Optimal transport for machine learning

Learning with optimal transport

Rémi Flamary

April 8 2019

Tutorial ISBI 2019, Venice, Italy

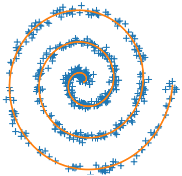
http://tinyurl.com/otml-isbi

Introduction

Three aspects of Machine Learning

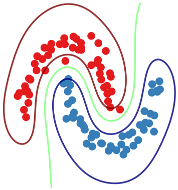
Unsupervised learning

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



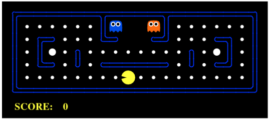
Supervised Learning

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

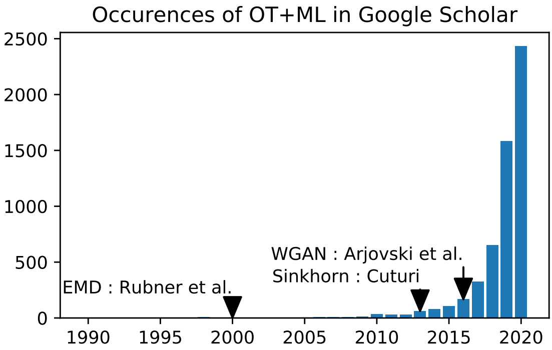


Reinforcement Learning

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



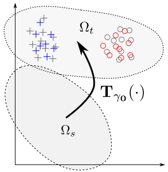
Optimal transport for machine learning

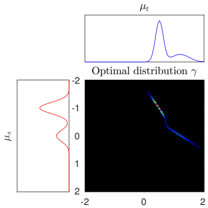


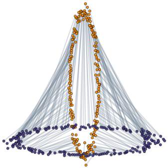
Short history of OT for ML

- 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







Transporting with optimal transport

- Color adaptation in image [Ferradans et al., 2014].
- Domain adaptation [Courty et al., 2016].
- OT mapping estimation [Perrot 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]
- Loss for spectral unmixing [Flamary et al., 2016b].

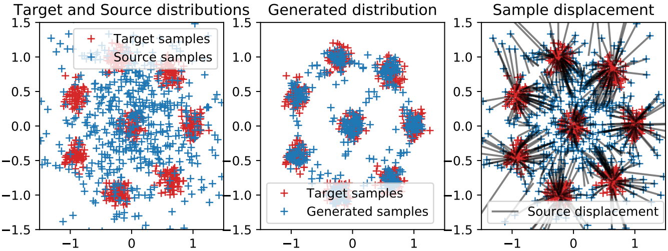
Divergence between empirical distributions

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

Table of content

Mapping with optimal transport

Mapping with optimal transport



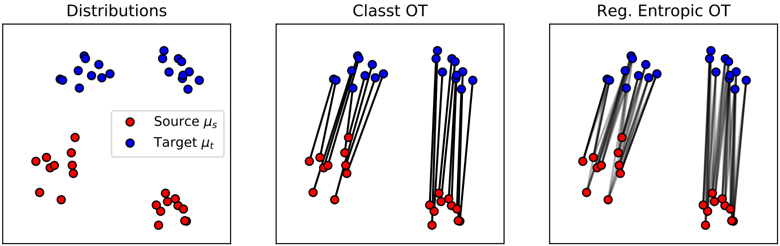
Mapping estimation

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

Why map ?

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

Transporting the discrete samples

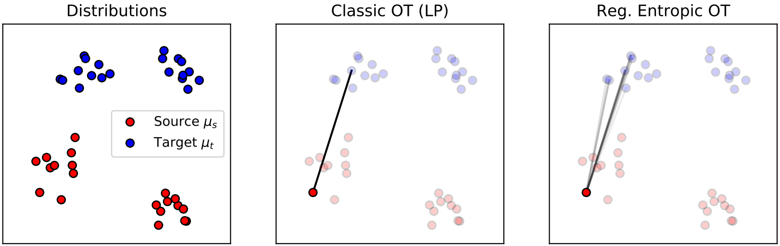


Barycentric mapping [Ferradans et al., 2014]

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

- 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

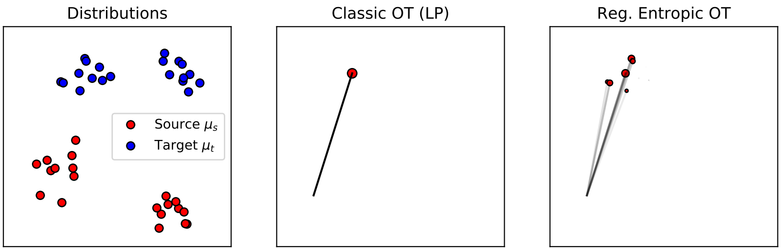


Barycentric mapping [Ferradans et al., 2014]

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

- 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

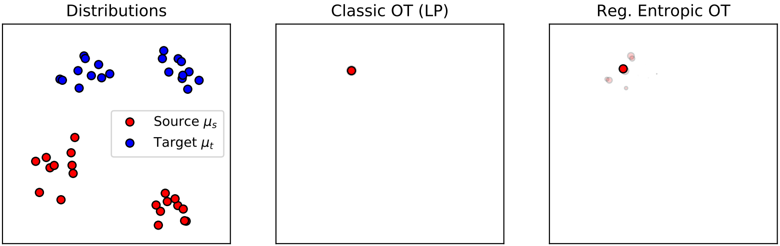


Barycentric mapping [Ferradans et al., 2014]

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

- 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.

Transporting the discrete samples

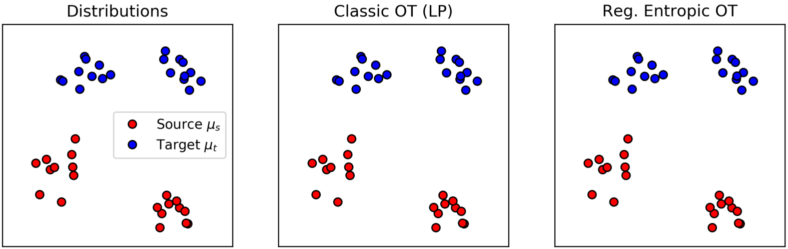


Barycentric mapping [Ferradans et al., 2014]

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

- 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.

Transporting the discrete samples



Barycentric mapping [Ferradans et al., 2014]

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

- 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.

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

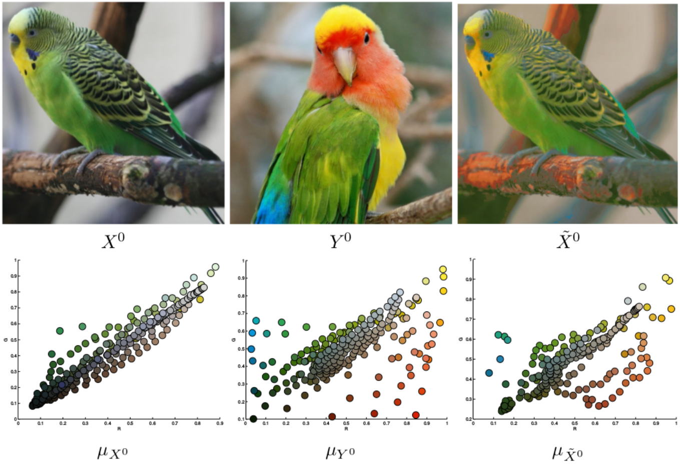
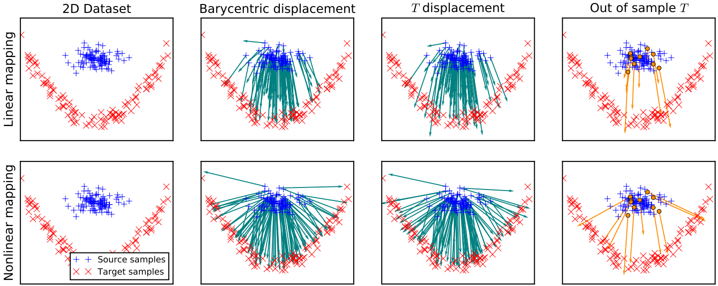


Image colorization [Ferradans et al., 2014]

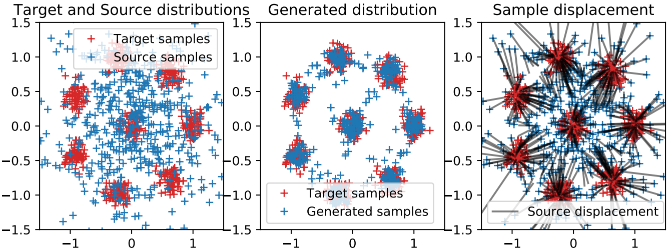
Joint OT and mapping estimation



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

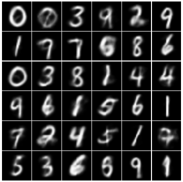
$$\min_{T,\boldsymbol{\gamma}\in\mathcal{P}} \quad \langle \boldsymbol{\gamma}, \mathbf{C}\rangle_F + \sum_i \|T(\mathbf{x}_i^s) - \hat{T}_{\boldsymbol{\gamma}}(\mathbf{x}_i^s)\|^2 + \lambda\|T\|^2$$

- 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.

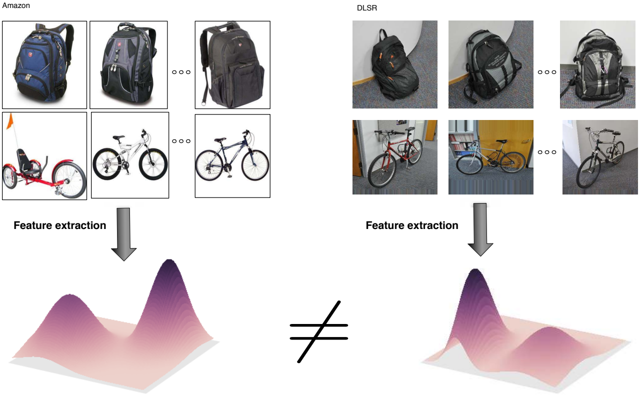


Large scale mapping estimation [Seguy et al., 2017]

- 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 OT and mapping for small regularization.



Domain Adaptation problem

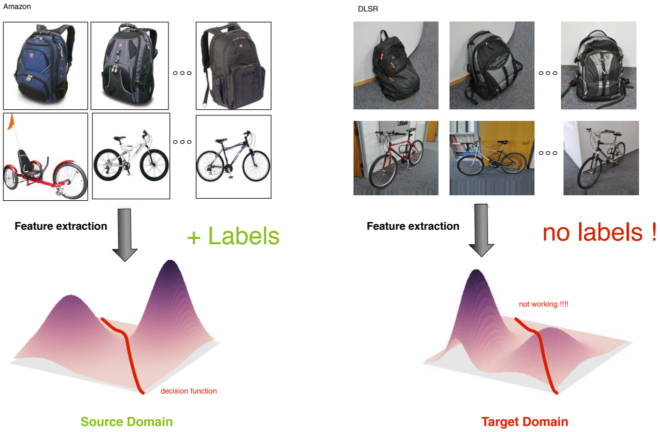


Probability Distribution Functions over the domains

Our context

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

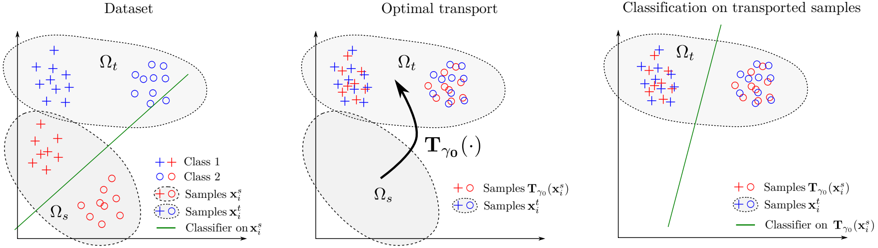
Unsupervised domain adaptation problem



Problems

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

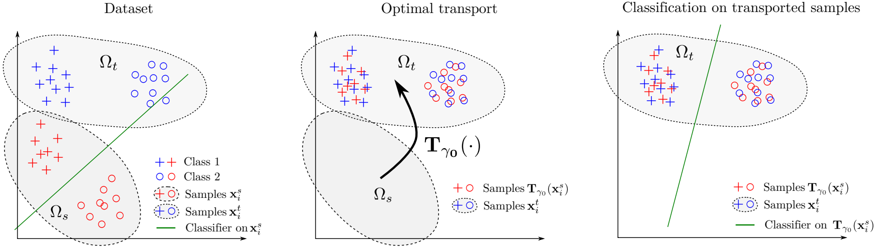
OT for domain adaptation : Step 1



Step 1 : Estimate optimal transport between distributions.

- Choose the ground metric (squared euclidean in our experiments).
- Using regularization allows
  - Large scale and regular OT with entropic regularization [Cuturi, 2013].
  - Class labels in the transport with group lasso [Courty et al., 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



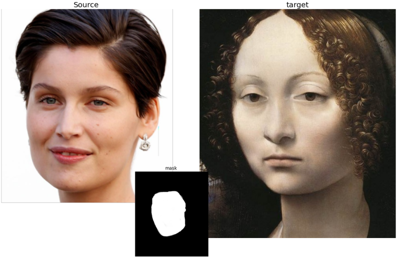
Step 2 : Transport the training samples onto the target distribution.

- 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.

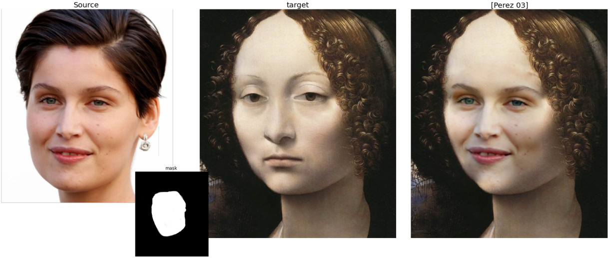
Seamless copy in images



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



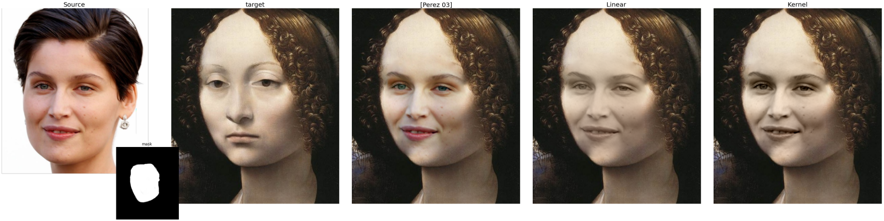
Poisson image editing [Pérez et al., 2003]

- 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]

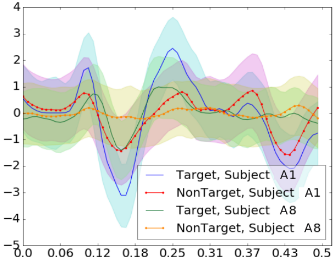
- Use the color gradient from the source image.
- Use color border conditions on the target image.
- Solve Poisson equation to reconstruct the new image.

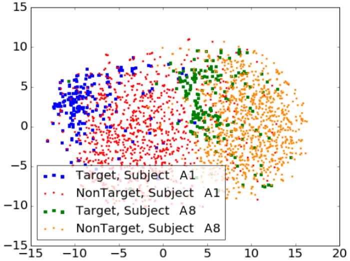
Seamless copy with gradient adaptation [Perrot et al., 2016]

- 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.

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

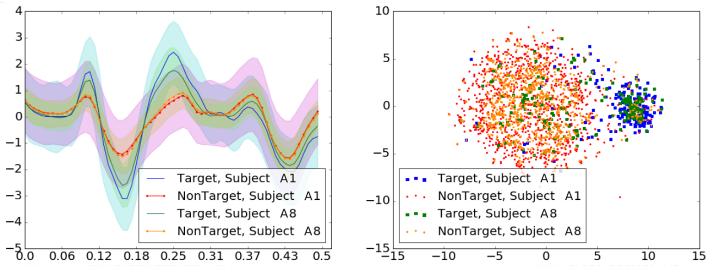
OTDA for biomedical data (1)





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.



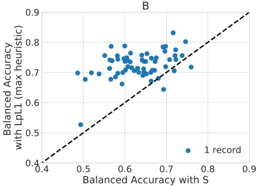
Multi-subject P300 classification [Gayraud et al., 2017]

- 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)

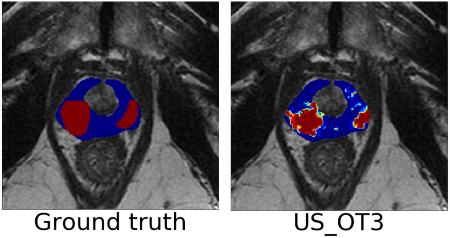
EEG sleep stage classification [Chambon et al., 2018]

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

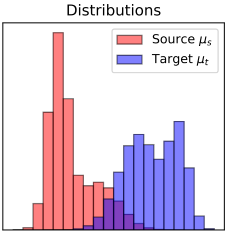


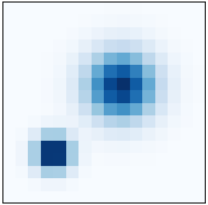
Prostace cancer classification [Gautheron et al., 2017]

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



Learning from histograms with Optimal Transport

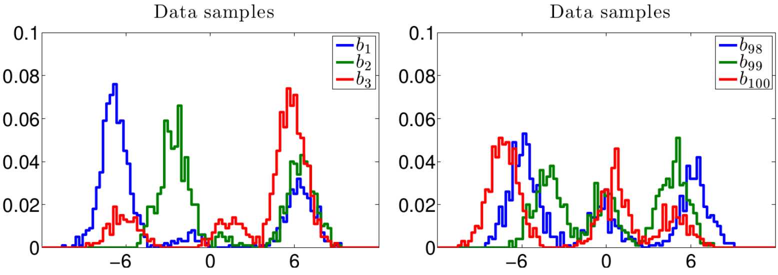




Data as histograms

- 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 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

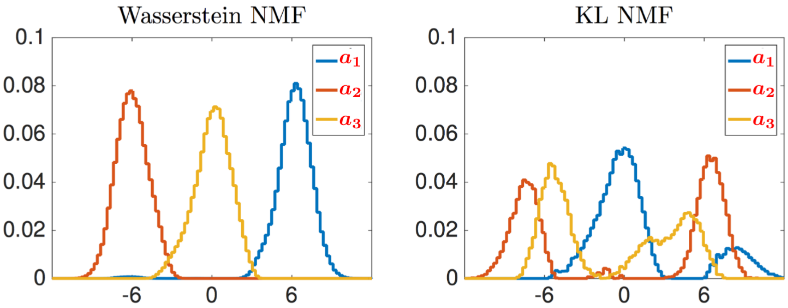


DL with Wasserstein distance [Sandler and Lindenbaum, 2011]

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

- 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

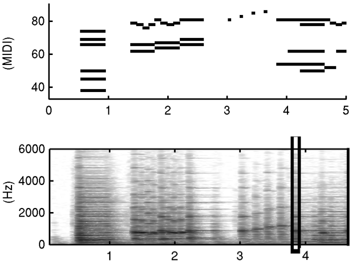


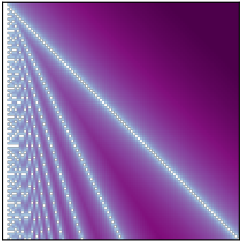
DL with Wasserstein distance [Sandler and Lindenbaum, 2011]

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

- 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].

Optimal Spectral Transportation (OST)



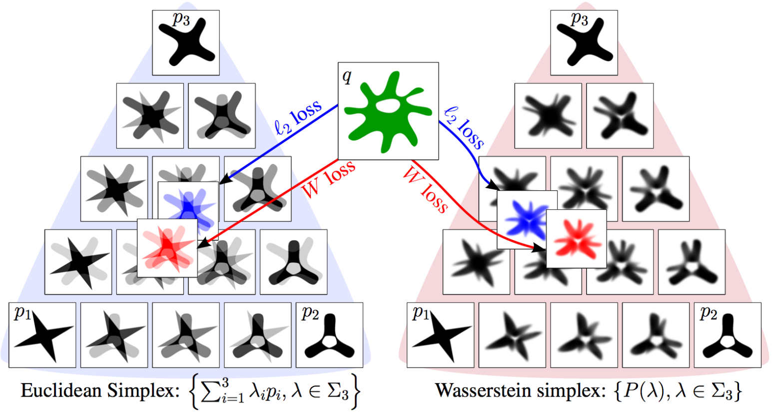


OT linear spectral unmixing of musical data [Flamary et al., 2016b]

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

- 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`

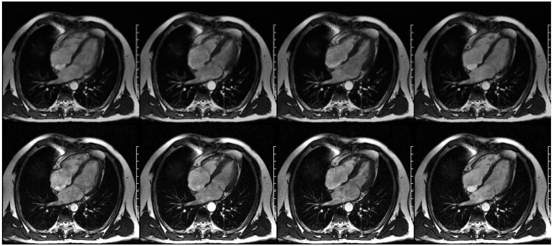


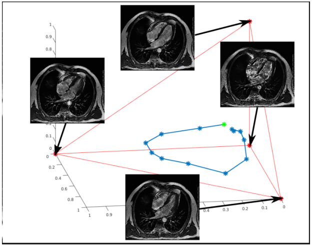
Nonlinear unmixing with Wasserstein simplex [Schmitz et al., 2017]

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

with $WB(\mathbf{D}, \mathbf{h}) = \arg\min_{\mathbf{a}} \sum_i h_i W_{\mathbf{C}}(\mathbf{d_i}, \mathbf{a})$

Wasserstein dictionary learning (2)





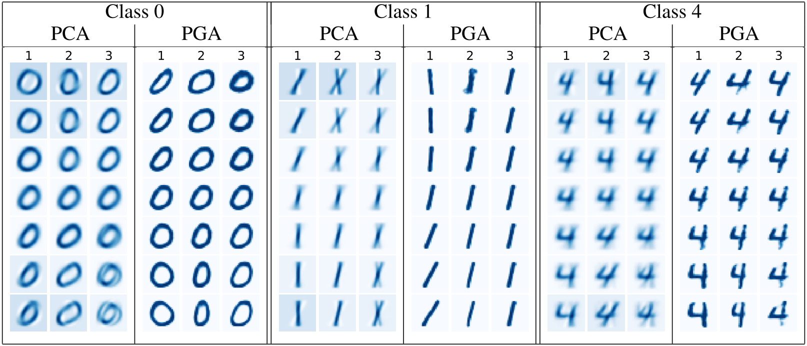
Nonlinear unmixing with Wasserstein simplex [Schmitz et al., 2017]

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

with $WB(\mathbf{D}, \mathbf{h}) = \arg\min_{\mathbf{a}} \sum_i h_i W_{\mathbf{C}}(\mathbf{d_i}, \mathbf{a})$

- 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.

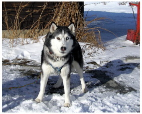
Principal Geodesics Analysis



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

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

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.

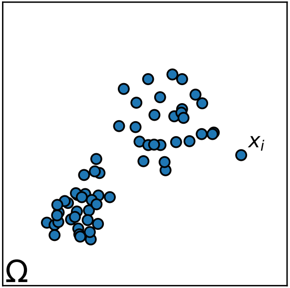
Learning from empirical distributions with Optimal Transport

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.

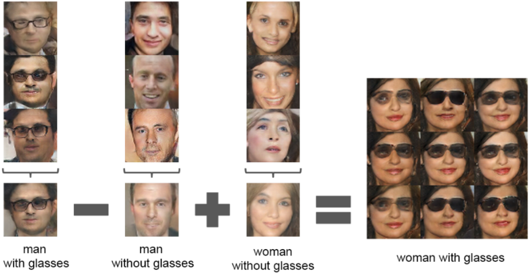


Generative Adversarial Networks (GAN)

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]).

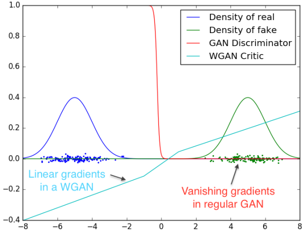


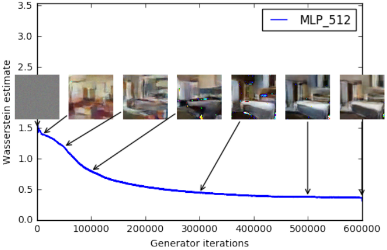
Generative Adversarial Networks (GAN) [Goodfellow et al., 2014]

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

- 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]

$$\min_{G} \quad W_1^1(G\#\mu_z, \mu_d), \tag{3}$$

- 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).

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

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

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

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

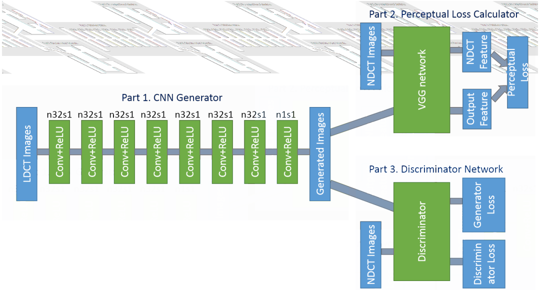
- Actually **not** equivalent to solve the optimal transport, but gradients are aligned.

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



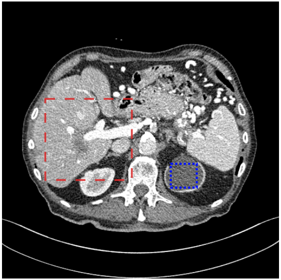
Reconstructing low dose CT images [Yang et al., 2018]

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

- 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]

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

- 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

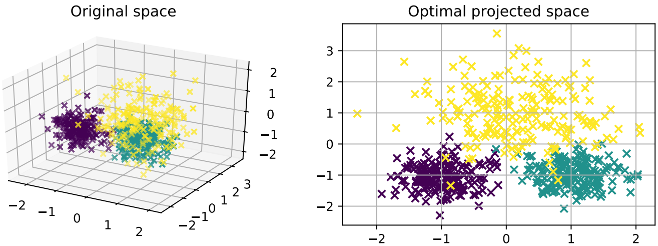
Full dose | Quarter dose | WGAN-VGG rec.

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

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

- 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.

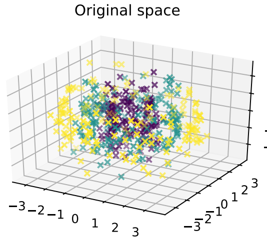
Wasserstein Discriminant Analysis (WDA)

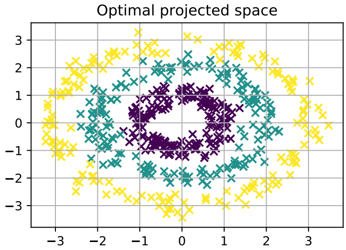


$$\max_{\mathbf{P}\in\mathcal{S}} \quad \frac{\sum_{c,c'>c} W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^{c'})}{\sum_c W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^c)} \quad (5)$$

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

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



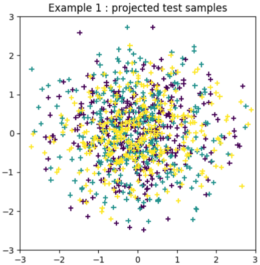


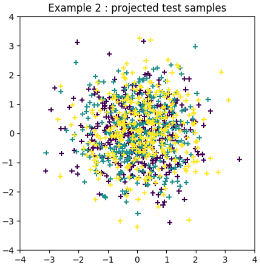
$$\max_{\mathbf{P}\in\mathcal{S}} \quad \frac{\sum_{c,c'>c} W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^{c'})}{\sum_c W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^c)} \tag{5}$$

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

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

Wasserstein Discriminant Analysis (WDA)



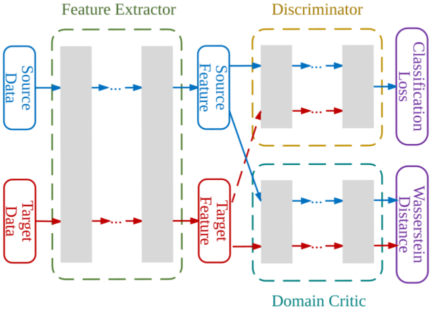


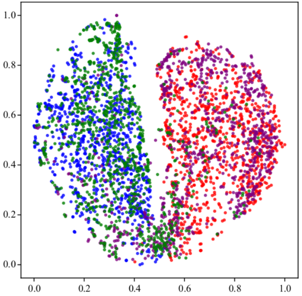
$$\max_{\mathbf{P}\in\mathcal{S}} \quad \frac{\sum_{c,c'>c} W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^{c'})}{\sum_c W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^c)} \quad (5)$$

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

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

Domain adaptation with Wasserstein distance



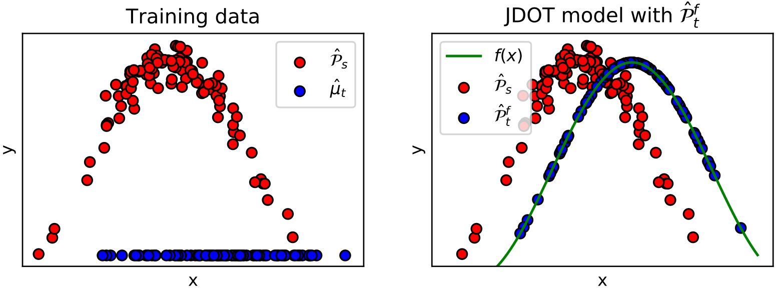


(d) t-SNE of WDGRL features

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].

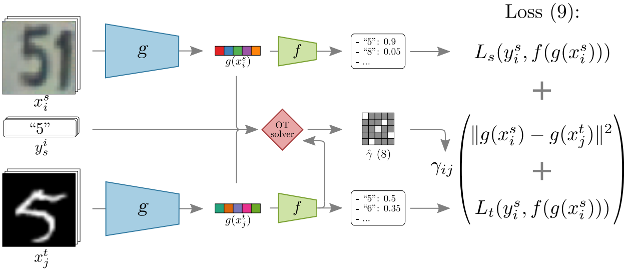
Joint Distribution Optimal Transport for DA



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

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

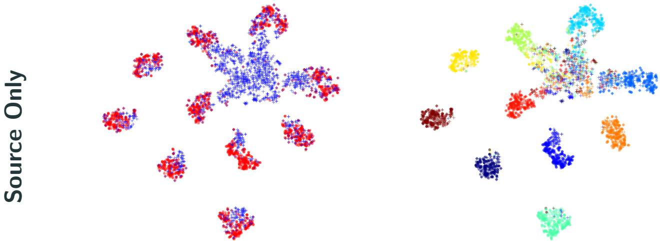
- $\hat{\mathcal{P}}_t{}^f = \frac{1}{N_t}\sum_{i=1}^{N_t} \delta_{\mathbf{x}_i^t, f\mathbf{x}_i^t}$ is the proxy joint feature/label distribution.
- $\mathcal{D}(\mathbf{x}_i^s, y_i^s; \mathbf{x}_j^t, f(\mathbf{x}_j^t)) = \alpha\|\mathbf{x}_i^s - \mathbf{x}_j^t\|^2 + \mathcal{L}(y_i^s, f(\mathbf{x}_j^t))$ 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.



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.
- Scales to large datasets and estimate a representation for both domains.
- TSNE projections of embeddings (MNIST→MNIST-M).

JDOT for large scale deep learning

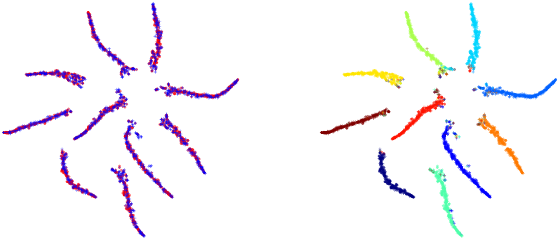


DeepJDOT [Damodaran et al., 2018]

- 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.
- 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

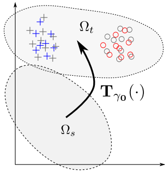


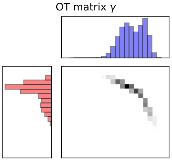
DeepJDOT [Damodaran et al., 2018]

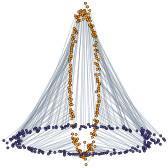
- Learn simultaneously the embedding g and the classifier f.
- JDOT performed in the joint embedding/label space.
- Use minibatch to estimate OT and update g, f at each iterations.
- Scales to large datasets and estimate a representation for both domains.
- TSNE projections of embeddings (MNIST$\rightarrow$MNIST-M).

Conclusion

Optimal transport for machine learning







Mapping with optimal transport

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

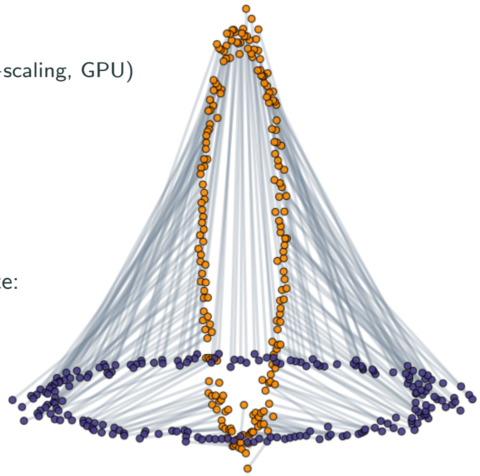
Learning with optimal transport

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

Thank you

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

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



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

Post docavailable in Nice (France)

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

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. (2015).
Iterative Bregman projections for regularized transportation problems.
SISC.

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

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

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

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

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

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

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

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

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

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

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, L., Lartizien, C., and Redko, I. (2017).
Domain adaptation using optimal transport: application to prostate cancer mapping.

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.

References vi

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, 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.

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

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

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

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, 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.

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.

Schmitz, M. A., Heitz, M., Bonneel, N., Mboula, F. M. N., Coeurjolly, D., Cuturi, M., Peyré, G., and Starck, J.-L. (2017).
Wasserstein dictionary learning: Optimal transport-based unsupervised non-linear dictionary learning.
arXiv preprint arXiv:1708.01955.

Seguy, V., Bhushan Damodaran, B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. (2017).
Large-scale optimal transport and mapping estimation.

Seguy, V. and Cuturi, M. (2015).
Principal geodesic analysis for probability measures under the optimal transport metric.
In *Advances in Neural Information Processing Systems*, pages 3312–3320.

References x

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

Sun, B. and Saenko, K. (2016).
Deep CORAL: Correlation Alignment for Deep Domain Adaptation, pages 443–450.
Springer International Publishing, Cham.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. (2014).
Deep domain confusion: Maximizing for domain invariance.
arXiv preprint arXiv:1412.3474.

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