
In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 630–638.

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.

References VI

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.

References VII

Zen, G., Ricci, E., and Sebe, N. (2014).
Simultaneous ground metric learning and matrix factorization with earth mover's distance.
In *ICPR*, pages 3690–3695.

Zhao, J., Mathieu, M., and LeCun, Y. (2016).
Energy-based generative adversarial network.
arXiv preprint arXiv:1609.03126.

Collaborators

N. Courty

R. Flamary

D. Tuia

A. Rakotomamonjy

Barycenters

L2 Barycenter

L1 Barycenter

KL Barycenter

Wass. Barycenter