Optimal Transport for Machine Learning

R. Flamary - Lagrange, OCA, CNRS, Université Côte d'Azur

Habilitation à Diriger des Recherches
November 29 2019

Table of content

Curriculum Vitae

Education and current position

2007 - 2008 — **Engineer + Master degrees**, Electrical Engineering, *INSA de Lyon*
Major : Signal and image processing.

2008 - 2012 — **PhD + Assistant Professor (1/2 ATER)**, *Université de Rouen*
UFR des Sciences et Techniques, Laboratoire LITIS EA 4108.
Subject : *Machine learning for signal processing : applications to Brain-Computer Interfaces*.

2012 - — **Associate Professor (MCF)**, *Université de Nice Sophia Antipolis*
UFR des Sciences, Département of Electronics,
Observatoire de la Côte d'Azur
Laboratoire Lagrange
Université Côte d'Azur

Teaching activities

- $\approx$ 192h EDTD / year since 2014.
- Creation of courses slides and practical sessions.
- All support available on website.
- Organization of Kaggle competitions.

Courses (2012 - 2019)

- Signals and Systems (L)
- Random processes (L)
- Numerical methods in C (L)
- Statistical learning and BCI (M)
- Signal processing and applications (M)
- Theory of Machine Learning (M)

Administrative tasks

- **Coordinator of License 3 Electronics,** 2017-2019, Resp: Planning, Jury, Admission in L3
- **Coordinator of Competency-based learning**, Since 2017, Resp: Define, write and evaluate competencies for L and M in Electronics.

Machine learning and numerical optimization

- Large scale sparse numerical optimization for machine learning
- Multi-task and transfer learning
- Optimal transport for machine learning (since 2014)

Applications

- Biomedical data processing (BCI, CAD, Spike sorting)
- Remote sensing (image classification, label noise)
- Astronomy (image processing, coronagraphy)

Research projects

- **Chair 3IA Côte d'Azur**, 2019-2023
- **OATMIL**, ANR Project 2017-2020, Local PI
  Optimal transport for machine learning.
- **AMOR**, Young researcher project GDR ISIS 2013-2014, PI

\+ Magellan, ON FIRE, TOPASE, DESTOPT, HYPANEMA

PhD Students

- **Kilian Fatras**, with N. Courty, Université Bretagne Sud, 2018-2021.
  Optimal Transport and deep learning,
- **Laurent Dragoni**, with K. Lounici and P. Bouret, UCA, 2017-2020.
  Spike sorting for massive neurophysiological data sets,
- **Raphael Rougeot**, with D. Mary and C. Aime, UCA / ESA, 2017-2020.
  Modeling and computation of diffraction effects for end-to-end performance of hight-contrast space optical instruments,
- **Ibrahim El Khalil Harrane**, with C. Richard, UCA, 2015-2019 (June 21).
  Distributed estimation over multitask networks,

Other collaborations

- 4 Master's internships supervision.
- Past and current collaboration with other PhD students:
  R. Turrisi, T. Vayer, M. Ducoffe, P. Hartley, L. Laporte.

Publications (since 2013)

- 17 International Journal papers (4 A&A, 2 ML, 2 TNNLS, 1 TPAMI).
- 32 International Conference papers (4 NeurIPS, 2 ICLR, 1 ICML).
- 3 Book chapters, 1 book as editor.

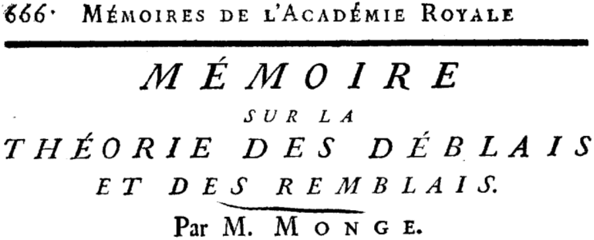
Organized scientific events

- **Optimal Tranport for Machine Learning Workshop**, NeurIPS 2019.
- **Basmati CNRS Summer School**, 2015 and 2018.
- **GDR ISIS**, 2 meetings, leader of specific action for Theme A.

Reproducible research

- POT Python Optimal Transport Toolbox (100k+ downloads).
- More than 35 publications with provided open source code.

Introduction


666. MÉMOIRES DE L'ACADÉMIE ROYALE

MÉMOIRE

SUR LA

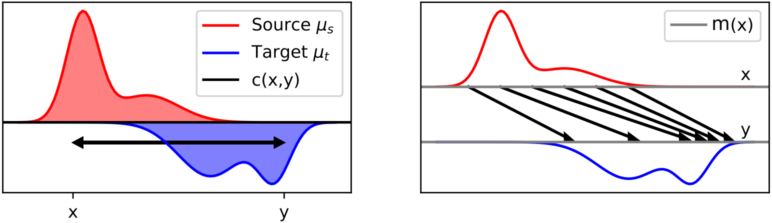
THÉORIE DES DÉBLAIS

ET DES REMBLAIS.

Par M. MONGE.

Problem [Monge, 1781]

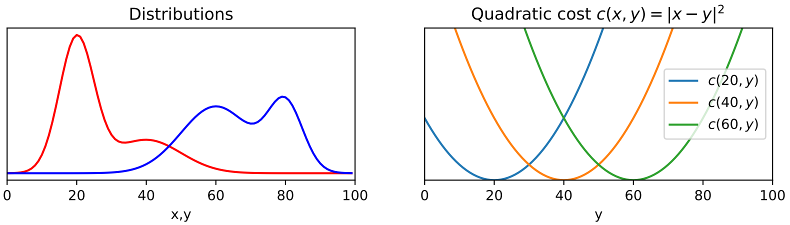
- How to move dirt from one place (déblais) to another (remblais) while minimizing the effort ?
- Find a mapping m between the two distributions of mass (transport).
- Optimize with respect to a displacement cost $c(x, y)$ (optimal).



Problem [Monge, 1781]

- How to move dirt from one place (déblais) to another (remblais) while minimizing the effort ?
- Find a mapping m between the two distributions of mass (transport).
- Optimize with respect to a displacement cost $c(x, y)$ (optimal).

Optimal transport (Monge formulation)



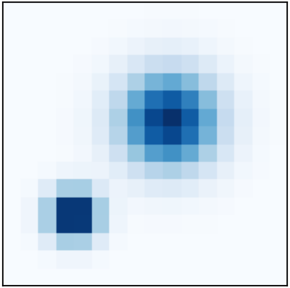
- Probability measures μ_s and μ_t on and a cost function $c : \Omega_s \times \Omega_t \rightarrow \mathbb{R}^+$.
- The Monge formulation [Monge, 1781] aim at finding a mapping $m : \Omega_s \rightarrow \Omega_t$

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

- Non convex problem because of the constraint $m\#\mu_s = \mu_t$.
- [Brenier, 1991] proved existence and unicity of the Monge map for $c(x,y) = \|x-y\|^2$ and distributions with densities.
- What about discrete distribution?

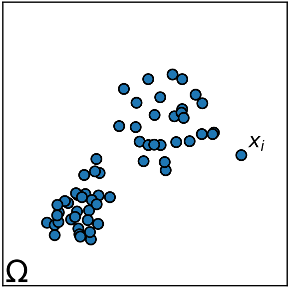
Discrete measure: $\mu = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}, \quad \mathbf{x}_i \in \Omega, \quad \sum_{i=1}^{n} a_i = 1$

Eulerian (histograms)



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

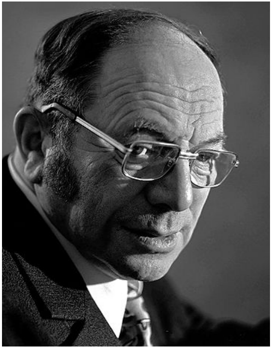
Lagrangian (point clouds, empirical)

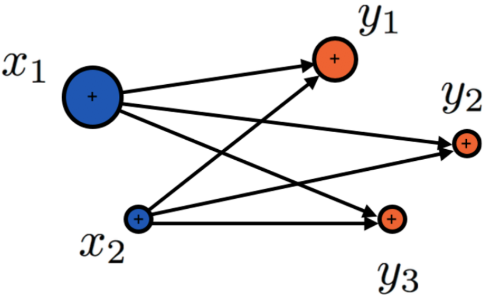


- Constant weight: $a_i = \frac{1}{n}$
- Quotient space: Ω^n, Σ_n

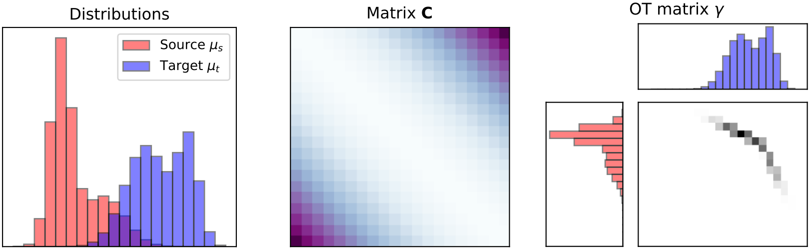
Inspired from Gabriel Peyré

Kantorovich relaxation





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



Kantorovitch formulation : OT Linear Program

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

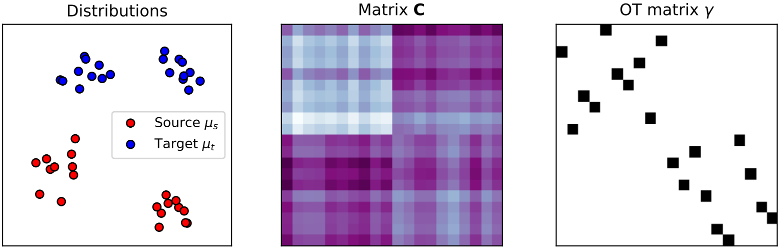
$$\boldsymbol{T}_0 = \underset{\boldsymbol{T} \in \Pi(\mu_s, \mu_t)}{\operatorname{argmin}} \left\{ \langle \boldsymbol{T}, \mathbf{C} \rangle_F = \sum_{i,j} T_{i,j} c_{i,j} \right\}$$

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

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

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

Optimal transport with discrete distributions



Kantorovitch formulation : OT Linear Program

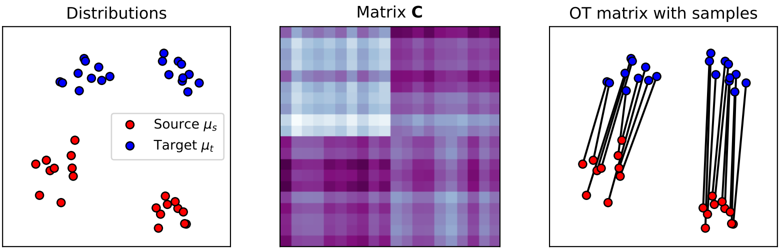
When $\mu_s = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i^s}$ and $\mu_t = \sum_{i=1}^{n} b_i \delta_{\mathbf{x}_i^t}$

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

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

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

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



Kantorovitch formulation : OT Linear Program

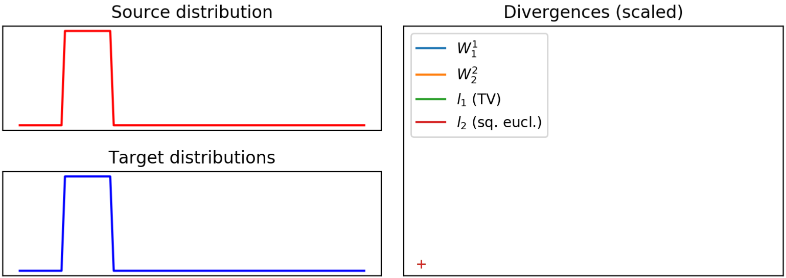
When $\mu_s = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i^s}$ and $\mu_t = \sum_{i=1}^{n} b_i \delta_{\mathbf{x}_i^t}$

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

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

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

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



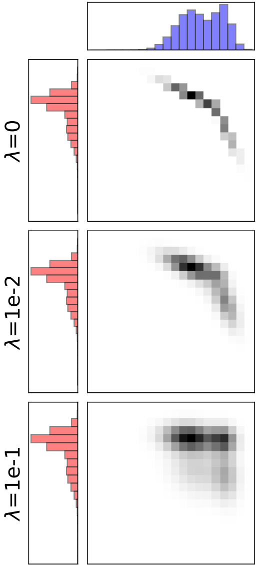
Wasserstein distance

$$W_p^p(\mu_s, \mu_t) = \min_{\gamma \in \Pi(\mu_s, \mu_t)} \int_{\Omega_s \times \Omega_t} \|\mathbf{x} - \mathbf{y}\|^p \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \tag{2}$$

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

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

$$\boldsymbol{T}_0^\lambda = \underset{\boldsymbol{T}\in\mathcal{P}}{\text{argmin}} \quad \langle \boldsymbol{T}, \mathbf{C}\rangle_F + \lambda\Omega(\boldsymbol{T}), \tag{3}$$

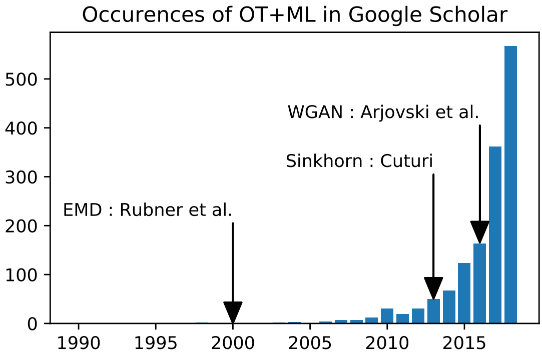


Regularization term $\Omega(\boldsymbol{T})$

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

Why regularize?

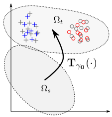
- Smooth the "distance" estimation:
  $$W_\lambda(\mu_s, \mu_t) = \langle \boldsymbol{T}_0^\lambda, \mathbf{C}\rangle_F$$
- Encode prior knowledge on the data.
- Better posed problem (strict convexity, stability).
- Better statistical property (sample complexity).
- Fast algorithms to solve the OT problem (Sinkhorn).



Short history of OT for ML

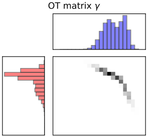
- Recently reintroduced to ML (well known in image processing since 2000s).
- Computational OT allow numerous applications (regularization).
- Deep learning boost (numerical optimization and GAN).

Four aspects of Optimal Transport for Machine Learning



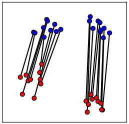
Mapping with optimal transport

- Continuous mapping estimation [Perrot et al., 2016, Flamary et al., 2019].
- Domain adaptation [Courty et al., 2016].



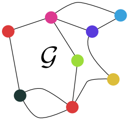
Divergence between histograms

- Invariant ground metric [Flamary et al., 2016].
- Wasserstein embeddings [Courty et al., 2018]



Divergence between empirical distributions

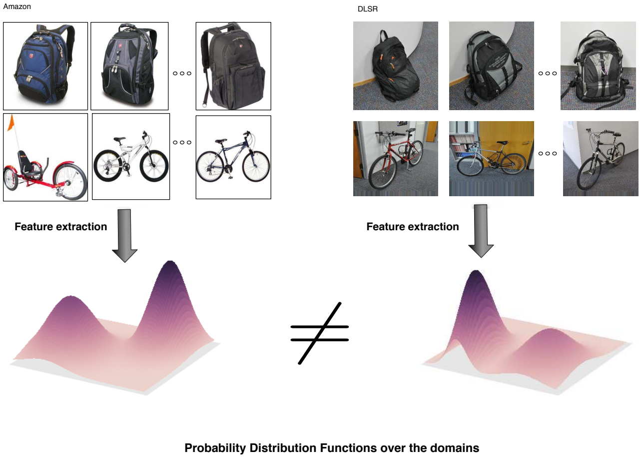
- Estimate discriminant subspace [Flamary et al., 2018].
- Domain adaptation [Courty et al., 2017].



Divergence between structured data

- Modeling labeled graphs as distributions.
- Fused Gromov-Wasserstein divergence [Vayer et al., 2018a].

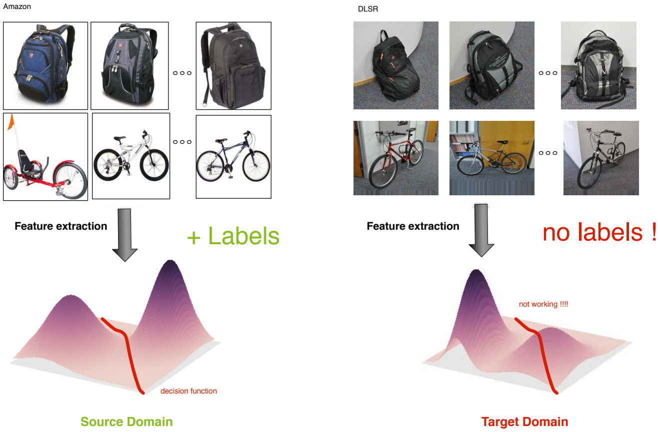
Domain Adaptation problem



Domain adaptation context

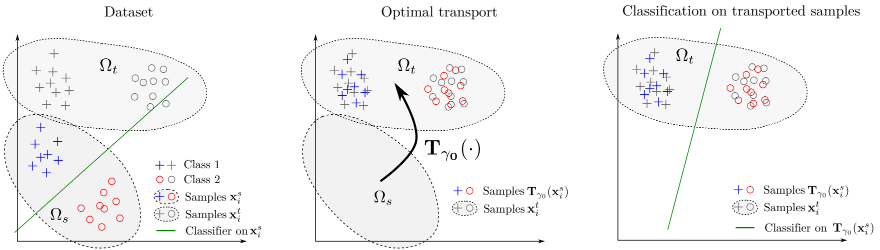
- Classification problem with data coming from different sources (domains).
- Distributions are different but related.

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



Assumptions

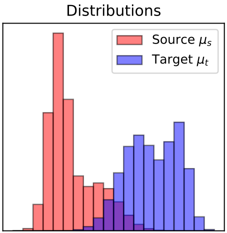
- There exists an OT mapping T in the feature space between the two domains.
- The transport preserves the joint distributions:

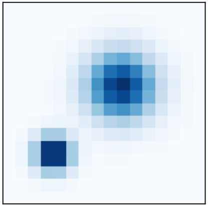
$$\mathcal{P}_s(\mathbf{x}_s, y) = \mathcal{P}_t(T(\mathbf{x}_s), y).$$

3-step strategy [Courty et al., 2016]

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

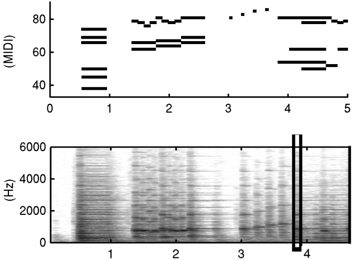
Generalization results under assumptions above [Flamary et al., 2019].

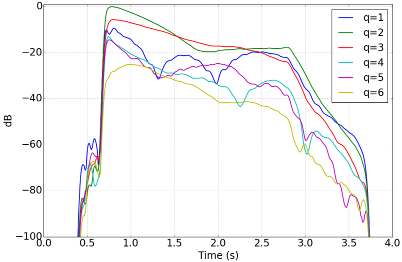




Data as histograms

- 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)!



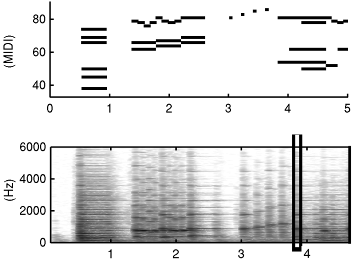


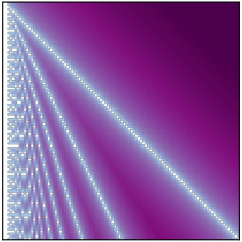
OT linear spectral unmixing of musical data [Flamary et al., 2016]

$$\min_{\mathbf{h}\in\Delta} \quad W_{\mathbf{C}}(\mathbf{v}, \mathbf{D}\mathbf{h}) \tag{4}$$

- 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`



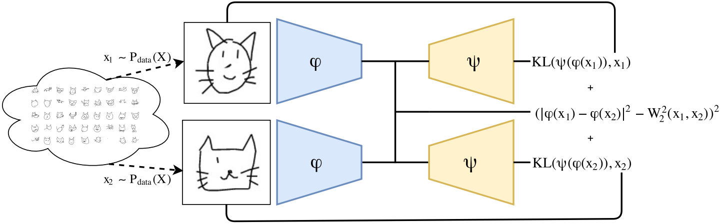


OT linear spectral unmixing of musical data [Flamary et al., 2016]

$$\min_{\mathbf{h}\in\Delta} \quad W_{\mathbf{C}}(\mathbf{v}, \mathbf{D}\mathbf{h}) \tag{4}$$

- 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.

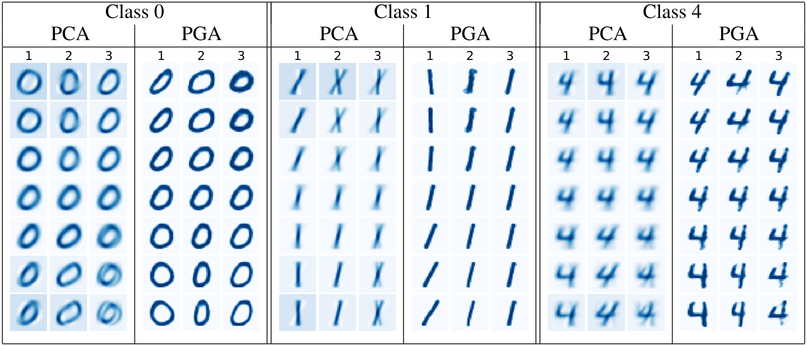
Demo : `https://github.com/rflamary/OST`



Deep Wasserstein Embeddings [Courty et al., 2018]

- Learn a deep embedding φ and decoder ψ for histograms with fixed support.
- Siamese network for Wasserstein metric learning.
- The embedding mimics the behavior of Wasserstein in the original histograms.
- Train a decoder to reconstruct the original histogram.
- Very fast computation of approximate Wasserstein distance and barycenters, PGA.

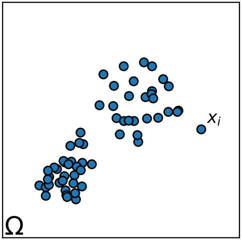
Learning Wasserstein embeddings

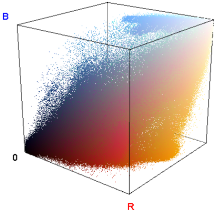


Deep Wasserstein Embeddings [Courty et al., 2018]

- Learn a deep embedding φ and decoder ψ for histograms with fixed support.
- Siamese network for Wasserstein metric learning.
- The embedding mimics the behavior of Wasserstein in the original histograms.
- Train a decoder to reconstruct the original histogram.
- Very fast computation of approximate Wasserstein distance and barycenters, PGA.

Empirical distributions A.K.A datasets

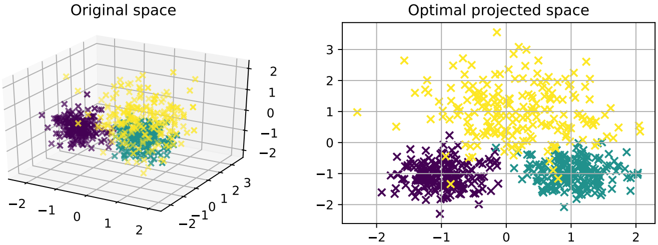




Empirical distribution

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

- Training set of all machine learning approaches.
- Two realizations never overlap.
- How to measure discrepancy?
- Wasserstein distance.

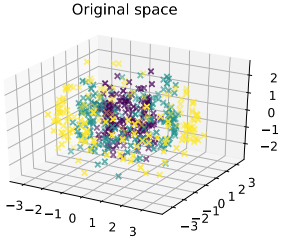


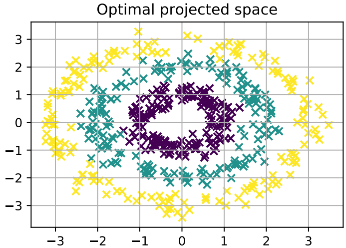
$$\max_{\mathbf{P}\in\Delta} \quad \frac{\sum_{c,c'>c} W_\lambda(\mathbf{P}\#\mu^c, \mathbf{P}\#\mu^{c'})}{\sum_c W_\lambda(\mathbf{P}\#\mu^c, \mathbf{P}\#\mu^c)} \tag{5}$$

- W_λ in entropic reg. OT loss.
- μ^c is distrib. from class c.
- $\mathbf{P}$ is an orthogonal projection;

- Converges toward 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)

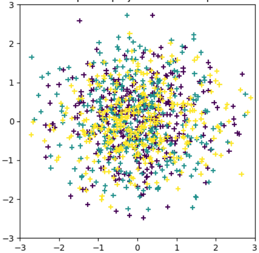


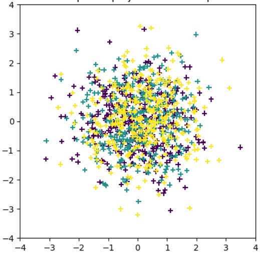


$$\max_{\mathbf{P}\in\Delta} \quad \frac{\sum_{c,c'>c} W_\lambda(\mathbf{P}\#\mu^c, \mathbf{P}\#\mu^{c'})}{\sum_c W_\lambda(\mathbf{P}\#\mu^c, \mathbf{P}\#\mu^c)} \tag{5}$$

- W_λ in entropic reg. OT loss.
- μ^c is distrib. from class c.
- $\mathbf{P}$ is an orthogonal projection;

- Converges toward 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.

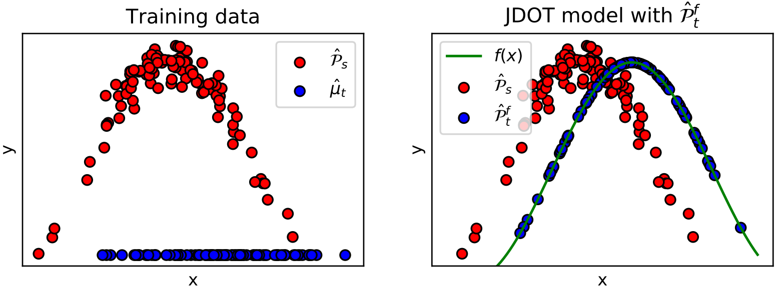




$$\max_{\mathbf{P}\in\Delta} \quad \frac{\sum_{c,c'>c} W_\lambda(\mathbf{P}\#\mu^c, \mathbf{P}\#\mu^{c'})}{\sum_c W_\lambda(\mathbf{P}\#\mu^c, \mathbf{P}\#\mu^c)} \tag{5}$$

- W_λ in entropic reg. OT loss.
- μ^c is distrib. from class c.
- $\mathbf{P}$ is an orthogonal projection;

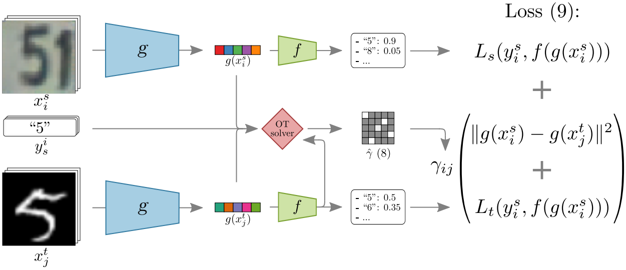
- Converges toward Fisher Discriminant when $\lambda \to \infty$.
- Non parametric method that allows nonlinear discrimination.
- Problem solved with gradient ascent in the Stiefel manifold $\mathcal{S}$.
- Gradient computed using automatic differentiation of Sinkhorn algorithm.



Learning with JDOT [Courty et al., 2017]

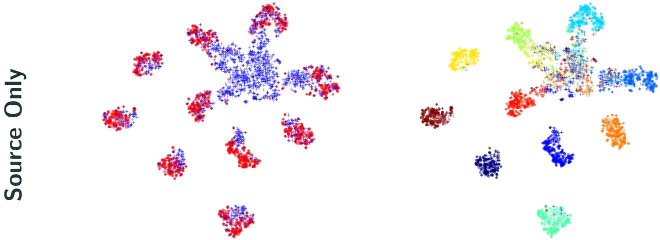
$$\min_f \quad \left\{ W_1(\hat{\mathcal{P}}_s, \hat{\mathcal{P}}_t{}^f) = \inf_{\boldsymbol{T}\in\Pi} \sum_{ij} \mathcal{D}(\mathbf{x}_i^s, \mathbf{y}_i^s; \mathbf{x}_j^t, f(\mathbf{x}_j^t)) T_{i,j} \right\} \tag{6}$$

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



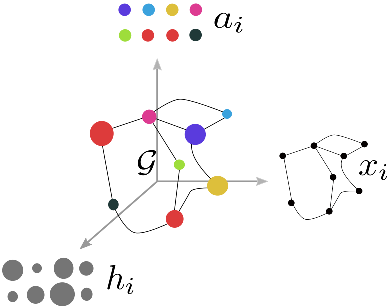
DeepJDOT [Damodaran et al., 2018]

- 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).

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).

Structured data as distributions



$$\mu = \sum_i h_i \delta_{(x_i, a_i)}$$

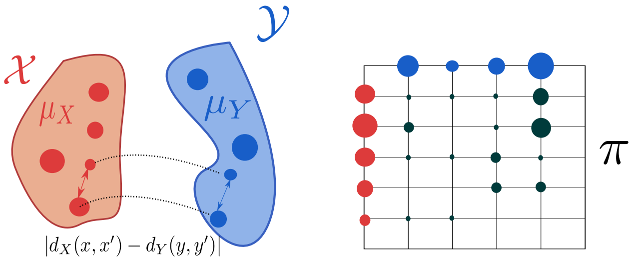
$$\mu_A = \sum_i h_i \delta_{a_i}$$

$$\mu_X = \sum_i h_i \delta_{x_i}$$

Graph data representation

$$\color{red}{\mu_s} = \sum_{i=1}^{n} h_i \delta_{(x_i a_i)} \qquad \color{blue}{\mu_t} = \sum_{j=1}^{m} g_j \delta_{y_j, b_j}$$

- Nodes are weighted by their mass h_i and g_j.
- Features values a_i and b_j can be compared through the common metric
- Relationship between nodes is encoded through $\|x_i - x_j\|$ (shortest path).
- But no common between the structure points x_i and y_j across graphs.

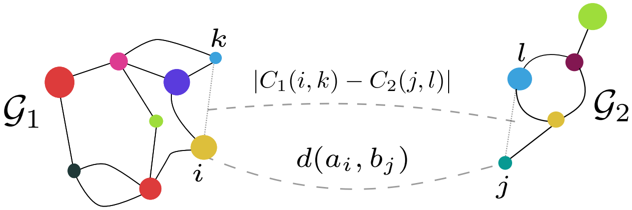


Inspired (again) from Gabriel Peyré

GW distance [Mémoli, 2011]
$\mathcal{X} = (X, d_X, \mu_X)$ and $\mathcal{Y} = (Y, d_Y, \mu_Y)$, two mesurable metric spaces.

$$\mathcal{GW}_{p,\alpha}(\mu_X, \mu_Y) = \left(\min_{\boldsymbol{T} \in \Pi(\mu_X, \mu_Y)} \sum_{i,j,k,l} |C_{i,k} - C'_{j,l}|^p T_{i,j} T_{k,l} \right)^{\frac{1}{p}}$$

- $C_{i,k} = \|x_i - x_j\|$ and $C'_{j,l} = \|y_j - y_l\|$ distances in the structures.
- Distance over measures with no common ground space.
- Compares the intrinsic distances in each space (with matrices C and C').
- Invariant to rotations and translation in either spaces.

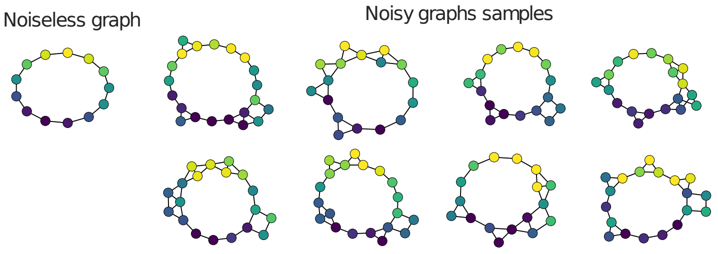


Fused Gromov Wasserstein distance [Vayer et al., 2018b]

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

$$\mathcal{FGW}_{p,q,\alpha}(\mu_s, \mu_t)^p = \min_{\pi \in \Pi(\mu_s, \mu_t)} \sum_{i,j,k,l} \left((1-\alpha) M_{i,j}^q + \alpha |C_{i,k} - C'_{j,l}|^q\right)^p \pi_{i,j}\, \pi_{k,l}$$

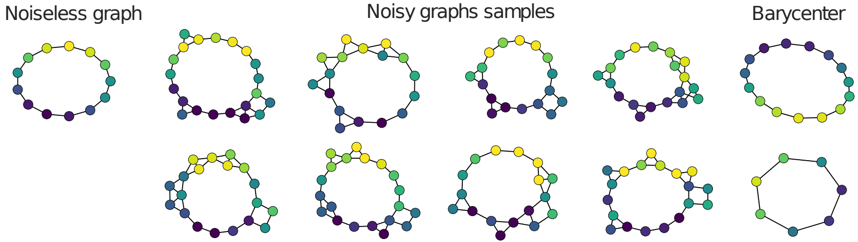
- $M_{i,j} = d(a_i, b_j)$ is the distance betweens the features.
- $\alpha \in [0,1]$ is a trade off parameter between structure and features.
- $\mathcal{FGW}$ is a metric for $q = 1$ a semi metric for $q > 1$, $\forall p \geq 1$.



Barycenter an graph compression

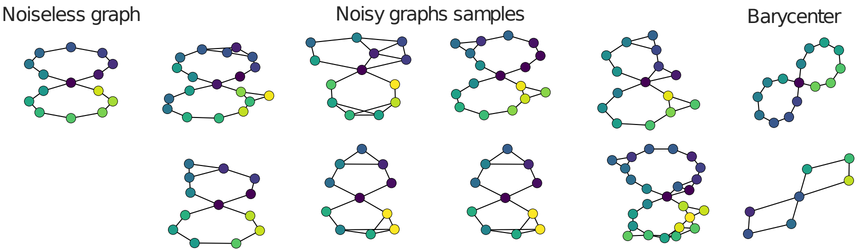
- We compute the barycenter of several graphs on $n = 15$ and $n = 7$ nodes.
- Barycenter graph is obtained through thresholding of the D matrix.
- Community clustering:
  - Approximate a graph with a small number of nodes (clusters)
  - OT matrix give the clustering affectation.

FGW barycenter on labeled graphs



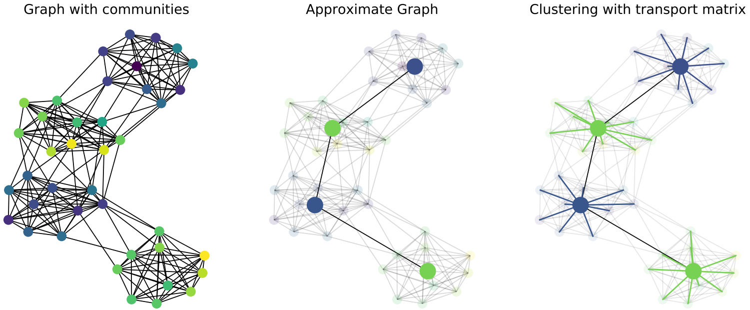
Barycenter an graph compression

- We compute the barycenter of several graphs on $n = 15$ and $n = 7$ nodes.
- Barycenter graph is obtained through thresholding of the D matrix.
- Community clustering:
    - Approximate a graph with a small number of nodes (clusters)
    - OT matrix give the clustering affectation.



Barycenter an graph compression

- We compute the barycenter of several graphs on $n = 15$ and $n = 7$ nodes.
- Barycenter graph is obtained through thresholding of the D matrix.
- Community clustering:
  - Approximate a graph with a small number of nodes (clusters)
  - OT matrix give the clustering affectation.

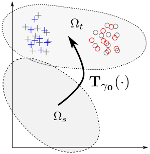


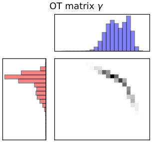
Barycenter an graph compression

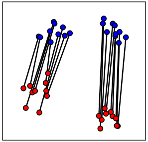
- We compute the barycenter of several graphs on $n = 15$ and $n = 7$ nodes.
- Barycenter graph is obtained through thresholding of the D matrix.
- Community clustering:
  - Approximate a graph with a small number of nodes (clusters)
  - OT matrix give the clustering affectation.

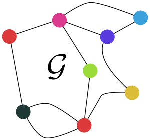
Conclusion and discussion

Future works and open questions









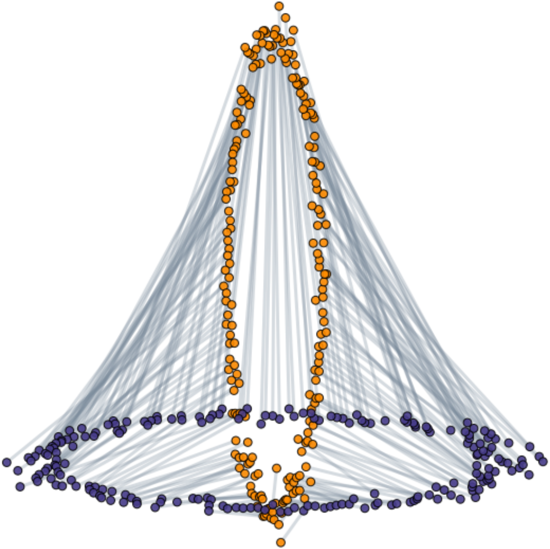
Future works

- Monge mapping estimation (non linear, statistical properties).
- Minibatch Wasserstein (geometrical regularization).
- Adversarial Wasserstein regularization (pairwise regularization between classes).
- OT on graphs (dictionary learning)

The big questions

- Large scale optimization (solving is OK, optimizing still hard).
- Wasserstein distance and regularization (keep geometry, lose complexity).
- Learning the ground metric.

Thank you



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

Courty, N., Flamary, R., and Ducoffe, M. (2018).
Learning wasserstein embeddings.
In *International Conference on Learning Representation (ICMR)*.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017).
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., Kellenberger, B., Flamary, R., Tuia, D., and Courty, N. (2018).
Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation.
In *European Conference in Computer Vision (ECCV)*.

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

Flamary, R., Cuturi, M., Courty, N., and Rakotomamonjy, A. (2018).
Wasserstein discriminant analysis.
Machine Learning.

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, R., Lounici, K., and Ferrari, A. (2019).
Concentration bounds for linear monge mapping estimation and optimal transport domain adaptation.

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

Mémoli, F. (2011).
Gromov-Wasserstein distances and the metric approach to object matching.
Foundations of Computational Mathematics, pages 1–71.

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

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

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000).
The earth mover's distance as a metric for image retrieval.
International journal of computer vision, 40(2):99–121.

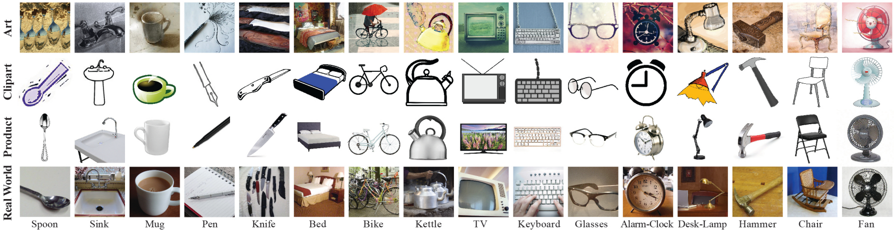
Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. (2018a).
Fused gromov-wasserstein distance for structured objects: theoretical foundations and mathematical properties.

Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. (2018b).
Optimal transport for structured data.

DeepJDOT [Damodaran et al., 2018]

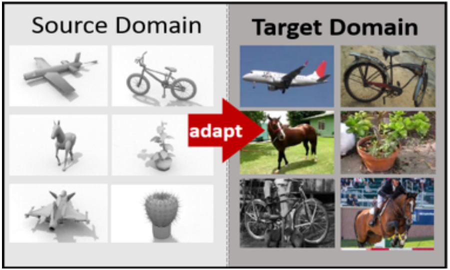
- Evaluation of DeepJDOT on visual classification tasks.
- Digit adaptation between MNIST, USPS, SVHN, MNIST-M.
- Home-office [?] and VisDA 2017 [?] dataset.
- Ablation study : all terms are important.
- TSNE projections of embeddings (MNIST$\rightarrow$MNIST-M).

DeepJDOT in action



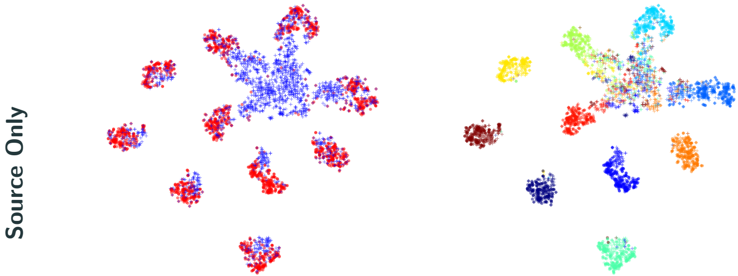
DeepJDOT [Damodaran et al., 2018]

- Evaluation of DeepJDOT on visual classification tasks.
- Digit adaptation between MNIST, USPS, SVHN, MNIST-M.
- Home-office [?] and VisDA 2017 [?] dataset.
- Ablation study : all terms are important.
- TSNE projections of embeddings (MNIST$\rightarrow$MNIST-M).



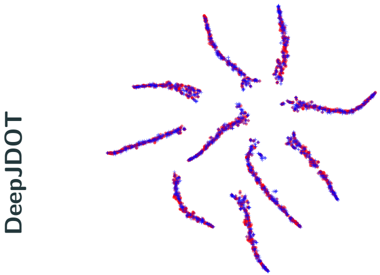
DeepJDOT [Damodaran et al., 2018]

- Evaluation of DeepJDOT on visual classification tasks.
- Digit adaptation between MNIST, USPS, SVHN, MNIST-M.
- Home-office [?] and VisDA 2017 [?] dataset.
- Ablation study : all terms are important.
- TSNE projections of embeddings (MNIST$\rightarrow$MNIST-M).



DeepJDOT [Damodaran et al., 2018]

- Evaluation of DeepJDOT on visual classification tasks.
- Digit adaptation between MNIST, USPS, SVHN, MNIST-M.
- Home-office [?] and VisDA 2017 [?] dataset.
- Ablation study : all terms are important.
- TSNE projections of embeddings (MNIST→MNIST-M).



DeepJDOT [Damodaran et al., 2018]

- Evaluation of DeepJDOT on visual classification tasks.
- Digit adaptation between MNIST, USPS, SVHN, MNIST-M.
- Home-office [?] and VisDA 2017 [?] dataset.
- Ablation study : all terms are important.
- TSNE projections of embeddings (MNIST→MNIST-M).