3iA Côte d'Azur
Institut interdisciplinaire d'intelligence artificielle

UNIVERSITÉ CÔTE D'AZUR

Adapter une IA à de nouvelles données

La stratégie du moindre effort

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

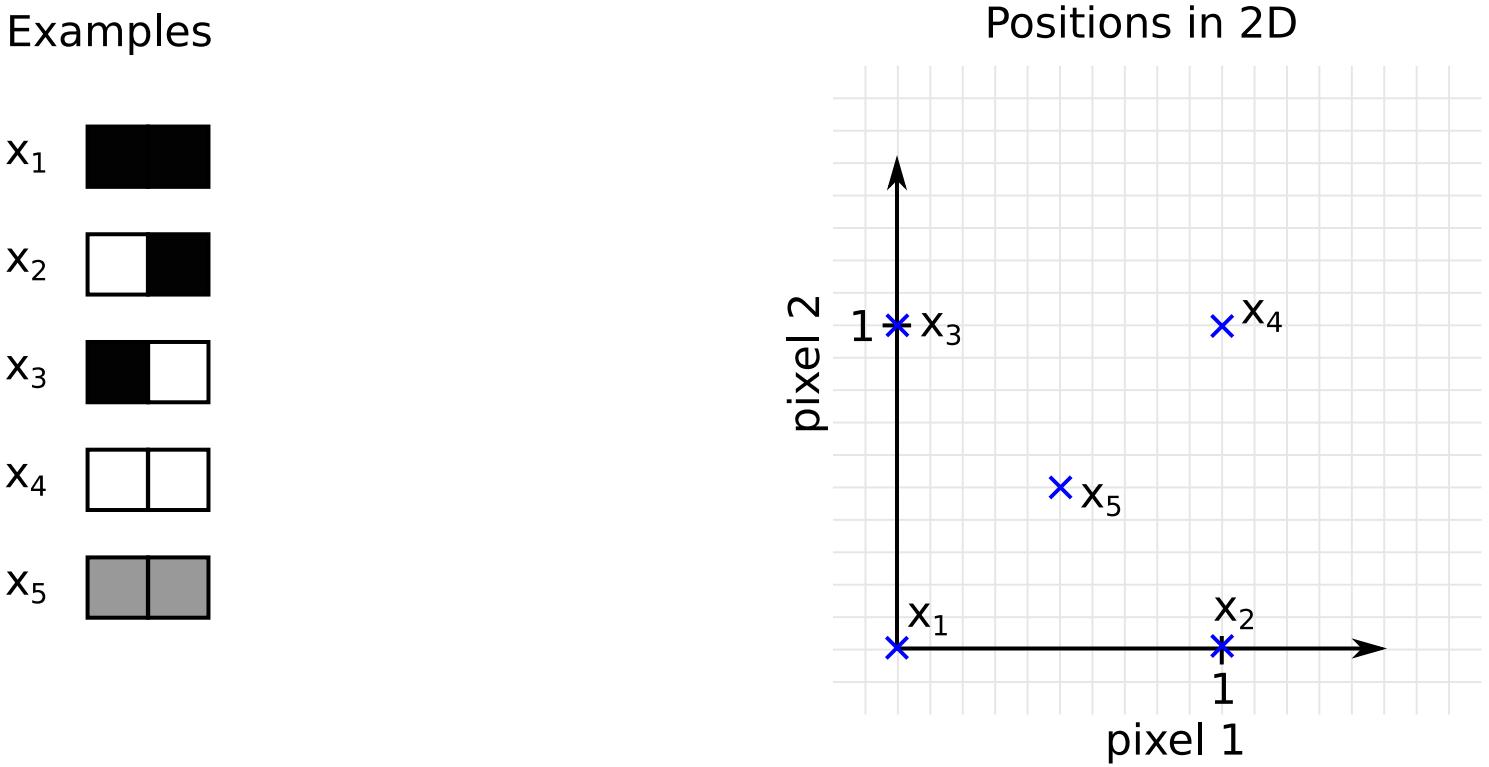
December 12 2019

Vancouver, Canada

Table of content

Introduction

The 2 pixels image

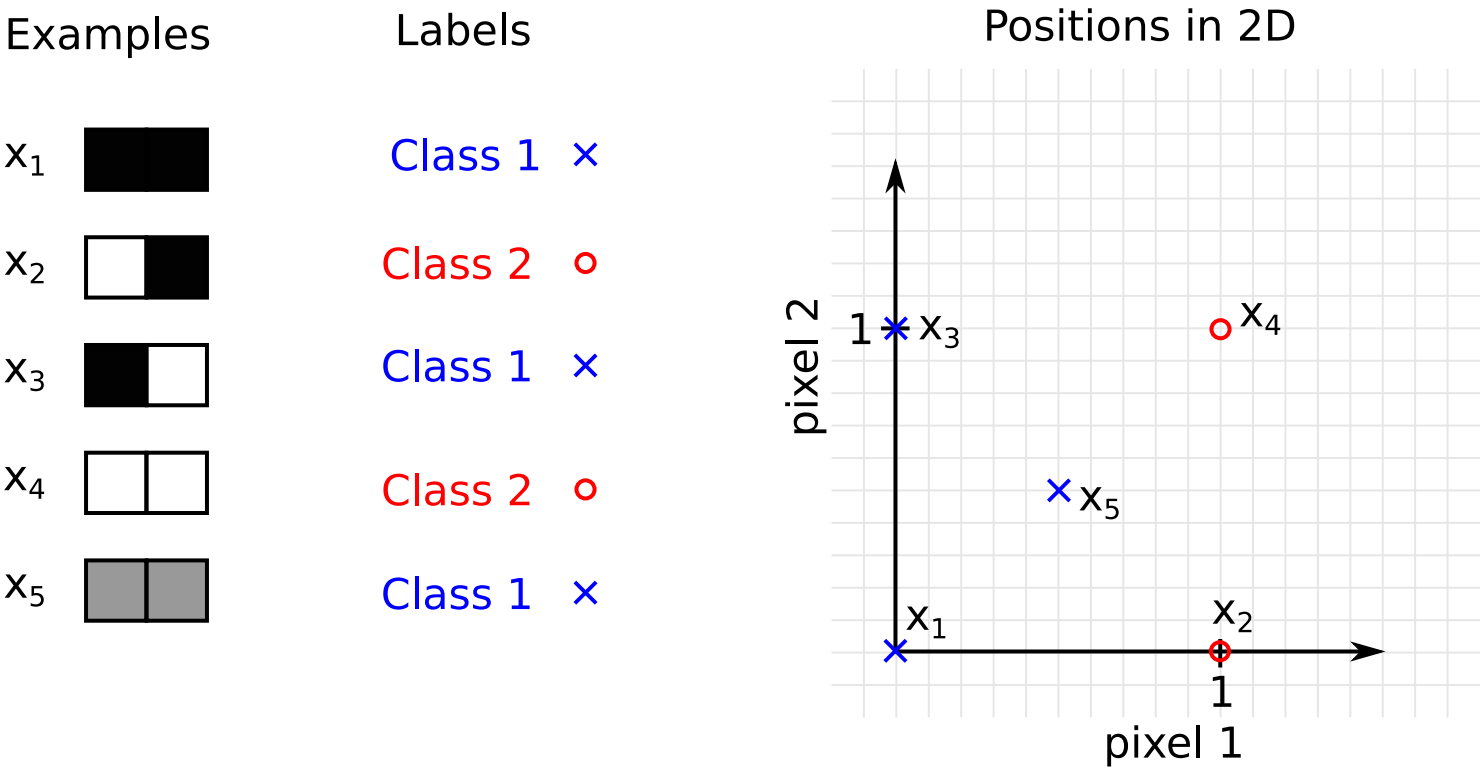


Data as vector in high dimensional space

- Most datasets can be expressed as samples in a vector space.
- The samples correspond to position in this space.
- When examples have a label, one may want to perform classification.
- Training a classifier f corresponds to finding a partition of this space.

Demo

The 2 pixels image

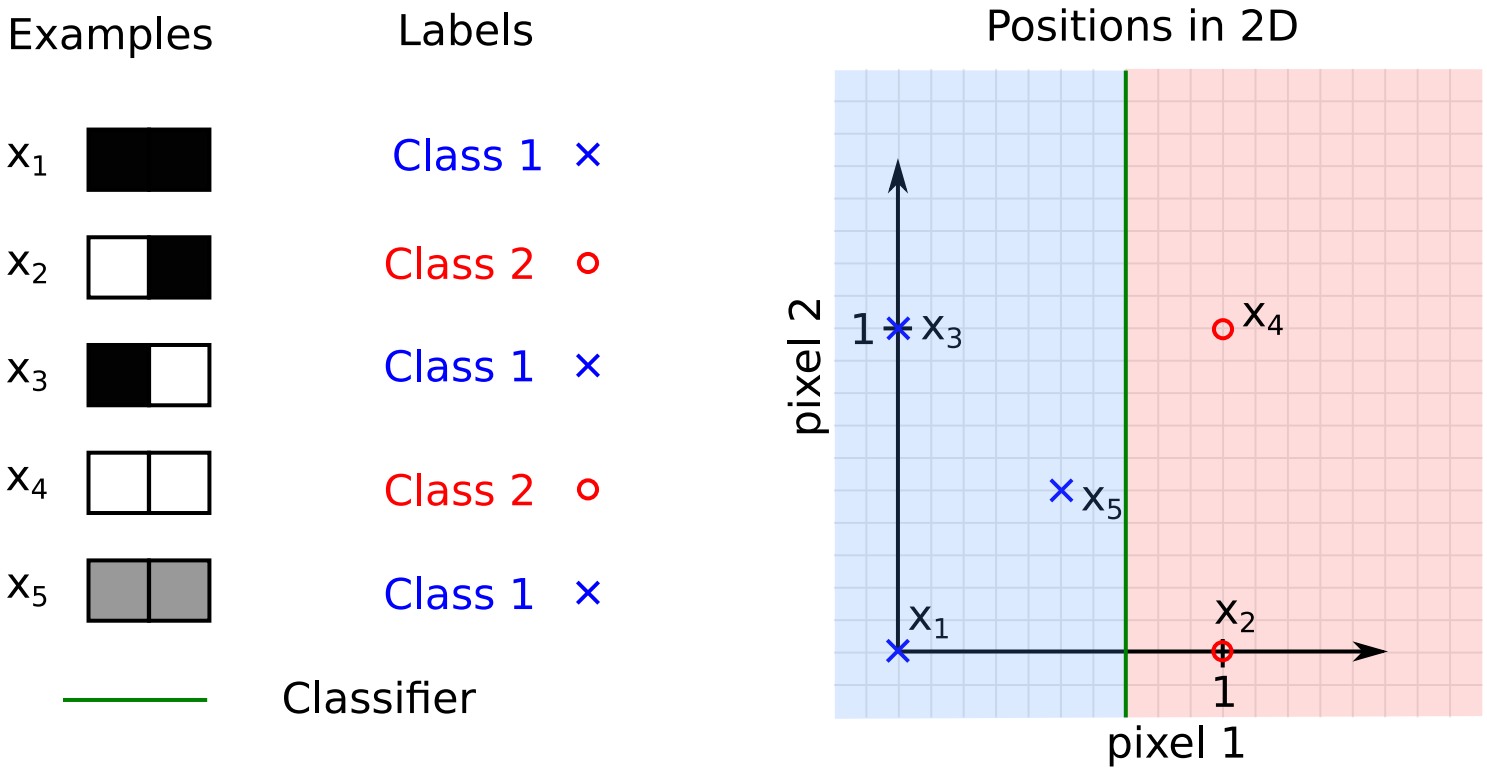


Data as vector in high dimensional space

- Most datasets can be expressed as samples in a vector space.
- The samples correspond to position in this space.
- When examples have a label, one may want to perform classification.
- Training a classifier f corresponds to finding a partition of this space.

Demo

The 2 pixels image



Data as vector in high dimensional space

- Most datasets can be expressed as samples in a vector space.
- The samples correspond to position in this space.
- When examples have a label, one may want to perform classification.
- Training a classifier f corresponds to finding a partition of this space.

Demo

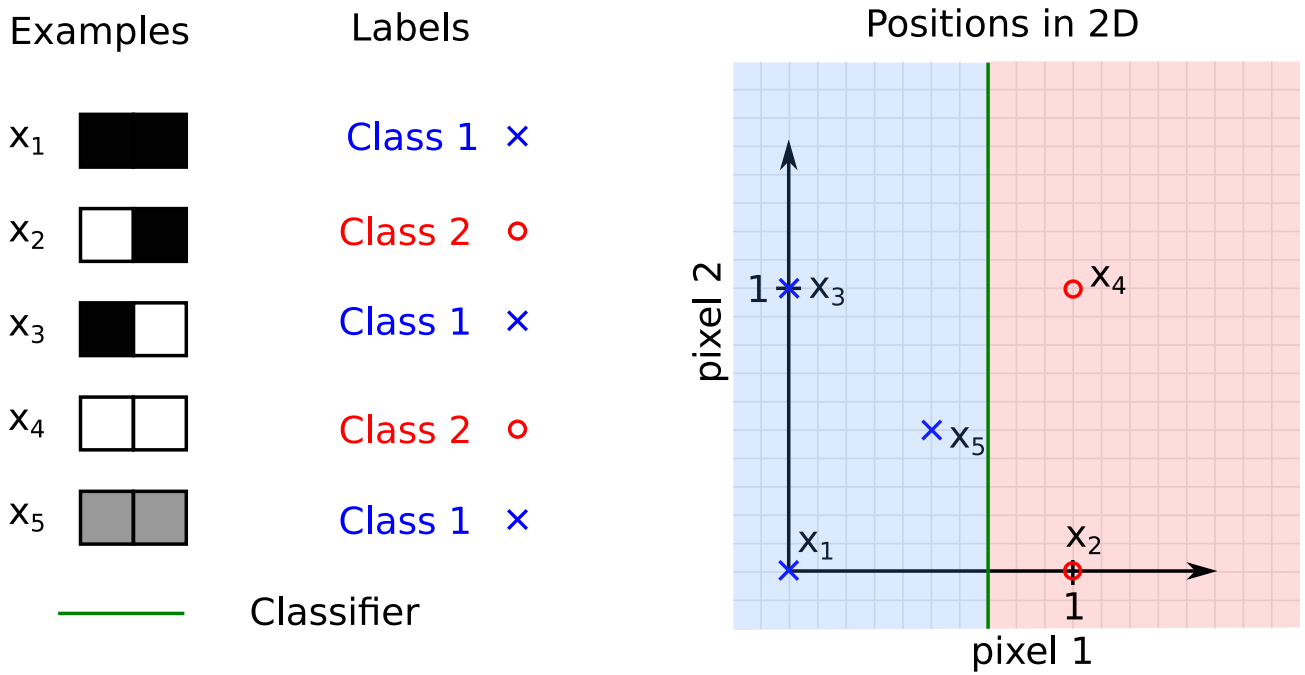
The 2 pixels image

Data as vector in high dimensional space

- Most datasets can be expressed as samples in a vector space.
- The samples correspond to position in this space.
- When examples have a label, one may want to perform classification.
- Training a classifier f corresponds to finding a partition of this space.

Demo

Supervised learning

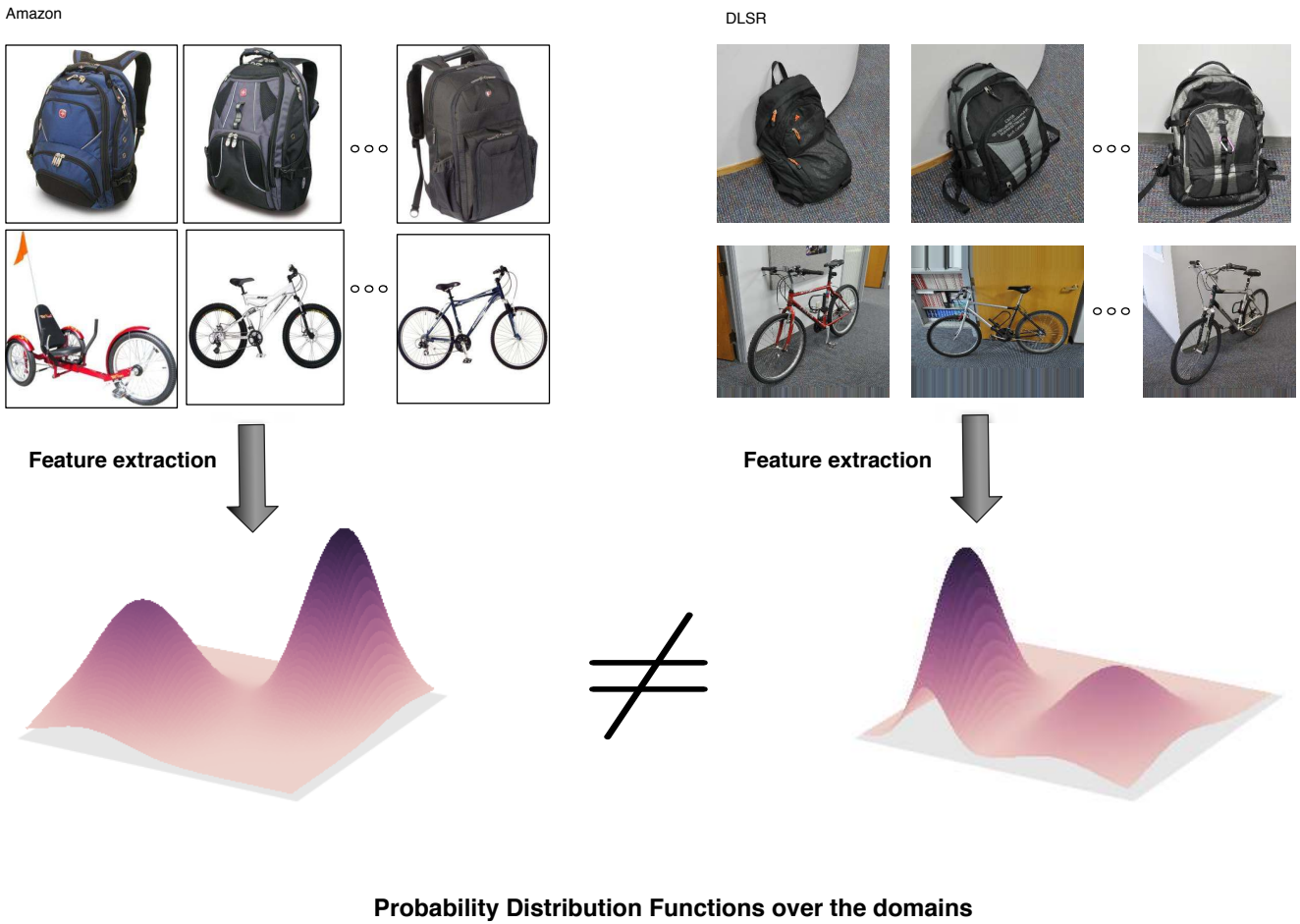


Traditional supervised learning : empirical risk minimization
We week for a predictor f minimizing

$$\min_f \left\{ \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\widehat{\mathcal{P}}} \mathcal{L}(y, f(\mathbf{x})) = \sum_j \mathcal{L}(y_j, f(\mathbf{x}_j)) \right\} \tag{1}$$

- Well known generalization results for predicting on new data.
- Loss is usually $\mathcal{L}(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$ for least square regression and is $\mathcal{L}(y, f(\mathbf{x})) = \max(0, 1 - yf(\mathbf{x}))^2$ for squared Hinge loss SVM.

Domain Adaptation problem

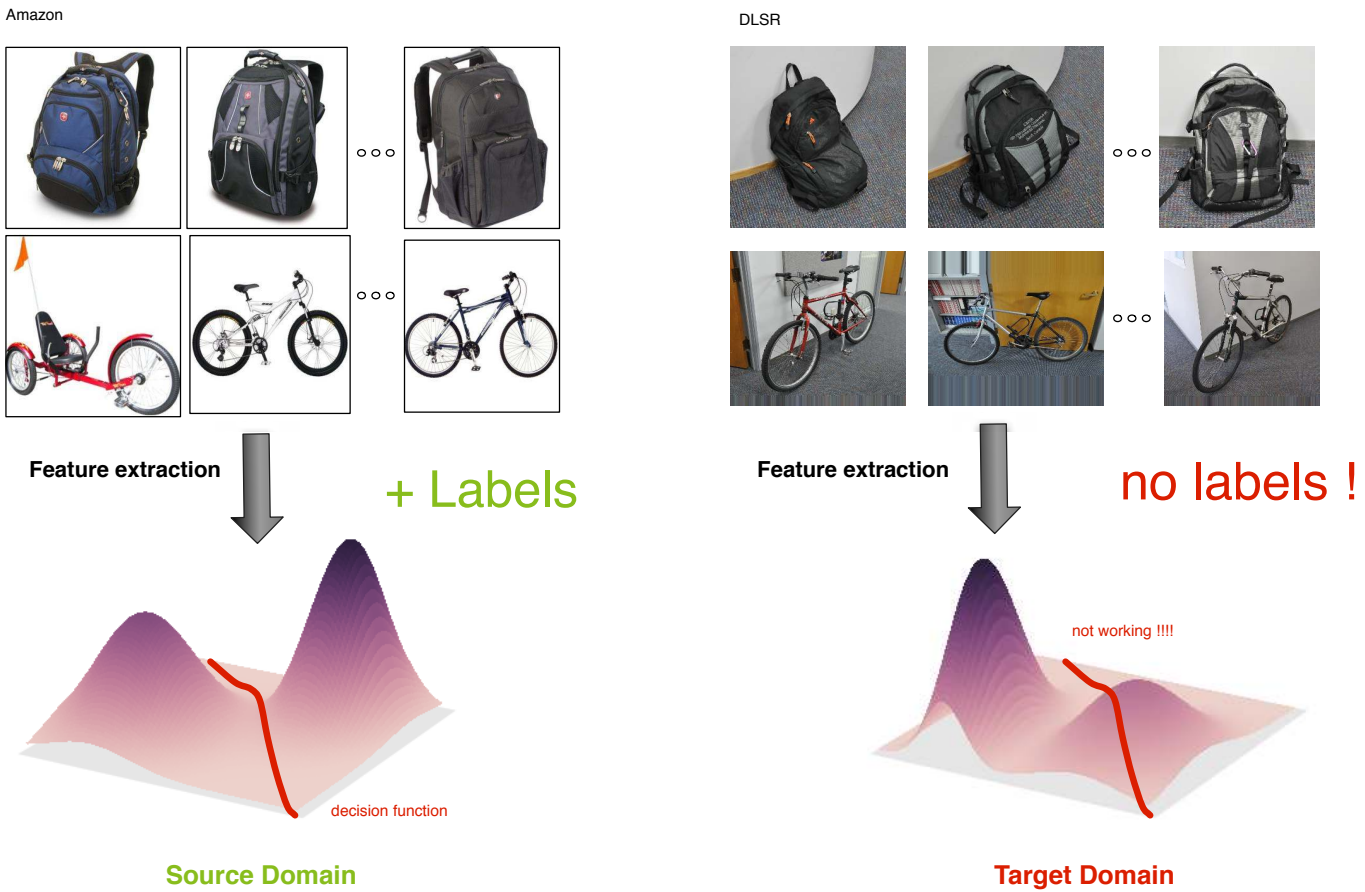


Probability Distribution Functions over the domains

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

Optimal Transport for Domain Adaptation

The origins of optimal transport

666 · MÉMOIRES DE L'ACADÉMIE ROYALE

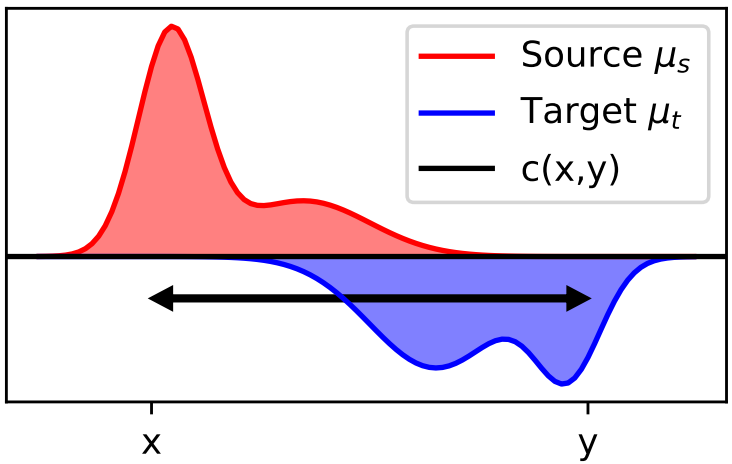
MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.

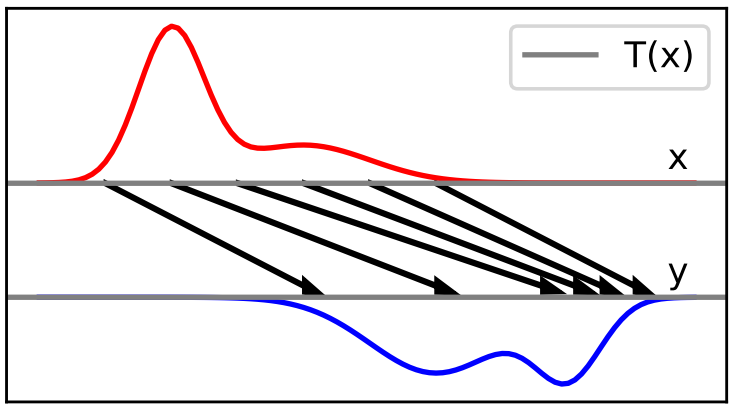
Par M. MONGE.

Problem [Monge, 1781]

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

The origins of optimal transport

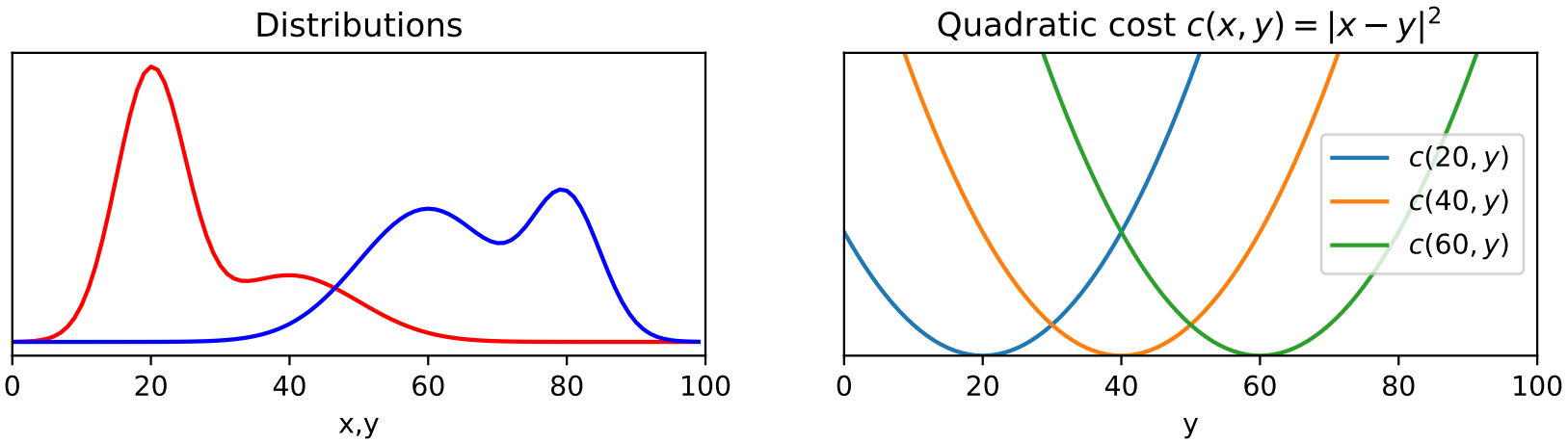




Problem [Monge, 1781]

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

Optimal transport (Monge formulation)

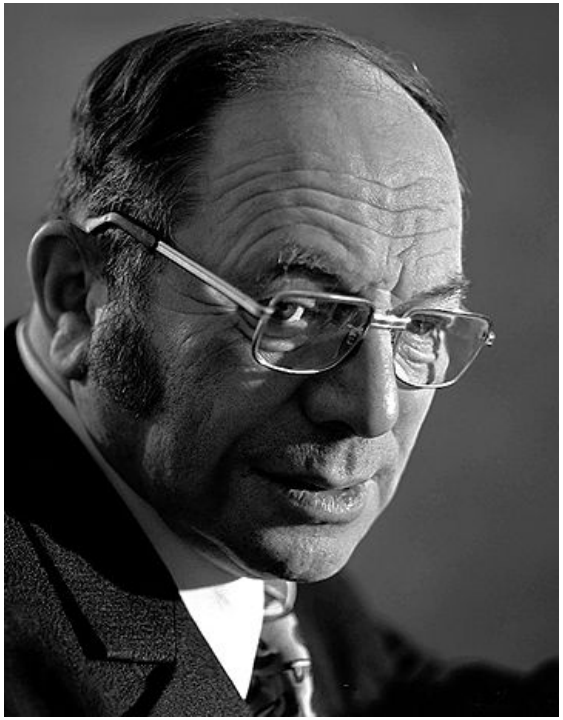


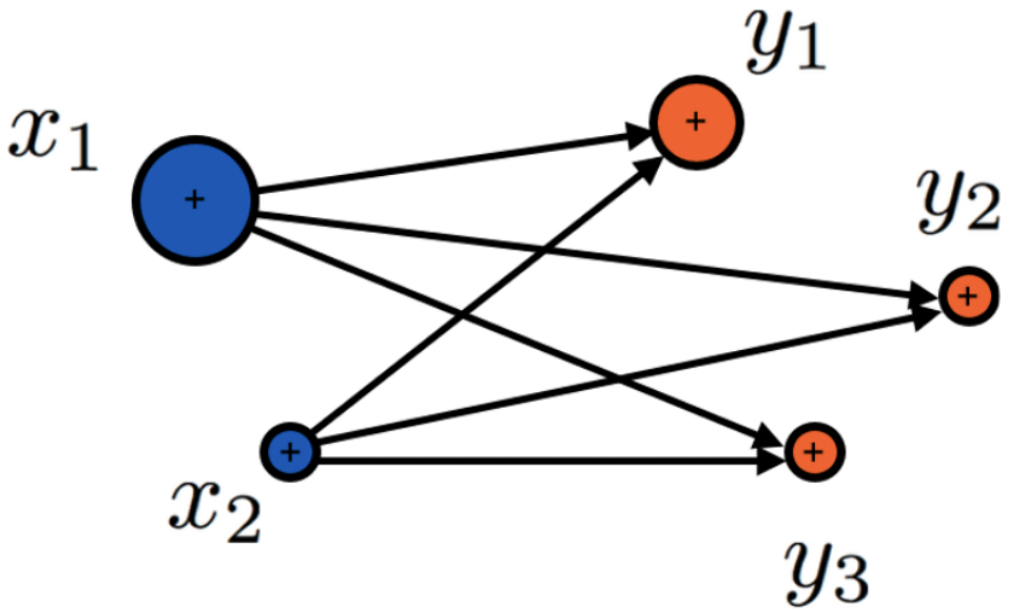
- Probability measures μ_s and μ_t on and a cost function $c : \Omega_s \times \Omega_t \to \mathbb{R}^+$.
- The Monge formulation [Monge, 1781] aim at finding a mapping $T : \Omega_s \to \Omega_t$

$$\inf_{T\#\mu_s = \mu_t} \quad \int_{\Omega_s} c(\mathbf{x}, T(\mathbf{x}))\mu_s(\mathbf{x})d\mathbf{x} \tag{2}$$

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

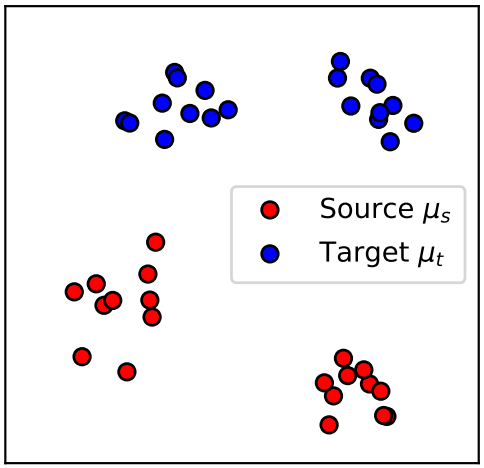
Kantorovich relaxation

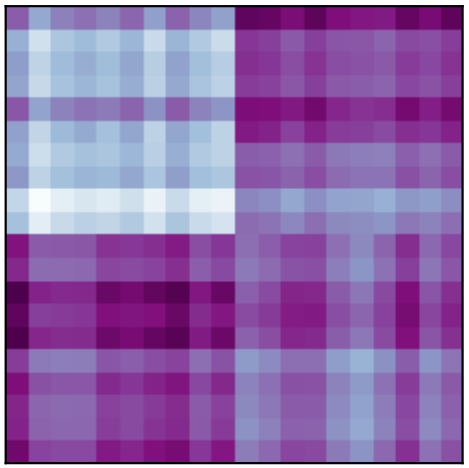


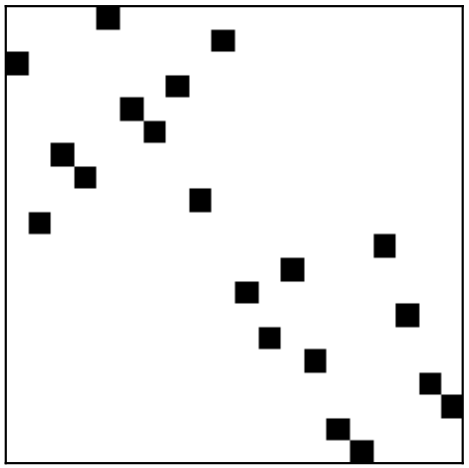


- Leonid Kantorovich (1912–1986), Economy nobelist in 1975
- Focus on where the mass goes, allow splitting [Kantorovich, 1942].
- Applications mainly for resource allocation problems

Optimal transport with discrete distributions







Kantorowitch formulation : OT Linear Program

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

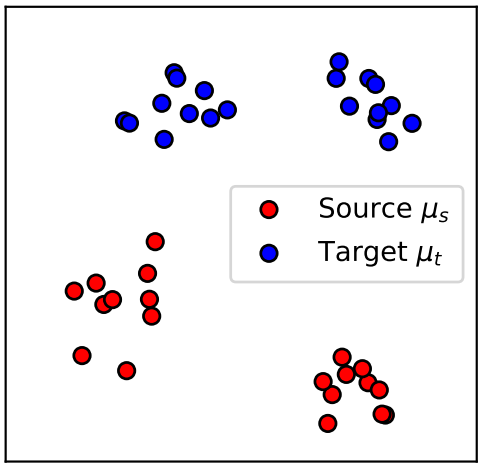
$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma} \in \mathcal{P}}{\operatorname{argmin}} \quad \left\{ \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F = \sum_{i,j} \gamma_{i,j} c_{i,j} \right\}$$

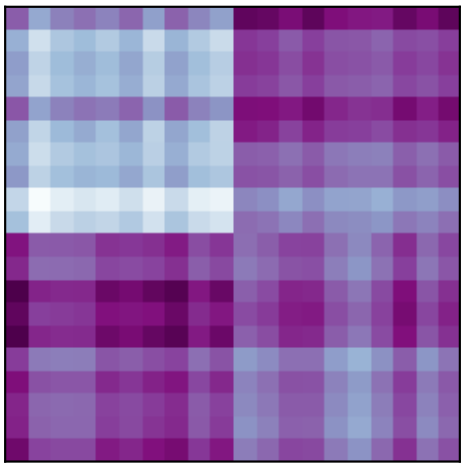
where $\mathbf{C}$ is a cost matrix with $c_{i,j} = c(\mathbf{x}_i^s, \mathbf{x}_j^t)$ and the marginals constraints are

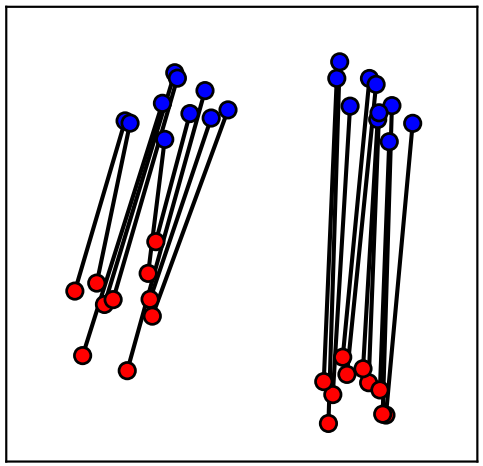
$$\mathcal{P} = \left\{ \boldsymbol{\gamma} \in (\mathbb{R}^+)^{\mathbf{n_s} \times \mathbf{n_t}} \,|\, \boldsymbol{\gamma} \mathbf{1}_{\mathbf{n_t}} = \mathbf{a}, \boldsymbol{\gamma}^{\mathbf{T}} \mathbf{1}_{\mathbf{n_s}} = \mathbf{b} \right\}$$

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

Optimal transport with discrete distributions







Kantorowitch formulation : OT Linear Program

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

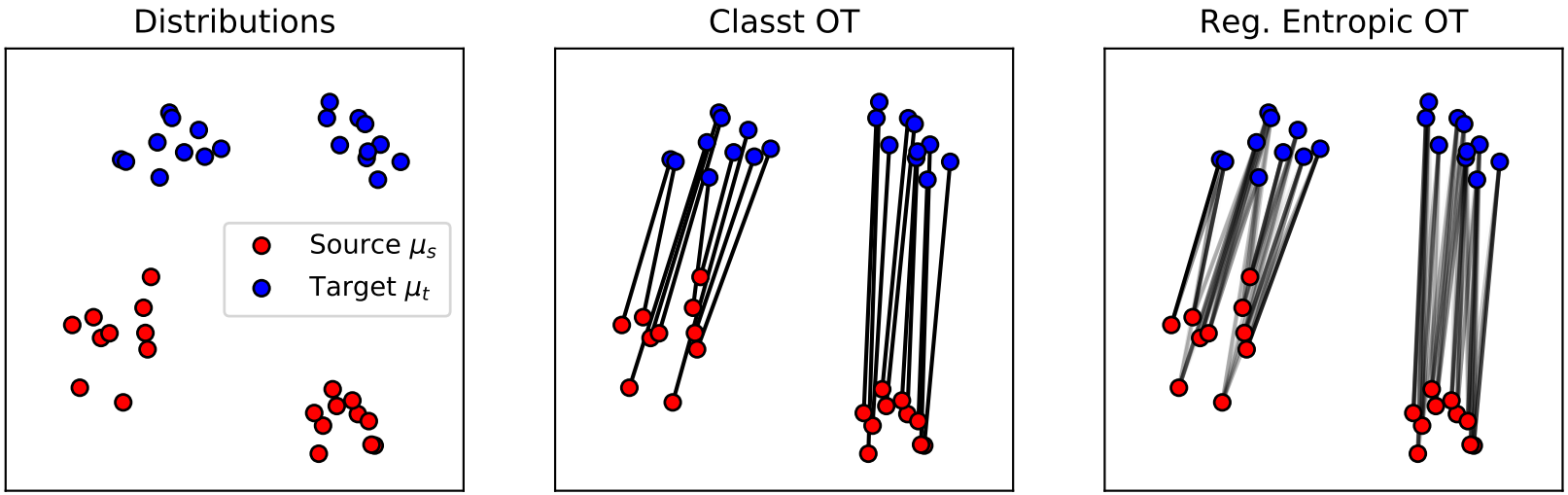
$$\boldsymbol{\gamma}_0 = \underset{\boldsymbol{\gamma} \in \mathcal{P}}{\operatorname{argmin}} \quad \left\{ \langle \boldsymbol{\gamma}, \mathbf{C} \rangle_F = \sum_{i,j} \gamma_{i,j} c_{i,j} \right\}$$

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

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

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

Transporting the discrete samples

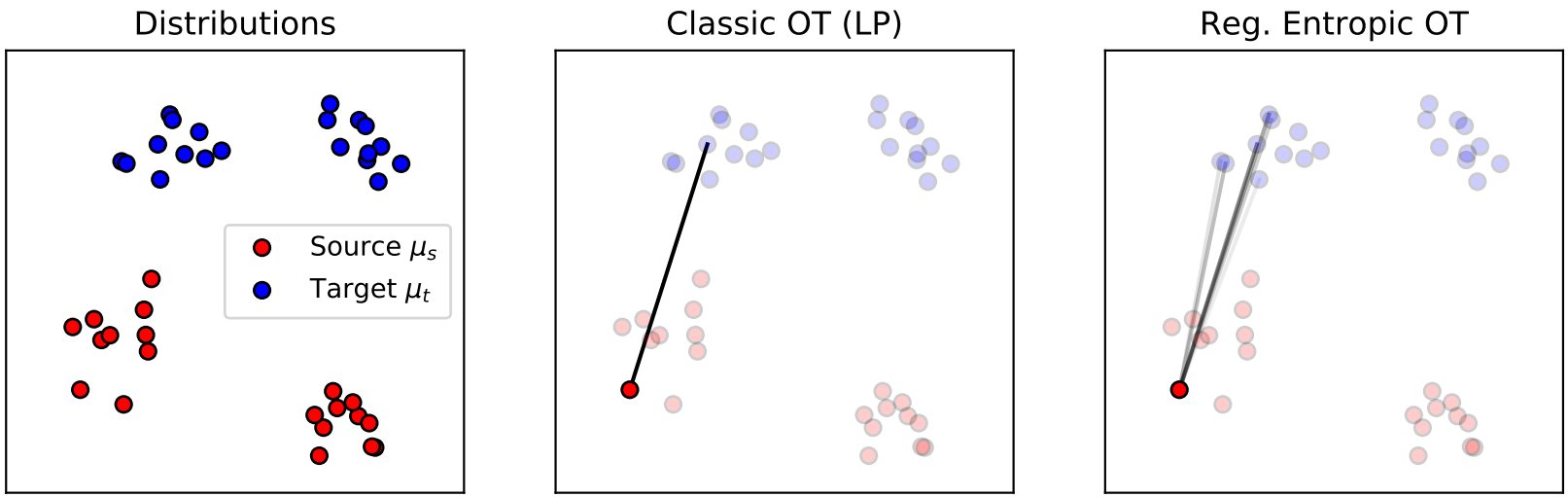


Barycentric mapping [Ferradans et al., 2014]

$$\widehat{T}_{\boldsymbol{\gamma}_0}(\mathbf{x}_i^s) = \underset{\mathbf{x}}{\operatorname{argmin}} \quad \sum_j \boldsymbol{\gamma}_0(i,j) c(\mathbf{x}, \mathbf{x}_j^t). \tag{3}$$

- The mass of each source sample is spread onto the target samples (line of $\boldsymbol{\gamma}_0$).
- The mapping is the barycenter of the target samples weighted by $\boldsymbol{\gamma}_0$
- 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

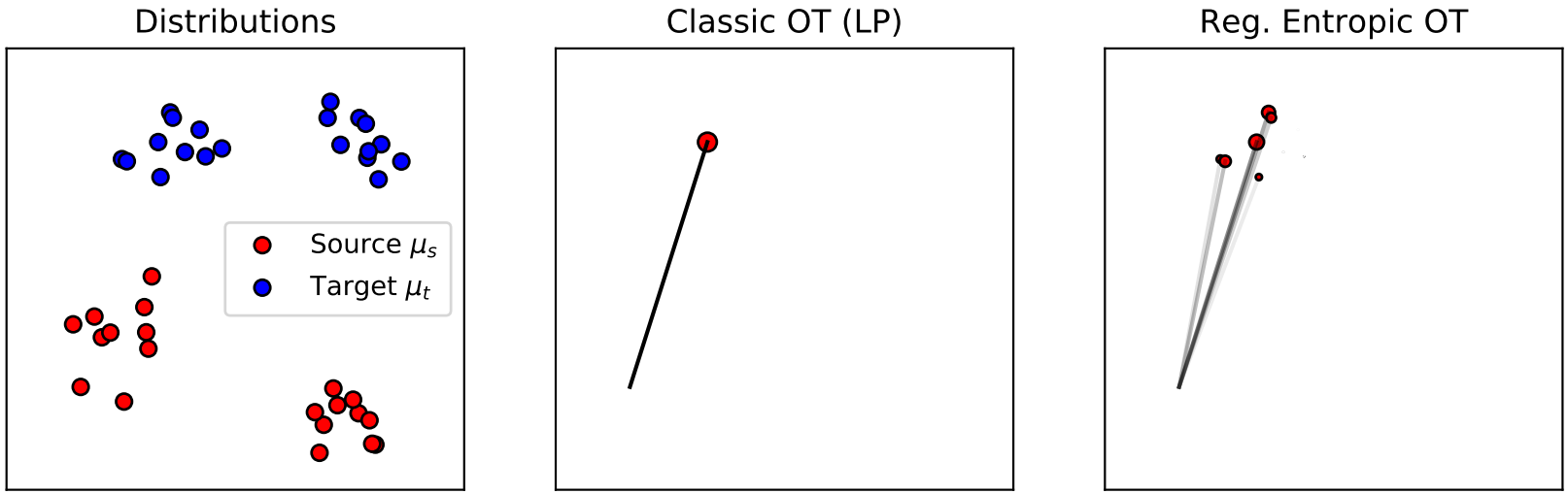


Barycentric mapping [Ferradans et al., 2014]

$$\widehat{T}_{\boldsymbol{\gamma}_0}(\mathbf{x}_i^s) = \underset{\mathbf{x}}{\operatorname{argmin}} \quad \sum_j \boldsymbol{\gamma}_0(i,j)\|\mathbf{x} - \mathbf{x}_j^t\|^2. \tag{3}$$

- The mass of each source sample is spread onto the target samples (line of $\boldsymbol{\gamma}_0$).
- The mapping is the barycenter of the target samples weighted by $\boldsymbol{\gamma}_0$
- 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

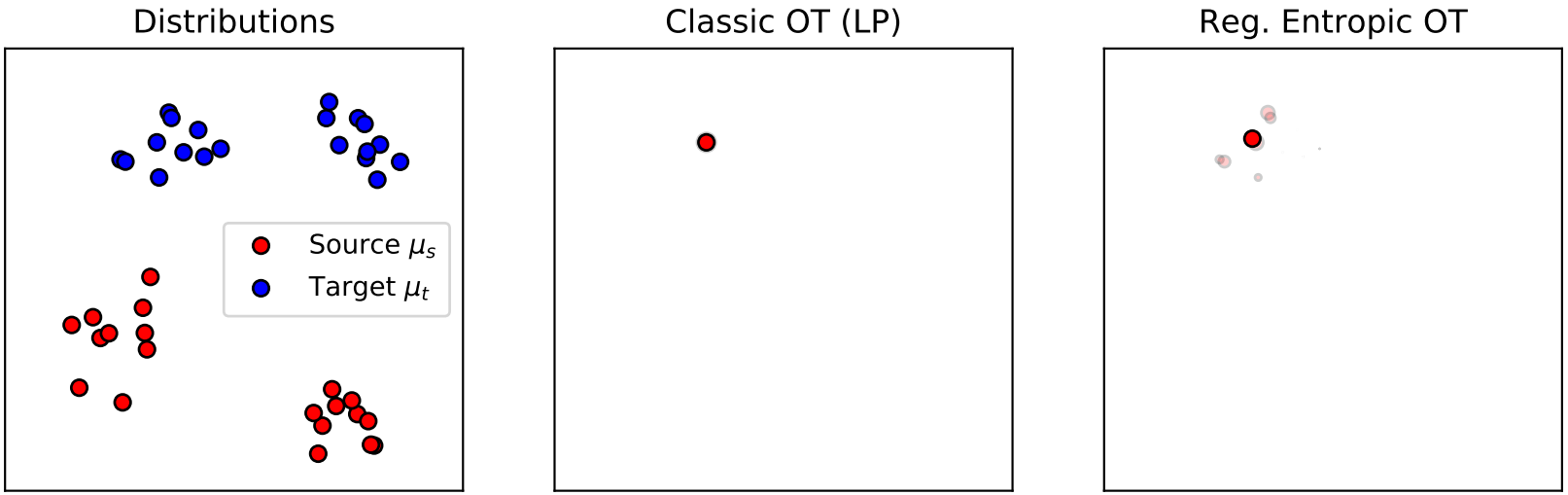


Barycentric mapping [Ferradans et al., 2014]

$$\widehat{T}_{\boldsymbol{\gamma}_0}(\mathbf{x}_i^s) = \frac{1}{\sum_j \boldsymbol{\gamma}_0(i,j)} \sum_j \boldsymbol{\gamma}_0(i,j)\mathbf{x}_j^t. \tag{3}$$

- The mass of each source sample is spread onto the target samples (line of $\boldsymbol{\gamma}_0$).
- The mapping is the barycenter of the target samples weighted by $\boldsymbol{\gamma}_0$
- 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

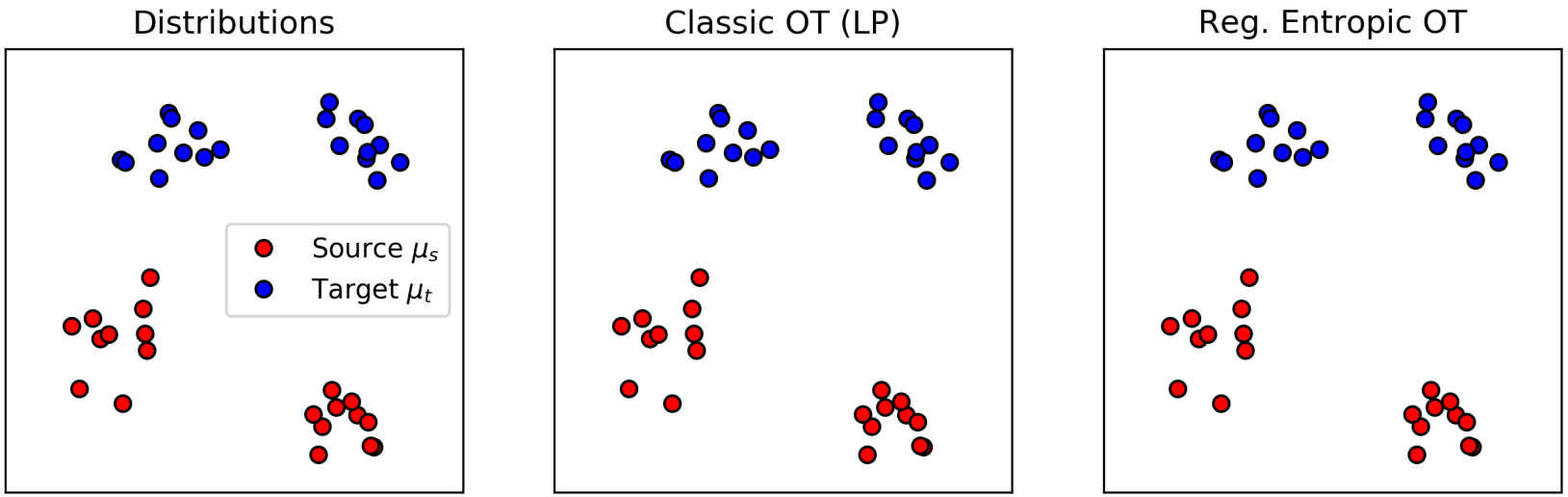


Barycentric mapping [Ferradans et al., 2014]

$$\widehat{T}_{\gamma_0}(\mathbf{x}_i^s) = \frac{1}{\sum_j \gamma_0(i,j)} \sum_j \gamma_0(i,j)\mathbf{x}_j^t. \tag{3}$$

- 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]

$$\widehat{T}_{\gamma_0}(\mathbf{x}_i^s) = \frac{1}{\sum_j \gamma_0(i,j)} \sum_j \gamma_0(i,j)\mathbf{x}_j^t. \tag{3}$$

- 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.

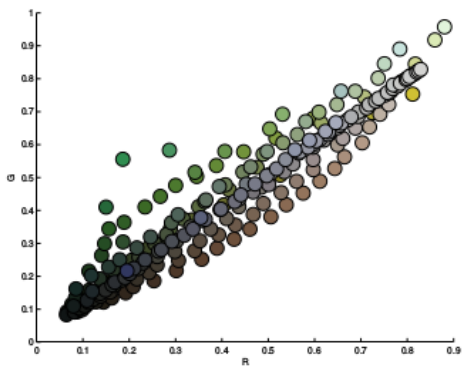
Histogram matching in images

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

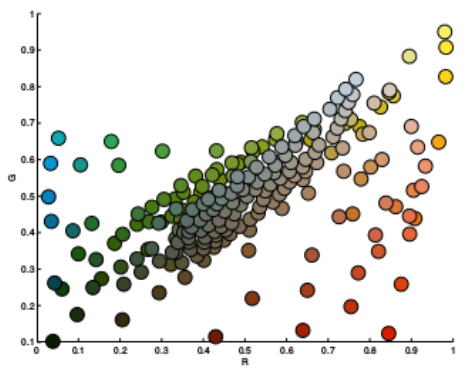
X^0

Y^0

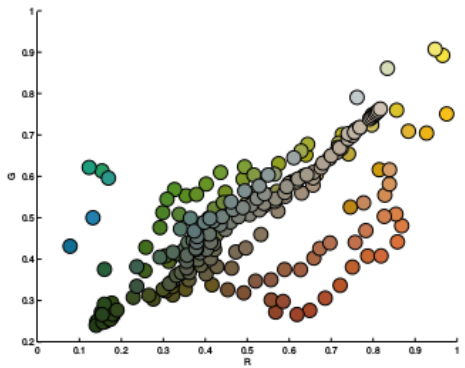
$\tilde{X}^0$



μ_{X^0}



μ_{Y^0}

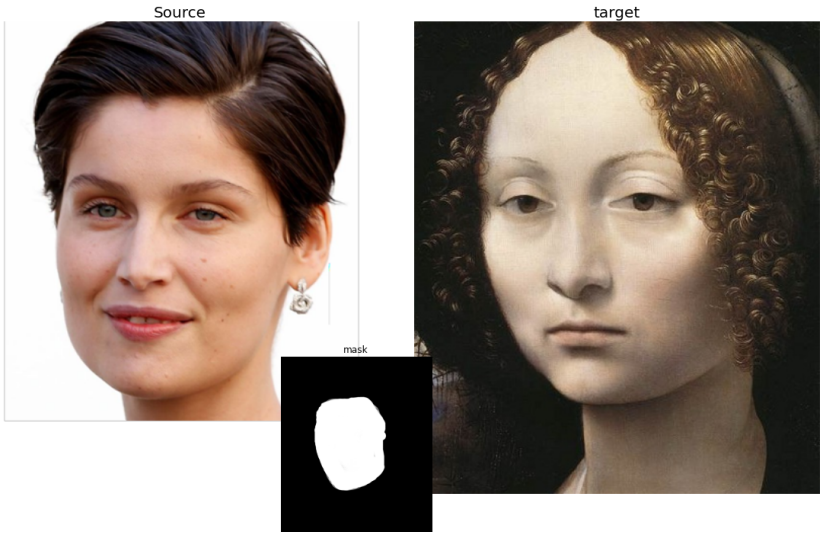


$\mu_{\tilde{X}^0}$

Histogram matching in images

Image colorization [Ferradans et al., 2014]

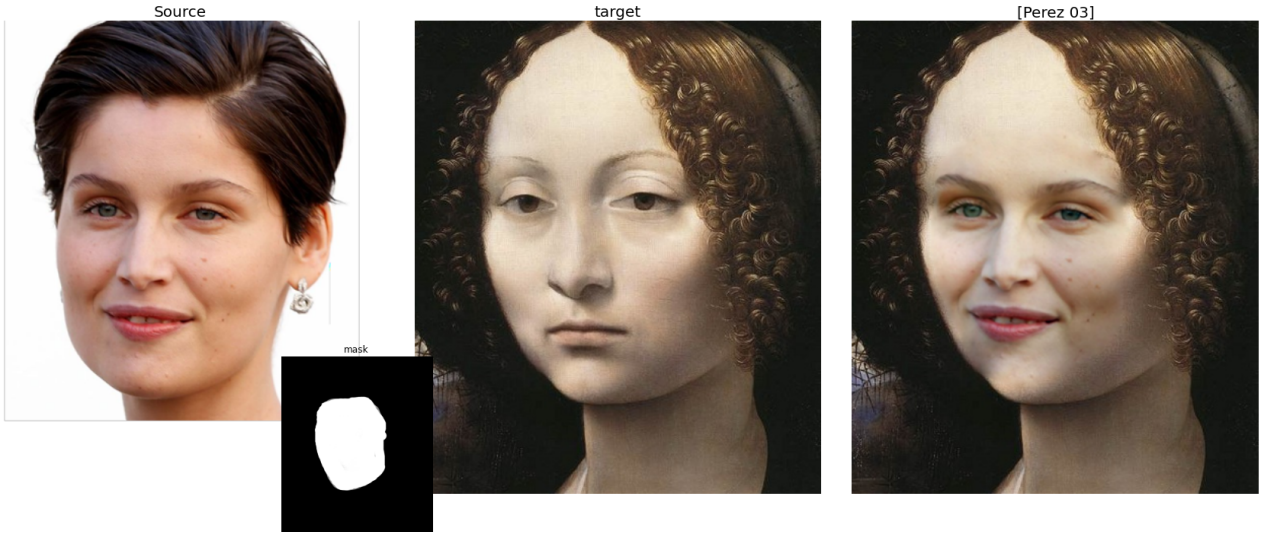
Seamless copy in images



Poisson image editing [Pérez et al., 2003]

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

Seamless copy in images



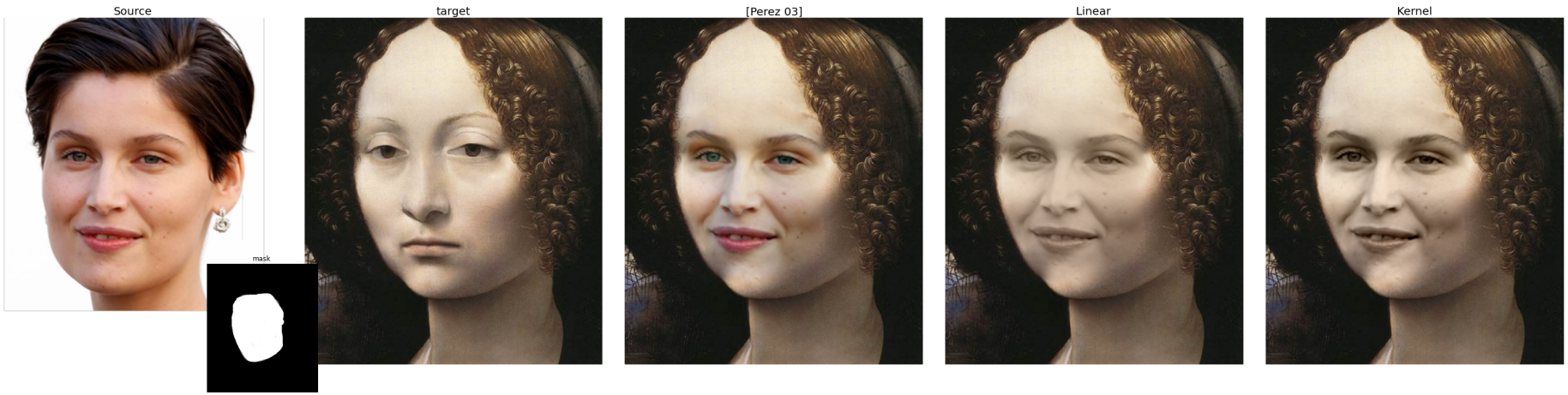
Poisson image editing [Pérez et al., 2003]

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

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

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

Seamless copy in images



Poisson image editing [Pérez et al., 2003]

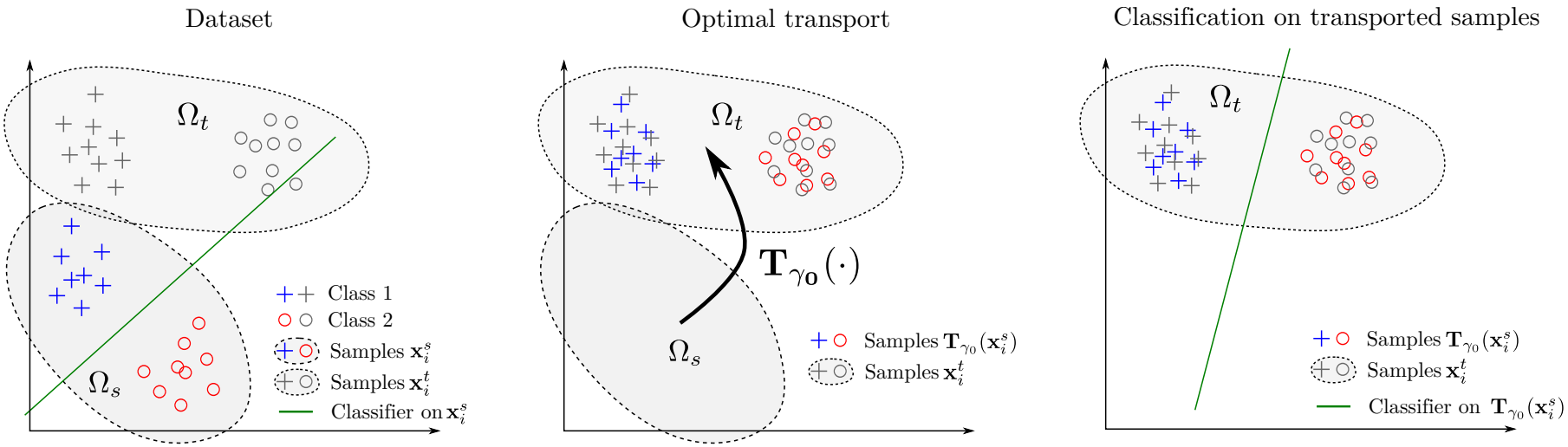
- Use the color gradient from the source image.
- Use color border conditions on the target image.
- Solve Poisson equation to reconstruct the new image.

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

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

Seamless copy with gradient adaptation

Optimal transport for domain adaptation



Assumptions

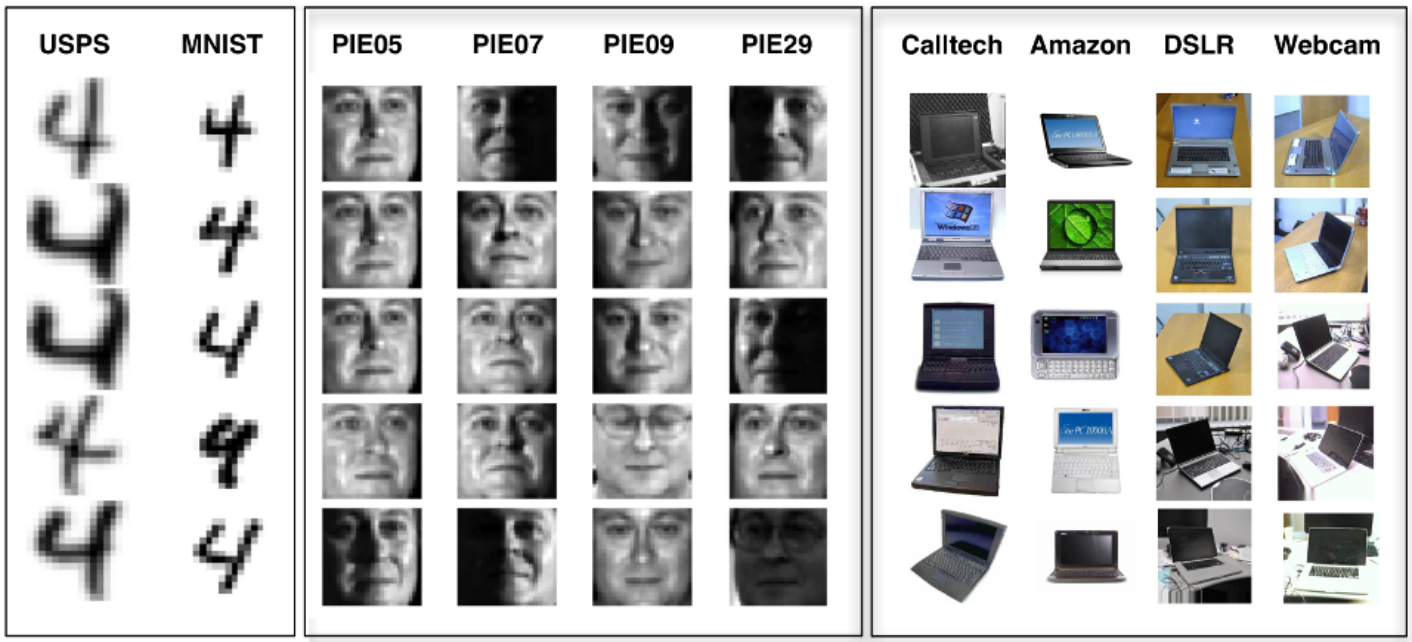
- There exist a transport in the feature space $\mathbf{T}$ between the two domains.
- The transport preserves the conditional distributions:

$$P_s(y|\mathbf{x}_s) = P_t(y|\mathbf{T}(\mathbf{x}_s)).$$

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

1. Estimate optimal transport between distributions.
2. Transport the training samples with barycentric mapping .
3. Learn a classifier on the transported training samples.

Visual adaptation datasets



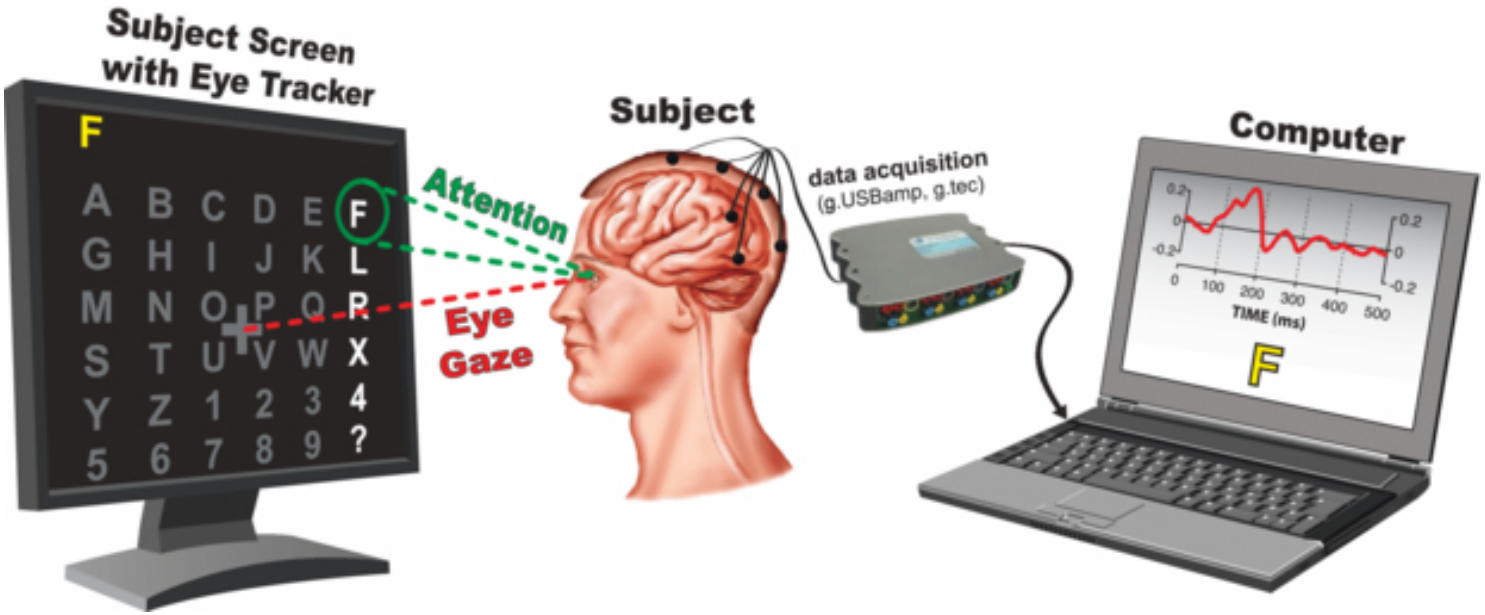
Datasets

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

Numerical experiments

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

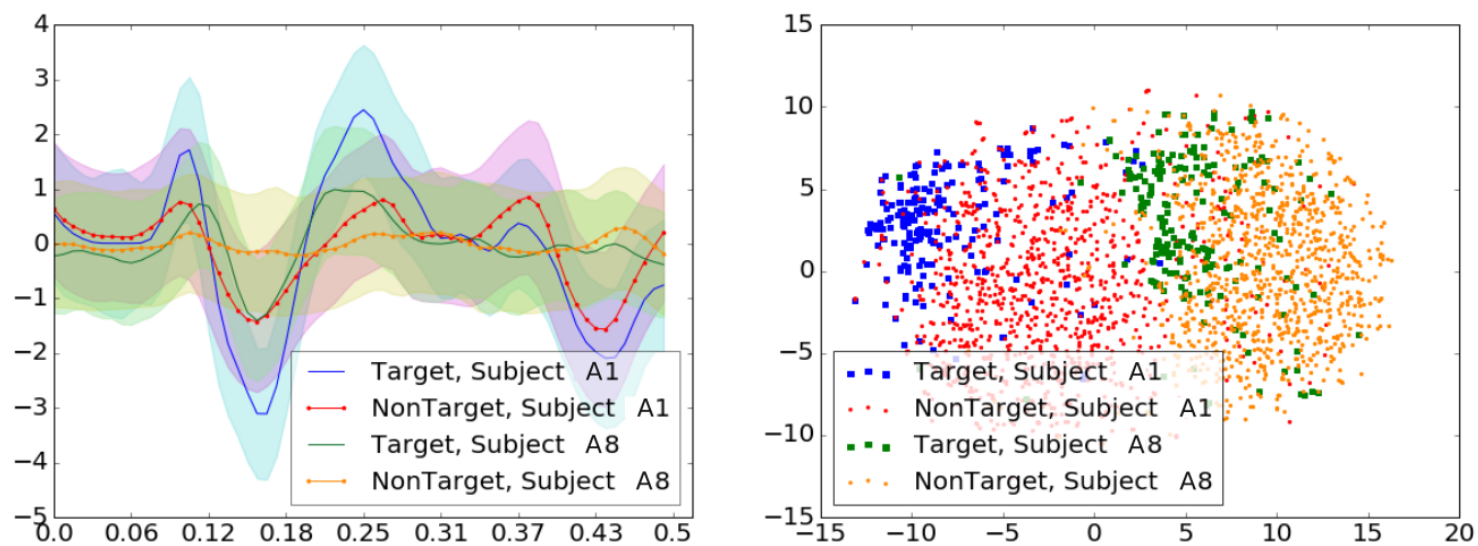
OTDA for biomedical data (1)



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.

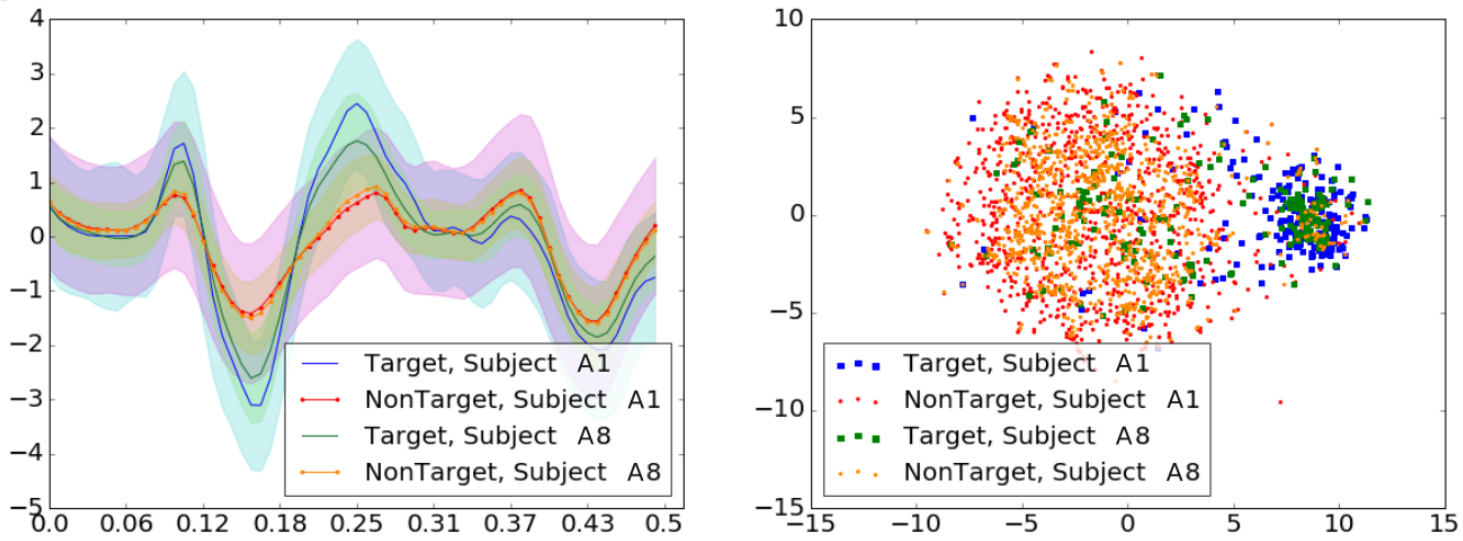
OTDA for biomedical data (1)



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 (1)



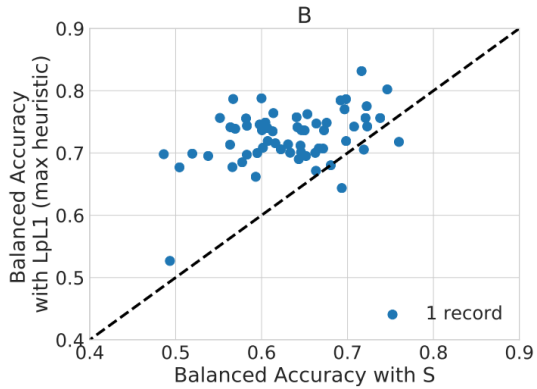
Multi-subject P300 classification [Gayraud et al., 2017]

- 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)

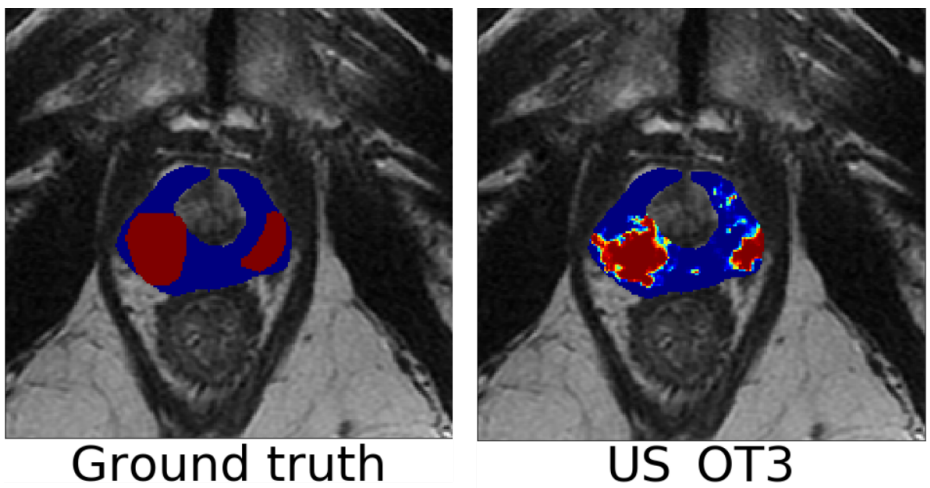
EEG sleep stage classification [Chambon et al., 2018]

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



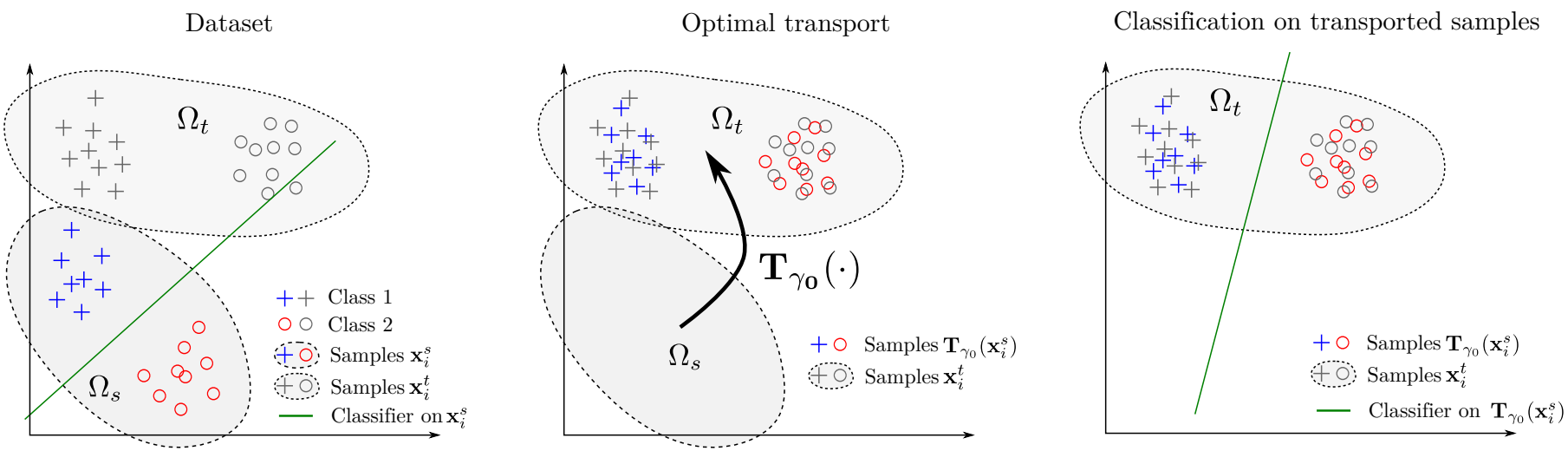
Prostace cancer classification [Gautheron et al., 2017]

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



Conclusion

Conclusion



Optimal Transport for Domin Adaptation

- OT is a tool to estimate least effort mapping.
- The OT mapping can be used to adapt data.
- When continuous mapping is available the adaptation can be done the other way around with $f \circ T^{-1}$ [Flamary et al., 2019].
- Other variants of OT for DA rely on transporting the features/label simultaneously [Courty et al., 2017, Damodaran et al., 2018].

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.

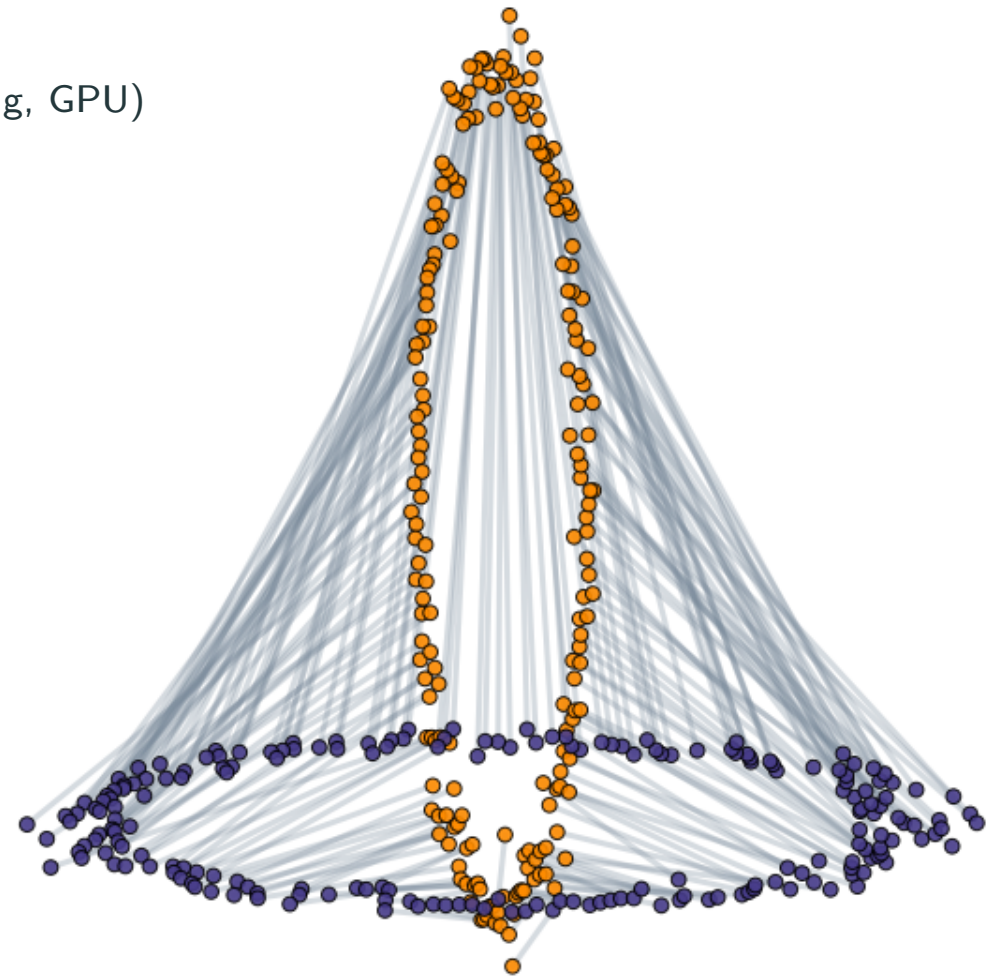
Tutorial on OT for ML:

`http://tinyurl.com/otml-isbi`

Papers available on my website:

`https://remi.flamary.com/`

Post docs available in: Nice (France)



References i

Brenier, Y. (1991).

Polar factorization and monotone rearrangement of vector-valued functions.

Communications on pure and applied mathematics, 44(4):375–417.

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., 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.

References ii

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., Lounici, K., and Ferrari, A. (2019).

Concentration bounds for linear monge mapping estimation and optimal transport domain adaptation.

arXiv preprint arXiv:1905.10155.

Gautheron, L., Lartizien, C., and Redko, I. (2017).

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

References iii

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.

Kantorovich, L. (1942).

On the translocation of masses.

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

Monge, G. (1781).

Mémoire sur la théorie des déblais et des remblais.

De l'Imprimerie Royale.

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)*.


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.