



**LABORATOIRE**  
**HUBERT CURIEN**  
UMR • CNRS • 5516 • SAINT-ETIENNE

# Contributions to Optimal Transport for Machine Learning: Ground Metric and Generalized Framework

Thesis defense of Tanguy KERDONCUFF

Élisa FROMONT Professeure, Université de Rennes 1 Présidente  
Marianne CLAUSEL Professeure, Université de Lorraine Rapporteuse  
Nicolas COURTY Professeur, Université Bretagne Sud Rapporteur

Marc SEBBAN Professeur, Université de Saint-Étienne Directeur  
Rémi EMONET Maître de conférences, Université de Saint-Étienne Co-encadrant

November 15, 2021

# Basic Machine Learning examples



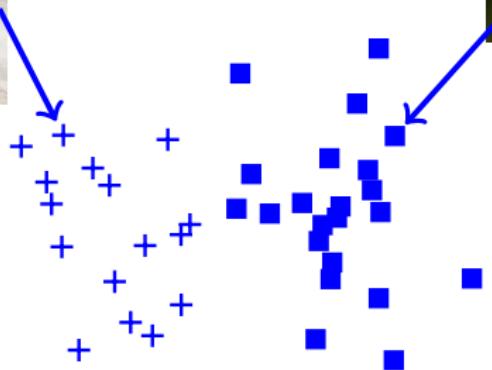
# Basic Machine Learning examples



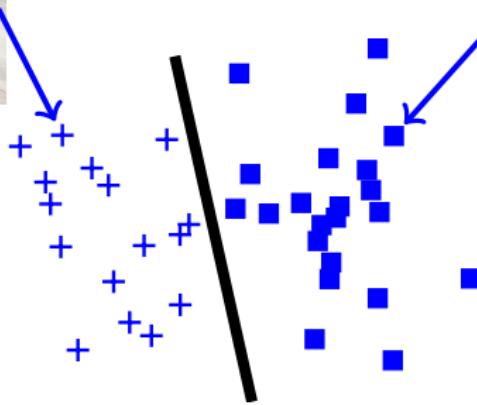
+



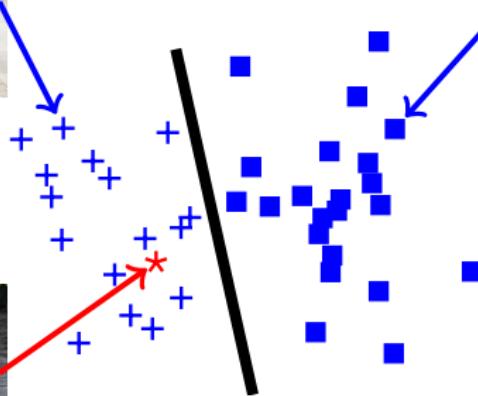
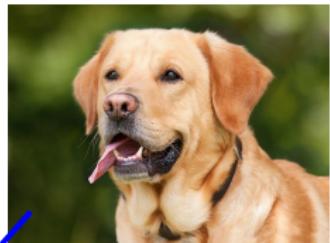
# Basic Machine Learning examples



# Basic Machine Learning examples



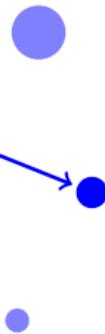
# Basic Machine Learning examples



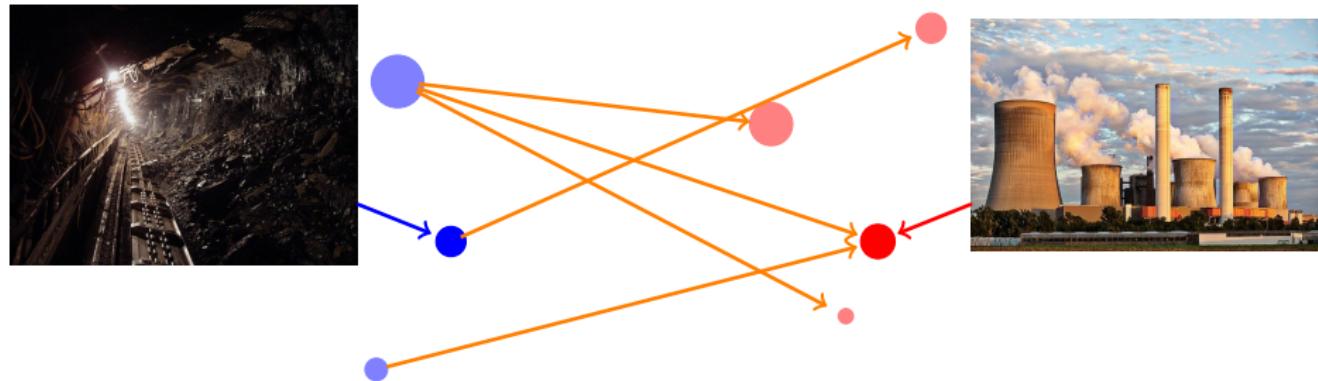
# Intuition of the Optimal Transport



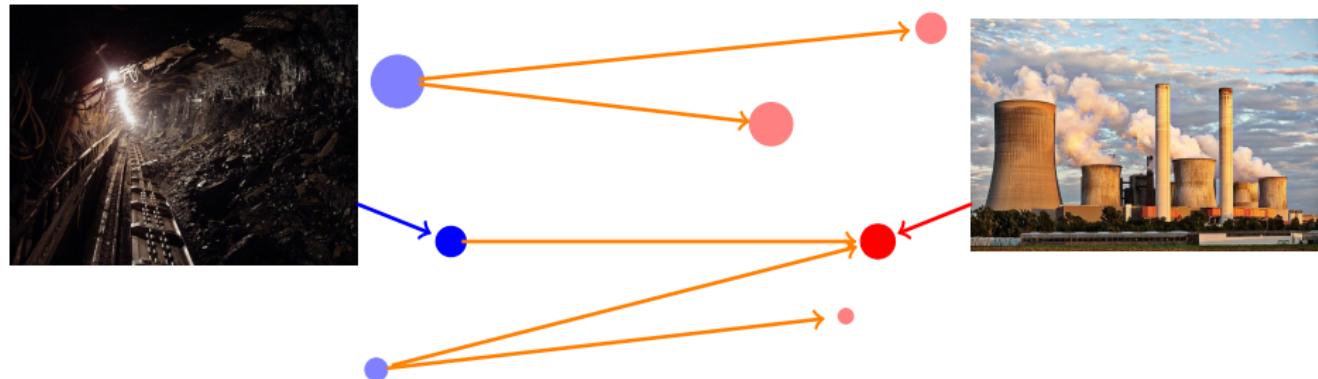
# Intuition of the Optimal Transport



# Intuition of the Optimal Transport

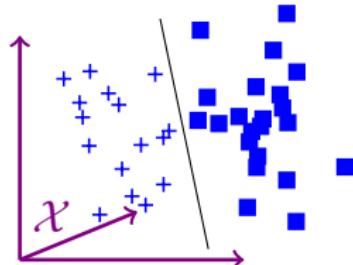


# Intuition of the Optimal Transport

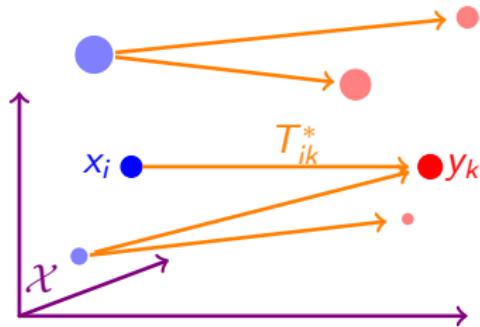


# Overview

## Machine Learning

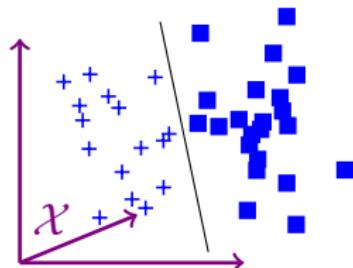


## Optimal Transport

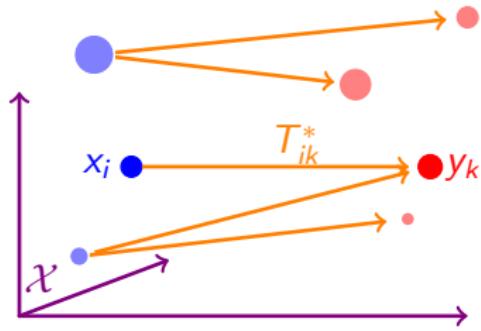


# Overview

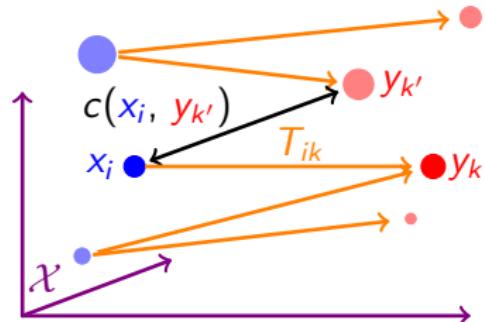
## Machine Learning



## Optimal Transport

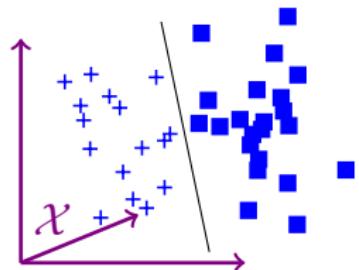


## Ground Metric

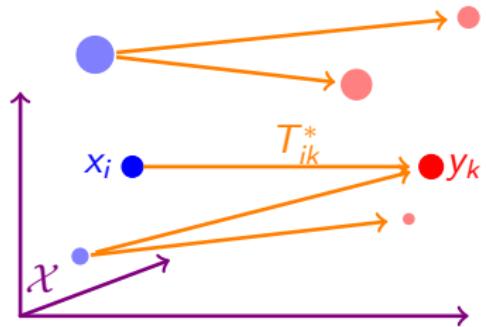


# Overview

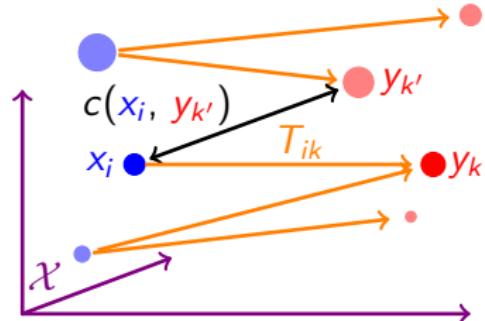
## Machine Learning



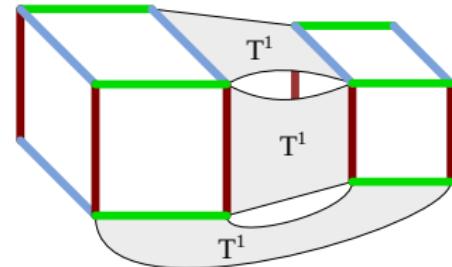
## Optimal Transport



## Ground Metric



## Generalized Framework

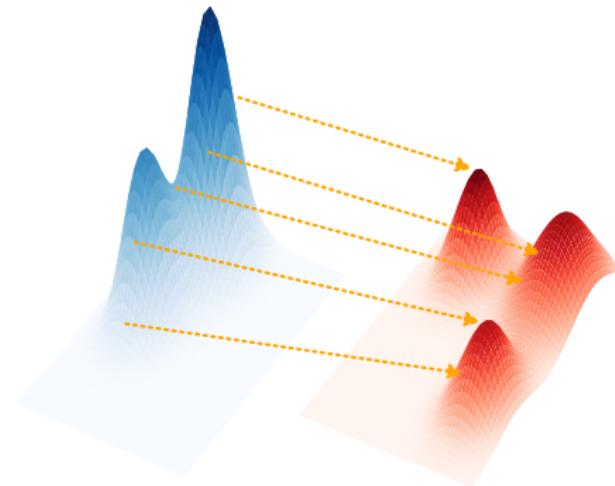


# Table of Contents

- 1 Background on Optimal Transport
  - Optimal Transport
- 2 Metric Learning for Optimal Transport
  - DA, OTDA, ML and MLOT
  - Illustration and experiments
- 3 Minimax OT
  - Intuition of Minimax problem and Cutting set algorithm
  - Stability and Experiments
- 4 Sampled Gromov Wasserstein
  - The Gromov Wasserstein Problem and how to approximate it
  - Comparison of the GW distance approximation
- 5 Optimal Tensor Transport
  - Optimal Tensor Transport formulation
  - Interest of such a formulation: Domain Adaptation

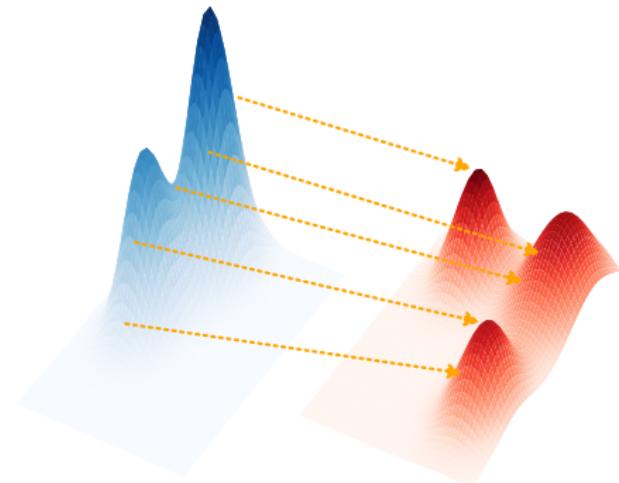
# Optimal Transport (OT)

Figure 1: Optimal Transport  
between two continuous  
distributions  $\mu \in \mathcal{P}(\mathcal{X})$  and  
 $\nu \in \mathcal{P}(\mathcal{Y})$ .



# Optimal Transport (OT)

Figure 1: Optimal Transport  
between two continuous  
distributions  $\mu \in \mathcal{P}(\mathcal{X})$  and  
 $\nu \in \mathcal{P}(\mathcal{Y})$ .



Optimal Transport [Kantorovich, 1942; Villani, 2008]

- Set of transport plans:  $\Pi(\mu, \nu)$  containing all  $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$  with marginal  $\mu$  and  $\nu$ .

$$\min_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) \quad (1)$$

# Discrete Optimal Transport

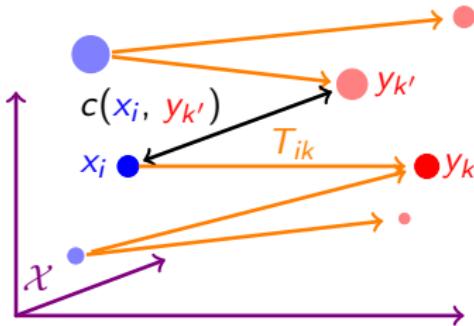


Figure 2: Optimal Transport between two distributions using the ground cost  $c$  to compare points.

# Discrete Optimal Transport

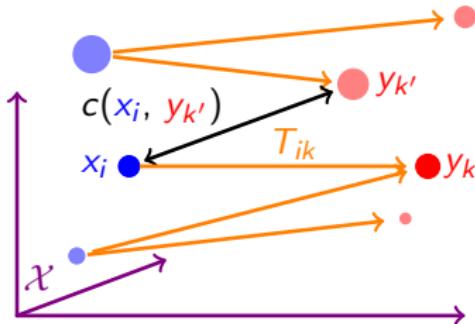


Figure 2: Optimal Transport between two distributions using the ground cost  $c$  to compare points.

Optimal Transport [Monge, 1781; Kantorovich, 1942; Villani, 2008]

- Two empirical distributions:  $\mu = \sum_{i=1}^I a_i \delta_{x_i}$  and  $\nu = \sum_{k=1}^K b_k \delta_{y_k}$ .
- Set of transport plans:  $\Pi(\mu, \nu) = \{T \in \mathbb{R}_+^{I \times K} \mid T^\top \mathbf{1}_I = b, T \mathbf{1}_K = a\}$

$$\min_{T \in \Pi(\mu, \nu)} \sum_{i,k=1}^{I,K} c(x_i, y_k) T_{ik} = \min_{T \in \Pi(\mu, \nu)} \langle C, T \rangle \quad (2)$$

# Polytope of all transport plans

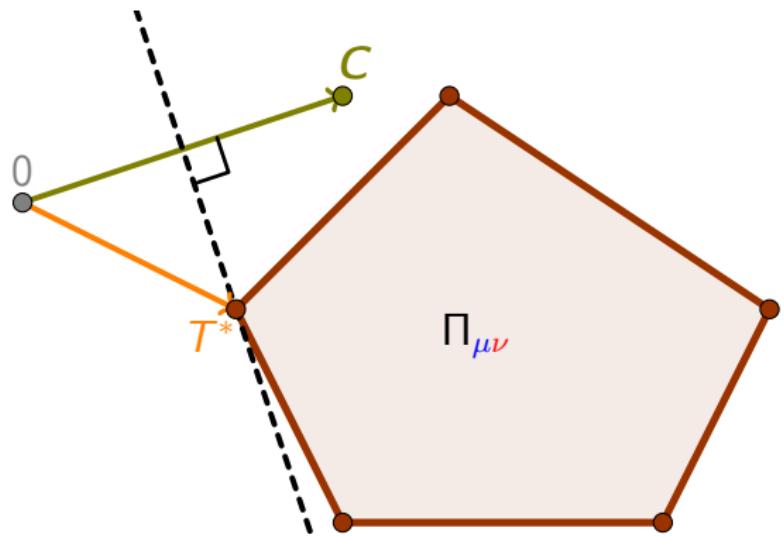


Figure 3: Representation of the OT problem induce by  $C$  and its solution  $T^*$  on the polytope  $\Pi_{\mu\nu}$ .

# Entropy regularization

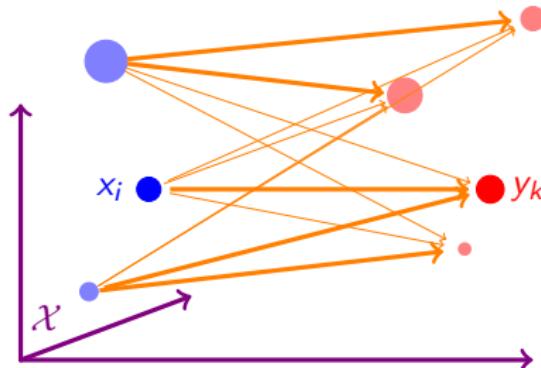


Figure 4: Solution of the entropy regularized Optimal Transport problem. The transport plan  $T$  is not sparse anymore.

# Entropy regularization

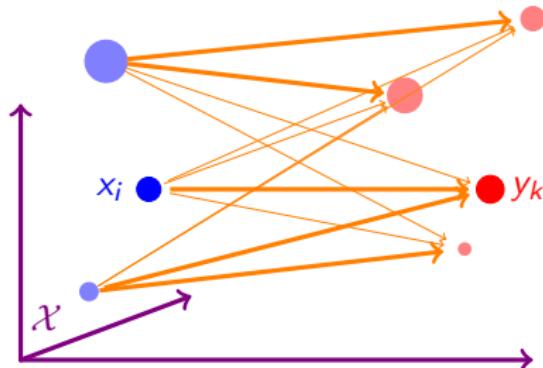


Figure 4: Solution of the entropy regularized Optimal Transport problem. The transport plan  $T$  is not sparse anymore.

# Entropy regularization

- Unique solution
- Solution not anymore on the edge of the transport plan
- Can take into account a prior on the transport plan
- Fast and simple solver available ( $P \times N^2$ )

Regularized OT Cuturi [2013]; Sinkhorn and Knopp [1967]

• Kullback-Leibler regularization  $KL(\textcolor{orange}{T} || T') = \sum_{i,k=1}^{I,K} \textcolor{orange}{T}_{ik} \log\left(\frac{\textcolor{orange}{T}_{ik}}{\bar{T}'_{ik}}\right)$ :

$$\min_{T \in \Pi(\mu, \nu)} \sum_{i,k=1}^{I,K} c(x_i, y_k) \textcolor{orange}{T}_{ik} + \epsilon KL(\textcolor{orange}{T} || ab^\top) \quad (3)$$

## Other solvers

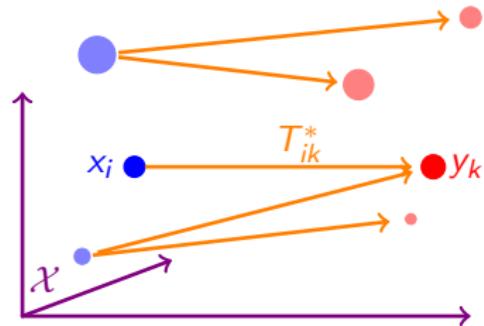
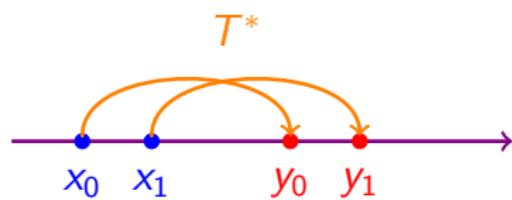


Figure 4: EMD: Based on the min cost flow problem [Ahuja et al., 1988] and  $N^3 \log(N)$  in the worst case.



$$c(x, y) = (x - y)^2$$

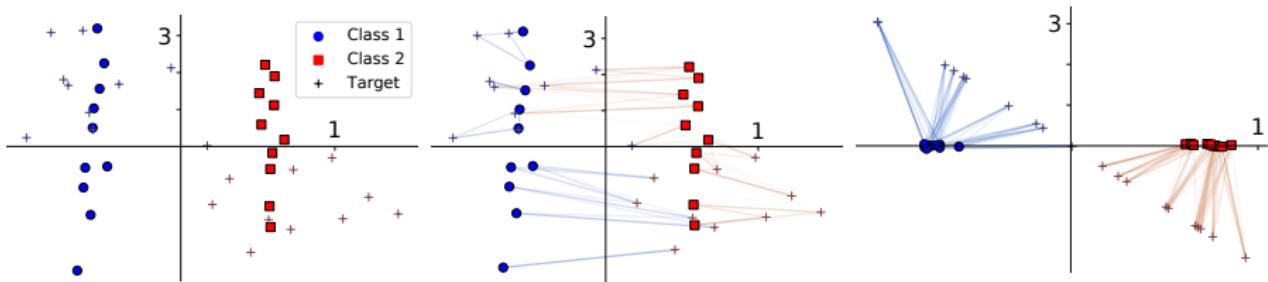
Figure 5: 1D Optimal Transport: Can be computed in  $N \log(N)$  by sorting the two lists.

# Table of Contents

- 1 Background on Optimal Transport
  - Optimal Transport
- 2 Metric Learning for Optimal Transport
  - DA, OTDA, ML and MLOT
  - Illustration and experiments
- 3 Minimax OT
  - Intuition of Minimax problem and Cutting set algorithm
  - Stability and Experiments
- 4 Sampled Gromov Wasserstein
  - The Gromov Wasserstein Problem and how to approximate it
  - Comparison of the GW distance approximation
- 5 Optimal Tensor Transport
  - Optimal Tensor Transport formulation
  - Interest of such a formulation: Domain Adaptation

# Metric Learning For Optimal Transport

Based on a published paper at the International Joint Conference on Artificial Intelligence (IJCAI) [Kerdoncuff et al., 2020]



# Domain Adaptation (DA): Intuition



- Internship supervision of Thibaud Leteno, in collaboration with DGFiP, on companies **fraud detection**.

- ① Usual Goal: Predict the companies that might defraud
- ② Goal in DA context: Predict the companies that might defraud in a **Covid crisis** context

# Intuition of Optimal Transport for Domain Adaptation

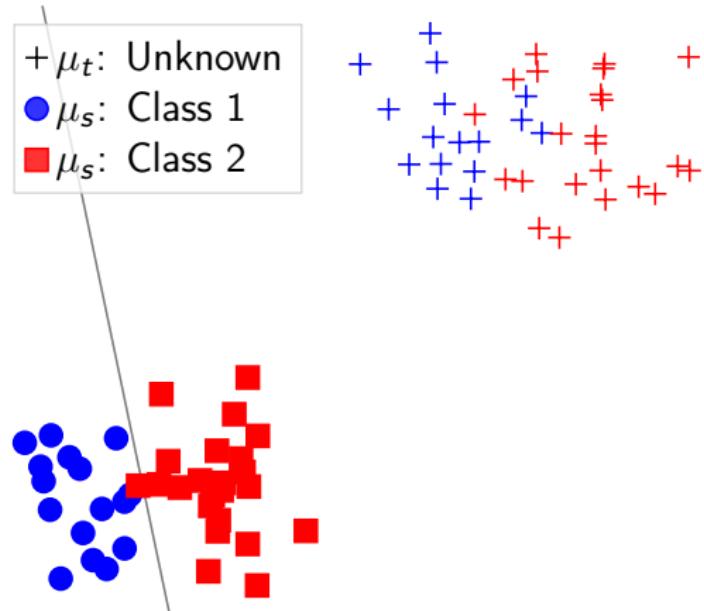


Figure 6: Optimal Transport for Domain Adaptation

# Intuition of Optimal Transport for Domain Adaptation

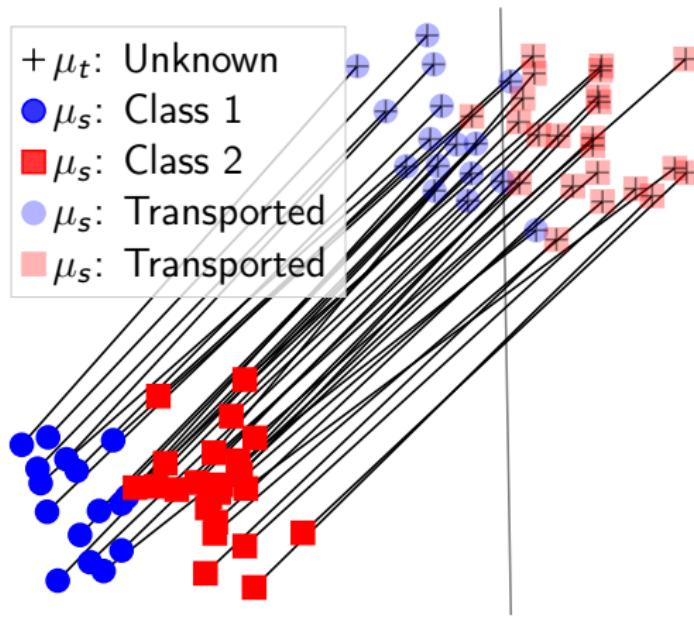


Figure 7: Optimal Transport for Domain Adaptation

# Intuition of Optimal Transport for Domain Adaptation

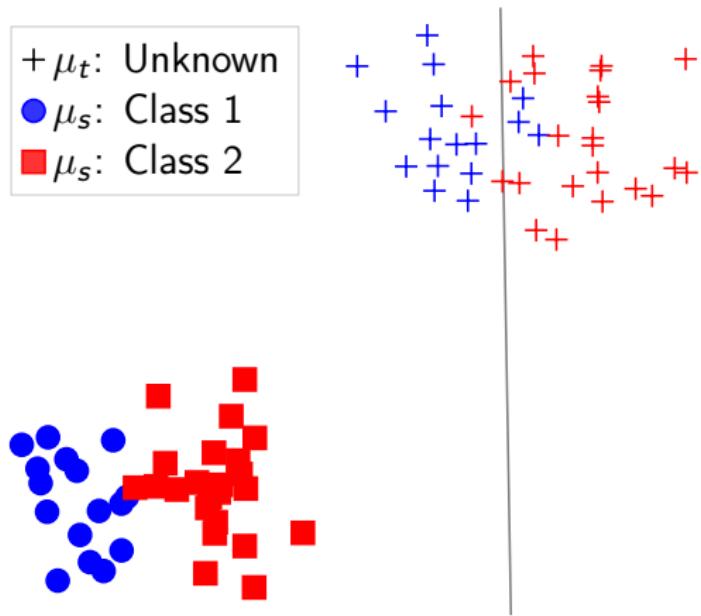


Figure 8: Optimal Transport for Domain Adaptation

# Optimal Transport for Domain Adaptation [Courty et al., 2017]

## Optimal Transport for Domain Adaptation

$$\min_{\mathcal{T} \in \Pi(\hat{\mu}_s, \hat{\mu}_t)} \langle \mathcal{T}, \mathcal{C} \rangle + \epsilon KL(\mathcal{T} || \mathbf{a} \times \mathbf{b}) + \lambda_c \left( \sum_{k=1}^K \sum_{cl=1}^{\#classes} \|\mathcal{T}(\mathcal{I}_{cl}, k)\|_2 \right) \quad (4)$$

- ✓ Points of different classes should **not** be sent to same location

# Metric learning

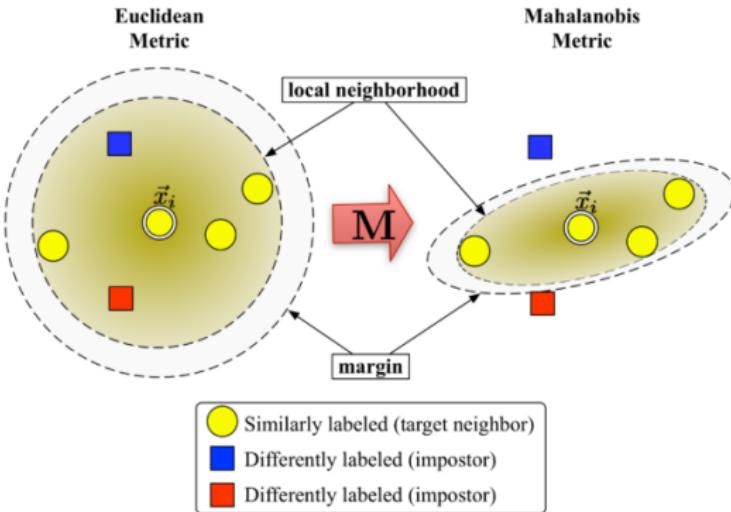


Figure 9: LMNN [Weinberger and Saul, 2009] CC-BY Mlguy (wikipedia)

## Mahalanobis metric

- $C_M^2(x, x') = (x - x')M(x - x')^T = \|Lx - Lx'\|_2^2$  with  $L^T L = M$

# Bound on the Target Risk

Classical form of Target Risk Bound [Ben-David et al., 2007]

$$\begin{aligned} \text{target risk} &\leq \text{source risk} \\ &+ \text{distance}(\mu_s, \mu_t) \\ &+ \lambda \text{ (error of the best classifier)} \end{aligned} \tag{5}$$

# Bound on the Target Risk

## Classical form of Target Risk Bound [Ben-David et al., 2007]

$$\begin{aligned} \text{target risk} &\leq \text{source risk} \\ &+ \text{distance}(\mu_s, \mu_t) \\ &+ \lambda \text{ (error of the best classifier)} \end{aligned} \tag{5}$$

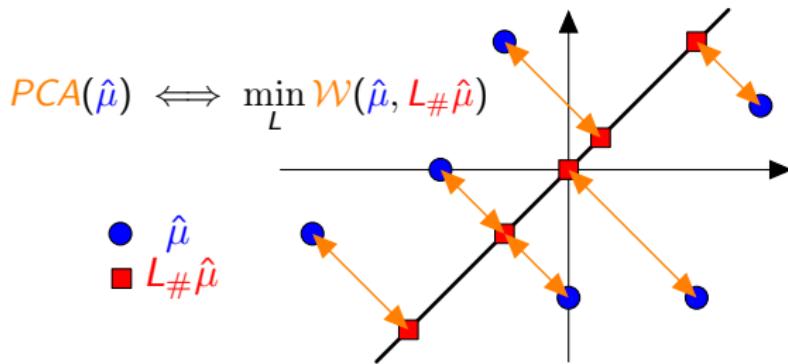
## Our Target Risk Bound (based on [Redko et al., 2017])

$$\begin{aligned} \text{target risk} &\leq \text{source risk} + \mathcal{W}(L_{s\#}\hat{\mu}_s, L_{t\#}\hat{\mu}_t) \\ &+ \mathcal{W}(\hat{\mu}_s, L_{s\#}\hat{\mu}_s) + \mathcal{W}(L_{t\#}\hat{\mu}_t, \hat{\mu}_t) \\ &+ \mathcal{W}(\mu_s, \hat{\mu}_s) + \mathcal{W}(\hat{\mu}_t, \mu_t) + \lambda \end{aligned} \tag{6}$$

# Equivalence between PCA and Wasserstein Distance

## Our Target Risk Bound

$$\begin{aligned} \text{target risk} &\leq \text{source risk} + \mathcal{W}(L_s \# \hat{\mu}_s, L_t \# \hat{\mu}_t) \\ &\quad + \mathcal{W}(\hat{\mu}_s, L_s \# \hat{\mu}_s) + \mathcal{W}(\hat{\mu}_t, L_t \# \hat{\mu}_t) \\ &\quad + \mathcal{W}(\mu_s, \hat{\mu}_s) + \mathcal{W}(\hat{\mu}_t, \mu_t) + \lambda \end{aligned} \tag{7}$$



# Illustration of the usefulness of Metric Learning

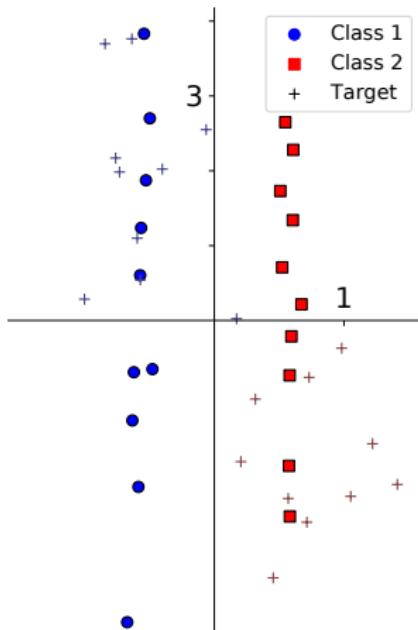


Figure 10: Dataset

# Illustration of the usefulness of Metric Learning

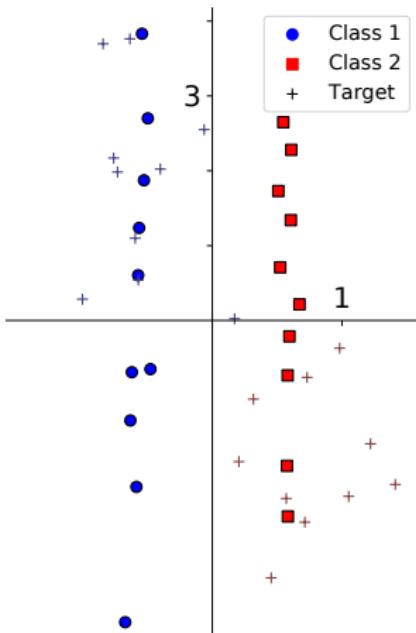


Figure 10: Dataset

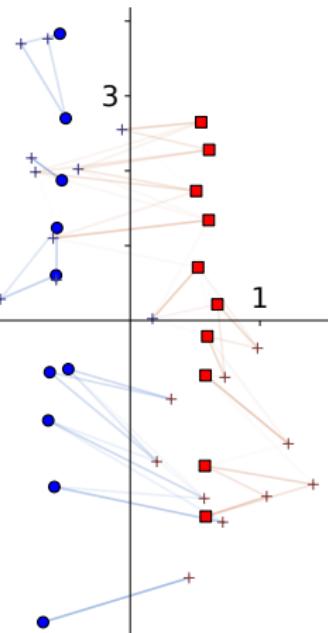


Figure 11: OTDA

# Illustration of the usefulness of Metric Learning

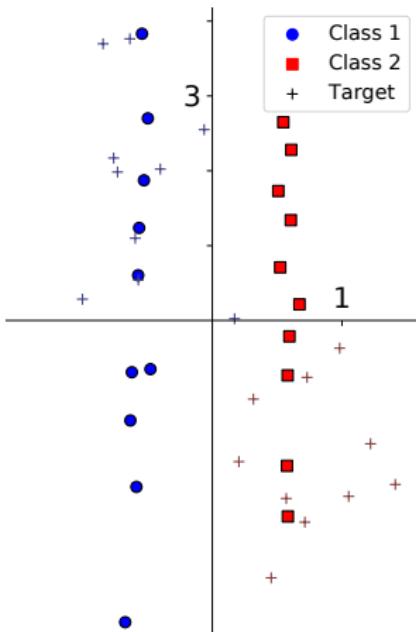


Figure 10: Dataset

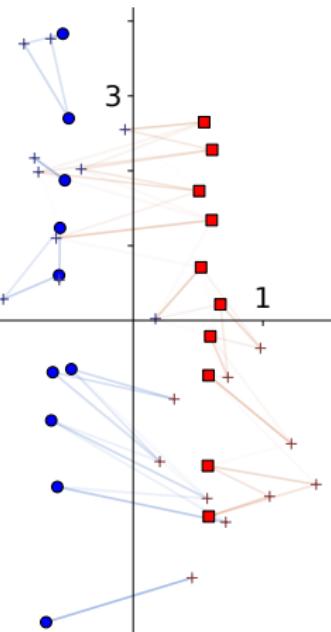


Figure 11: OTDA

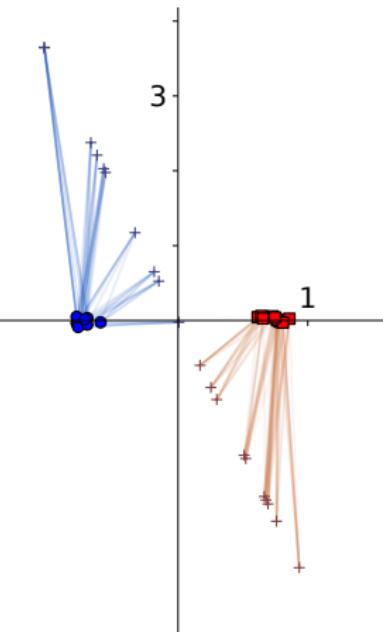


Figure 12: MLOT

# Cross validation of hyperparameters

- In **Unsupervised Domain Adaptation** the target labels are **not available**
- We use the Reverse Validation of Zhong et al. [2010]

Method	OT	TCA	LMNN	SA	JDOT	OTDA	OTDA <sub>p</sub>	MLOT
Classical Validation	45.3	45.3	47.9	47.4	48.5	52.8	54	<b>55.1</b>
Reverse Validation	42.2	44.5	44.8	45.8	47.0	48.2	48.8	<b>49.7</b>

Table 1: Accuracy on Office-Caltech dataset [Gong et al., 2012] with SURF features.

# OTDA vs MLOT

Figure 13: MLOT - OTDA

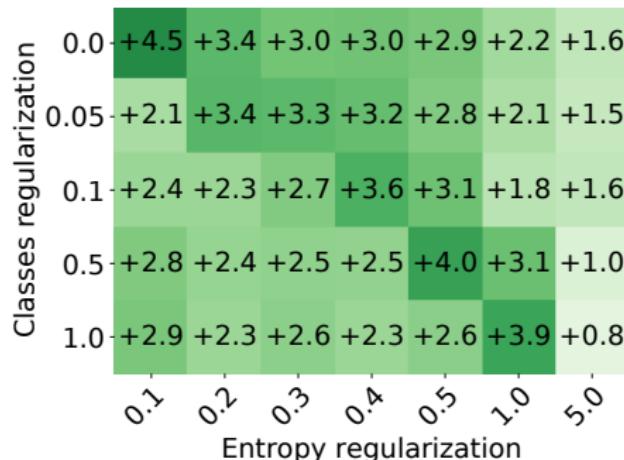


Figure 14: MLOT

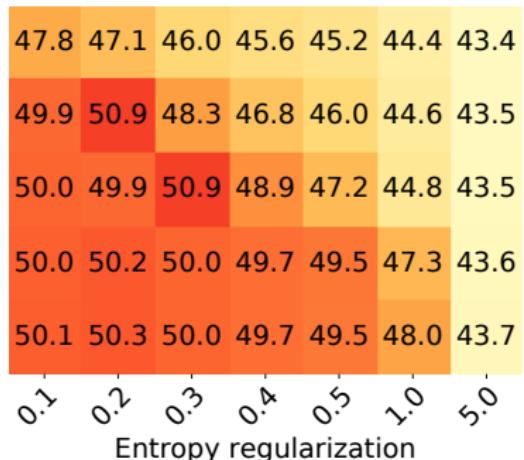


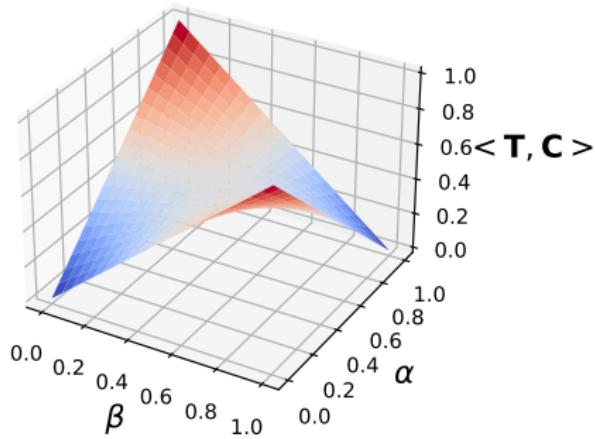
Figure 15: Accuracy of OTDA and MLOT on Office-Caltech dataset [Gong et al., 2012] with SURF features.

# Table of Contents

- 1 Background on Optimal Transport
  - Optimal Transport
- 2 Metric Learning for Optimal Transport
  - DA, OTDA, ML and MLOT
  - Illustration and experiments
- 3 Minimax OT
  - Intuition of Minimax problem and Cutting set algorithm
  - Stability and Experiments
- 4 Sampled Gromov Wasserstein
  - The Gromov Wasserstein Problem and how to approximate it
  - Comparison of the GW distance approximation
- 5 Optimal Tensor Transport
  - Optimal Tensor Transport formulation
  - Interest of such a formulation: Domain Adaptation

# A Swiss Army Knife for Minimax Optimal Transport

Based on a published paper at International Conference on Machine Learning (ICML) 2020 [Dhouib et al., 2020]



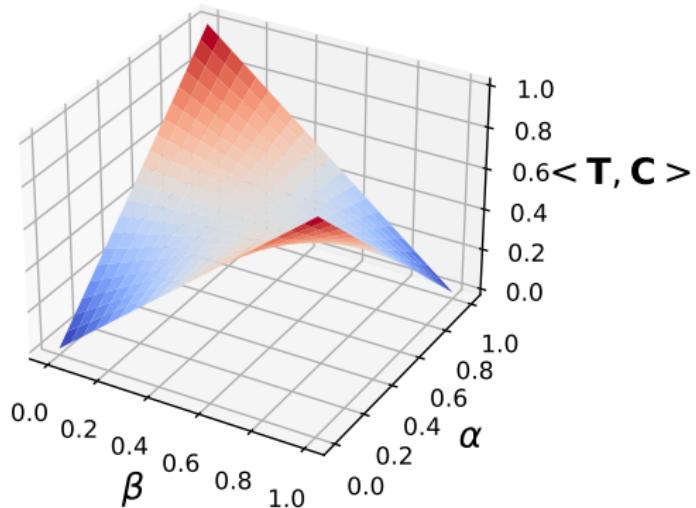
# Intuition of the OT Minimax problem

## Cost matrices

- $C_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
- $C_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
- $C = \alpha C_1 + (1 - \alpha) C_2$

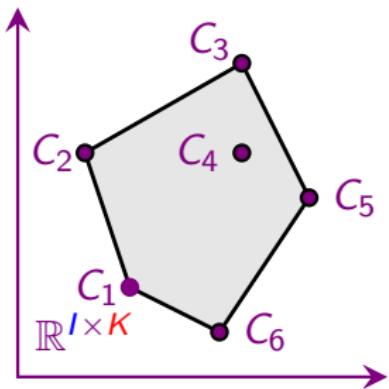
## Associated Transport Plan

- $T_1^* = \begin{pmatrix} 0 & 0.5 \\ 0.5 & 0 \end{pmatrix}$
- $T_2^* = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$
- $T = \beta T_1^* + (1 - \beta) T_2^*$



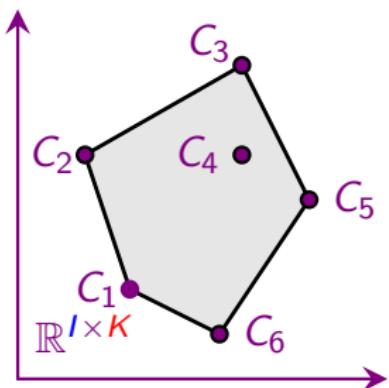
# Different type of set

- Convex hull of finite set of cost matrices:

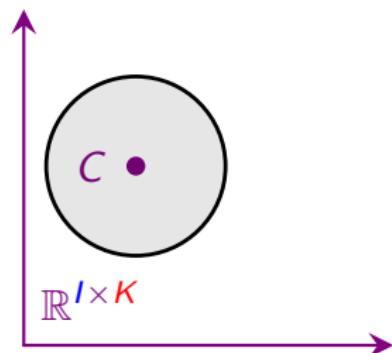


# Different type of set

- Convex hull of finite set of cost matrices:



- Mahalanobis ball centered at a matrix  $C$ :



# Cutting set algorithm [Mutapcic and Boyd, 2009]

## Proper definition of the Minimax problem

$$\mathcal{W}_{\mathcal{C}} = \min_{T \in \text{Conv}(\mathcal{P})} \max_{C \in \mathcal{C}} \langle T, C \rangle$$

# Cutting set algorithm [Mutapcic and Boyd, 2009]

## Proper definition of the Minimax problem

$$\begin{aligned}\mathcal{W}_{\mathcal{C}} = \min_{T \in \text{Conv}(\mathcal{P})} \max_{C \in \mathcal{C}} \langle T, C \rangle &= C^* \in \operatorname{argmax}_{C \in \mathcal{C}, \omega \geq 0} \omega \\ \text{s.t. } \langle T, C \rangle &\geq \omega, \forall T \in \mathcal{P}\end{aligned}\quad (8)$$

---

### Algorithm 2 Cutting set method for RKP( $\Pi, \mathcal{C}$ )

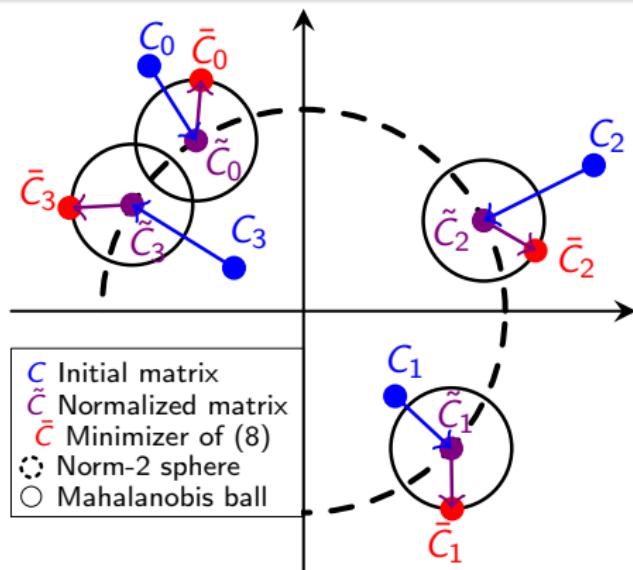
---

- 1: **Input:**  $\mathcal{C}, \mathcal{P}_0 \subset \Pi$ , thd
- 2: **while**  $err > \text{thd}$  **do**
- 3:     Solve (8) to obtain  $(\omega_t, C_t)$  and  $(q_0, \dots, q_{|\mathcal{P}_t|-1})$
- 4:     Find  $T_t \in \operatorname{argmin}_{T \in \Pi} \langle T, C_t \rangle$
- 5:      $err \leftarrow (\omega_t - \langle T_t, C_t \rangle) / \langle T_t, C_t \rangle$
- 6:      $\mathcal{P}_{t+1} = \mathcal{P}_t \cup \{T_t\}$
- 7:      $t \leftarrow t + 1$
- 8: **end while**
- 9: **return**  $\sum_{l=0}^{|\mathcal{P}_t|-1} q_l T_l, C_t$

# Wasserstein stability

Wasserstein stability of a matrix  $C$  induce by  $\mu$  and  $\nu$

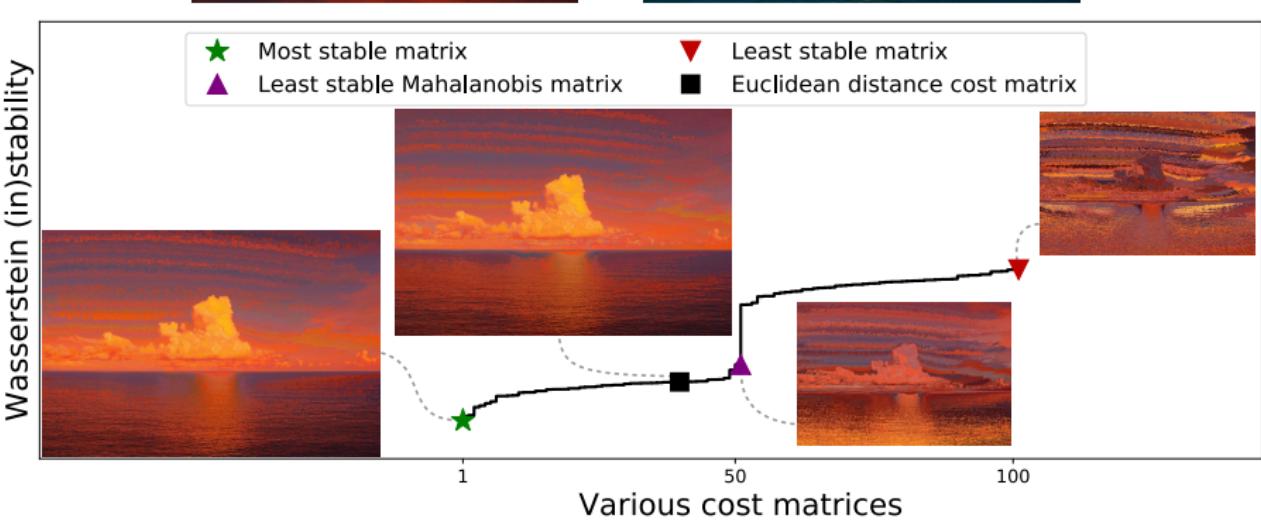
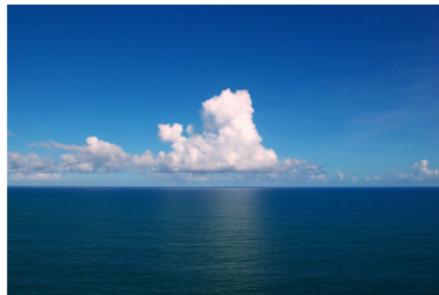
$$\mathcal{WS}_C = \mathcal{W}_{C_C}(\mu, \nu) - \mathcal{W}_C(\mu, \nu)$$



# Usefulness of the Wasserstein stability (Color transfer)



# Usefulness of the Wasserstein stability (Color transfer)

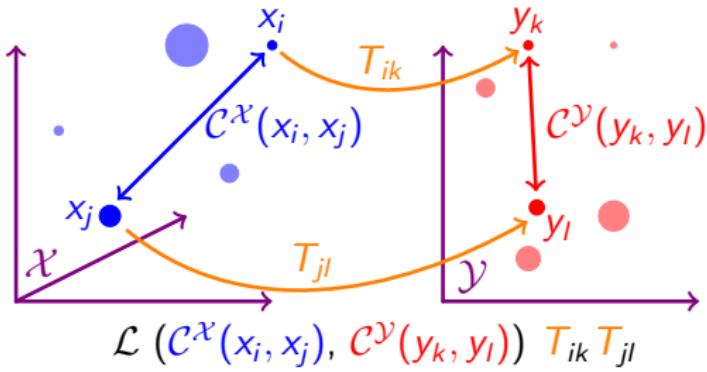


# Table of Contents

- 1 Background on Optimal Transport
  - Optimal Transport
- 2 Metric Learning for Optimal Transport
  - DA, OTDA, ML and MLOT
  - Illustration and experiments
- 3 Minimax OT
  - Intuition of Minimax problem and Cutting set algorithm
  - Stability and Experiments
- 4 Sampled Gromov Wasserstein
  - The Gromov Wasserstein Problem and how to approximate it
  - Comparison of the GW distance approximation
- 5 Optimal Tensor Transport
  - Optimal Tensor Transport formulation
  - Interest of such a formulation: Domain Adaptation

## Sampled Gromov Wasserstein

Based on a published paper in the Machine Learning Journal (MLJ) [Kerdoncuff et al., 2021] and presented at the ECML-PKDD 2021 conference.



# The Gromov Wasserstein problem

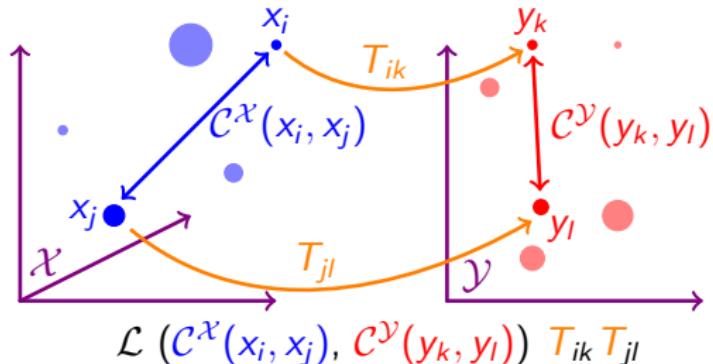


Figure 16: Gromov Wasserstein between two distributions in two vector spaces  $\mathcal{X}$  and  $\mathcal{Y}$ , with one term from Equation (9).

Gromov Wasserstein (GW) Memoli [2007]; Peyré et al. [2016]

- The GW distance rely only on intra-pairwise distance,

$$\min_{\mathbf{T} \in \Pi(\mu, \nu)} \mathcal{E}(\mathbf{T}, \mathbf{T}) = \min_{\mathbf{T} \in \Pi(\mu, \nu)} \sum_{ijkl=1}^{I,I,K,K} \mathcal{L}(\mathcal{C}^{\mathcal{X}}(x_i, x_j), \mathcal{C}^{\mathcal{Y}}(y_k, y_l)) T_{ik} T_{jl}. \quad (9)$$

# The mostly used solver: Entropy Gromov Wasserstein

---

## Algorithm 3 Entropy Gromov Wasserstein (EGW) Peyré et al. [2016]

---

**Require:**  $a, b, \mathcal{C}^x, \mathcal{C}^y, \mathcal{L}, \epsilon$

1:  $T_0 = ab^\top$

2: **for**  $s = 0$  **to**  $S-1$  **do**

3:      $\Lambda = \nabla_{\mathcal{T}} \mathcal{E}(\mathcal{T}, \mathcal{T}) = \sum_{jl=1}^{I, K} \mathcal{L}(\mathcal{C}^x(\cdot, x_j), \mathcal{C}^y(\cdot, y_l)) \mathcal{T}_{jl}$

4:      $T'_s = \min_{T' \in \Pi(\mu, \nu)} \langle \Lambda, T' \rangle + \epsilon KL(T' || ab^\top)$

5:      $\mathcal{T}_{s+1} = (1 - 1) \mathcal{T}_s + 1 T'_s$

6: **end for**

---

# The mostly used solver: Entropy Gromov Wasserstein

---

## Algorithm 4 Entropy Gromov Wasserstein (EGW) Peyré et al. [2016]

---

**Require:**  $a, b, \mathcal{C}^{\mathcal{X}}, \mathcal{C}^{\mathcal{Y}}, \mathcal{L}, \epsilon$

1:  $T_0 = ab^\top$

2: **for**  $s = 0$  **to**  $S-1$  **do**

3:      $\Lambda = \nabla_{\mathcal{T}} \mathcal{E}(\mathcal{T}, \mathcal{T}) = \sum_{jl=1}^{I, K} \mathcal{L}(\mathcal{C}^{\mathcal{X}}(\cdot, x_j), \mathcal{C}^{\mathcal{Y}}(\cdot, y_l)) \mathcal{T}_{jl}$

4:      $T'_s = \min_{T' \in \Pi(\mu, \nu)} \langle \Lambda, T' \rangle + \epsilon KL(T' || ab^\top)$

5:      $\mathcal{T}_{s+1} = (1 - 1) \mathcal{T}_s + 1 T'_s$

6: **end for**

---

Relation with the Frank-Wolfe (FW) algorithm [Frank et al., 1956]

- EGW with no entropy ( $\epsilon = 0$ )  $\iff$  FW with full step size ( $\alpha = 1$ ).

## Existing limitations

- Time complexity for computing the gradient  $\nabla_{\mathcal{T}} \mathcal{E}(\mathcal{T}, \mathcal{T})$  is  $O(N^4)$ .
- Time complexity can be reduced to  $O(N^3)$  but only for specific loss functions  $\mathcal{L}(x, y) = f_1(x) + h_1(x)h_2(y) + f_2(y)$ .
- The convergence to stationary points of such methods has not been proven.

## Contributions

- ✓ Approximate the gradient  $\nabla_{\mathcal{T}} \mathcal{E}(\mathcal{T}, \mathcal{T})$  in  $O(M \times N^2)$ .
- ✓ Propose a convergence bound to stationary points.
- ✓ Explore a very fast variant with  $M = 1$  using the 1D OT solver.

## Sampled Gromov Wasserstein idea (SaGroW)

We want to minimize:

$$\min_{T' \in \Pi(\mu, \nu)} \langle \nabla_{\mathcal{T}} \mathcal{E}(\mathcal{T}, \mathcal{T}), T' \rangle = \min_{T' \in \Pi(\mu, \nu)} \left\langle \sum_{j,l} \mathcal{T}_{jl} \mathbf{L}_{j.l}, T' \right\rangle. \quad (10)$$

We defined  $\mathbf{C}$  the (matrix) random variable,  $\mathbb{P}(\mathbf{C} = L_{j.l}) = \mathcal{T}_{jl}$ ,

$$\min_{T' \in \Pi(\mu, \nu)} \langle \nabla_{\mathcal{T}} \mathcal{E}(\mathcal{T}, \mathcal{T}), T' \rangle = \min_{T' \in \Pi_{\mu\nu}} \langle \mathbb{E}(\Lambda), T' \rangle. \quad (11)$$

Then sample  $M$  times this random variable,

$$\min_{T' \in \Pi(\mu, \nu)} \langle \nabla_{\mathcal{T}} \mathcal{E}(\mathcal{T}, \mathcal{T}), T' \rangle \approx \min_{T' \in \Pi_{\mu\nu}} \left\langle \frac{1}{M} \sum_{m=1}^M \mathbf{C}^m, T' \right\rangle. \quad (12)$$

# Summary of SaGroW algorithm.

---

## Algorithm 5 SaGroW

---

**Require:**  $a, b, \mathcal{C}^{\mathcal{X}}, \mathcal{C}^{\mathcal{Y}}, \mathcal{L}, M, \epsilon, \alpha$

1:  $T_0 = ab^\top$

2: **for**  $s = 0$  **to**  $S-1$  **do**

3:  $(j_m, l_m) \sim \text{Sample}(T_s)$   $\forall m \in [1, M]$

4:  $\widehat{\Lambda} = \frac{1}{M} \sum_{m=1}^M \mathcal{L}(\mathcal{C}^{\mathcal{X}}(\cdot, x_{j_m}), \mathcal{C}^{\mathcal{Y}}(\cdot, y_{l_m}))$

5:  $T'_s = \min_{T' \in \Pi(\mu, \nu)} \langle \widehat{\Lambda}, T' \rangle + \epsilon KL(T' || ab^\top)$

6:  $T_{s+1} = (1 - \alpha) T_s + \alpha T'_s$

7: **end for**

---

## Bound for convergence to a stationary point

- The FW gap:  $G(\textcolor{orange}{T}) = \mathcal{E}(\textcolor{orange}{T}, \textcolor{orange}{T}) - \min_{T' \in \Pi(\mu, \nu)} \mathcal{E}(\textcolor{orange}{T}, T').$
- $G(\textcolor{orange}{T}) = 0 \iff \textcolor{orange}{T}$  is a stationary point of GW.

## Bound for convergence to a stationary point

- The FW gap:  $G(\textcolor{orange}{T}) = \mathcal{E}(\textcolor{orange}{T}, \textcolor{orange}{T}) - \min_{T' \in \Pi(\mu, \nu)} \mathcal{E}(\textcolor{orange}{T}, T').$
- $G(\textcolor{orange}{T}) = 0 \iff \textcolor{orange}{T}$  is a stationary point of GW.

Theorem (Based on Reddi et al. [2016])

For any  $L_{ijkl} \in [0, B]$ , for any distributions  $\mu$  and  $\nu$  with uniform weights  $a$  and  $b$  respectively, for any optimal solution  $\textcolor{orange}{T}^*$  of Problem (9), on average for the transport plan  $\bar{T}$  uniformly sampled from  $(\textcolor{orange}{T}_s)_{s \in [0, S-1]}$ , on average over all the samplings, the following bound holds:

$$\mathbb{E}(G(\bar{T})) \leq \sqrt{\frac{2B(\mathcal{E}(\textcolor{orange}{T}_0) - \mathcal{E}(\textcolor{orange}{T}^*))N}{S}} + B\sqrt{\frac{2N}{M}} + \epsilon \log(N).$$

## Kullback Leiber regularization with previous iteration [Xu et al., 2019]

$$\langle \widehat{\Lambda}, T' \rangle + \epsilon KL(T' || T) = \langle \widehat{\Lambda} - \epsilon \log(T), T' \rangle + \epsilon \sum_{ik} T'_{ik} \log(T'_{ik}) \quad (13)$$

---

### Algorithm 6 SaGroW with KL regularization

---

**Require:**  $a, b, \mathcal{C}^x, \mathcal{C}^y, \mathcal{L}, M, \epsilon, \alpha$

- 1:  $T_0 = ab^\top$
- 2: **for**  $s = 0$  **to**  $S-1$  **do**
- 3:      $(j_m, l_m) \sim \text{Sample}(\textcolor{brown}{T}_s) \forall m \in \llbracket 1, M \rrbracket$
- 4:      $\widehat{\Lambda} = \frac{1}{M} \sum_{m=1}^M \mathcal{L}(\mathcal{C}^x(\cdot, x_{j_m}), \mathcal{C}^y(\cdot, y_{l_m}))$
- 5:      $\textcolor{brown}{T}_{s+1} = \min_{T' \in \Pi(\mu, \nu)} \langle \widehat{\Lambda} - \epsilon \log(\textcolor{brown}{T}_s), T' \rangle - \epsilon \mathcal{H}(T')$
- 6: **end for**

# Pointwise Gromov Wasserstein (PoGroW): sample $M = 1$

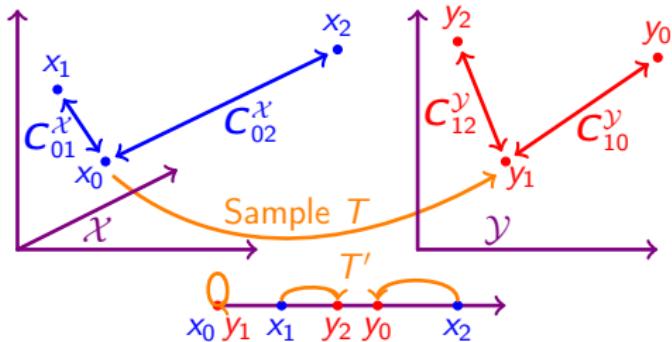


Figure 17: Intuition behind PoGroW when  $j, l = 0, 1$  are sampled from  $T$ : only the distances to  $x_0$  in  $\mathcal{X}$  (on the left) and to  $y_1$  in  $\mathcal{Y}$  (on the right) characterize a pair, and then  $T'$  can be computed in  $O(N \log N)$  like in 1D OT.

# Speed and performances

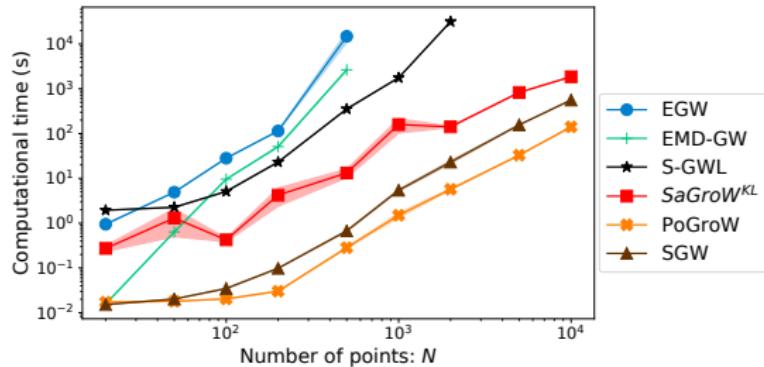


Figure 18: Computational time of various methods to compute the distance between samples from two mixtures of gaussians. The mean and the standard deviation over 10 runs are reported.

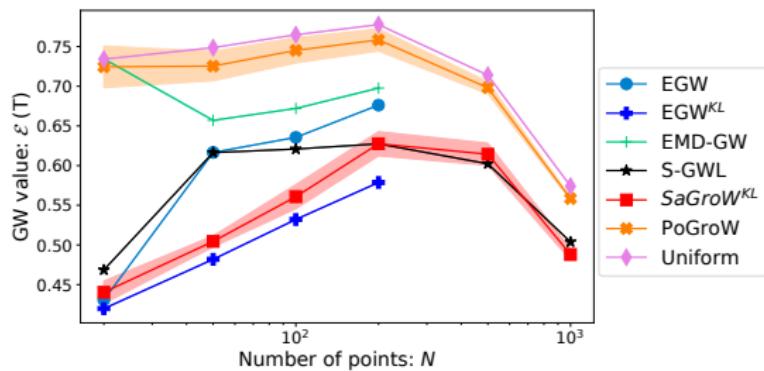


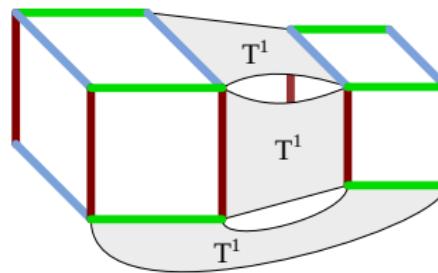
Figure 19: W distance estimation between synthetic graphs Brandes et al. [2003]. The mean and standard deviation over 10 runs are reported for the stochastic methods.

# Table of Contents

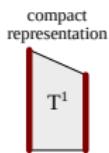
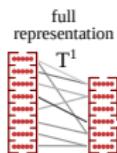
- 1 Background on Optimal Transport
  - Optimal Transport
- 2 Metric Learning for Optimal Transport
  - DA, OTDA, ML and MLOT
  - Illustration and experiments
- 3 Minimax OT
  - Intuition of Minimax problem and Cutting set algorithm
  - Stability and Experiments
- 4 Sampled Gromov Wasserstein
  - The Gromov Wasserstein Problem and how to approximate it
  - Comparison of the GW distance approximation
- 5 Optimal Tensor Transport
  - Optimal Tensor Transport formulation
  - Interest of such a formulation: Domain Adaptation

# Optimal Tensor Transport

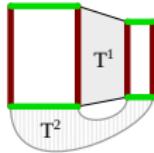
Based on an unpublished paper under submission at the AAAI conference.



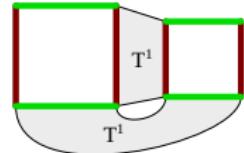
# Motivation of Optimal Tensor Transport



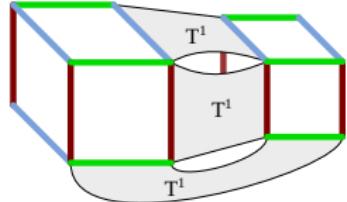
a)  $\text{OTT}_{11}$  (OT) ( $F=5$  features)



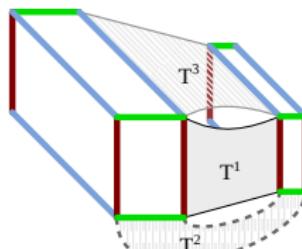
b)  $\text{OTT}_{12}$  (Co-OT)



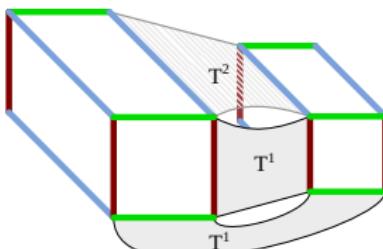
c)  $\text{OTT}_{11}$  (GW)



d)  $\text{OTT}_{111}$  (triplets)



e)  $\text{OTT}_{123}$  (triCo-OT)



f)  $\text{OTT}_{112}$  (GW collections)

## Optimal Transport

$$\text{OT} = \min_{\substack{\mathcal{T}^1 \in \Pi \\ a^1 b^1}} \sum_{i_1=1}^{l_1} \sum_{k_1=1}^{K_1} \mathcal{L}(X_{i_1}, Y_{k_1}) \mathcal{T}_{i_1 k_1}^1 \quad (14)$$

# Mathematical formulations

## Optimal Transport

$$\text{OT} = \min_{\substack{\mathcal{T}^1 \in \Pi \\ a^1 b^1}} \sum_{i_1=1}^{I_1} \sum_{k_1=1}^{K_1} \mathcal{L}(X_{i_1}, Y_{k_1}) \mathcal{T}_{i_1 k_1}^1 \quad (14)$$

## Co-Optimal Transport [Redko et al., 2020]

$$\text{Co-OT} = \min_{\substack{\mathcal{T}^1 \in \Pi \\ a^1 b^1}} \min_{\substack{\mathcal{T}^2 \in \Pi \\ a^2 b^2}} \sum_{i_1, i_2=1}^{I_1, I_2} \sum_{k_1, k_2=1}^{K_1, K_2} \mathcal{L}(X_{i_1 i_2}, Y_{k_1 k_2}) \mathcal{T}_{i_1 k_1}^1 \mathcal{T}_{i_2 k_2}^2 \quad (15)$$

# Mathematical formulations

## Optimal Transport

$$\text{OT} = \min_{\substack{T^1 \in \Pi \\ a^1 b^1}} \sum_{i_1=1}^{I_1} \sum_{k_1=1}^{K_1} \mathcal{L}(X_{i_1}, Y_{k_1}) T^1_{i_1 k_1} \quad (14)$$

## Co-Optimal Transport [Redko et al., 2020]

$$\text{Co-OT} = \min_{\substack{T^1 \in \Pi \\ a^1 b^1}} \min_{\substack{T^2 \in \Pi \\ a^2 b^2}} \sum_{i_1, i_2=1}^{I_1, I_2} \sum_{k_1, k_2=1}^{K_1, K_2} \mathcal{L}(X_{i_1 i_2}, Y_{k_1 k_2}) T^1_{i_1 k_1} T^2_{i_2 k_2} \quad (15)$$

## Gromov Wasserstein

$$\text{GW} = \min_{\substack{T^1 \in \Pi \\ a^1 b^1}} \sum_{i_1, i_2=1}^{I_1, I_2} \sum_{k_1, k_2=1}^{K_1, K_2} \mathcal{L}(X_{i_1 i_2}, Y_{k_1 k_2}) T^1_{i_1 k_1} T^1_{i_2 k_2} \quad (16)$$

# Mathematical formulations

## Optimal Tensor Transport with (1, 2, 2)

$$\min_{\substack{\mathcal{T}^1 \in \Pi_{a^1 b^1} \\ \mathcal{T}^2 \in \Pi_{a^2 b^2}}} \sum_{i_1, \dots, i_D=1} \sum_{k_1, \dots, k_D=1} \mathcal{L}(X_{i_1 \dots i_D}, Y_{k_1 \dots k_D}) \mathcal{T}_{i_1 k_1}^1 \mathcal{T}_{i_2 k_2}^2 \mathcal{T}_{i_3 k_3}^2$$

# Mathematical formulations

## Optimal Tensor Transport with (1, 2, 2)

$$\min_{T^1 \in \Pi_{a^1 b^1}, T^2 \in \Pi_{a^2 b^2}} \sum_{i_1, \dots, i_D=1} \sum_{k_1, \dots, k_D=1} \mathcal{L}(X_{i_1 \dots i_D}, Y_{k_1 \dots k_D}) T^1_{i_1 k_1} T^2_{i_2 k_2} T^2_{i_3 k_3}$$

## Optimal Tensor Transport

$$\text{OTT}_f(X, Y, (a^e)_{e \in [1, E]}, (b^e)_{e \in [1, E]}) = \min_{\forall e \quad T^e \in \Pi_{a^e b^e}} \mathcal{E}_f(X, Y, (T^e)_{e \in [1, E]})$$
$$\mathcal{E}_f(X, Y, (T^e)_{e \in [1, E]}) = \sum_{i_1, \dots, i_D=1} \sum_{k_1, \dots, k_D=1} \mathcal{L}(X_{i_1 \dots i_D}, Y_{k_1 \dots k_D}) \prod_{d=1}^D T^e_{i_d k_d}$$

# Stochastic Mirror Descent to solve the OTT problem

The gradient require  $N^{2D}$  operation

$$\nabla_{\mathcal{T}^e} \mathcal{E}_f = \sum_{\{d' | f(d') = e\}} \mathbb{E} (\Lambda^{d'})$$

---

## Algorithm 7 Alternated Stochastic Mirror Descent algorithm

---

**Require:**  $(a^e)_{e \in [1, E]}, (b^e)_{e \in [1, E]}, X, Y, \mathcal{L}, M, \epsilon$

1:  $\forall e \in [1, E], \mathcal{T}^e = a^e b^{e\top}$

2: **for**  $s = 0$  **to**  $S-1$  **do**

3:     **for**  $e = 1$  **to**  $E$  **do**

4:          $\widehat{\nabla_{\mathcal{T}^e} \mathcal{E}_f} = M$  samples of the gradient

5:          $\mathcal{T}^e = \min_{T' \in \Pi(a^e, b^e)} \left\langle \widehat{\nabla_{\mathcal{T}^e} \mathcal{E}_f}, T' \right\rangle + \epsilon KL(T' || \mathcal{T}^e)$

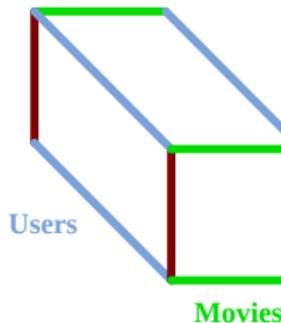
6:     **end for**

7: **end for**

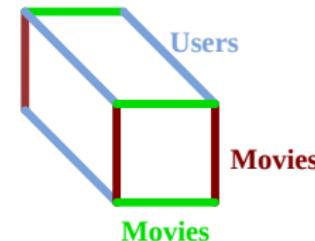
---

# Interest of such formulation: Domain Adaptation

## Domain Adaptation: predict the type of movies



Old movies



New movies

Datasets	Rdm	SVM	S-GWL	GW	Co-OT	OTT
AVG	$50.0 \pm 3.3$	58.6	63.8	64.9	68.2	<b><math>77.3 \pm 3.1</math></b>
AVG (best)	$50.0 \pm 3.3$	58.6	66.3	71.0	70.7	<b><math>78.9 \pm 2.9</math></b>

# Conclusion

## ① MLOT

- ✓ Interest of **learning a ground metric** for a DA task.

## ② Minimax

- ✓ General **cutting set** method for minimax OT problems
- ✓ **Wasserstein stability** definition

## ③ Sampled GW

- ✓ Fast approximation of the GW distance with **any losses**
- ✓ Theoretical **convergence bound** to stationary points
- ✓ **Very fast approximation** of the GW distance using the 1D OT fast solver

## ④ OTT

- ✓ New OT formulation to handle **tensors** of arbitrary orders
- ✓ Adaptation of the previous solver to solve it in a reasonable time

# Further works

## Related to the contributions

- ① Deep version of MLOT
- ② Extend the use of stability with min minimax
- ③ Convergence bound of SaGroW<sup>KL</sup> to stationary points
- ④ Fused OTT

## Some perspectives for the OT's application to ML

- ① GW  $\Leftrightarrow$  Linear GAN
- ② Under which condition on  $\mathcal{L}$ , GW has a sparse solution ?

- Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1988). Network flows.
- Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2007). Analysis of representations for domain adaptation. In *NIPS*.
- Brandes, U., Gaertler, M., and Wagner, D. (2003). Experiments on graph clustering algorithms. In *European Symposium on Algorithms*.
- Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2017). Optimal transport for domain adaptation. *PAMI*.
- Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*.
- Dhouib, S., Redko, I., Kerdoncuff, T., Emonet, R., and Sebban, M. (2020). A swiss army knife for minimax optimal transport. In *International Conference on Machine Learning*.
- Frank, M., Wolfe, P., et al. (1956). An algorithm for quadratic programming. *Naval research logistics quarterly*.
- Gong, B., Shi, Y., Sha, F., and Grauman, K. (2012). Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*.

- Kantorovich, L. (1942). On the translocation of masses. *Doklady of the Academy of Sciences of the USSR*, 37:199–201.
- Kerdoncuff, T., Emonet, R., and Sebban, M. (2020). Metric learning in optimal transport for domain adaptation. In Bessiere, C., editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2162–2168. International Joint Conferences on Artificial Intelligence Organization. Main track.
- Kerdoncuff, T., Emonet, R., and Sebban, M. (2021). Sampled gromov wasserstein. *Machine Learning*.
- Memoli, F. (2007). On the use of Gromov-Hausdorff Distances for Shape Comparison. In Botsch, M., Pajarola, R., Chen, B., and Zwicker, M., editors, *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association.
- Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie royale des sciences de Paris*.
- Mutapcic, A. and Boyd, S. P. (2009). Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods and Software*.
- Peyré, G., Cuturi, M., and Solomon, J. (2016). Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*.

## Biblio III

- Reddi, S. J., Sra, S., Póczos, B., and Smola, A. (2016). Stochastic frank-wolfe methods for nonconvex optimization. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*.
- Redko, I., Habrard, A., and Sebban, M. (2017). Theoretical analysis of domain adaptation with optimal transport. In *ECML PKDD*.
- Redko, I., Vayer, T., Flamary, R., and Courty, N. (2020). Co-optimal transport. In *NeurIPS 2020-Thirty-four Conference on Neural Information Processing Systems*.
- Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*.
- Villani, C. (2008). *Optimal transport: old and new*.
- Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *JMLR*.
- Xu, H., Luo, D., and Carin, L. (2019). Scalable gromov-wasserstein learning for graph partitioning and matching. In *Advances in neural information processing systems*.
- Zhong, E., Fan, W., Yang, Q., Verschueren, O., and Ren, J. (2010). Cross validation framework to choose amongst models and datasets for transfer learning. In *ECML PKDD*.

# Conclusions

- Problem:

- We tackle the challenging unsupervised DA problem
- OT for DA always relies on the euclidean distance

- ✓ Solutions:

- We provide a metric dependent bound on the target error
- We use labeled sources to learn a class-informed metric
- We project, independently, the source and target point into subspaces
- We show a relation between OT and PCA, and use PCA as initialization

- ✓ Results:

- Competitive against various baselines with a Reverse Validation method
- Improve the performance of OTDA for different set of hyperparameters

# Continuous GW

