Optimal transport for machine learning

Learning with optimal transport

https://bit.ly/2VSjgWJ

Rémi Flamary, CMAP, École Polytechnique

July 19 2021

SIAM 2021, Mini-tutorial

Outline of the tutorial

Part 1 : Introduction to optimal transport

Monge, Kantorovitch and Wasserstein distance

Barycenters and geometry of optimal transport

Computational aspects of optimal transport and regularization

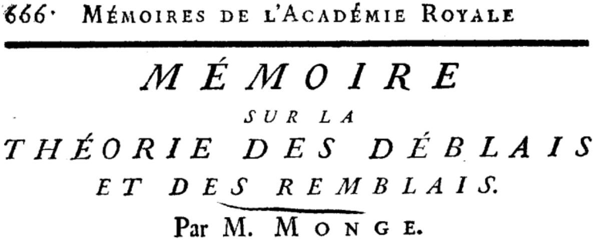
Part 2 : Optimal transport in machine learning applications

Mapping with optimal transport

Learning from histograms with OT

Learning from empirical distributions with Optimal Transport

The origins of optimal transport


666· MÉMOIRES DE L'ACADÉMIE ROYALE

MÉMOIRE

SUR LA

THÉORIE DES DÉBLAIS

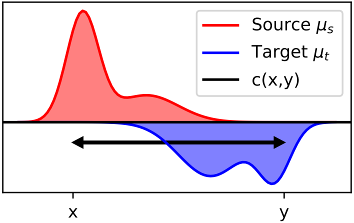
ET DES REMBLAIS.

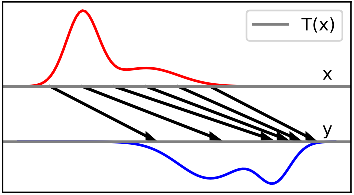
Par M. MONGE.

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

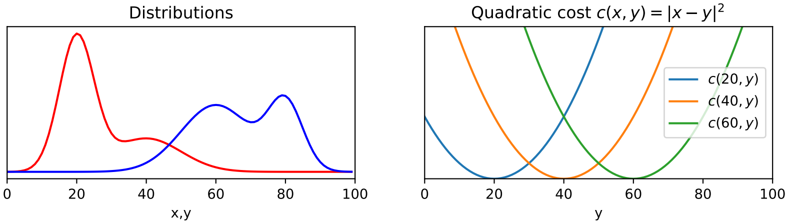




Problem [Monge, 1781]

- 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)

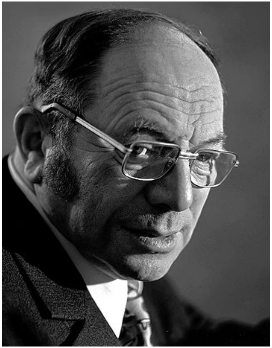


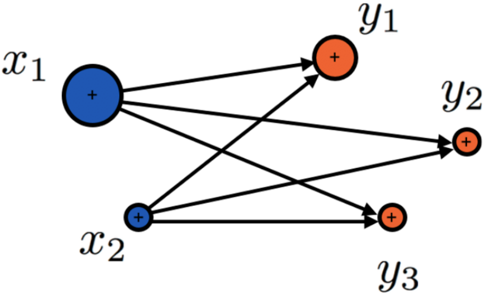
- Probability measures μ_s and μ_t on and a cost function $c:\Omega_s\times\Omega_t\rightarrow\mathbb{R}^+$.
- The Monge formulation [Monge, 1781] aim at finding a mapping $T:\Omega_s\rightarrow\Omega_t$

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

- Non convex problem because of the constraint $T\#\mu_s=\mu_t$.
- A mapping T might not exist especially for discrete distributions.

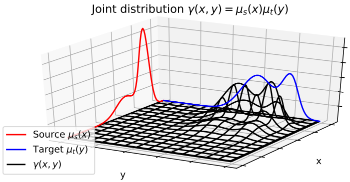
Kantorovich relaxation

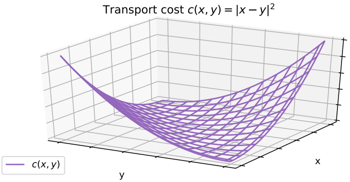




- Leonid Kantorovich (1912–1986), Economy nobelist in 1975
- Focus on where the mass goes, allow splitting [Kantorovich, 1942].
- Applications mainly for resource allocation problems

Optimal transport (Kantorovich formulation)





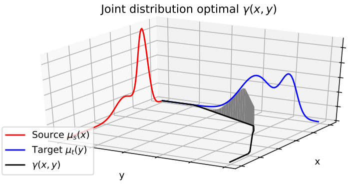
- The Kantorovich formulation [Kantorovich, 1942] seeks for a probabilistic coupling $\gamma \in \mathcal{P}(\Omega_s \times \Omega_t)$ between Ω_s and Ω_t:

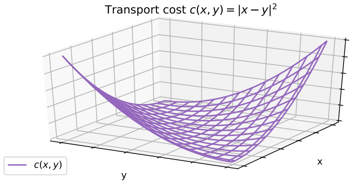
$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \int_{\Omega_s \times \Omega_t} c(\mathbf{x}, \mathbf{y}) \boldsymbol{\gamma}(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}, \tag{2}$$

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

- γ is a joint probability measure with marginals μ_s and μ_t.
- Linear Program that always has a solution.
- Relation with Monge problem proved for smooth distributions by [Brenier, 1991].

Optimal transport (Kantorovich formulation)





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

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

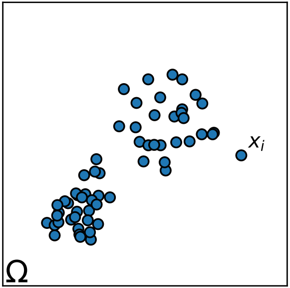
$$\text{s.t.} \quad \gamma \in \mathcal{P} = \left\{ \boldsymbol{\gamma} \geq \mathbf{0}, \int_{\boldsymbol{\Omega_t}} \boldsymbol{\gamma}(\mathbf{x}, \mathbf{y}) \mathbf{dy} = \boldsymbol{\mu_s}, \int_{\boldsymbol{\Omega_s}} \boldsymbol{\gamma}(\mathbf{x}, \mathbf{y}) \mathbf{dx} = \boldsymbol{\mu_t}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

- γ is a joint probability measure with marginals μ_s and μ_t.
- Linear Program that always has a solution.
- Relation with Monge problem proved for smooth distributions by [Brenier, 1991].

Discrete distributions: Empirical vs Histogram

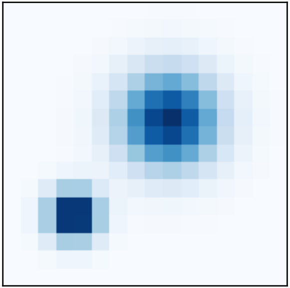
Discrete measure: $\mu = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}, \quad \mathbf{x}_i \in \Omega, \quad \sum_{i=1}^{n} a_i = 1$

Lagrangian (point clouds)



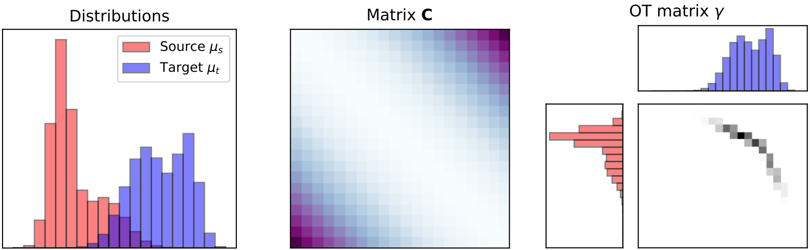
- Constant weight: $a_i = \frac{1}{n}$
- Quotient space: Ω^n, Σ_n

Eulerian (histograms)



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

Optimal transport with discrete distributions



OT Linear Program

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

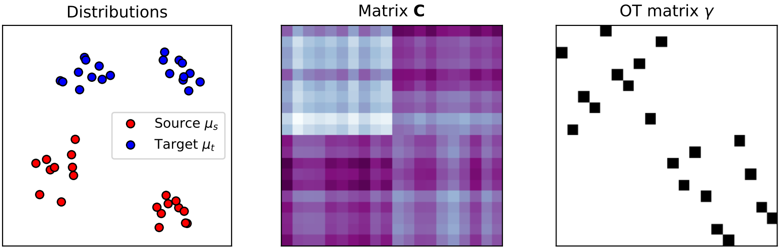
$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma}\in\mathcal{P}}{\operatorname{argmin}} \quad \left\{ \langle \boldsymbol{\gamma}, \mathbf{C}\rangle_F = \sum_{i,j} \gamma_{i,j} c_{i,j} \right\}$$

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

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

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 $\mu_t = \sum_{i=1}^{n} b_i \delta_{\mathbf{x}_i^t}$

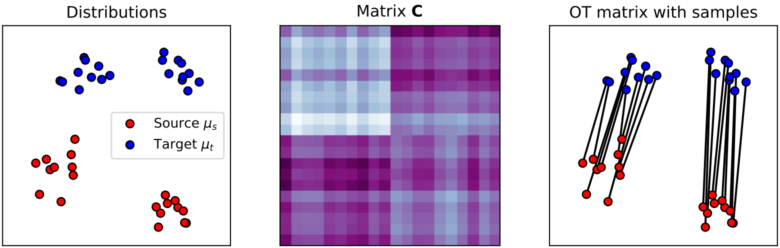
$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma}\in\mathcal{P}}{\text{argmin}} \quad \left\{ \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F = \sum_{i,j} \gamma_{i,j} c_{i,j} \right\}$$

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

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

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 $\mu_t = \sum_{i=1}^{n} b_i \delta_{\mathbf{x}_i^t}$

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

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

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

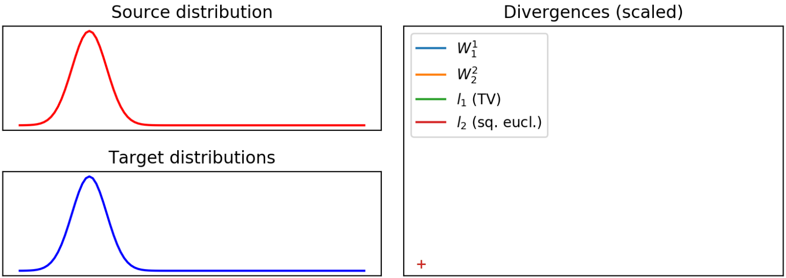
Linear program with $n_s n_t$ variables and $n_s + n_t$ constraints. Demo

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



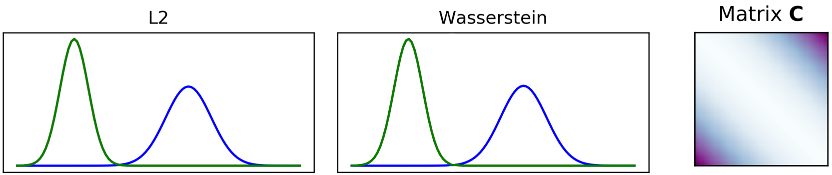
Wasserstein distance

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

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

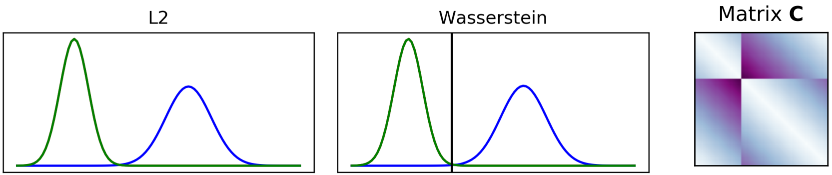


Barycenters [Agueh and Carlier, 2011]

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

- $\lambda_i > 0$ 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

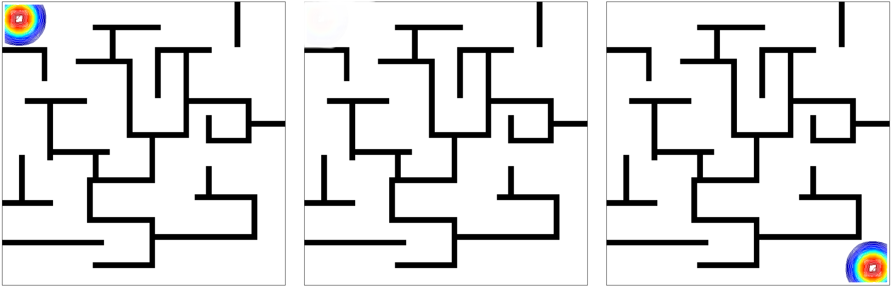


Barycenters [Agueh and Carlier, 2011]

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

- $\lambda_i > 0$ 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

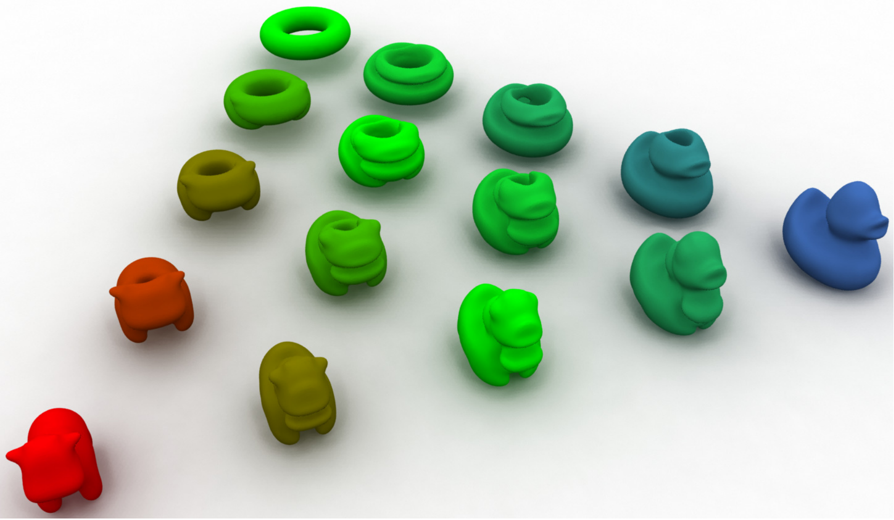


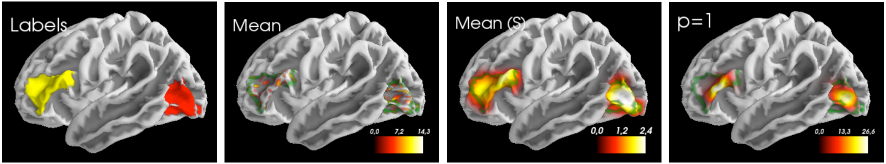
- The space of probability distribution equipped with the Wasserstein metric $(\mathcal{P}_p(X), W_2^2(X))$ defines a geodesic space with a Riemannian structure [Santambrogio, 2014].
- Geodesics are shortest curves on $\mathcal{P}_p(X)$ that link two distributions
- Cost between two pixels is the shortest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

Illustration from [Papadakis et al., 2014]

3D Wasserstein barycenter

Shape interpolation [Solomon et al., 2015]





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

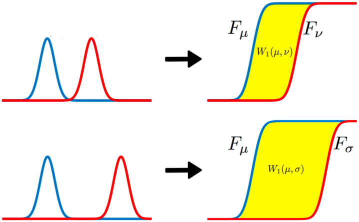
- Average fMRI activation maps on voxels or cortical surface (natural metric).
- Classical average across subjects and gaussian blur loose information.
- OT averaging recover central activation areas with better precision.
- Can encode both geometrical (3D position) or anatomical connectivity information.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

Special cases for OT

Solving OT in 1D

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

- $F_\mu^{-1}(q)$ is the quantile function of μ.
- Very fast $O(n\log(n))$ computation on discrete distributions.
- Used for sliced Radon Wasserstein in high dimension[Bonneel et al., 2015].

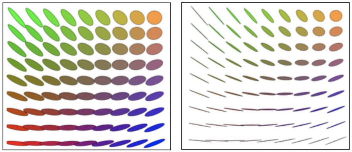


Solving OT between Gaussian distributions

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

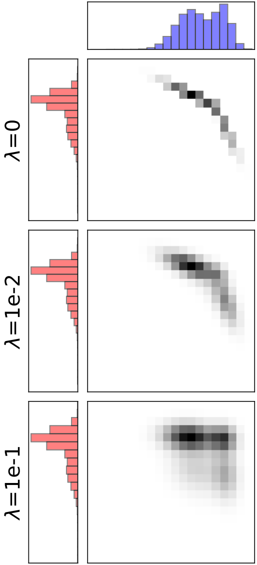
- When $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$, $\mu_s \sim \mathcal{N}(\mathbf{m}_1, \Sigma_1)$ and $\mu_t \sim \mathcal{N}(\mathbf{m}_2, \Sigma_2)$
- $\mathbb{B}(,)$ is the Bures metric:

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



Regularized optimal transport

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



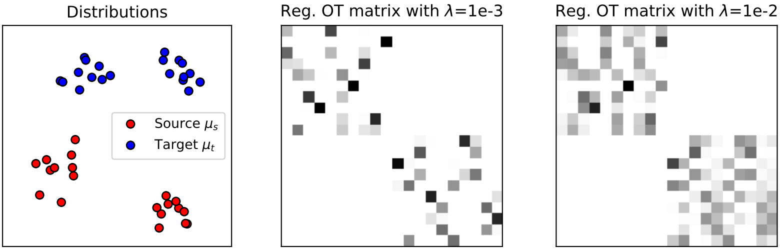
Regularization term $\Omega(\gamma)$

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

Why regularize?

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

Entropic regularized optimal transport



Entropic regularization [Cuturi, 2013]

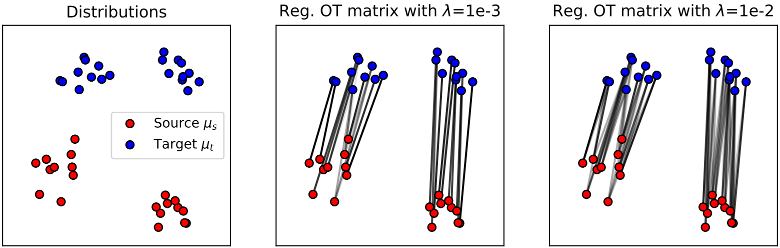
$$W_\lambda(\mu_s, \mu_t) = \min_{\gamma \in \mathcal{P}} \quad \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F + \lambda \sum_{i,j} \boldsymbol{\gamma}(i,j) \log \boldsymbol{\gamma}(i,j)$$

- Looses sparsity, but strictly convex optimization problem [Benamou et al., 2015].
- Sinkhorn divergence [Genevay et al., 2017, Genevay et al., 2018]:

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

- Loss and OT matrix are differentiable and have better statistical properties [Genevay et al., 2018].

Entropic regularized optimal transport



Entropic regularization [Cuturi, 2013]

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

- Looses sparsity, but strictly convex optimization problem [Benamou et al., 2015].
- Sinkhorn divergence [Genevay et al., 2017, Genevay et al., 2018]:

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

- Loss and OT matrix are differentiable and have better statistical properties [Genevay et al., 2018].

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

- 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].
- Speedup for kernel and OT matrix approximations [Altschuler et al., 2018, Scetbon and Cuturi, 2020].

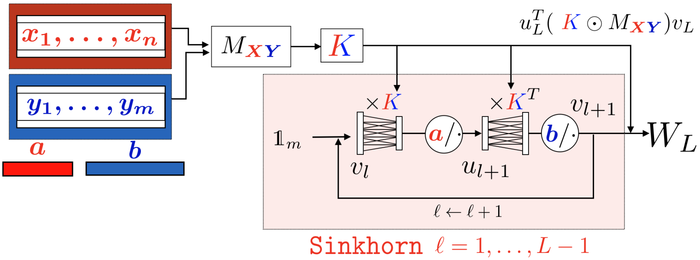


Image from Marco Cuturi

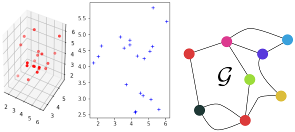
Sinkhorn Autodiff [Genevay et al., 2017]

- 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 (`logsumexp`).

Extensions for all you OT needs

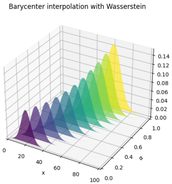
OT between different metric spaces

- Gromov-Wasserstein [Memoli, 2011, Peyré et al., 2016] and Co-OT [Redko et al., 2020].
- Can be used on graph data [Peyré et al., 2016, Vayer et al., 2018].



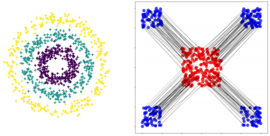
OT between unbalanced distributions

- Partial OT [Figalli, 2010] and Unbalanced OT [Chizat et al., 2018].
- Principle of UOT : relax the marginal constraints and penalize their violation.



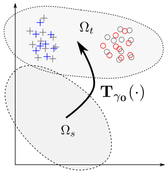
Learning the OT metric c

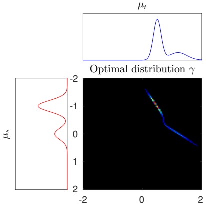
- Adversarially [Genevay et al., 2017].
- Discriminant metric [Flamary et al., 2018]
- Robust metric [Paty and Cuturi, 2019]).

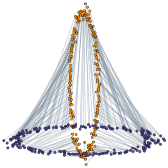


Outline

Three aspects of optimal transport for ML







Transporting with optimal transport

- Color adaptation in image [Ferradans et al., 2014].
- Style transfer [Mroueh, 2019].
- Domain adaptation [Courty et al., 2016b].

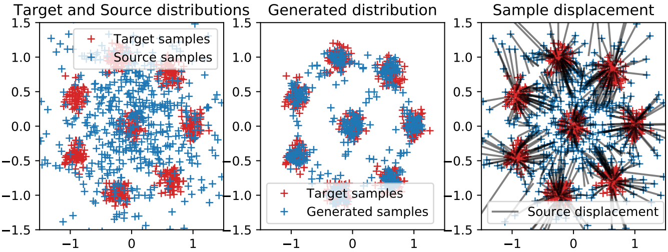
Divergence between histograms

- Use the ground metric to encode complex relations between the bins.
- Loss for multilabel classifier [Frogner et al., 2015]
- Adversarial regularization [Fatras et al., 2021a].

Divergence between empirical distributions

- Non parametric divergence between non overlapping distributions.
- Generative modeling [Arjovsky et al., 2017].
- Data imputation [Muzellec et al., 2020].

Mapping with optimal transport

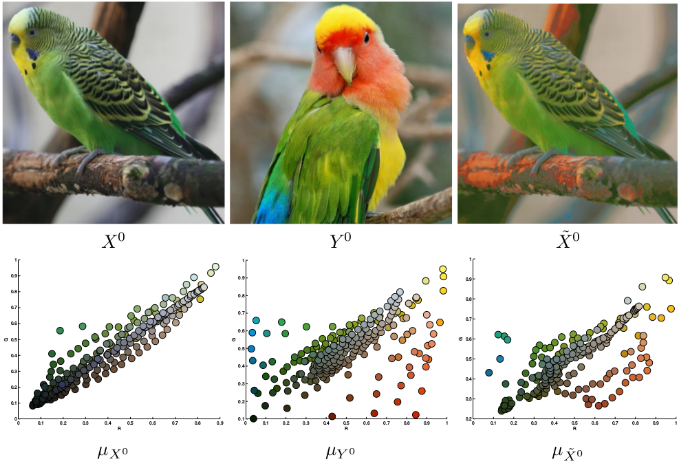


Mapping estimation

- Barycentric mapping using the OT matrix [Ferradans et al., 2014].
- Linear Monge mapping when data supposed Gaussian [Flamary et al., 2019].
- Smooth mapping estimation
  [Perrot et al., 2016, Seguy et al., 2017, Paty et al., 2020].
- Estimation for W_2 using input convex neural networks [Makkuva et al., 2020].
- Can be used to linearize the Wasserstein space [Mérigot et al., 2020]

Histogram matching in images

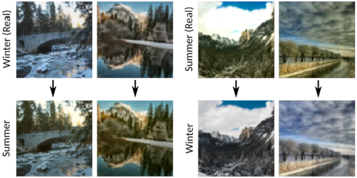
Pixels as empirical distribution [Ferradans et al., 2014]



Histogram matching in images

Image colorization [Ferradans et al., 2014]

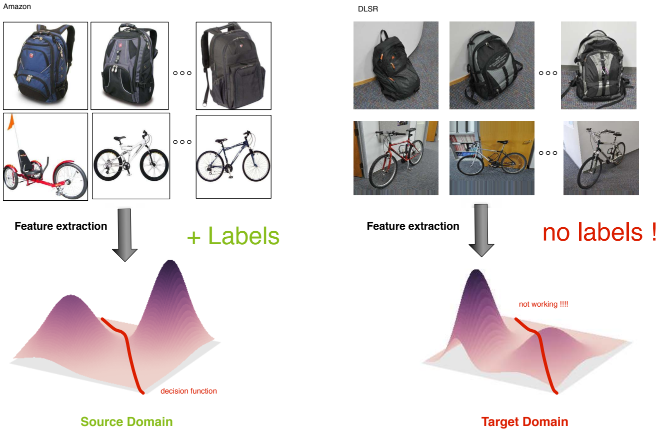
Monge mapping for Image-to-Image translation



Principle

- Encode image as a distribution in a DNN embedding.
- Transform between images using estimated Monge mapping.
- Linear Monge Mapping (Wasserstein Style Transfer [Mroueh, 2019]).
- Nonlinear Monge Mapping using input Convex Neural Networks [Korotin et al., 2019].
- Allows for transformation between two images but also style interpolation with Wasserstein barycenters.

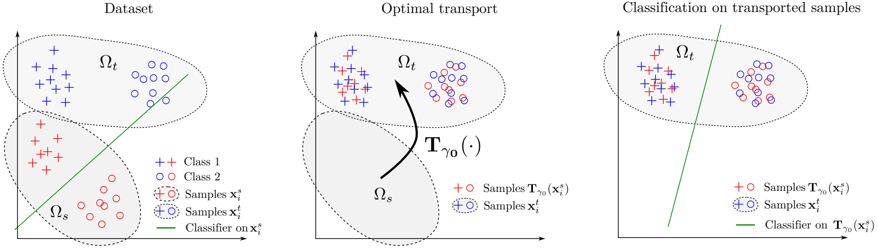
Unsupervised domain adaptation problem



Problems

- Classification problem with data coming from different sources (domains).
- Labels available in the **source domain**, and classification done in the **target domain**.
- Classifier trained on the source domain data performs badly in the target domain

Optimal Transport for Domain Adaptation



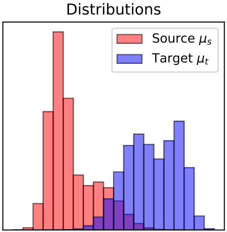
Mapping the samples [Courty et al., 2016b, Flamary et al., 2019]

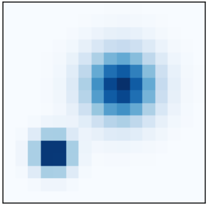
1. Estimate optimal transport between distributions.
2. Transport the training samples on target domain.
3. Learn a classifier on the transported training samples.

Mapping the labels

- Label Propagation using OT matrix [Solomon et al., 2014, Redko et al., 2019].
- Optimize the target classifier [Courty et al., 2017b, Damodaran et al., 2018].
- Change in proportion of classes (target shift) [Redko et al., 2019, Rakotomamonjy et al., 2020, Fatras et al., 2021b].

Learning from histograms

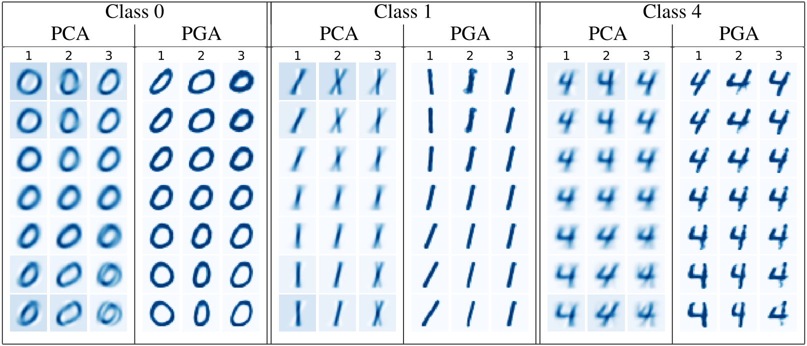




Data as histograms

- Fixed bin positions $\mathbf{x}_i$ e.g. grid, simplex $\Delta = \left\{(\mu_i)_i \geq 0; \sum_i \mu_i = 1\right\}$
- A lot of datasets comes under the form of histograms.
- Images are photon counts (black and white), text as word counts.
- Output of a softmax in a neural network is a histogram.
- Natural divergence is Kullback–Leibler.
- Wasserstein distance can encode complex relationship between the bins in the histograms (through ground metric $\mathbf{C}$).

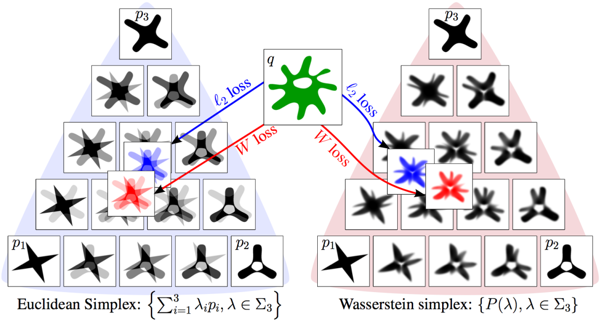
Principal Geodesics Analysis



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

- Generalization of Principal Component Analysis to the Wasserstein manifold.
- Regularized OT [Seguy and Cuturi, 2015].
- Approximation using Wasserstein embedding [Courty et al., 2017a].

Dictionary Learning with OT

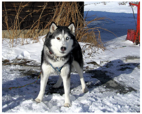


DL with OT [Sandler and Lindenbaum, 2011, Schmitz et al., 2017]

$$\min_{\mathbf{D},\mathbf{H}} \sum_i W_{\mathbf{C}}(\mathbf{v}_i, \mathbf{D}\mathbf{h}_i) \qquad \text{or} \qquad \min_{\mathbf{D},\mathbf{H}} \sum_i W_{\mathbf{C}}(\mathbf{v}_i, WB(\mathbf{D}, \mathbf{h}_i))$$

- NMF: columns of $\mathbf{D}$ and $\mathbf{H}$ are on the simplex, WB is Wasserstein barycenter $WB(\mathbf{D}, \mathbf{h}) = \operatorname{argmin}_{\mathbf{a}} \sum_i h_i W_{\mathbf{C}}(\mathbf{d_i}, \mathbf{a})$.
- Ground metric learning [Zen et al., 2014], regularized OT [Rolet et al., 2016].
- Unmixing when D is fixed [Nakhostin et al., 2016, Flamary et al., 2016].

Multi-label learning with Wasserstein Loss



Siberian husky

Eskimo dog

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

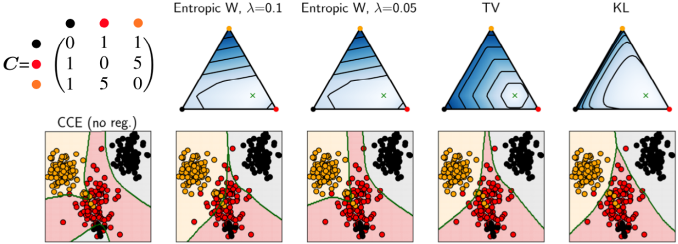
Flickr : water, boat, ref ection, sun-shine
Prediction : water, river, lake, summer;

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.

Wasserstein Adversarial Regularization



Principle [Fatras et al., 2021a]

$$R_{\mathbf{C}}(f, \mathbf{x}) = \max_{\|\mathbf{r}\| \leq \epsilon} W_{\mathbf{C}}(f(\mathbf{x} + \mathbf{r}), f(\mathbf{x}))$$

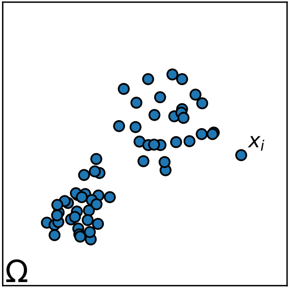
- Use (virtual) adversarial examples to promote a better generalization of DNN (close samples should have close predictions) [Miyato et al., 2018].
- The ground metric $\mathbf{C}$ encodes pairwise class relations and will promote smooth/complex between them.
- State of the art performance for learning with label noise when using semantic relations between the classes for $\mathbf{C}$ (word2vec).

Empirical distributions A.K.A datasets

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

Empirical distribution

- Two realizations never overlap.
- Training base of all machine learning approaches.
- How to measure discrepancy?
- Maximum Mean Discrepancy (ℓ_2 after convolution).
- Wasserstein distance.



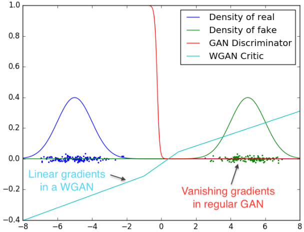
Generative Adversarial Networks (GAN)

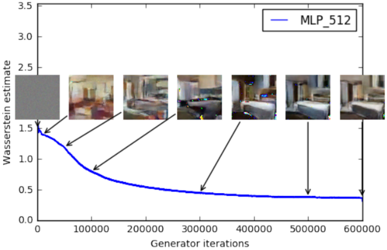
Generative Adversarial Networks (GAN) [Goodfellow et al., 2014]

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

- 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 (WGAN)





Wasserstein GAN [Arjovsky et al., 2017]

$$\min_G \quad W_1^1(G\#\mu_z, \mu_d), \tag{5}$$

- Minimizes the Wasserstein distance between the data μ_d and the generated data $G\#\mu_z$ when $\mu_z = \mathcal{N}(0, \mathbf{I})$.
- Wasserstein in the dual (separable w.r.t. the samples).

$$\min_G \sup_{\phi \in \mathrm{Lip}^1} \quad \mathbb{E}_{\mathbf{x}\sim\mu_d}[\phi(\mathbf{x})] - \mathbb{E}_{\mathbf{z}\sim\mu_z}[\phi(G(\mathbf{z}))]$$

- ϕ is a neural network that acts as an *actor critic*.
- Lipschitz constant forced or penalized in the loss [Gulrajani et al., 2017].

Wasserstein GAN loss on Biomedical images

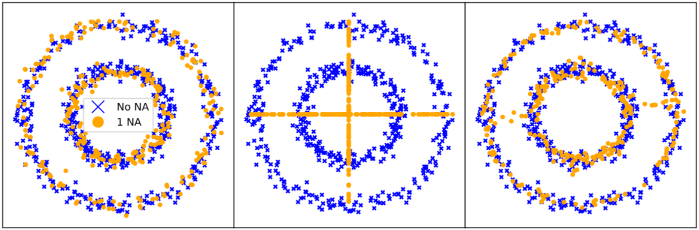
Full dose | Quarter dose | WGAN-VGG rec.

Reconstructing from low dose CT images [Yang et al., 2018]

$$\min_{G} \quad W_1^1(G\#\mu_l, \mu_f) + \lambda_1 E_{\mathbf{x}\sim\mu_l}[\|VGG(\mathbf{x}_l) - VGG(G(\mathbf{x}_l))\|^2], \tag{6}$$

- Use Wasserstein to make reconstruction of quarter dose CT images (μ_l) similar to high dose (resolution) CT images (μ_f).
- Perceptual loss based on VGG [Simonyan and Zisserman, 2014] embedding to keep image information.

Data imputation with Optimal Transport

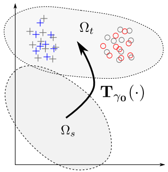


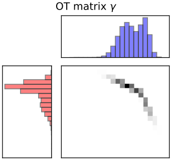
Missing Data imputation [Muzellec et al., 2020]

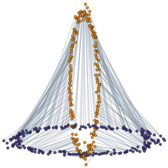
$$\min_{\mathbf{X}^{imp}} \quad \mathbb{E}[SD(\mu_m(\hat{\mathbf{X}}), \mu_m(\hat{\mathbf{X}}))]$$

- $\mathbf{X} \odot \mathbf{M}$ is the partially observed data with binary mask $\mathbf{M}$.
- $\hat{\mathbf{X}} = \mathbf{X} \odot \mathbf{M} + (1 - \mathbf{M}) \odot \mathbf{X}^{imp}$ is the data imputed by $\mathbf{X}^{imp}$
- $\mu_m(\mathbf{X})$ is a minibatch of $\mathbf{X}$, expectation is taken *w.r.t.* the minibatches.
- Out of sample imputation with model [Muzellec et al., 2020, Algo 2 & 3]
- Optimizing minibatch Wasserstein is a classical approach [Fatras et al., 2020].

Optimal transport for machine learning







Computational optimal transport

- Large linear program between discrete distributions.
- Entropic regularization for speedup and smoothness.
- Stochastic optimization deep learning.

Mapping with optimal transport

- Optimal displacement from one distribution to another.
- Can estimate smooth mapping.
- Domain, Image color and style adaptation.

Learning with optimal transport

- Natural divergence for machine learning and estimation.
- Cost encode complex relations in an histogram.
- Sensible loss between non overlapping distributions.
- Works with both histograms and empirical distributions.

Thank you

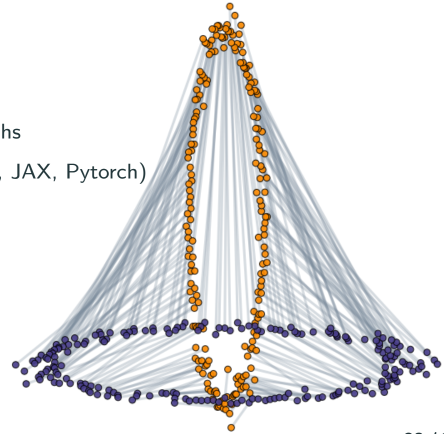
Python Optimal Transport (POT) [Flamary et al., 2021] :

`https://github.com/PythonOT/POT`

- OT LP solver, Sinkhorn (stabilized, $\epsilon-$scaling, GPU)
- Domain adaptation with OT.
- Barycenters, Wasserstein unmixing.
- Wasserstein Discriminant Analysis.
- Gromov-Wasserstein and variants for graphs
- New ! Different backend support (Numpy, JAX, Pytorch)

Slides and papers available on my website:
`https://remi.flamary.com/`

Post doc available in Paris/Saclay (France)



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.

[Altschuler et al., 2018] Altschuler, J., Bach, F., Rudi, A., and Niles-Weed, J. (2018).
Massively scalable sinkhorn distances via the nystr\" om method.
arXiv preprint arXiv:1812.05189.

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

[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

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

[Bonneel 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.

[Chizat et al., 2018] Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. (2018).
Scaling algorithms for unbalanced optimal transport problems.
Mathematics of Computation, 87(314):2563–2609.

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

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.

[Courty et al., 2017b] Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2017b).
Optimal transport for domain adaptation.
IEEETPAMI, 39(9):1853–1865.

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

[Damodaran et al., 2018] Damodaran, B. B., Kellenberger, B., Flamary, R., Tuia, D., and Courty, N. (2018).
Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation.

References iv

[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.

[Fatras et al., 2021a] Fatras, K., Bhushan Damodaran, B., Lobry, S., Flamary, R., Tuia, D., and Courty, N. (2021a).
Wasserstein adversarial regularization for learning with label noise.
Pattern Analysis and Machine Intelligence, IEEE Transactions on.

[Fatras et al., 2021b] Fatras, K., Séjourné, T., Courty, N., and Flamary, R. (2021b).
Unbalanced minibatch optimal transport; applications to domain adaptation.
In *International Conference on Machine Learning (ICML)*.

[Fatras et al., 2020] Fatras, K., Zine, Y., Flamary, R., Gribonval, R., and Courty, N. (2020).
Learning with minibatch wasserstein : asymptotic and gradient properties.
In *International Conference on Artificial Intelligence and Statistics (AISTAT)*.

References v

[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).

[Figalli, 2010] Figalli, A. (2010).
The optimal partial transport problem.
Archive for rational mechanics and analysis, 195(2):533–560.

[Flamary et al., 2021] Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021).
Pot: Python optimal transport.
Journal of Machine Learning Research, 22(78):1–8.

[Flamary et al., 2018] Flamary, R., Cuturi, M., Courty, N., and Rakotomamonjy, A. (2018).
Wasserstein discriminant analysis.
Machine learning, 107:1923–1945.

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

[Flamary et al., 2019] Flamary, R., Lounici, K., and Ferrari, A. (2019).
Concentration bounds for linear monge mapping estimation and optimal transport domain adaptation.
arXiv preprint arXiv:1905.10155.

[Frogner et al., 2015] 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.

[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., 2017] Genevay, A., Peyré, G., and Cuturi, M. (2017).
Sinkhorn-autodiff: Tractable wasserstein learning of generative models.
arXiv preprint arXiv:1706.00292.

[Goodfellow et al., 2014] 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.

[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.

[Gulrajani et al., 2017] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017).
Improved training of wasserstein gans.
In *Advances in Neural Information Processing Systems*, pages 5769–5779.

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

On the translocation of masses.

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

[Korotin et al., 2019] Korotin, A., Egiazarian, V., Asadulaev, A., Safin, A., and Burnaev, E. (2019).

Wasserstein-2 generative networks.

arXiv preprint arXiv:1909.13082.

[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.

[Makkuva et al., 2020] Makkuva, A., Taghvaei, A., Oh, S., and Lee, J. (2020).
Optimal transport mapping via input convex neural networks.
In *International Conference on Machine Learning*, pages 6672–6681. PMLR.

[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.

[Mérigot et al., 2020] Mérigot, Q., Delalande, A., and Chazal, F. (2020).
Quantitative stability of optimal transport maps and linearization of the 2-wasserstein space.
In *International Conference on Artificial Intelligence and Statistics*, pages 3186–3196. PMLR.

[Miyato et al., 2018] Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. (2018).
Virtual adversarial training: a regularization method for supervised and semi-supervised learning.
IEEE transactions on pattern analysis and machine intelligence, 41(8):1979–1993.

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

[Mroueh, 2019] Mroueh, Y. (2019).
Wasserstein style transfer.
arXiv preprint arXiv:1905.12828.

[Muzellec et al., 2020] Muzellec, B., Josse, J., Boyer, C., and Cuturi, M. (2020).
Missing data imputation using optimal transport.
In *International Conference on Machine Learning*, pages 7130–7140. PMLR.

[Nakhostin et al., 2016] Nakhostin, S., Courty, N., Flamary, R., Tuia, D., and Corpetti, T. (2016).

Supervised planetary unmixing with optimal transport.

In *Whorkshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*.

[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.

[Paty and Cuturi, 2019] Paty, F.-P. and Cuturi, M. (2019).

Subspace robust wasserstein distances.

arXiv preprint arXiv:1901.08949.

[Paty et al., 2020] Paty, F.-P., d'Aspremont, A., and Cuturi, M. (2020).

Regularity as regularization: Smooth and strongly convex brenier potentials in optimal transport.

In *International Conference on Artificial Intelligence and Statistics*, pages 1222–1232. PMLR.

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

[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.

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

[Rakotomamonjy et al., 2020] Rakotomamonjy, A., Flamary, R., Gasso, G., Alaya, M., Berar, M., and Courty, N. (2020).
Match and reweight strategy for generalized target shift.

[Redko et al., 2019] Redko, I., Courty, N., Flamary, R., and Tuia, D. (2019).
Optimal transport for multi-source domain adaptation under target shift.
In *International Conference on Artificial Intelligence and Statistics (AISTAT)*.

[Redko et al., 2020] Redko, I., Vayer, T., Flamary, R., and Courty, N. (2020).
Co-optimal transport.
In *Neural Information Processing Systems (NeurIPS)*.

[Rolet et al., 2016] 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 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.

References xiv

[Sandler and Lindenbaum, 2011] Sandler, R. and Lindenbaum, M. (2011).
Nonnegative matrix factorization with earth mover's distance metric for image analysis.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(8):1590–1602.

[Santambrogio, 2014] Santambrogio, F. (2014).
Introduction to optimal transport theory.
Notes.

[Scetbon and Cuturi, 2020] Scetbon, M. and Cuturi, M. (2020).
Linear time sinkhorn divergences using positive features.
arXiv preprint arXiv:2006.07057.

[Schmitz et al., 2017] 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 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.

[Seguy and Cuturi, 2015] 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.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014).
Very deep convolutional networks for large-scale image recognition.
arXiv preprint arXiv:1409.1556.

[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., 2014] Solomon, J., Rustamov, R., Guibas, L., and Butscher, A. (2014).
Wasserstein propagation for semi-supervised learning.
In *International Conference on Machine Learning*, pages 306–314. PMLR.

[Vayer et al., 2018] Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. (2018).
Fused gromov-wasserstein distance for structured objects: theoretical foundations and mathematical properties.

[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.

[Yang et al., 2018] Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M. K., Zhang, Y., Sun, L., and Wang, G. (2018).
Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss.
IEEE transactions on medical imaging, 37(6):1348–1357.

[Zen et al., 2014] Zen, G., Ricci, E., and Sebe, N. (2014).
Simultaneous ground metric learning and matrix factorization with earth mover's distance.
In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 3690–3695.

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