

Optimal Transport for Machine Learning

Rémi Flamary, Laetitia Chapel, Nicolas Courty

June 22, 2022

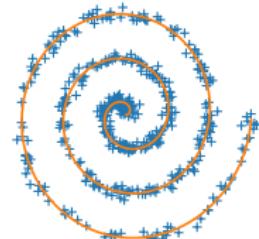
Ecole de recherche : Divergences statistiques et géométriques pour l'apprentissage machine, 2022, Rennes

Introduction

Three aspects of Machine Learning

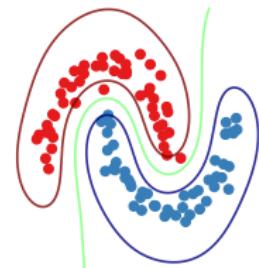
Unsupervised learning

- Extract information from unlabeled data
- Find labels (clustering) or subspaces/manifolds.
- Generate realistic data (GAN).



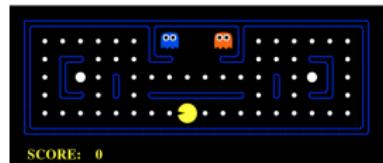
Supervised Learning

- Learning to predict from labeled dataset.
- Regression, Classification.
- Can use unsupervised information (DA, Semi-sup.)

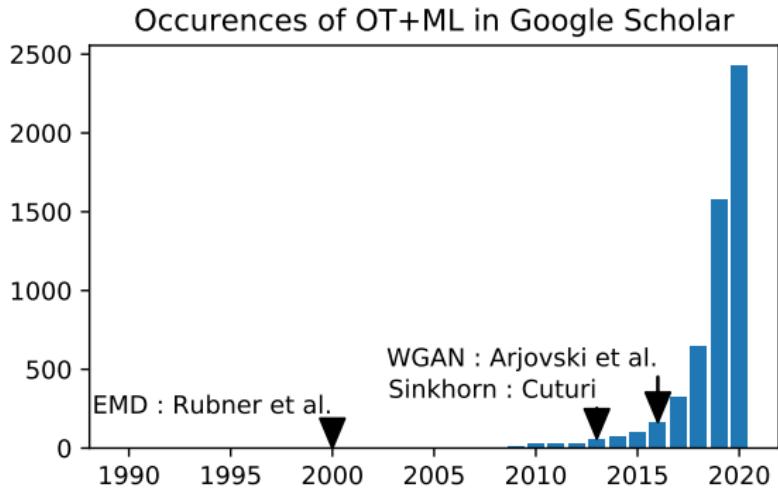


Reinforcement Learning

- Let the machine experiment.
- Learn from its mistakes.
- Framework for learning to play games.



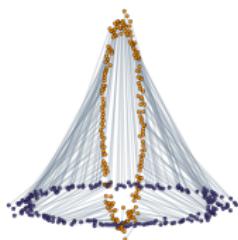
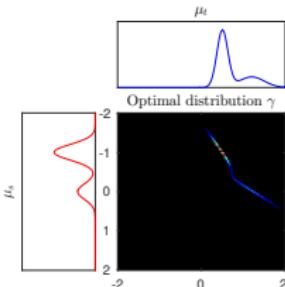
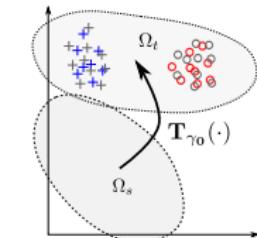
Optimal transport for machine learning



Short history of OT for ML

- Recently introduced to ML (well known in image processing since 2000s).
- Computationnal OT allow numerous applications (regularization).
- Deep learning boost (numerical optimization and GAN).

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., 2016].

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].

Table of content

Introduction

Mapping with optimal transport

Optimal transport mapping estimation

Optimal transport for domain adaptation

Learning from histograms with Optimal Transport

Unsupervised learning

Supervised learning

Learning from empirical distributions with Optimal Transport

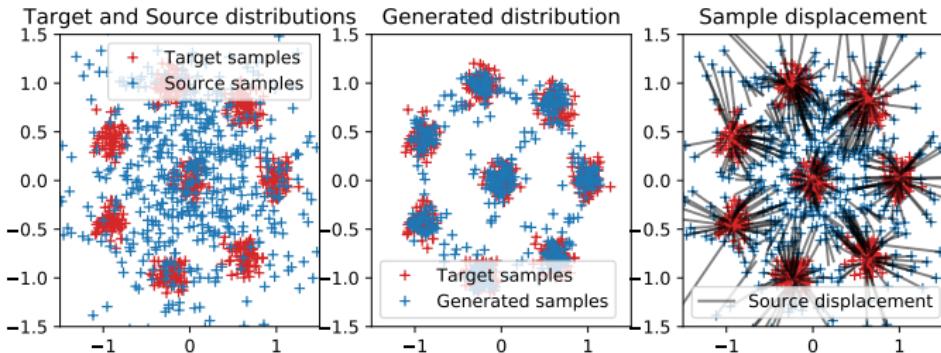
Unupervised learning

Supervised learning and domain adaptation

Conclusion

Mapping with optimal transport

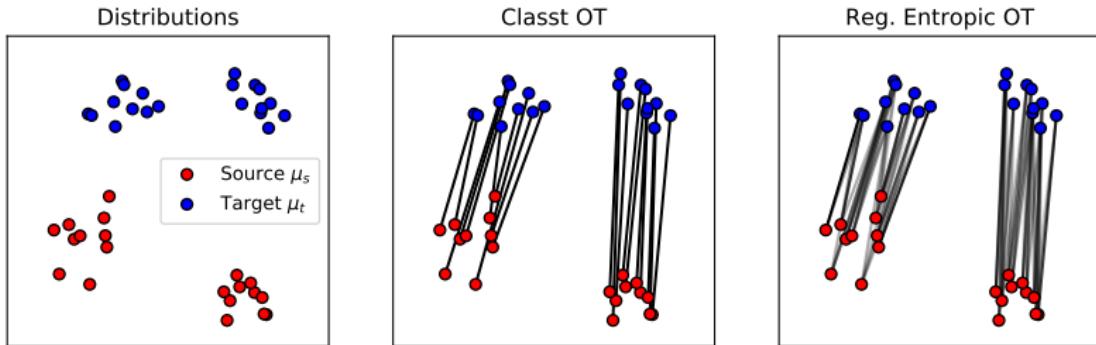
Mapping with optimal transport



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 [Makkluva et al., 2020].
- Can be used to linearize the Wasserstein space [Mérigot et al., 2020]

Transporting the discrete samples

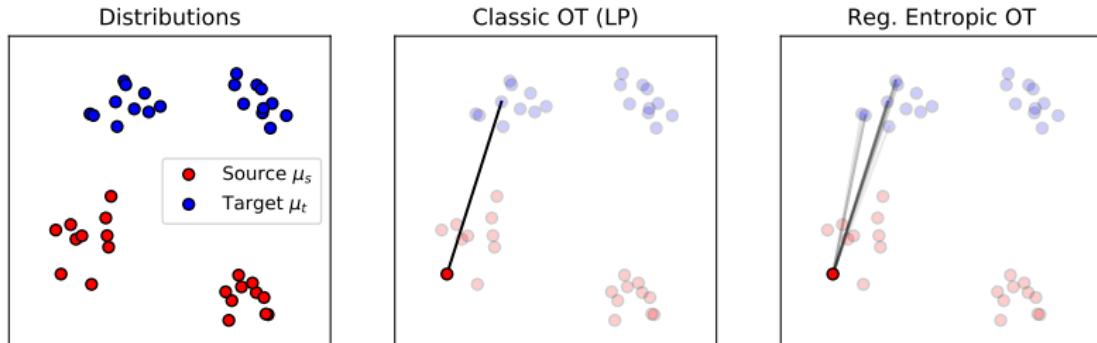


Barycentric mapping [Ferradans et al., 2014]

$$\widehat{T}_{\gamma_0}(\mathbf{x}_i^s) = \arg \min_{\mathbf{x}} \sum_j \gamma_0(i, j) c(\mathbf{x}, \mathbf{x}_j^t). \quad (1)$$

- 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).
- Trick: learn OT on few samples and apply displacement to the nearest point.

Transporting the discrete samples

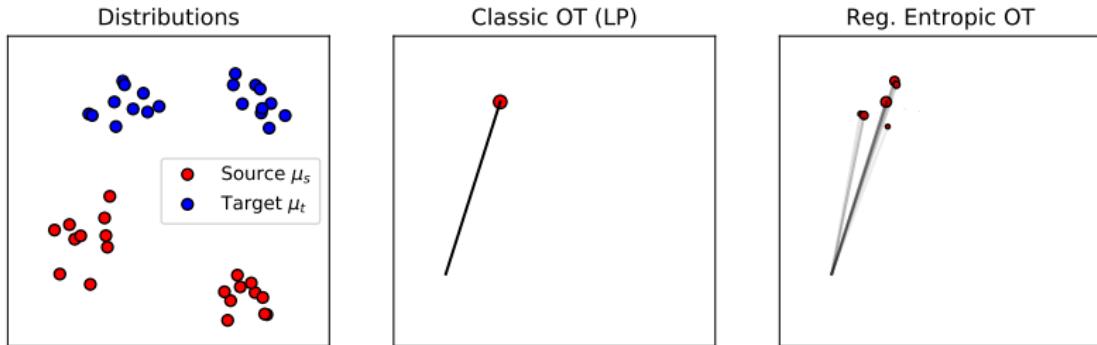


Barycentric mapping [Ferradans et al., 2014]

$$\widehat{T}_{\gamma_0}(\mathbf{x}_i^s) = \arg \min_{\mathbf{x}} \sum_j \gamma_0(i, j) \|\mathbf{x} - \mathbf{x}_j^t\|^2. \quad (1)$$

- 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).
- Trick: learn OT on few samples and apply displacement to the nearest point.

Transporting the discrete samples

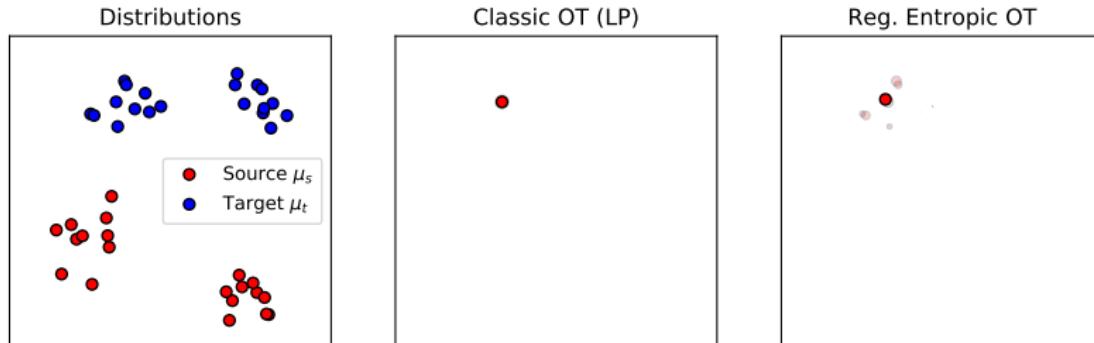


Barycentric mapping [Ferradans et al., 2014]

$$\hat{T}_{\gamma_0}(\mathbf{x}_i^s) = \frac{1}{\sum_j \gamma_0(i,j)} \sum_j \gamma_0(i,j) \mathbf{x}_j^t. \quad (1)$$

- 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).
- Trick: learn OT on few samples and apply displacement to the nearest point.

Transporting the discrete samples

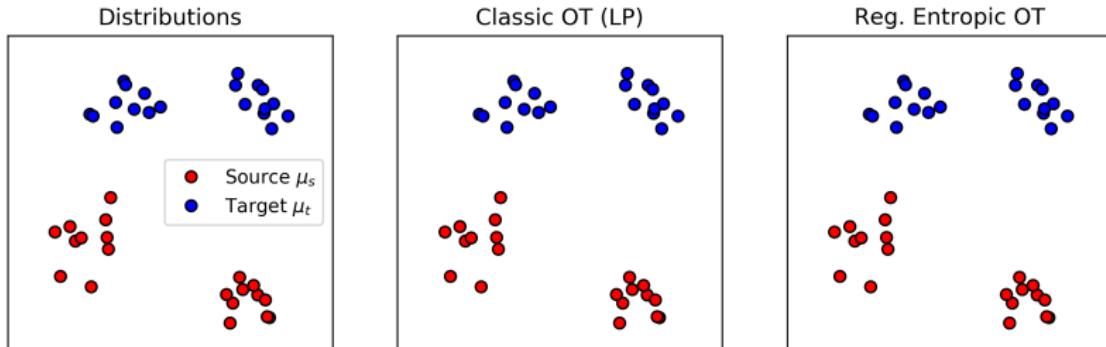


Barycentric mapping [Ferradans et al., 2014]

$$\hat{T}_{\gamma_0}(\mathbf{x}_i^s) = \frac{1}{\sum_j \gamma_0(i,j)} \sum_j \gamma_0(i,j) \mathbf{x}_j^t. \quad (1)$$

- 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).
- Trick: learn OT on few samples and apply displacement to the nearest point.

Transporting the discrete samples

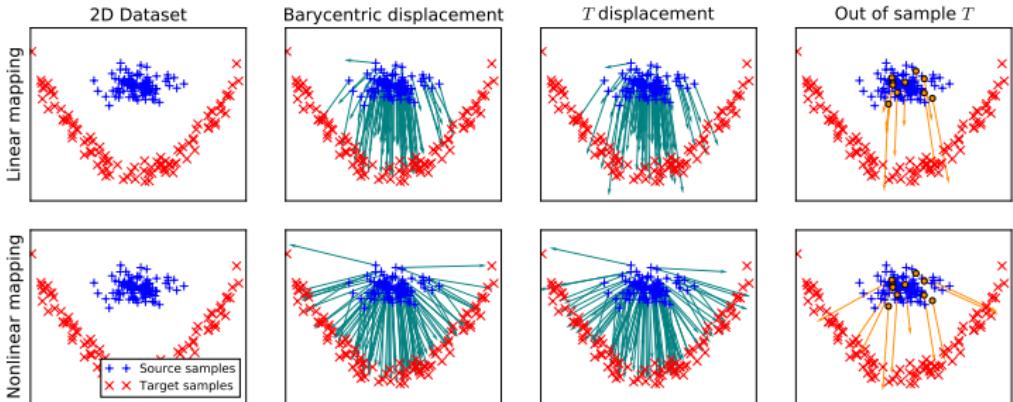


Barycentric mapping [Ferradans et al., 2014]

$$\widehat{T}_{\gamma_0}(\mathbf{x}_i^s) = \frac{1}{\sum_j \gamma_0(i,j)} \sum_j \gamma_0(i,j) \mathbf{x}_j^t. \quad (1)$$

- 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).
- Trick: learn OT on few samples and apply displacement to the nearest point.

Joint OT and mapping estimation

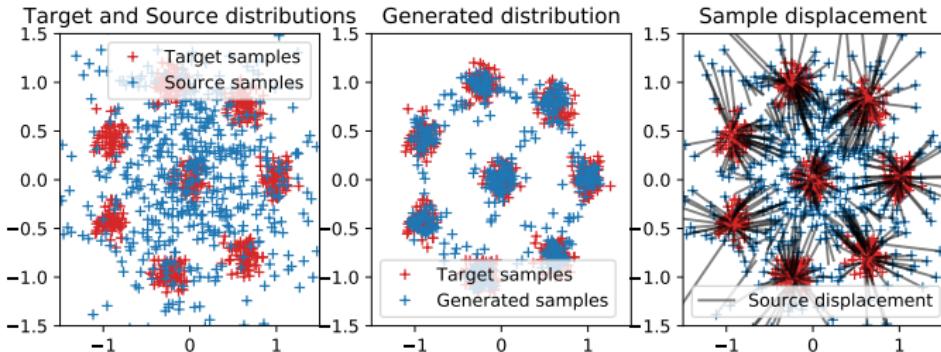


Simultaneous OT matrix and mapping [Perrot et al., 2016]

$$\min_{T, \gamma \in \mathcal{P}} \quad \langle \gamma, \mathbf{C} \rangle_F + \sum_i \|T(\mathbf{x}_i^s) - \hat{T}_\gamma(\mathbf{x}_i^s)\|^2 + \lambda \|T\|^2$$

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

Large scale optimal transport and mapping estimation

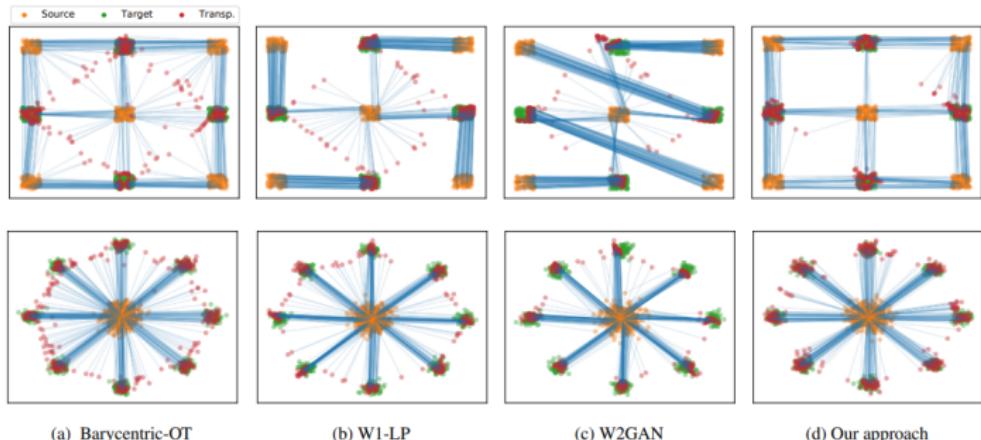


Large scale mapping estimation [Seguy et al., 2017]

- 2-step procedure:
 - 1 (Stochastic) estimation of regularized $\hat{\gamma}$.
 - 2 (Stochastic) estimation of T with a neural network.
- OT solved with Stochastic Gradient Ascent in the dual.
- Convergence to the true mapping for small regularization.
- Convergence to the smooth mapping for large n
[Pooladian and Niles-Weed, 2021].

0	0	3	9	2	9
1	7	7	6	8	6
0	3	8	1	4	4
9	6	1	5	6	1
7	2	4	5	1	7
5	3	6	6	9	1

Monge Mapping with input convex neural networks



Principle [Makkuva et al., 2020]

- For the quadratic cost OT between two smooth distribution Brenier theorem states that the Monge mapping is the gradient of a convex function.
- Neural network can be designed to be convex wrt their input (ICNN) [Amos et al., 2017].
- [Makkuva et al., 2020] proposed to estimate directly the Monge as a gradient of an ICNN from the empirical distributions. mapping usin

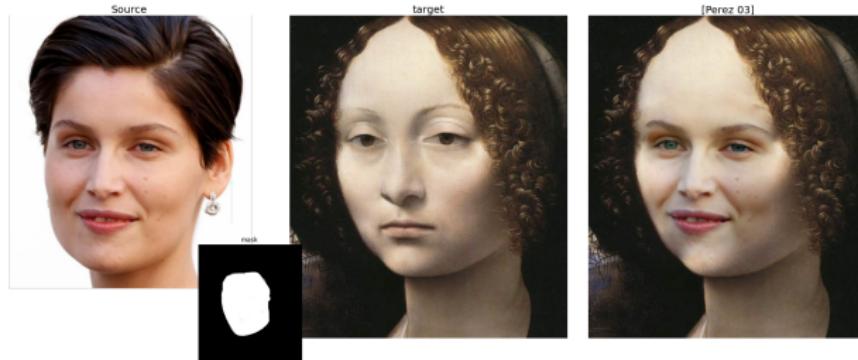
Seamless copy in images



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 in images



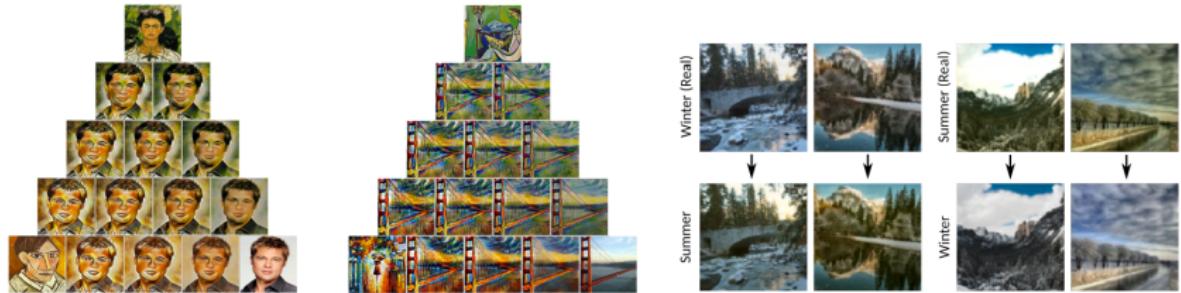
Poisson image editing [Pérez et al., 2003]

- 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.

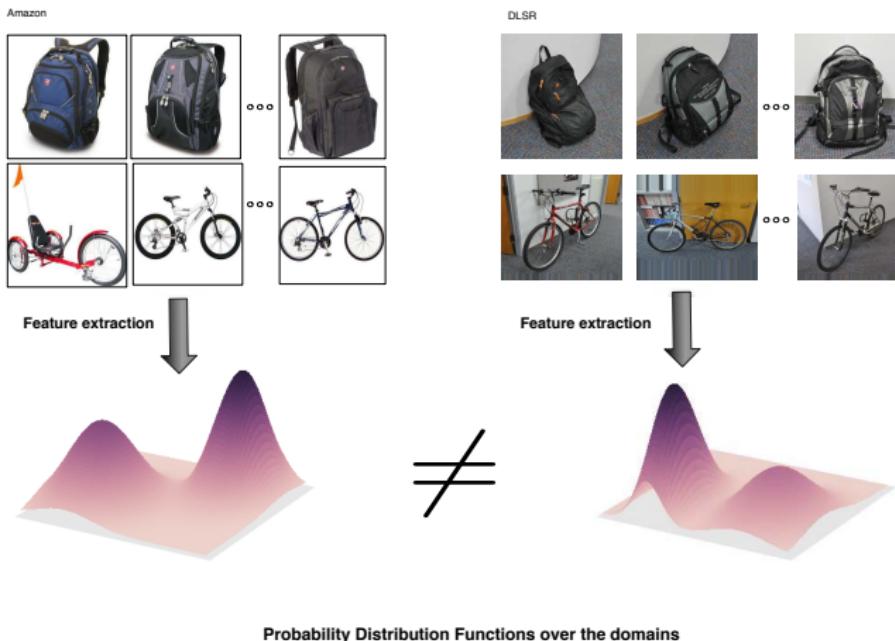
Monge mapping for Image-to-Image translation



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.

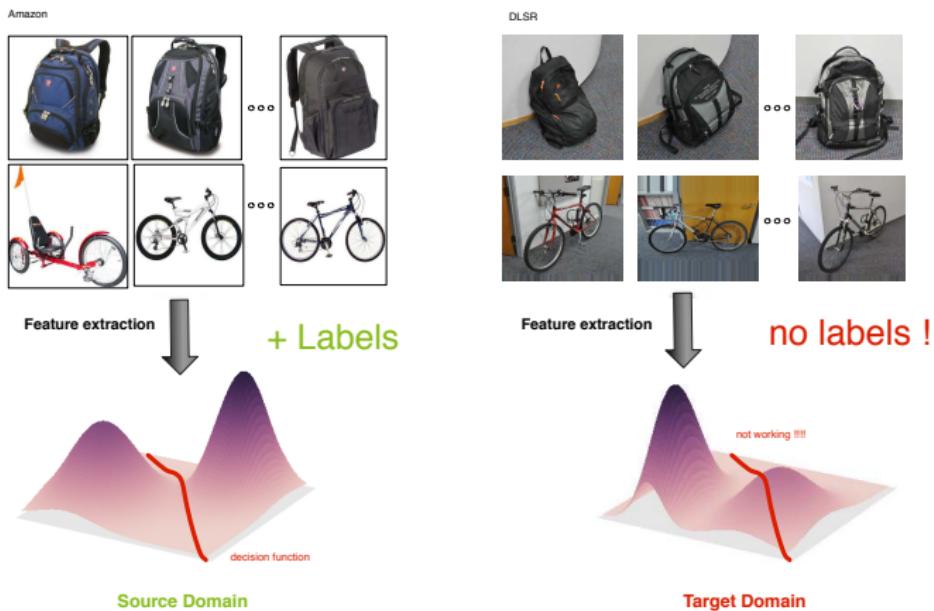
Domain Adaptation problem



Our context

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

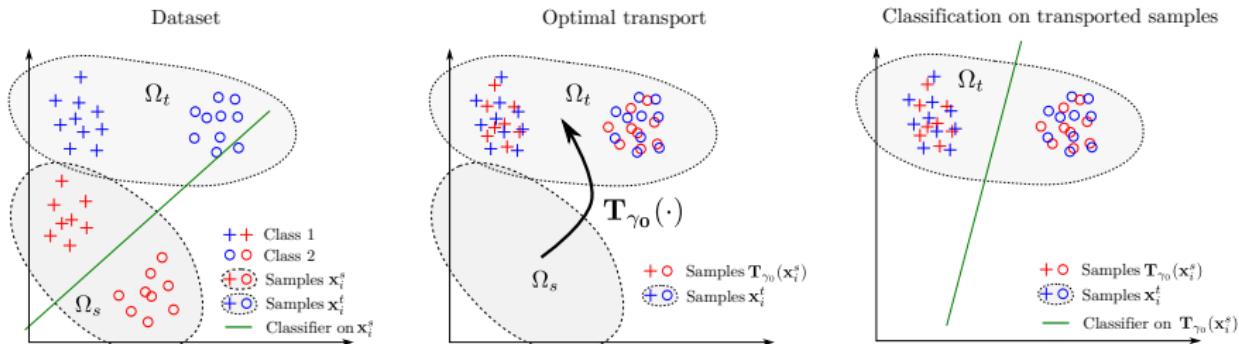
Unsupervised domain adaptation problem



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

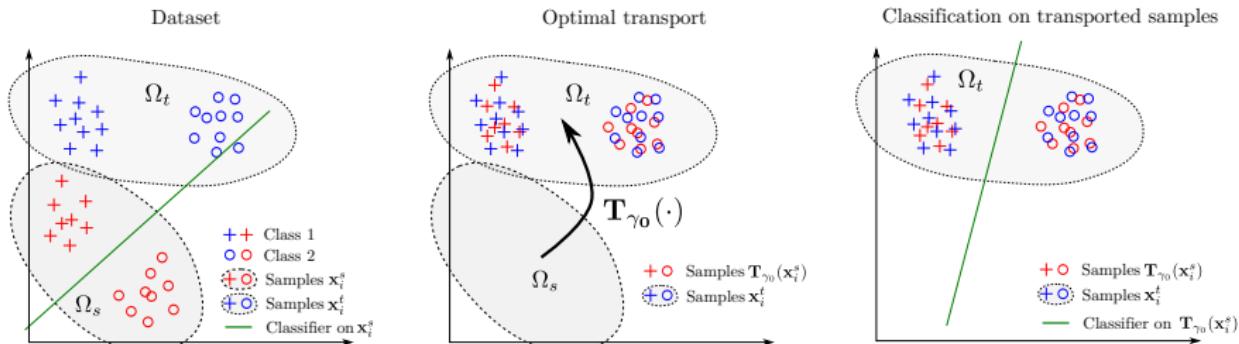
OT for domain adaptation : Step 1



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., 2016].
- 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



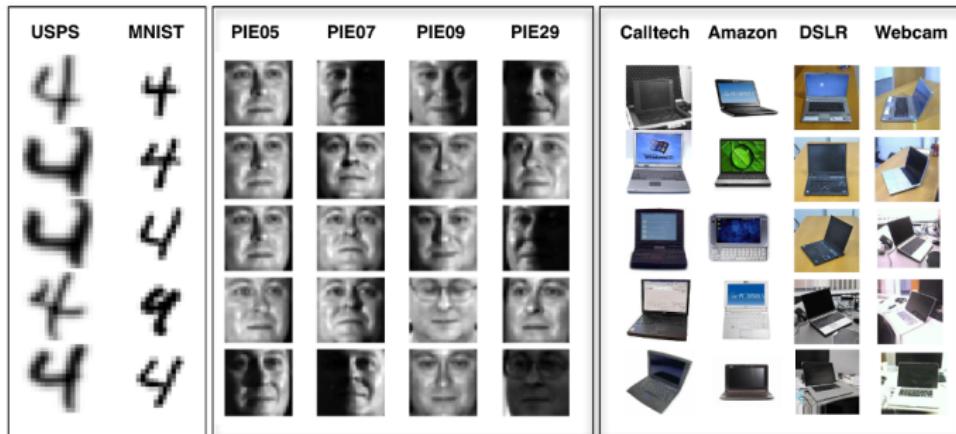
Step 2 : Transport the training samples onto the target distribution.

- 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



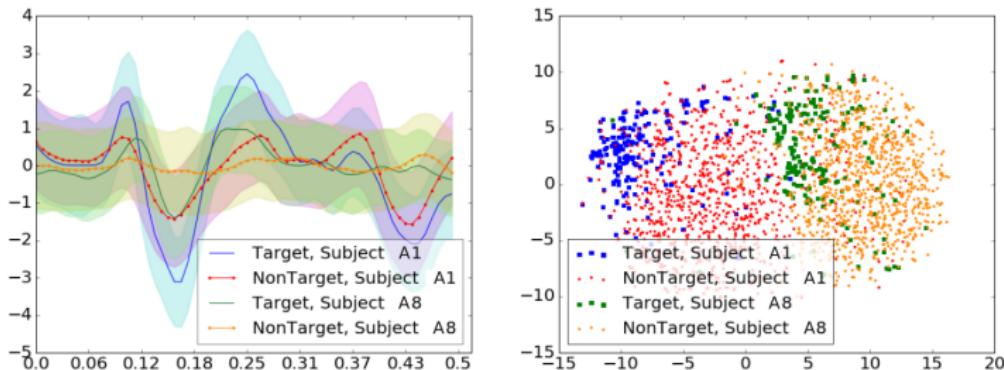
Datasets

- **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

- State of the art performances on the 3 datasets.
- Works well on deep features adaptation and extension to semi-supervised DA.

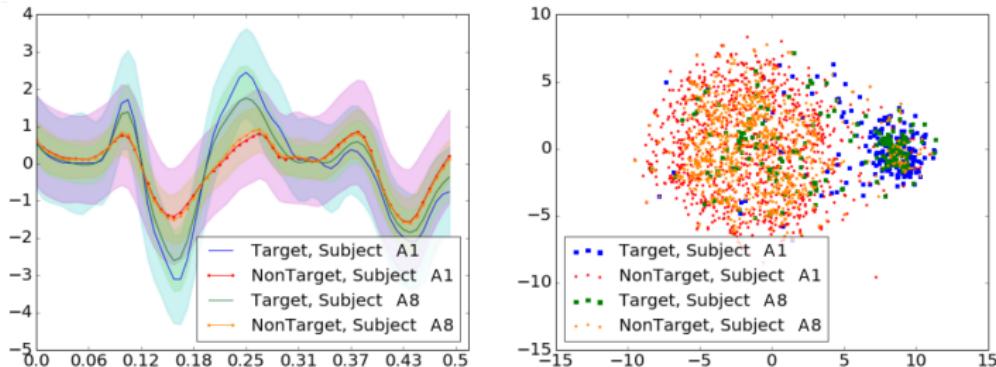
OTDA for biomedical data (1)



Multi-subject P300 classification [Gayraud et al., 2017]

- Objective : reduce calibration for BCI users.
- P300 signal is different accross subjects so adapting models is hard.
- Perform XDAWN [Rivet et al., 2009] as pre-processing.
- Use OTDA to adapt each subject in the dataset to a new subject.
- Train independent classifier on transported data and perform aggregation.

OTDA for biomedical data (1)



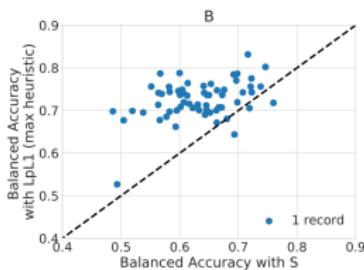
Multi-subject P300 classification [Gayraud et al., 2017]

- Objective : reduce calibration for BCI users.
- P300 signal is different accross subjects so adapting models is hard.
- Perform XDAWN [Rivet et al., 2009] as pre-processing.
- Use OTDA to adapt each subject in the dataset to a new subject.
- Train independent classifier on transported data and perform aggregation.

OTDA for biomedical data (2)

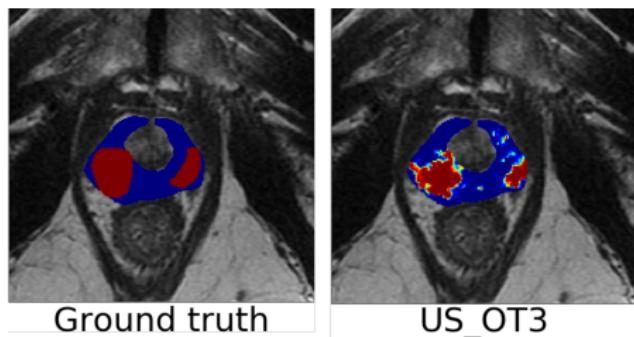
EEG sleep stage classification [Chambon et al., 2018]

- Use pre-trained neural network.
- Adapt with OTDA on the penultimate layer.
- OTDA best DA approach to adapt between EEG recordings.



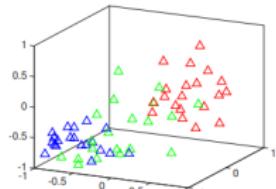
Prostate cancer classification [Gautheron et al., 2017]

- Adaptation of MRI voxel features from 1.5T to 3T.
- Achieve good performance across subjects and modality with no target labels.

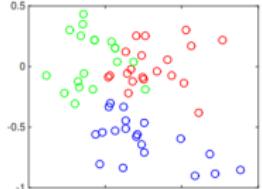


Heterogeneous Domain Adaptation with GW

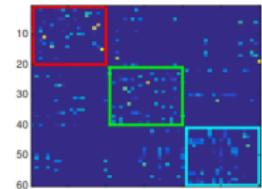
source data
 transported source data
 target data
 labeled target data in SGW



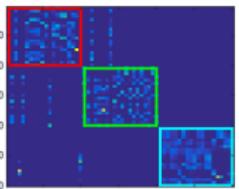
(a) source data



(b) target data



(c) T obtained by EGW



(e) T obtained by SGW

Semi-supervised Heterogeneous Domain Adaptation [Yan et al., 2018]

- OT for DA initially proposed by [Courty et al., 2016].
- Use the OT matrix to transfer labels or samples between datasets.
- GW find correspondences across spaces but very noisy.
- Semi-supervised strategy allows very good performances.
- Alternative : Co-optimal transport that find correspondances between the variables and samples simultaneously [Redko et al., 2020].

Outline

Introduction

Mapping with optimal transport

- Optimal transport mapping estimation

- Optimal transport for domain adaptation

Learning from histograms with Optimal Transport

- Unsupervised learning

- Supervised learning

Learning from empirical distributions with Optimal Transport

- Unupervised learning

- Supervised learning and domain adaptation

Conclusion

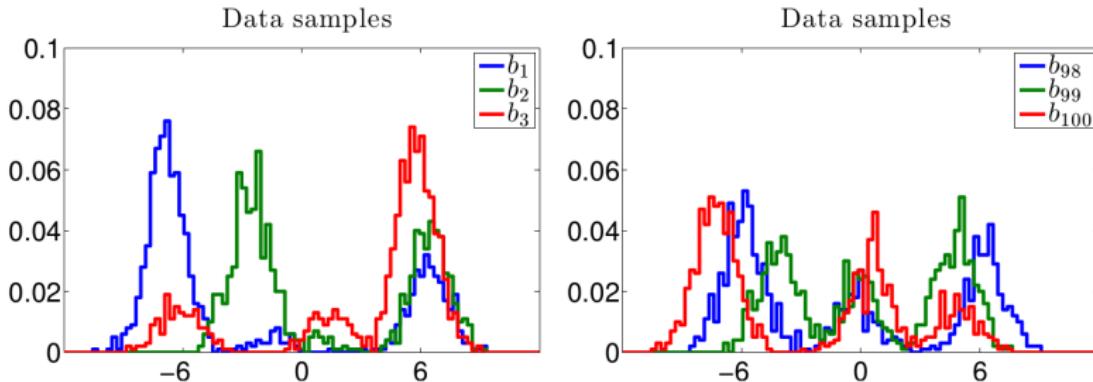
Learning from histograms



Data as histograms

- Fixed bin positions x_i e.g. grid, simplex $\Delta = \{(\mu_i)_i \geq 0; \sum_i \mu_i = 1\}$
- A lot of datasets comes under the form of histograms.
- Images are photo counts (black and white), text as word counts.
- Natural divergence is Kullback–Leibler.
- Not all data can be seen as histograms (positivity+constant mass)!

Dictionary learning on histograms

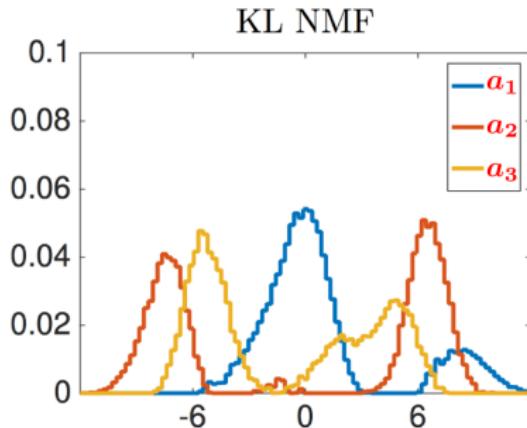
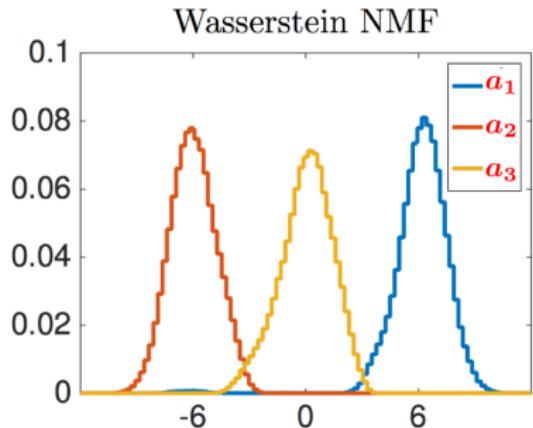


DL with Wasserstein distance [Sandler and Lindenbaum, 2011]

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

- NMF: columns of \mathbf{D} and \mathbf{H} are on the simplex.
- Metric \mathbf{C} can encode spatial relations between the bins of the histograms.
- Ground metric learning [Zen et al., 2014].
- Fast DL with regularized OT [Rolet et al., 2016].

Dictionary learning on histograms

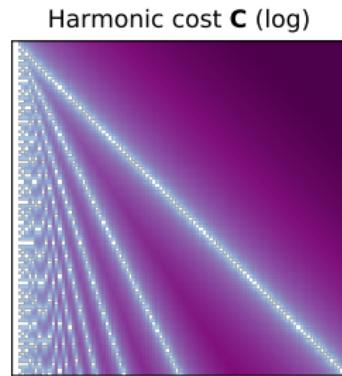
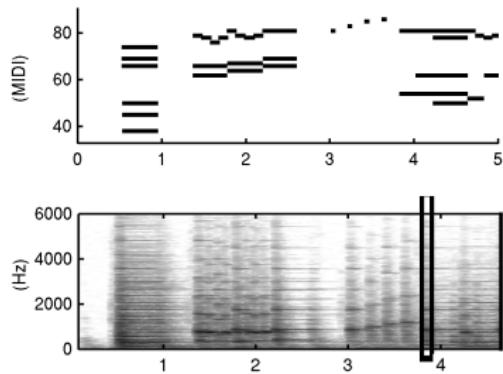


DL with Wasserstein distance [Sandler and Lindenbaum, 2011]

$$\min_{\mathbf{D}, \mathbf{H}} \sum_i W_C(\mathbf{v}_i, \mathbf{D}\mathbf{h}_i)$$

- NMF: columns of \mathbf{D} and \mathbf{H} are on the simplex.
- Metric C can encode spatial relations between the bins of the histograms.
- Ground metric learning [Zen et al., 2014].
- Fast DL with regularized OT [Rolet et al., 2016].

Optimal Spectral Transportation (OST)



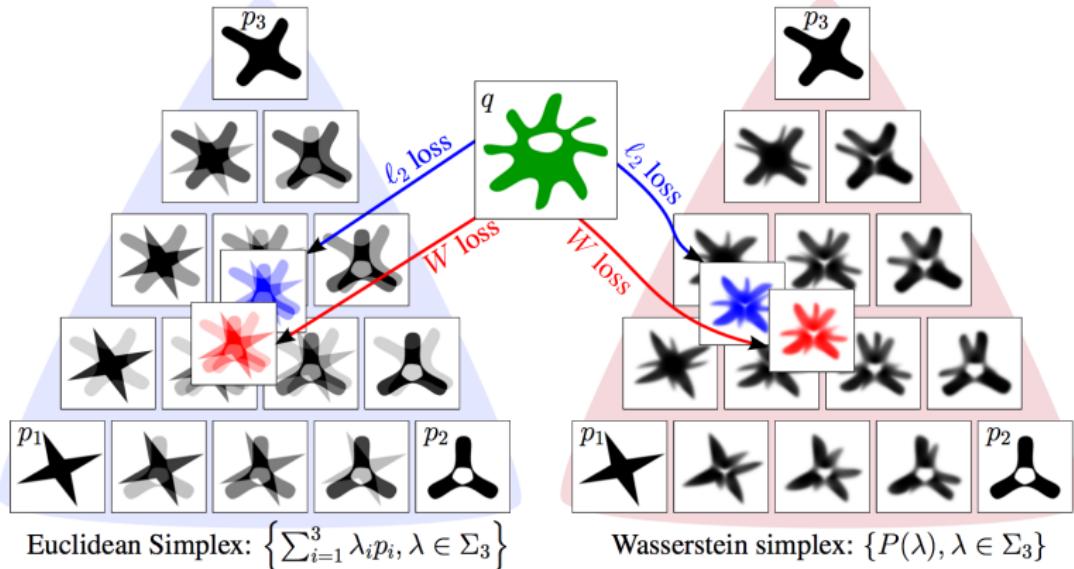
OT linear spectral unmixing of musical data [Flamary et al., 2016]

$$\min_{\mathbf{h} \in \Delta} W_C(\mathbf{v}, \mathbf{D}\mathbf{h}) \quad (2)$$

- Objective : robustness to harmonic magnitude and small frequency shift
- Encode harmonic structure in the cost matrix (harmonic robustness).
- Can use simple dictionary (diracs on fundamental frequency).
- Very fast solver for sparse and entropic regularization.

Demo : <https://github.com/rflamary/OST>

Wasserstein dictionary learning

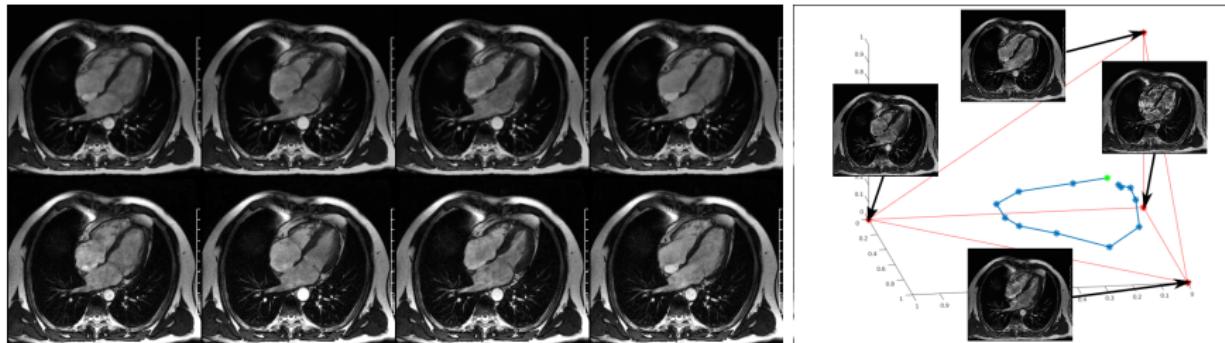


Nonlinear unmixing with Wasserstein simplex [Schmitz et al., 2017]

$$\min_{\mathbf{D}, \mathbf{H}} \quad \sum_i L(\mathbf{v}_i, WB(\mathbf{D}, \mathbf{h}_i))$$

with $WB(\mathbf{D}, \mathbf{h}) = \arg \min_{\mathbf{a}} \sum_i h_i W_C(\mathbf{d}_i, \mathbf{a})$

Wasserstein dictionary learning (2)



Nonlinear unmixing with Wasserstein simplex [Schmitz et al., 2017]

$$\min_{\mathbf{D}, \mathbf{H}} \quad \sum_i L(\mathbf{v}_i, WB(\mathbf{D}, \mathbf{h}_i))$$

with $WB(\mathbf{D}, \mathbf{h}) = \arg \min_{\mathbf{a}} \sum_i h_i W_C(\mathbf{d}_i, \mathbf{a})$

- Linear model is a barycenter for the squared ℓ_2 distance.
- Use Wasserstein barycenter for non-linear modeling.
- Application to cardiac sequence in MRI.
- One cardiac cycle is a trajectory in the simplex of the dictionary.

Principal Geodesics Analysis

Class 0			Class 1			Class 4		
PCA	PGA		PCA	PGA		PCA	PGA	
1	2	3	1	2	3	1	2	3
0 0 0	0 0 0	X X X	1 1 1	4 4 4	4 4 4			
0 0 0	0 0 0	X X X	1 1 1	4 4 4	4 4 4			
0 0 0	0 0 0	X X X	1 1 1	4 4 4	4 4 4			
0 0 0	0 0 0	X X X	1 1 1	4 4 4	4 4 4			
0 0 0	0 0 0	X X X	1 1 1	4 4 4	4 4 4			
0 0 0	0 0 0	X X X	1 1 1	4 4 4	4 4 4			
0 0 0	0 0 0	X X X	1 1 1	4 4 4	4 4 4			
0 0 0	0 0 0	X X X	1 1 1	4 4 4	4 4 4			

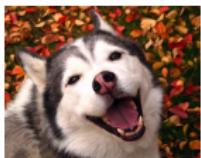
Geodesic PCA in the Wasserstein space [Bigot et al., 2017]

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

Multi-label learning with Wasserstein Loss



Siberian husky



Eskimo dog



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



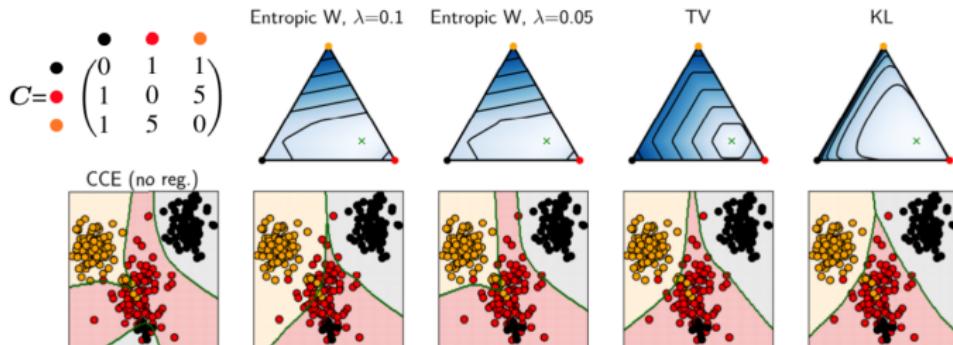
Flickr : water, boat, reflection, sun-shine
Prediction : water, river, lake, summer;

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]

$$R_C(f, \mathbf{x}) = \max_{\|\mathbf{v}\| \leq \epsilon} W_C(f(\mathbf{x} + \mathbf{v}), f(\mathbf{x}))$$

- 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} in regularization $R_C(f, \mathbf{x})$ 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).

Outline

Introduction

Mapping with optimal transport

- Optimal transport mapping estimation

- Optimal transport for domain adaptation

Learning from histograms with Optimal Transport

- Unsupervised learning

- Supervised learning

Learning from empirical distributions with Optimal Transport

- Unupervised learning

- Supervised learning and domain adaptation

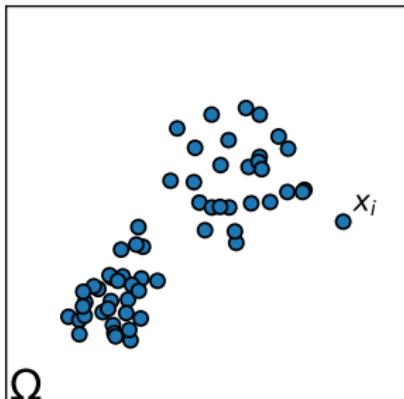
Conclusion

Empirical distributions A.K.A datasets

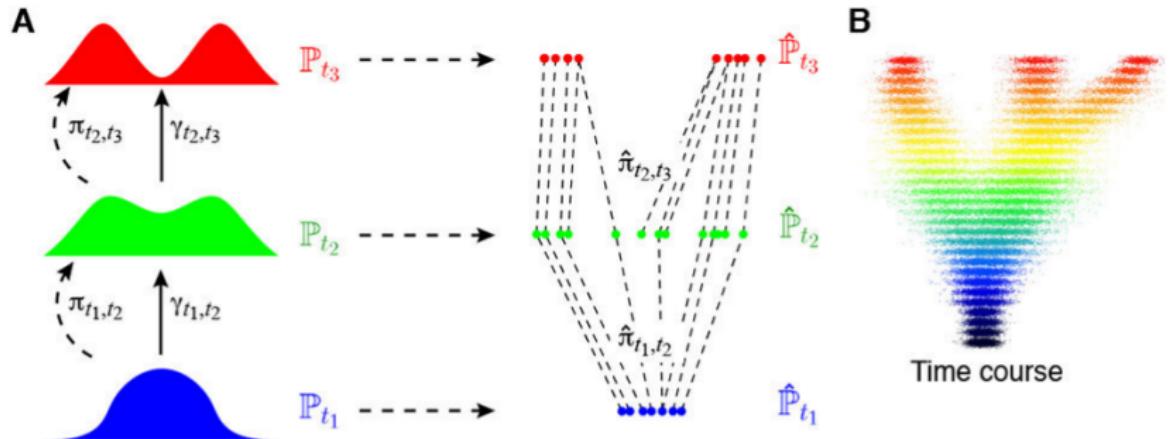
$$\mu = \sum_{i=1}^n a_i \delta_{\mathbf{x}_i}, \quad \mathbf{x}_i \in \Omega, \quad \sum_{i=1}^n a_i = 1$$

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.



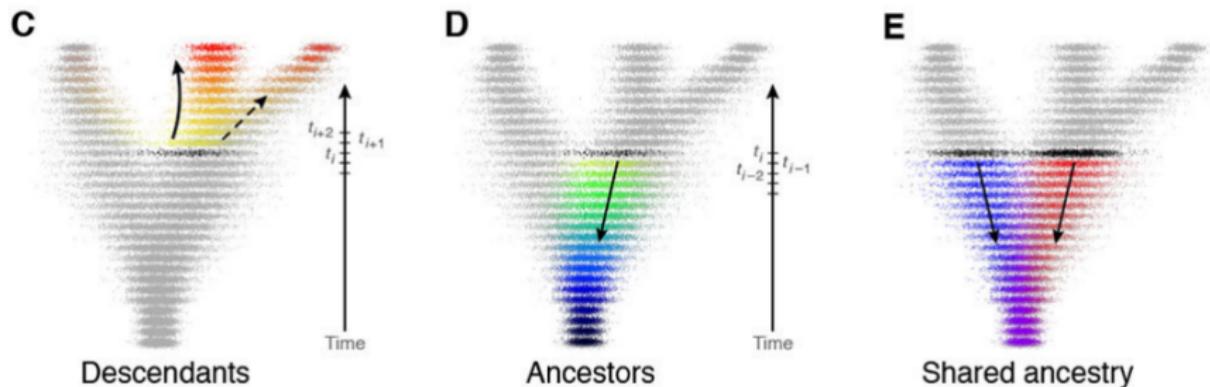
OT for modeling cell development



Principle [Schiebinger et al., 2019]

- Developmental trajectories of cells from stem cells to more specialized.
- Cell populations are samples at different times with scRNA-seq.
- Optimal transport can be used to find mapping/correspondances between across population measurements.
- Unbalanced OT is used to model cellular growth and death rates.

OT for modeling cell development



Principle [Schiebinger et al., 2019]

- Developmental trajectories of cells from stem cells to more specialized.
- Cell populations are samples at different times with scRNA-seq.
- Optimal transport can be used to find mapping/correspondances between across population measurements.
- Unbalanced OT is used to model cellular growth and death rates.

Generative Adversarial Networks (GAN)

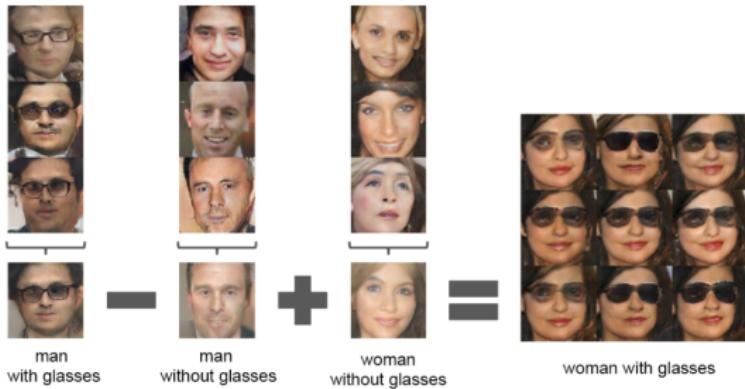


Generative Adversarial Networks (GAN) [Goodfellow et al., 2014]

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

- 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]).

Generative Adversarial Networks (GAN)

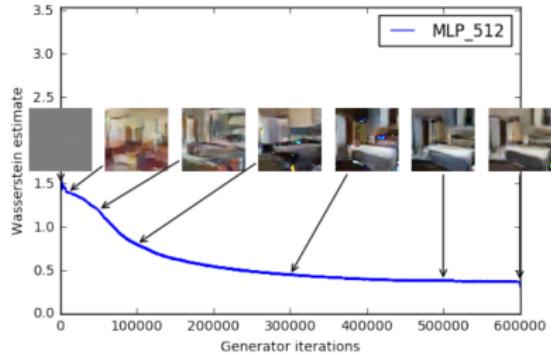
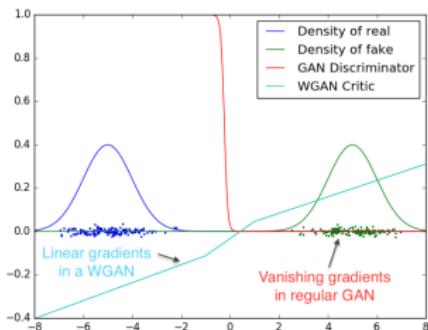


Generative Adversarial Networks (GAN) [Goodfellow et al., 2014]

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

- 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]

$$\min_G \quad W_1^1(G\#\mu_z, \mu_d), \quad (3)$$

- Minimizes the Wasserstein distance between the data μ_d and the generated data $G\#\mu_z$ whe $\mu_z = \mathcal{N}(0, \mathbf{I})$.
- No vanishing gradients ! Better convergence in practice.
- Wasserstein in the dual (separable w.r.t. the samples).

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

- ϕ is a neural network that acts as an *actor critic*

WGAN: the devil in the approximation

Neural network belonging to Lip^1 ?

- Not really! [Arjovsky et al., 2017] proposes to do weight clipping that force an upper bound on the Lipschitz constant.
- It is actually the supremum over K-Lipschitz functions that is approximated by a neural network

$$\max_{f \in \text{NN class}} L_{\text{WGAN}}(f, G) \leq \sup_{\|\phi\|_L \leq K} L_{\text{WGAN}}(\phi, G) = K \cdot W_1^1(G(\mathbf{z}), \mu_d)$$

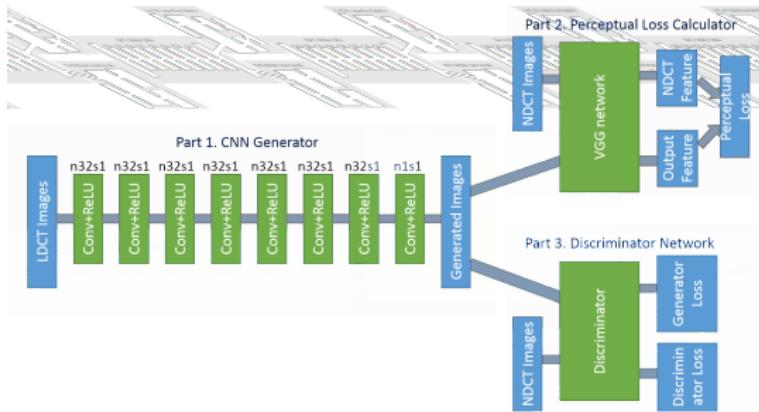
- Actually **not** equivalent to solve the optimal transport, but gradients are aligned.

Improved WGAN [Gulrajani et al., 2017]

$$\min_G \sup_{f \in \text{NN class}} \mathbb{E}_{\mathbf{x} \sim \mu_d}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{z} \sim \mu_z}[f(G(\mathbf{z}))] + \lambda \mathbb{E}_{\mathbf{x} \sim \mu_d}[(\|\nabla f(\mathbf{x})\|_2 - 1)^2]$$

Relaxation of the constraint (for W_1 the gradient of the potential is 1 almost everywhere).

Wasserstein GAN loss on Biomedical images



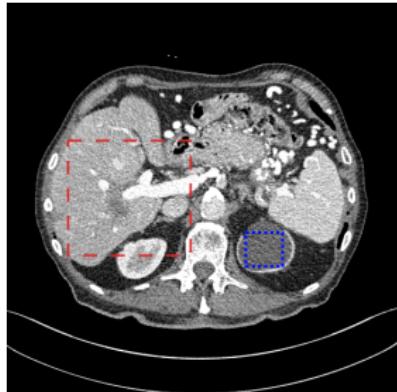
Reconstructing low dose CT images [Yang et al., 2018]

$$\min_G W_1^1(G\#\mu_l, \mu_f) + \lambda_1 E_{x \sim \mu_l} [\|VGG(x_l) - VGG(G(x_l))\|^2], \quad (4)$$

- 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.

Wasserstein GAN loss on Biomedical images

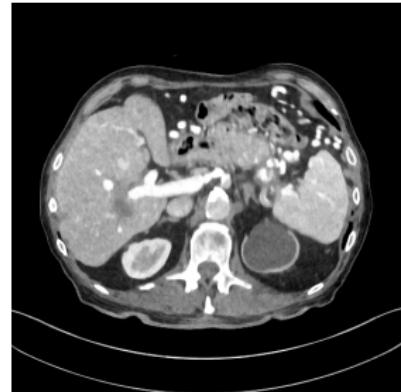
Full dose



Quarter dose



Dico rec.



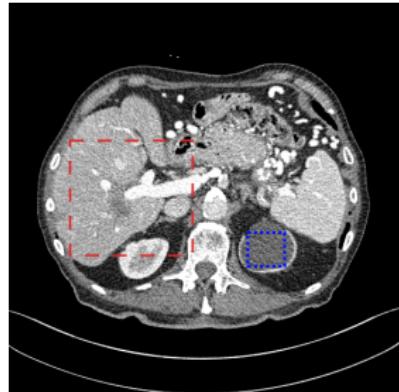
Reconstructing low dose CT images [Yang et al., 2018]

$$\min_G W_1^1(G\#\mu_l, \mu_f) + \lambda_1 E_{x \sim \mu_l} [\|VGG(x_l) - VGG(G(x_l))\|^2], \quad (4)$$

- 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.

Wasserstein GAN loss on Biomedical images

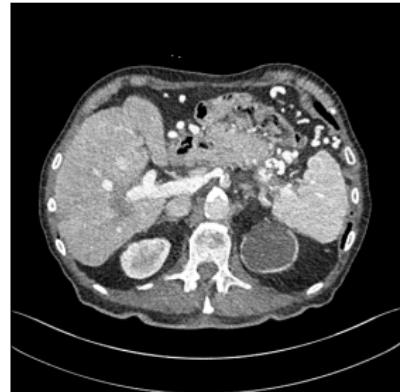
Full dose



Quarter dose



WGAN-VGG rec.

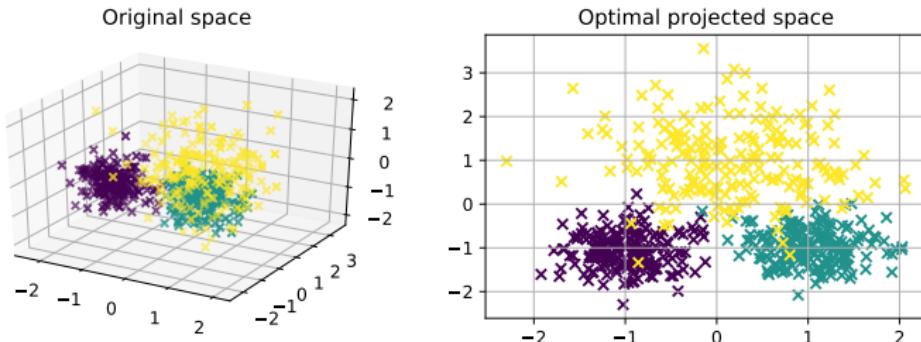


Reconstructing low dose CT images [Yang et al., 2018]

$$\min_G \quad W_1^1(G\#\mu_l, \mu_f) + \lambda_1 E_{\mathbf{x} \sim \mu_l} [\|VGG(\mathbf{x}_l) - VGG(G(\mathbf{x}_l))\|^2], \quad (4)$$

- 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.

Wasserstein Discriminant Analysis (WDA)

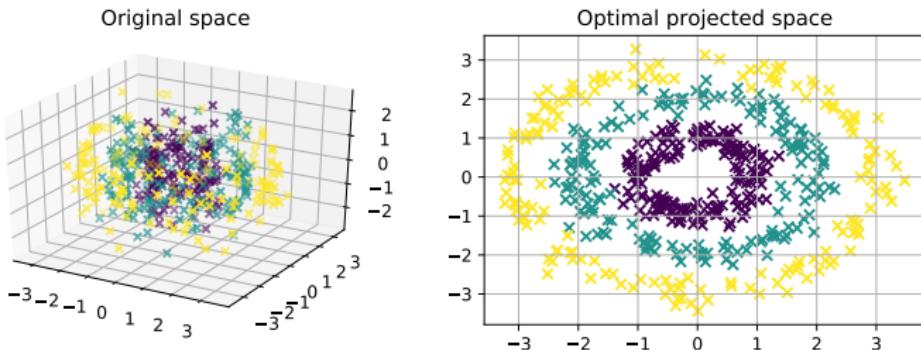


$$\max_{\mathbf{P} \in \mathcal{S}} \frac{\sum_{c,c' > c} W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^{c'})}{\sum_c W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^c)} \quad (5)$$

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

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

Wasserstein Discriminant Analysis (WDA)

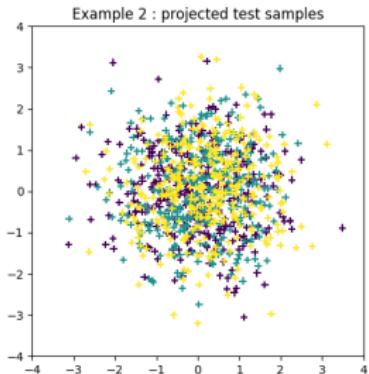
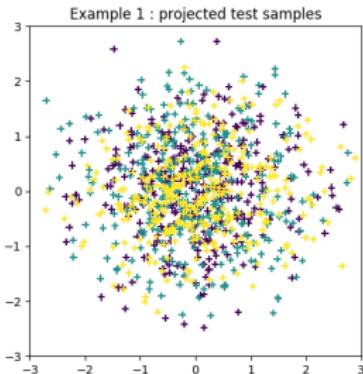


$$\max_{\mathbf{P} \in \mathcal{S}} \frac{\sum_{c, c' > c} W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^{c'})}{\sum_c W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^c)} \quad (5)$$

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

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

Wasserstein Discriminant Analysis (WDA)

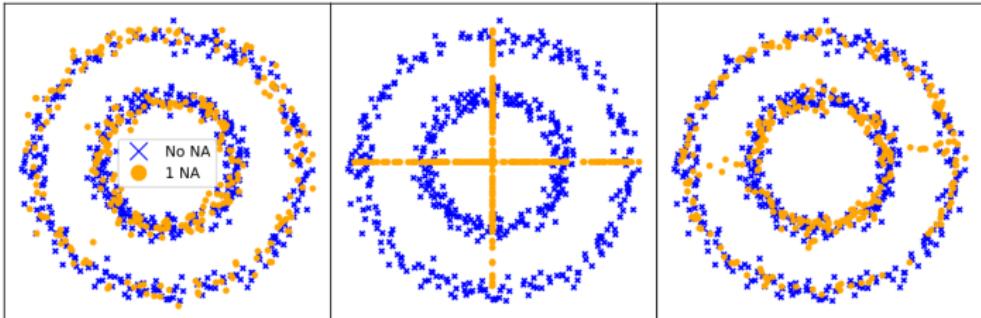


$$\max_{\mathbf{P} \in \mathcal{S}} \frac{\sum_{c,c' > c} W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^{c'})}{\sum_c W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^c)} \quad (5)$$

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

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

Data imputation with Optimal Transport

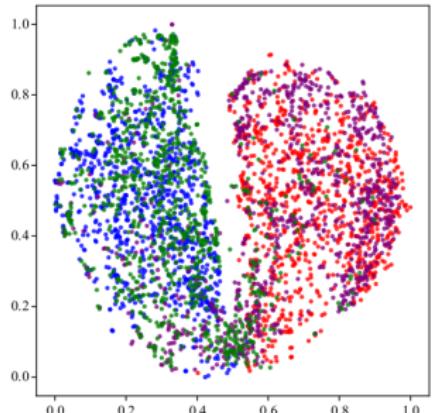
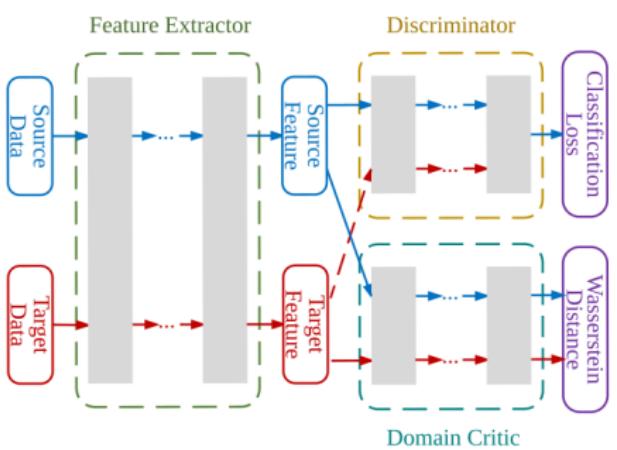


Missing Data imputation [Muzellec et al., 2020]

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

- $\mathbf{X} \odot \mathbf{M}$ 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].

Domain adaptation with Wasserstein distance

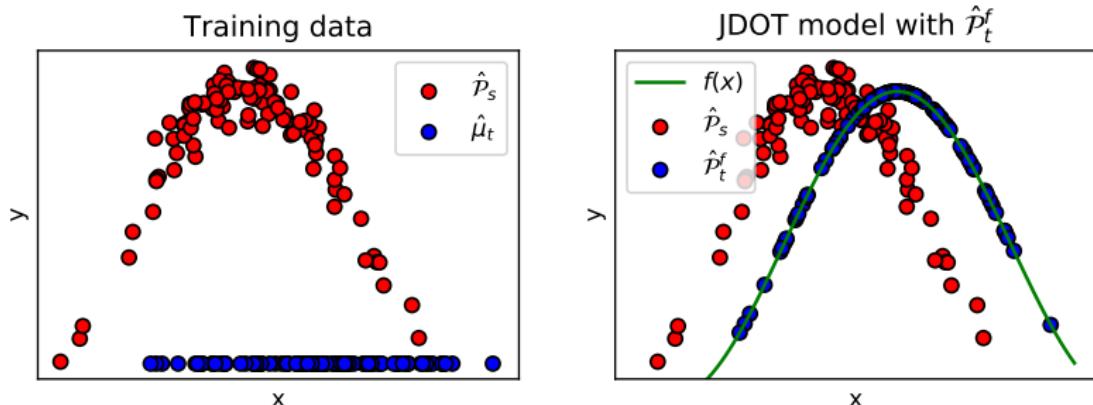


(d) t-SNE of WDGRL features

Domain adaptation for deep learning [Shen et al., 2018]

- Modern DA aim at aligning source and target in the deep representation : DANN [Ganin et al., 2016], MMD [Tzeng et al., 2014], CORAL [Sun and Saenko, 2016].
- Wasserstein distance (WGAN loss [Arjovsky et al., 2017]) used as objective for the adaptation [Shen et al., 2018].

Joint Distribution Optimal Transport for DA

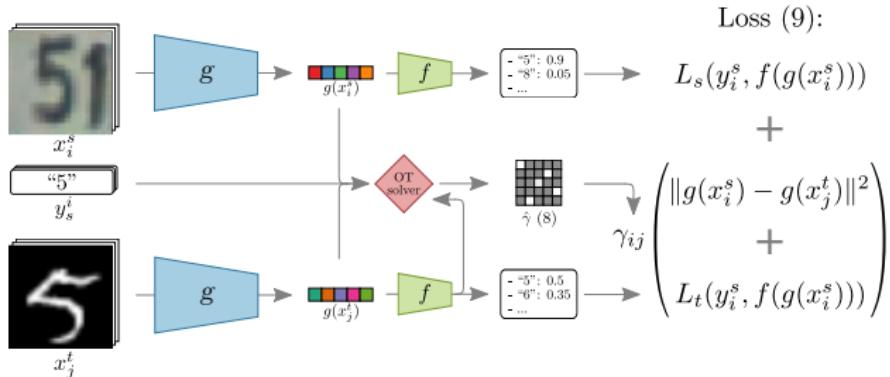


Learning with JDOT [Courty et al., 2017b]

$$\min_f \quad \left\{ W_1(\hat{\mathcal{P}}_s, \hat{\mathcal{P}}_t^f) = \inf_{\gamma \in \Pi} \sum_{ij} \mathcal{D}(\mathbf{x}_i^s, y_i^s; \mathbf{x}_j^t, f(\mathbf{x}_j^t)) \gamma_{ij} \right\} \quad (6)$$

- $\hat{\mathcal{P}}_t^f = \frac{1}{N_t} \sum_{i=1}^{N_t} \delta_{\mathbf{x}_i^t, f(\mathbf{x}_i^t)}$ is the proxy joint feature/label distribution.
- $\mathcal{D}(\mathbf{x}_i^s, y_i^s; \mathbf{x}_j^t, f(\mathbf{x}_j^t)) = \alpha \|\mathbf{x}_i^s - \mathbf{x}_j^t\|^2 + \mathcal{L}(y_i^s, f(\mathbf{x}_j^t))$ with $\alpha > 0$.
- We search for the predictor f that better align the joint distributions.
- OT matrix does the label propagation (no mapping).
- JDOT can be seen as minimizing a generalization bound.

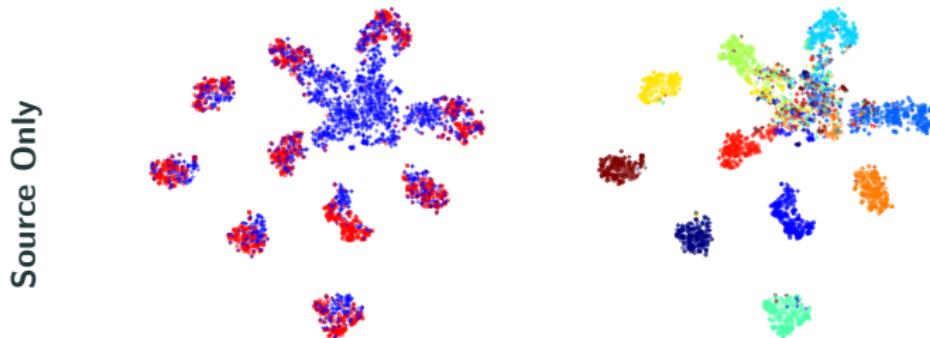
JDOT for large scale deep learning



DeepJDOT [Damodaran et al., 2018]

- Learn simultaneously the embedding g and the classifier f .
- JDOT performed in the joint embedding/label space.
- Use minibatch to estimate OT and update g, f at each iterations [Fatras et al., 2020].
- Scales to large datasets and estimate a representation for both domains.
- TSNE projections of embeddings (MNIST→MNIST-M).

JDOT for large scale deep learning



DeepJDOT [Damodaran et al., 2018]

- Learn simultaneously the embedding g and the classifier f .
- JDOT performed in the joint embedding/label space.
- Use minibatch to estimate OT and update g, f at each iterations [Fatras et al., 2020].
- Scales to large datasets and estimate a representation for both domains.
- TSNE projections of embeddings (MNIST→MNIST-M).

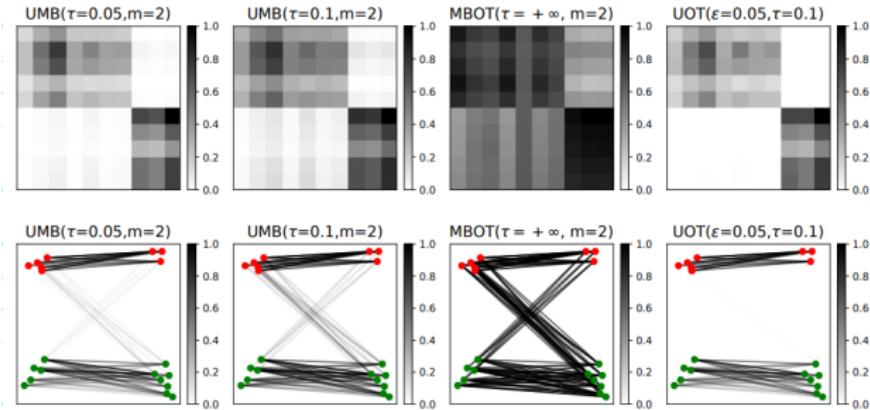
JDOT for large scale deep learning



DeepJDOT [Damodaran et al., 2018]

- Learn simultaneously the embedding g and the classifier f .
- JDOT performed in the joint embedding/label space.
- Use minibatch to estimate OT and update g, f at each iterations [Fatras et al., 2020].
- Scales to large datasets and estimate a representation for both domains.
- TSNE projections of embeddings (MNIST→MNIST-M).

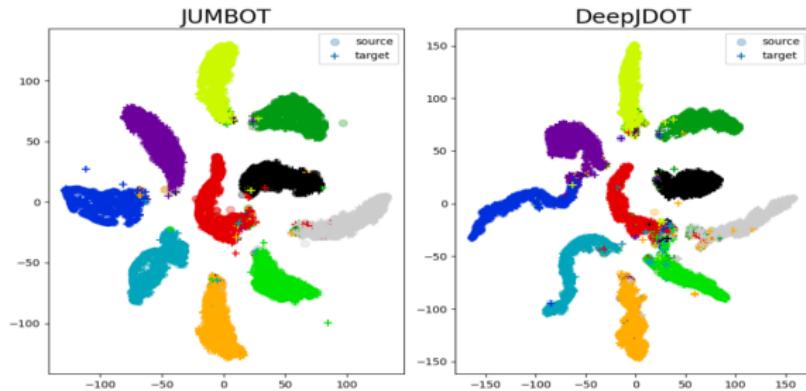
JUMBOT: DeepJDOT for unbalanced and noisy data



JUMBOT [Fatras et al., 2021b]

- Main idea : JDOT with minibatches and Unbalanced OT.
- Theoretical proof of robustness to outliers (UOT is upper bounded, not OT).
- Experiment on Partial DA (some classes are not in target) show robustness to different class proportions between domains.
- Better ability to handle sampling noise on minibatch because good performance on small minibatch size.

JUMBOT: DeepJDOT for unbalanced and noisy data



JUMBOT [Fatras et al., 2021b]

- Main idea : JDOT with minibatches and Unbalanced OT.
- Theoretical proof of robustness to outliers (UOT is upper bounded, not OT).
- Experiment on Partial DA (some classes are not in target) show robustness to different class proportions between domains.
- Better ability to handle sampling noise on minibatch because good performance on small minibatch size.

Outline

Introduction

Mapping with optimal transport

Optimal transport mapping estimation

Optimal transport for domain adaptation

Learning from histograms with Optimal Transport

Unsupervised learning

Supervised learning

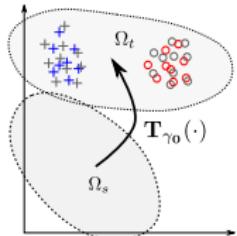
Learning from empirical distributions with Optimal Transport

Unupervised learning

Supervised learning and domain adaptation

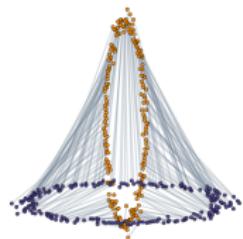
Conclusion

Three aspects of optimal transport



Transporting with optimal transport

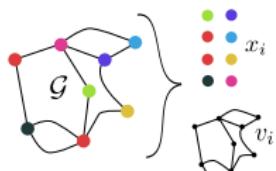
- Learn to map between distributions.
- Estimate a smooth mapping from discrete distributions.
- Applications in domain adaptation.



Divergence between histograms/empirical distributions

- Use the ground metric to encode complex relations between the bins of histograms for data fitting.
- OT losses are non-parametric divergences between non overlapping distributions.
- Used to train minimal Wasserstein estimators.

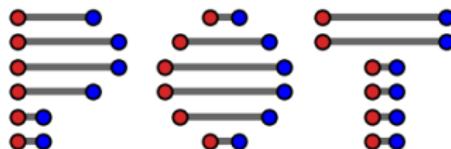
Divergence between structured objects and spaces



- Modeling of structured data and graphs as distribution.
- OT losses (Wass. or (F)GW) measure similarity between distributions/objects.
- OT find correspondance across spaces for adaptation.

Thank you

Python code available on GitHub:



<https://github.com/PythonOT/POT>

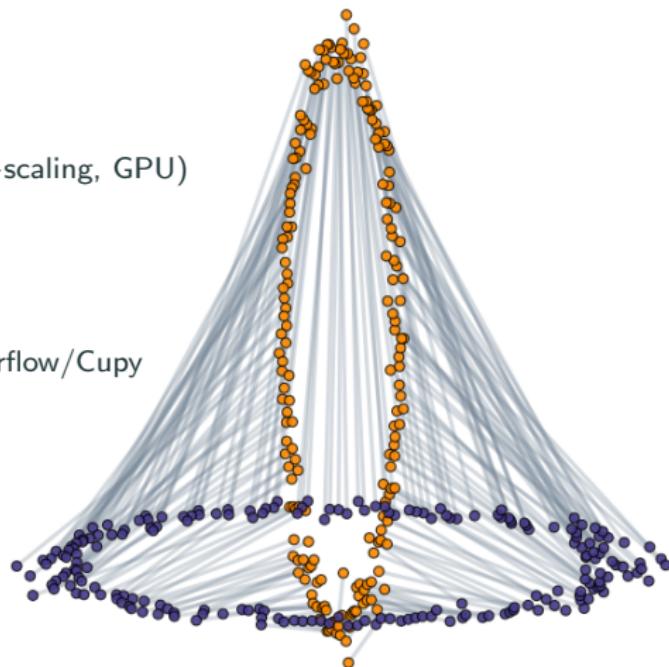
- OT LP solver, Sinkhorn (stabilized, ϵ -scaling, GPU)
- Domain adaptation with OT.
- Barycenters, Wasserstein unmixing.
- Gromov Wasserstein.
- Solvers for Numpy/Pytorch/Jax/tensorflow/Cupy

Tutorial on OT for ML:

<http://tinyurl.com/otml-isbi>

Papers available on my website:

<https://remi.flamary.com/>



References i

[Amos et al., 2017] Amos, B., Xu, L., and Kolter, J. Z. (2017).

Input convex neural networks.

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

[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.

[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.

References ii

[Chambon et al., 2018] Chambon, S., Galtier, M. N., and Gramfort, A. (2018).

Domain adaptation with optimal transport improves eeg sleep stage classifiers.

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

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

Learning wasserstein embeddings.

[Courty et al., 2017b] Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017b).

Joint distribution optimal transportation for domain adaptation.

In *Neural Information Processing Systems (NIPS)*.

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

Optimal transport for domain adaptation.

Pattern Analysis and Machine Intelligence, IEEE Transactions on.

References iii

[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.

[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)*.
- [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).
- [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.

References v

[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.

[Ganin et al., 2016] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016).

Domain-adversarial training of neural networks.

Journal of Machine Learning Research, 17(59):1–35.

[Gautheron et al., 2017] Gautheron, L., Lartizien, C., and Redko, I. (2017).

Domain adaptation using optimal transport: application to prostate cancer mapping.

[Gayraud et al., 2017] Gayraud, N. T., Rakotomamonjy, A., and Clerc, M. (2017).

Optimal transport applied to transfer learning for p300 detection.

In *BCI 2017-7th Graz Brain-Computer Interface Conference*, page 6.

[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.

[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.

[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.

[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.

- [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.
- [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.

References viii

[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.

[Pérez et al., 2003] Pérez, P., Gangnet, M., and Blake, A. (2003).

Poisson image editing.

ACM Trans. on Graphics, 22(3).

[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)*.

[Pooladian and Niles-Weed, 2021] Pooladian, A.-A. and Niles-Weed, J. (2021).

Entropic estimation of optimal transport maps.

arXiv preprint arXiv:2109.12004.

[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.

[Redko et al., 2020] Redko, I., Vayer, T., Flamary, R., and Courty, N. (2020).

Co-optimal transport.

In *Neural Information Processing Systems (NeurIPS)*.

[Rivet et al., 2009] Rivet, B., Souloumiac, A., Attina, V., and Gibert, G. (2009).

xdawn algorithm to enhance evoked potentials: application to brain-computer interface.

IEEE Transactions on Biomedical Engineering, 56(8):2035–2043.

[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.

References x

- [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.
- [Schiebinger et al., 2019] Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019).
Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming.
Cell, 176(4):928–943.
- [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.

References xi

- [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.
- [Shen et al., 2018] Shen, J., Qu, Y., Zhang, W., and Yu, Y. (2018).
Wasserstein distance guided representation learning for domain adaptation.
In *AAAI Conference on Artificial Intelligence*.
- [Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014).
Very deep convolutional networks for large-scale image recognition.
arXiv preprint arXiv:1409.1556.
- [Sun and Saenko, 2016] Sun, B. and Saenko, K. (2016).
Deep CORAL: Correlation Alignment for Deep Domain Adaptation, pages 443–450.
Springer International Publishing, Cham.
- [Tzeng et al., 2014] Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. (2014).
Deep domain confusion: Maximizing for domain invariance.
arXiv preprint arXiv:1412.3474.

- [Yan et al., 2018] Yan, Y., Li, W., Wu, H., Min, H., Tan, M., and Wu, Q. (2018).
Semi-supervised optimal transport for heterogeneous domain adaptation.
In *IJCAI*, pages 2969–2975.
- [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.