Partial Optimal Transport with Applications to Positive-Unlabeled (PU) Learning

Mokhtar Z. Alaya

SIAM Conference on Mathematics of Data Science

September 2022

Joint work with:

Laetitia Chapel

IRISA, UBS

Gilles Gasso

INSA, URN

Outline

1. Background on OT

2. Partial Wasserstein OT

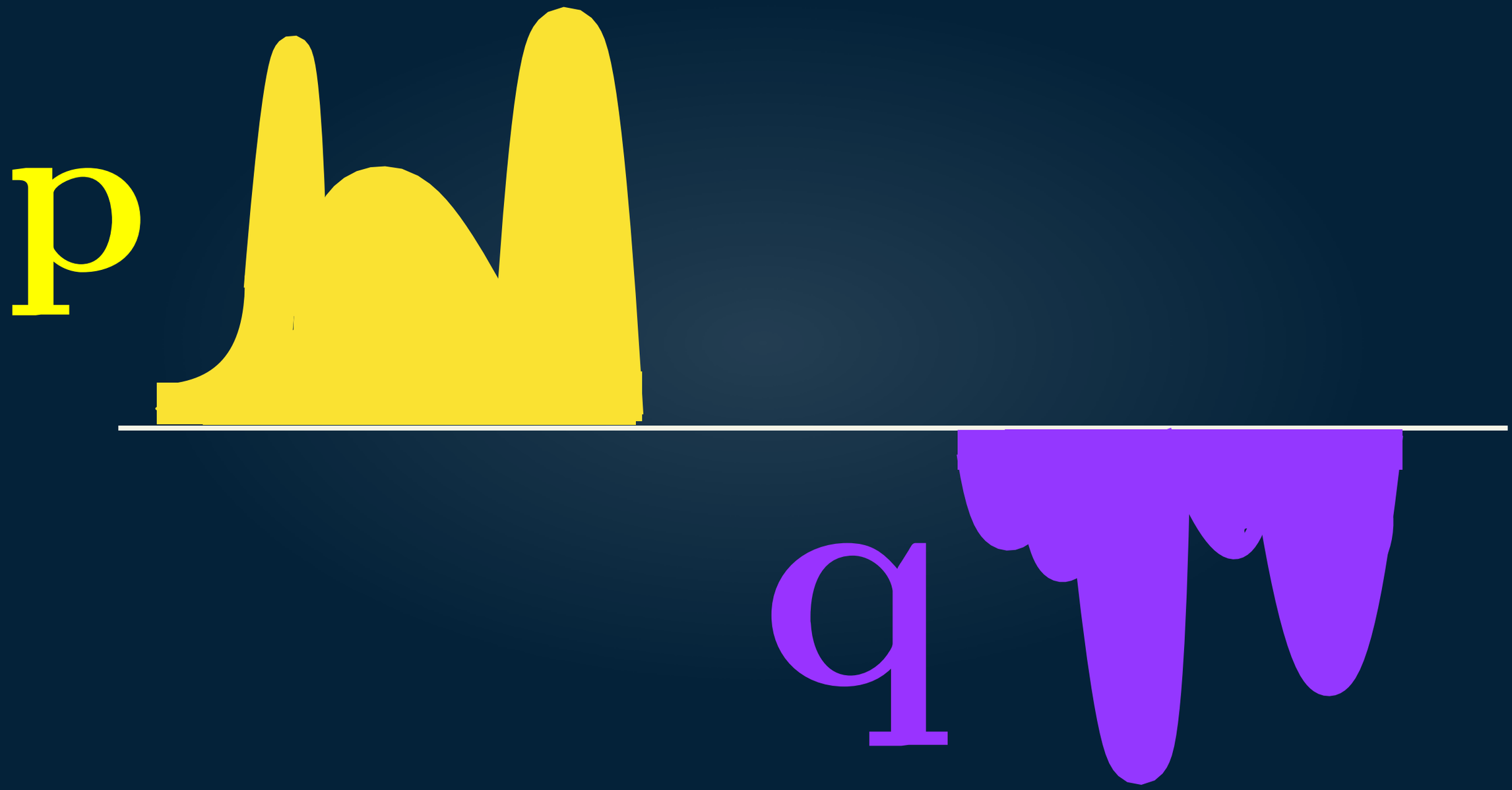
3. Partial OT for PU Learning

4. Numerical experiments

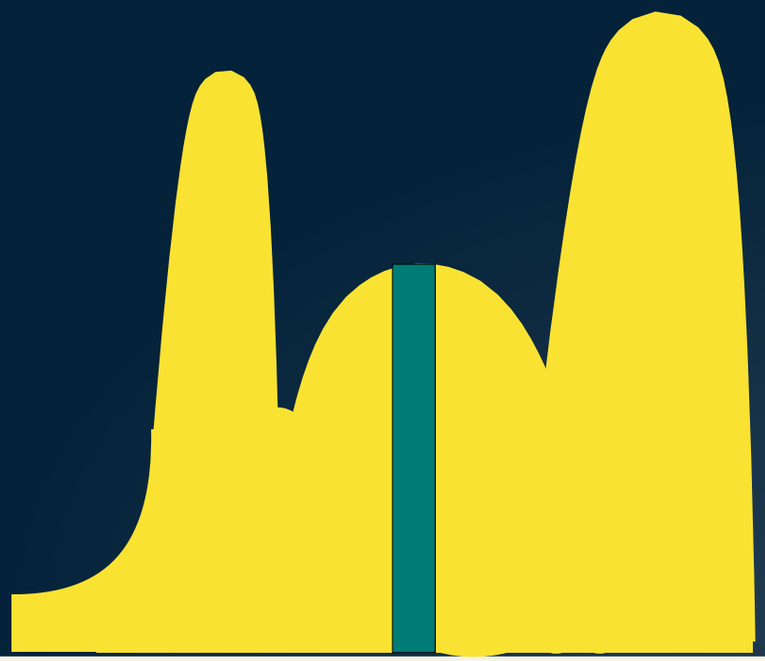
1. Background on OT

OT is …

- A method for comparing probability distributions with the ability to incorporate spatial information.

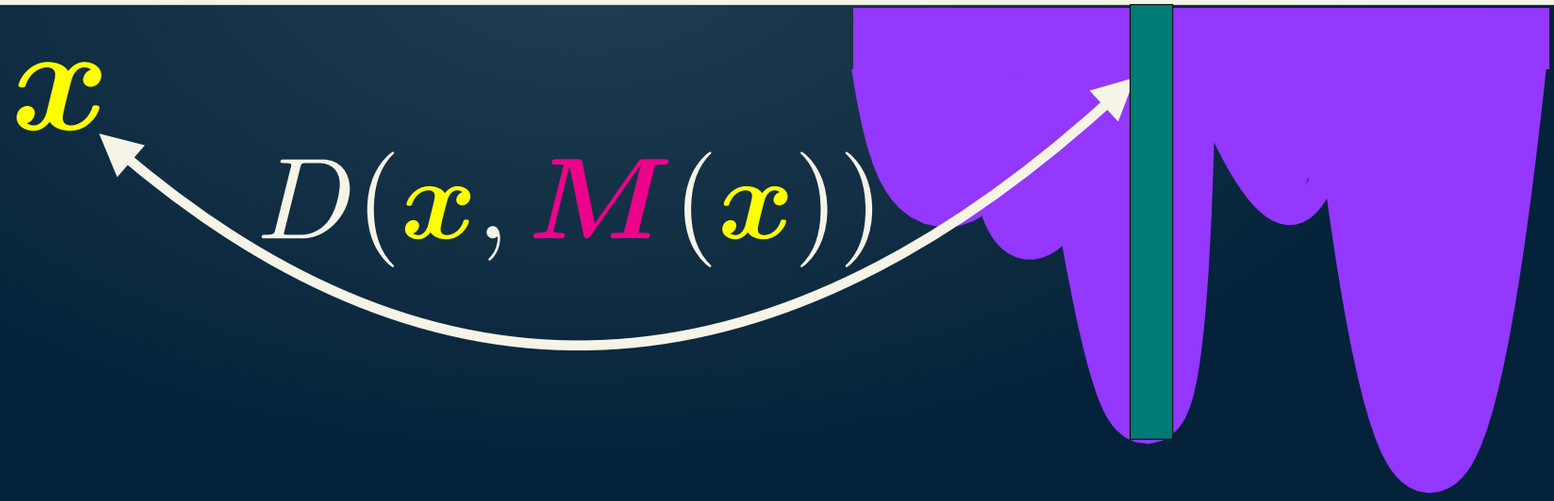


OT is …



$$M_{\#}\mathbf{p} = \mathbf{q}$$

$$M(x)$$



$$\min_{M_{\#}\mathbf{p}=\mathbf{q}} \int D(x, M(x))\mathbf{p}(dx)$$

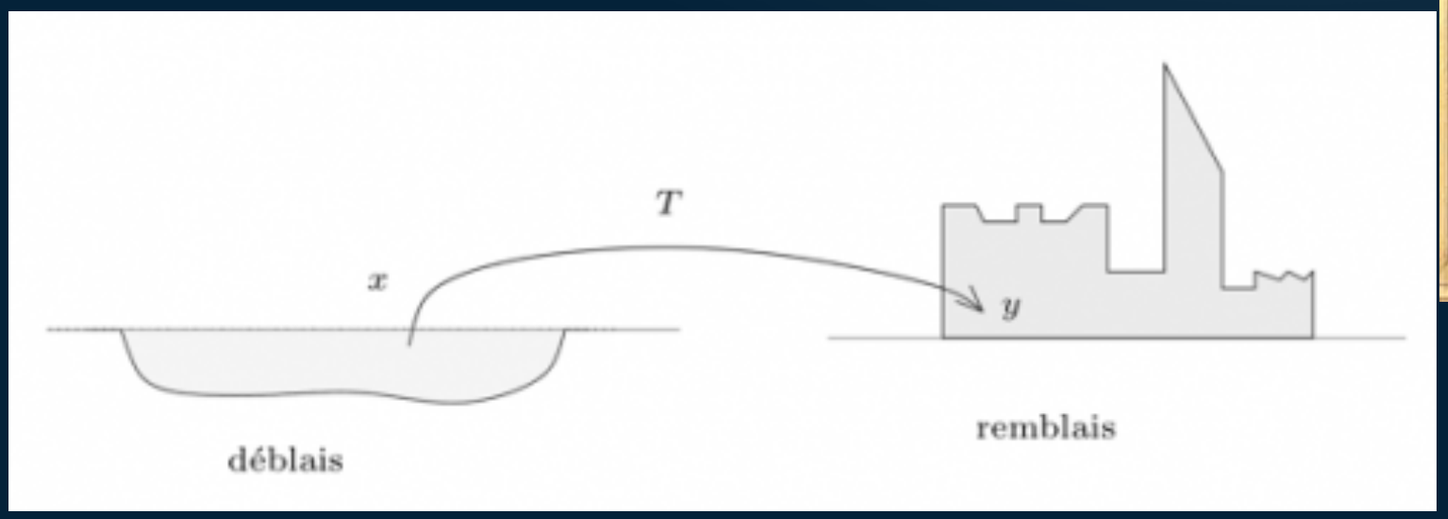
OT is …

666 · MÉMOIRES DE L'ACADÉMIE ROYALE

MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.

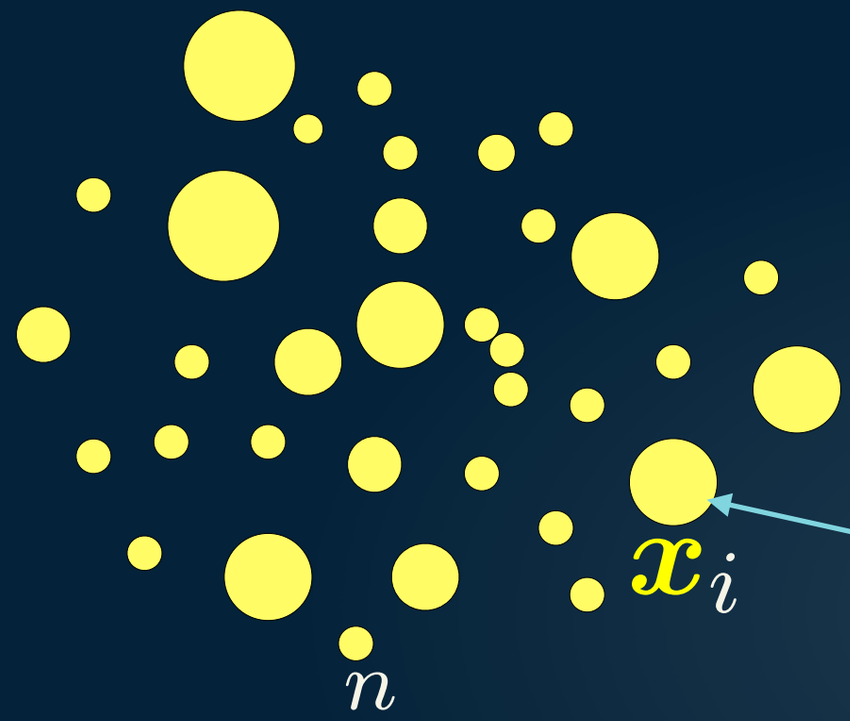
Par M. MONGE.

LORSQU'ON doit transporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit transporter, & le nom de *Remblai* à l'espace qu'elles doivent occuper après le transport.



Gaspard Monge
(1746 - 1818)

Discrete OT Framework



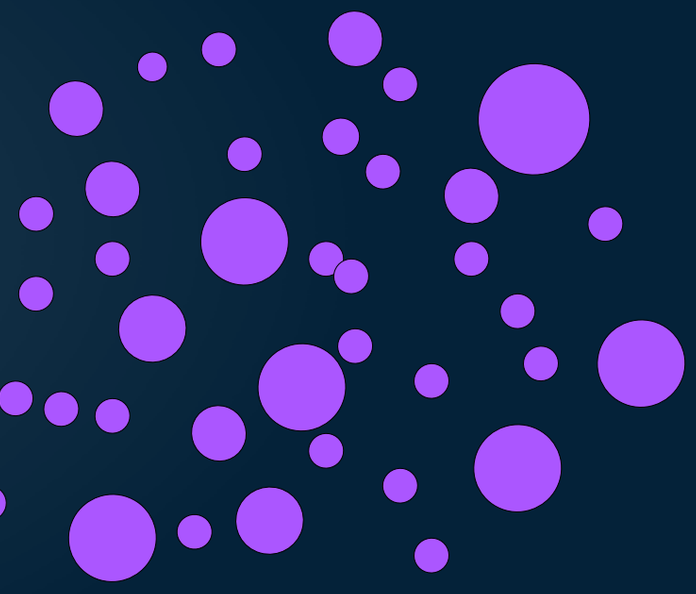
$\boldsymbol{x}_i$

n

$$\mathbf{p} = \sum_{i=1}^{n} \mathbf{p}_i \delta_{\boldsymbol{x}_i}$$

$$D^p(\boldsymbol{x}_i, \boldsymbol{y}_j)$$

$$\mathbf{q} = \sum_{j=1}^{m} \mathbf{q}_j \delta_{\boldsymbol{y}_j}$$



$\boldsymbol{y}_j$

m

$\boldsymbol{C}$

Cost Matrix

n

$$\boldsymbol{C}_{ij} = \left(D^p(\boldsymbol{x}_i, \boldsymbol{y}_j)\right)$$

Distance
(e.g. Euclidean distance)

Discrete OT: Monge's Formula

$$\min_{M_{\#}\mathbf{p}=\mathbf{q}} \sum_i D^p(\boldsymbol{x}_i, M(\boldsymbol{x}_i))\mathbf{p}_i$$

$$\forall i \in \{1,\dots,n\}, \mathbf{p}_i = \sum_{j:M(\boldsymbol{y}_j)=\boldsymbol{x}_i} \mathbf{q}_j$$

Strict: Deterministic Assignments

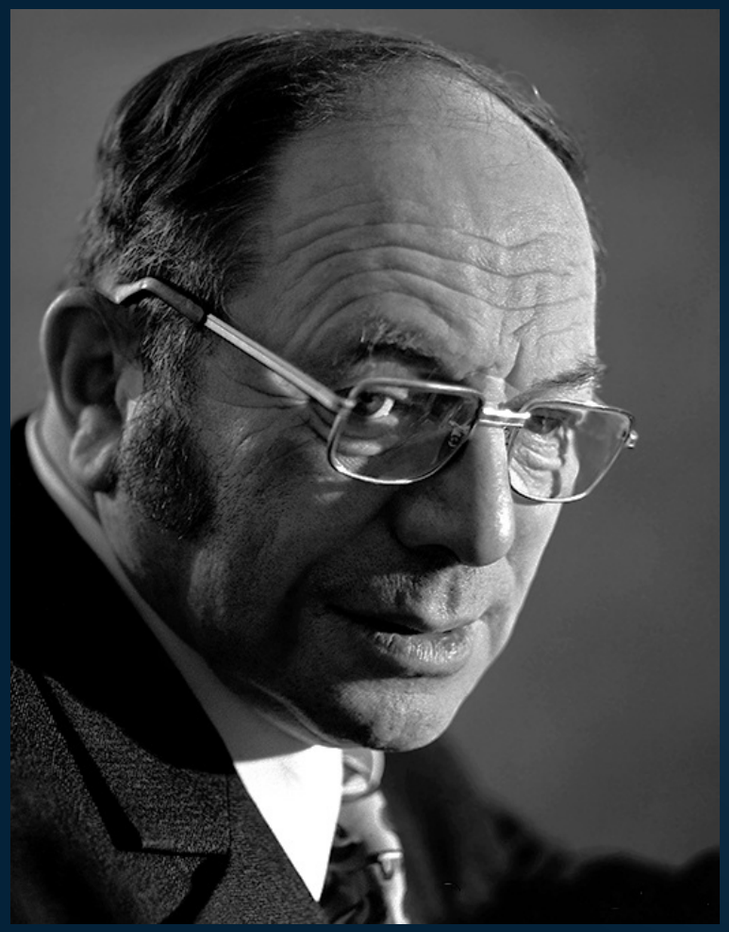
$$\forall j \in \{1,\dots,m\}, \mathbf{q}_j = \sum_{i:M(\boldsymbol{x}_i)=\boldsymbol{y}_j} \mathbf{p}_i$$

Uniform weights

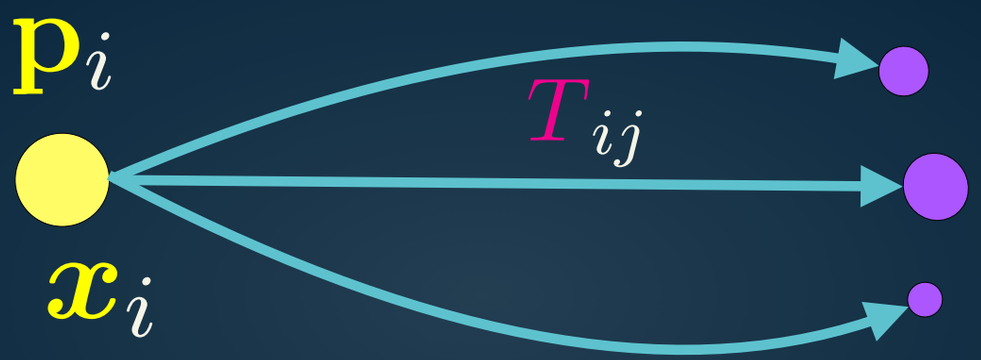
$$\min_{\sigma\in\mathfrak{S}_n} \frac{1}{n}\sum_{i=1}^{n} D^p\left(\boldsymbol{x}_i, y_{\sigma(i)}\right)$$

non-convex + combinatorial + non-existant

Discrete OT Framework: Kantorovich's Formula



Leonid Kantorovich
(1912-1986)



$$\mathbf{p}_i = \sum_{i=1}^{m} T_{ij}$$

Relaxed: Fractional Assignments

Probabilistic couplings set (Transport Polytope)

$$\boldsymbol{\Pi}(\mathbf{p}, \mathbf{q}) = \{\boldsymbol{T} \in \mathbb{R}_+^{n \times m}, \boldsymbol{T}\mathbf{1}_m = \mathbf{p}, \boldsymbol{T}^\top \mathbf{1}_n = \mathbf{q}\}$$

Mass conservation constraints

Discrete OT: Monge-Kantorovich / Wasserstein Distance

- Computing OT between $\mathbf{p}$ and $\mathbf{q}$ amounts to solving a linear problem:

Monge-Kantorovich / Wasserstein Distance

$$\mathcal{W}_p^p(\mathbf{p}, \mathbf{q}) = \min_{\boldsymbol{T} \in \boldsymbol{\Pi}(\mathbf{p}, \mathbf{q})} \{\langle \boldsymbol{C}, \boldsymbol{T} \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \boldsymbol{C}_{ij} \boldsymbol{T}_{ij}\}$$

- **Classical (balanced)** OT distances require that all the mass has to be transported and the two distributions have the same total probability mass i.e.:

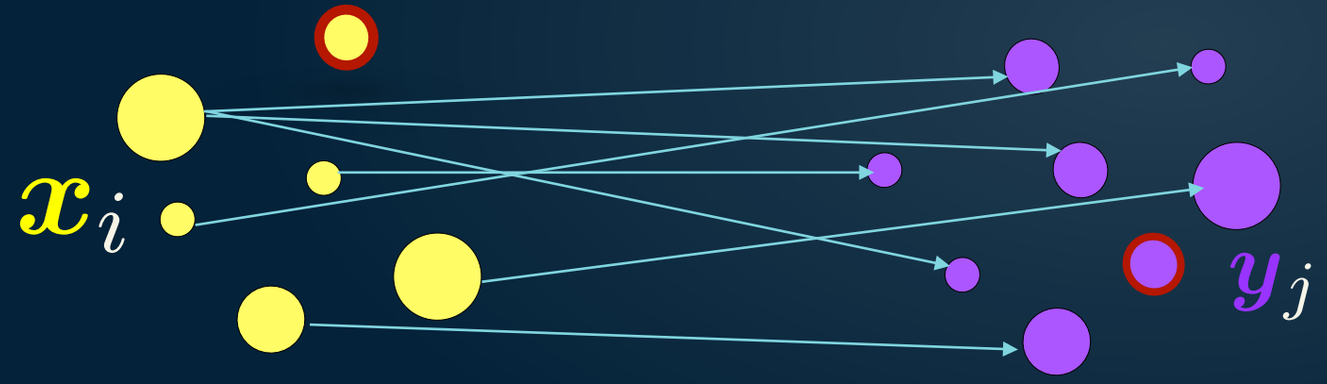
$$||\mathbf{p}||_1 = ||\mathbf{q}||_1$$

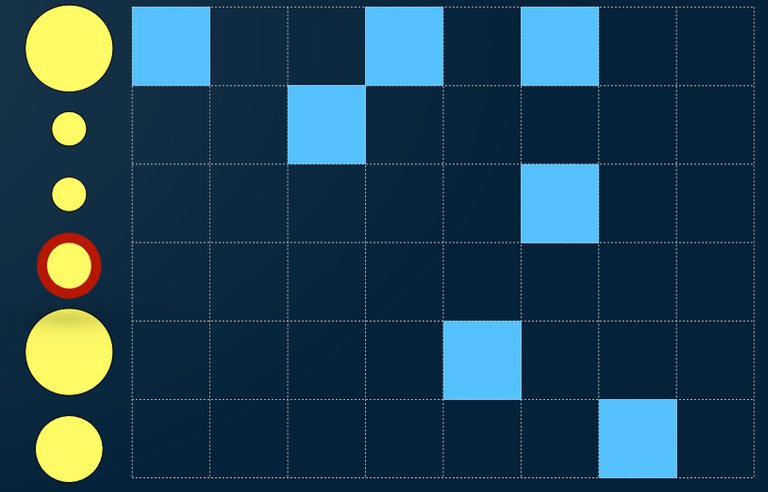
2. Partial Wasserstein OT

Partial OT:
Partial Wasserstein Distance

- Partial OT problem focuses on transporting only a fraction

$$0 \leq \boldsymbol{s} \leq \min(\|\mathbf{p}\|_1, \|\mathbf{q}\|_1)$$





- The set of admissible coupling becomes

$$\boldsymbol{\Pi}^u(\mathbf{p}, \mathbf{q}) = \{\boldsymbol{T} \in \mathbb{R}_+^{n \times m}, \boldsymbol{T}\mathbf{1}_m \leq \mathbf{p}, \boldsymbol{T}^\top \mathbf{1}_n \leq \mathbf{q}, \mathbf{1}_n^\top \boldsymbol{T} \mathbf{1}_m = \boldsymbol{s}\}$$

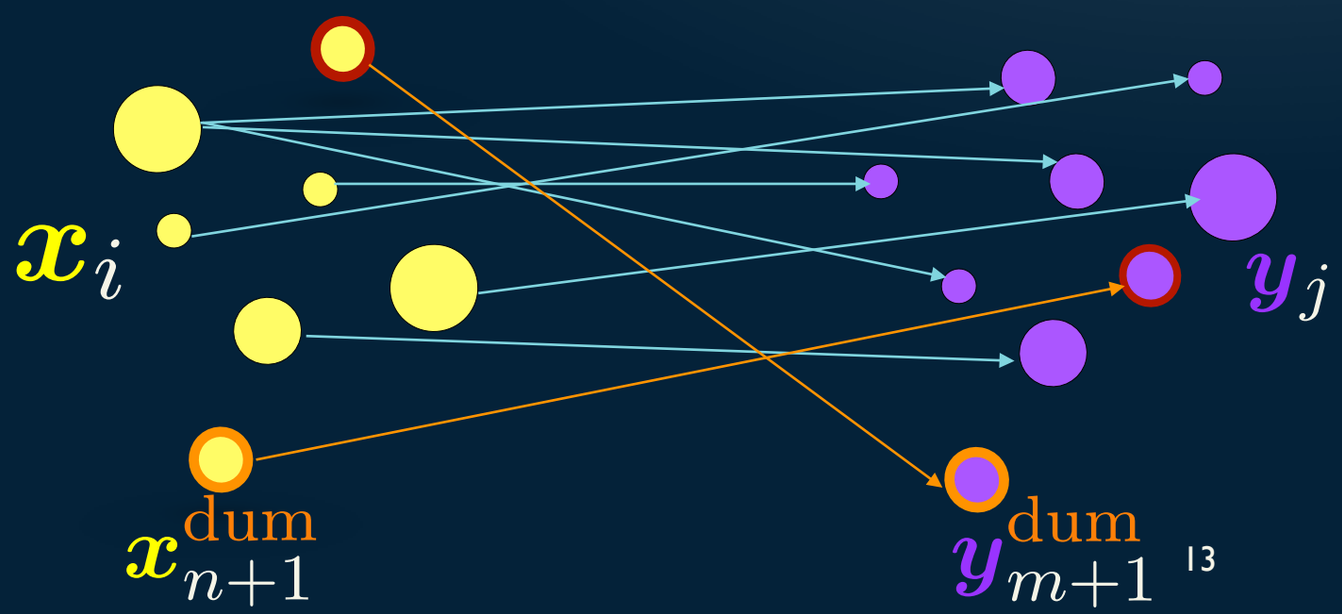
Partial OT: Partial Wasserstein Distance

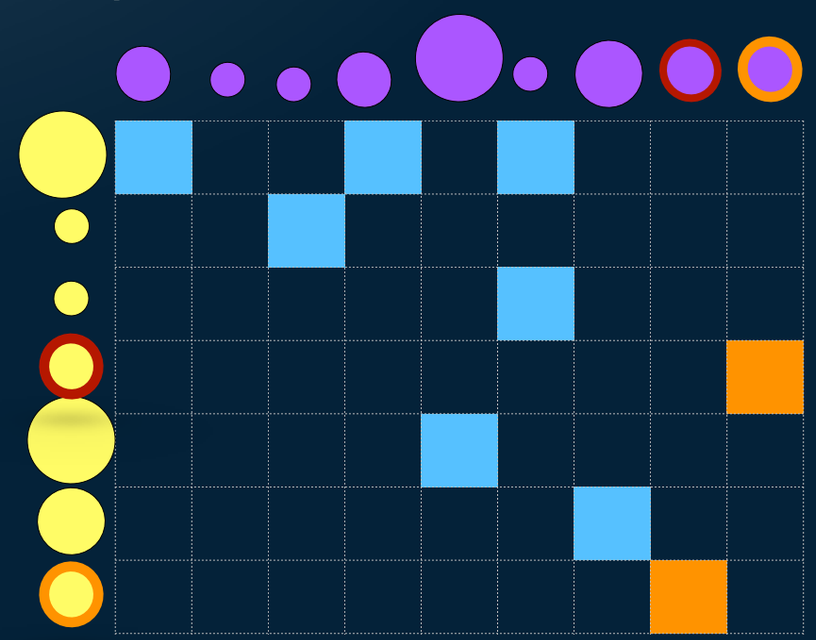
- The partial-Wasserstein distance reads as:

$$\mathcal{PW}_p^p(\mathbf{p}, \mathbf{q}) = \min_{\boldsymbol{T} \in \boldsymbol{\Pi}^u(\mathbf{p},\mathbf{q})} \{\langle \boldsymbol{C}, \boldsymbol{T} \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \boldsymbol{C}_{ij} \boldsymbol{T}_{ij}\}$$

[Caffareli and McCann, '10; Figalli, '10, Benamou et al., '15; Chizat et al., '18]

- **Solution:** we propose to directly solve the exact partial Wasserstein by adding **dummy** or **virtual** points $\boldsymbol{x}_{n+1}^{\text{dum}}$ and $\boldsymbol{y}_{m+1}^{\text{dum}}$:





Partial OT: Partial Wasserstein Distance

- We extend the cost matrix as follows:

$$\overline{C} = \begin{bmatrix} C & \xi \mathbf{1}_m \\ \xi \mathbf{1}_n^\top & 2\xi + A \end{bmatrix} \quad \text{for some } A > 0 \text{ and } \xi \geq 0,$$

and the mass probability vectors as:

$$\bar{\mathbf{p}} = [\mathbf{p}, \|\mathbf{q}\|_1 - s] \qquad \bar{\mathbf{q}} = [\bar{\mathbf{q}}, \|\mathbf{p}\|_1 - s]$$

- Hence,

$$\bar{\mathbf{p}} = \sum_{i=1}^{n} \mathbf{p}_i \delta_{\boldsymbol{x}_i} + (\|\mathbf{q}\|_1 - s)\delta_{\boldsymbol{x}_{n+1}^{\mathrm{dum}}}$$

$$\bar{\mathbf{q}} = \sum_{j=1}^{m} \mathbf{q}_j \delta_{\boldsymbol{y}_j} + (\|\mathbf{p}\|_1 - s)\delta_{\boldsymbol{y}_{m+1}^{\mathrm{dum}}}$$

Exact Partial Wasserstein Distance

- Let us define $\overline{T}^*$ the solution of the extended problem with $(\overline{C}, \bar{\mathbf{p}}, \bar{\mathbf{q}})$. Namely:

$$\overline{T}^* \in \mathcal{W}_p^p(\bar{\mathbf{p}}, \bar{\mathbf{q}}) := \min_{\bar{T} \in \mathbf{\Pi}(\bar{\mathbf{p}}, \bar{\mathbf{q}})} \langle \overline{C}, \overline{T} \rangle.$$

Proposition

Assume that $A > \max(C_{ij})$ and ξ is bounded, one has:

$$\mathcal{W}_p^p(\bar{\mathbf{p}}, \bar{\mathbf{q}}) - \mathcal{PW}_p^p(\mathbf{p}, \mathbf{q}) = \xi(\|\mathbf{p}\|_1 + \|\mathbf{q}\|_1 - 2\,\boldsymbol{s}).$$

The partial optimum transport plan T^* of the partial Wasserstein problem is the optimum transport plan $\overline{T}^*$ deprived from its last row and column.

3. OT for PU Learning

Overview of PU Learning

- PU learning is a variant of classical binary classification problem.
- Training data consists of only positive points **Pos** and testing data is composed of unlabeled positives and negatives **Unl**.

Unl = $\{\boldsymbol{x}_i^U\}_{i=1}^{n_U} = \{\boldsymbol{x}_i^{U,+}\} \cup \{\boldsymbol{x}_i^{U,-}\}$ **Pos** = $\{\boldsymbol{x}_i^P\}_{i=1}^{n_P}$

[Liu et al., '03; Denis et al., '05; Elkan and Noto, '08; Du Pelessis, '15; Bekker and Davis, '20]

- The true proportion of positives within **Unl**, called class prior, is assumed to be known and given by:

$$\pi = \mathbb{P}(y = +1 | o = 0)$$

unobserved positive sample

Assumptions to enable PU learning: Label mechanism assumptions

- The class prior plays an important role in PU learning and many PU learning methods require it as an input.

Selected completely at random (**SCAR**) assumption [Elkan and Noto, '08]

- **SCAR**: **Pos** samples are selected uniformly at random, independent from their features, from the positive distribution, i.e.

$$\{\boldsymbol{x}_i^P\} \quad i.i.d \sim \mathbb{P}[\boldsymbol{x}|\boldsymbol{y}=1]$$

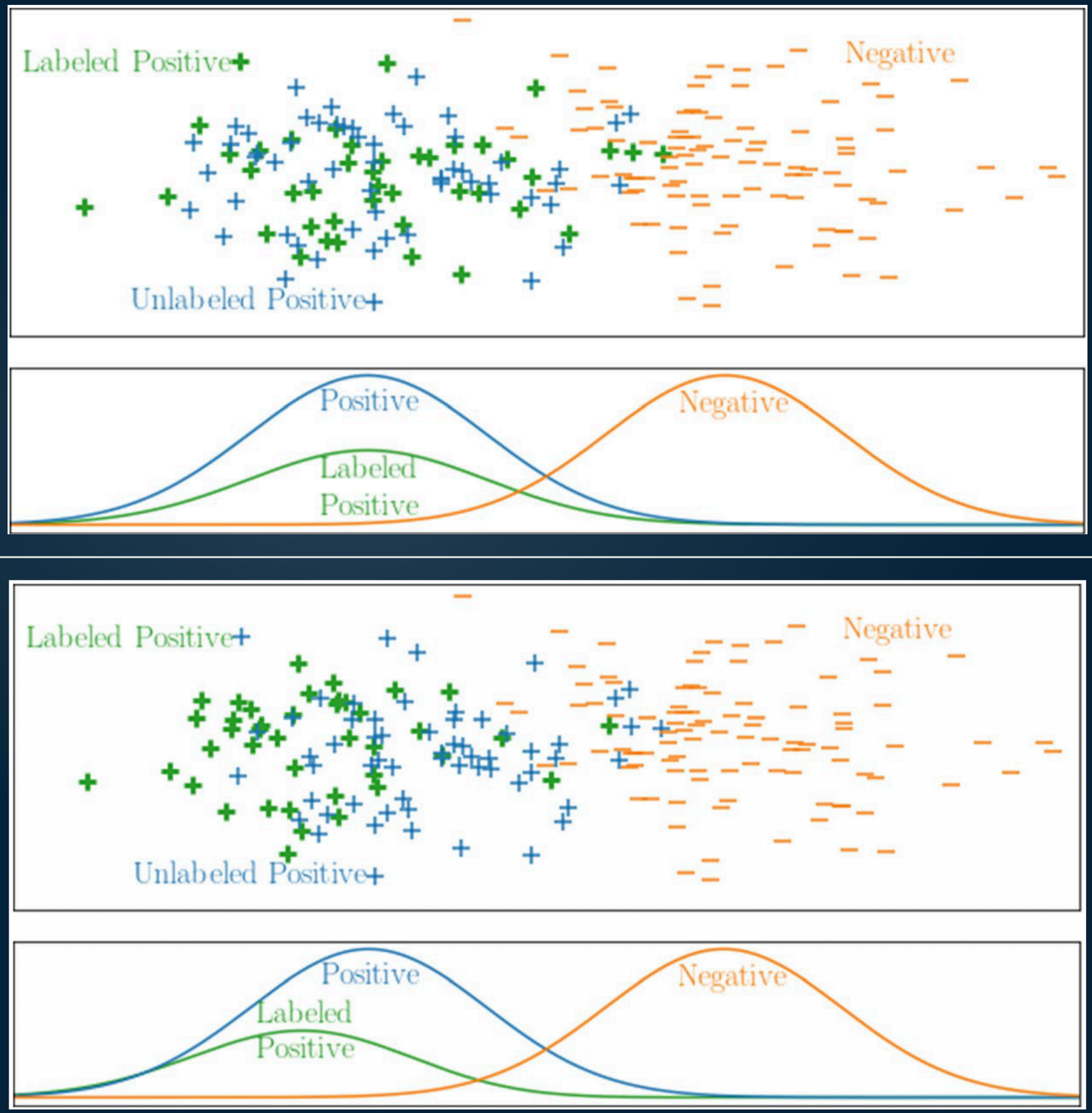
Selected at random (**SAR**) assumption [Bekker and Davis, '18]

- **SAR: Pos** samples are a biased sample from the positive distribution, where the bias completely depends on the features and it is defined by the propensity score

$$e(\boldsymbol{x}) = \mathbb{P}[o=1|\boldsymbol{x},\boldsymbol{y}=1]$$

Label mechanism assumptions

SCAR assumption



SAR assumption

Credit images: [Bekker and Davis, '20]

PU Learning as a Partial OT

- PU leaning can be broadcasted as a partial-like OT problem:
  - **Unl:** source distribution; **Pos**: target distribution.
  - Mass to be transported: $\boldsymbol{s} = \pi$
  - $n = n_U, m = n_P, \mathbf{p}_i = \frac{1}{n_U}, \mathbf{q}_j = \frac{\pi}{n_P}$
- We look for an optimal transport plan that belongs to the following set of couplings:

$$\boldsymbol{\Pi}^{PU}(\mathbf{p}, \mathbf{q}) = \{\boldsymbol{T} \in \mathbb{R}_+^{n_U \times n_P}, \boldsymbol{T}\mathbf{1}_{n_P} = \{\mathbf{p}, \mathbf{0}\}, \boldsymbol{T}^\top \mathbf{1}_{n_U} \leq \{\mathbf{q}, \mathbf{0}\}, \mathbf{1}_{n_U}^\top \boldsymbol{T}\mathbf{1}_{n_P} = \pi\}$$

- To avoid matching part of the mass of unlabeled negative with positive,

$\boldsymbol{T}\mathbf{1}_{n_P} = \{\mathbf{p}, \mathbf{0}\}$ means that $\sum_{j=1}^{n_P} \boldsymbol{T}_{ij} = \mathbf{p}_i, \forall i$ exactly or 0.

- We aim at solving:

$$\mathcal{PUW}_p^p(\mathbf{p}, \mathbf{q}) = \min_{\boldsymbol{T} \in \boldsymbol{\Pi}^{PU}(\mathbf{p}, \mathbf{q})} \{\langle \boldsymbol{C}, \boldsymbol{T} \rangle = \sum_{i=1}^{n_U} \sum_{j=1}^{n_P} \boldsymbol{C}_{ij} \boldsymbol{T}_{ij}\}$$

PU Learning as a Partial OT

- To enforce the condition $\boldsymbol{T}\mathbf{1}_{n_P} = \{\mathbf{p}, \mathbf{0}\}$ we adopt a regularised point of view of the partial OT problem: [Courty et al., '17], we then solve the following:

$$\overline{\boldsymbol{T}}^* \in \arg\min_{\overline{\boldsymbol{T}} \in \boldsymbol{\Pi}(\bar{\mathbf{p}}, \bar{\mathbf{q}})} \sum_{i=1}^{n_U+1} \sum_{j=1}^{n_P+1} \overline{\boldsymbol{C}}_{ij} \overline{\boldsymbol{T}}_{ij} + \eta \Omega(\overline{\boldsymbol{T}}),$$

where

$$\mathbf{p}_i = \frac{1-\alpha}{n_U}, \mathbf{q}_j = \frac{\boldsymbol{s}+\alpha}{n_P}, \eta \geq 0 \text{ (regularisation parameter).}$$

- $\alpha \in [0, 1 - \boldsymbol{s}]$ is the percentage of **Pos** that we assume to be noisy (that is to say we do not want to map them to point of **Unl**).

PU Learning as a Partial OT

- We choose

$$\Omega(\overline{\boldsymbol{T}}) = \sum_{i=1}^{n=n_U} (\|\overline{\boldsymbol{T}}_{i(:m)}\|_2 + (\overline{\boldsymbol{T}}_{i(m+1)})^2)$$

- This group Lasso regularisation leads to a sparse transportation map and enforces each of **Unl** samples to be mapped to only **Pos** sample or to the dummy point $\boldsymbol{x}^{\mathrm{dum}}_{n_P+1}$.

Proposition

Assume that $A > 0, \xi$ is a constant, there exits a large $\eta > 0$ such that

$$\mathcal{W}^{*p}_p(\bar{\mathbf{p}}, \bar{\mathbf{q}}) - \mathcal{PUW}^p_p(\mathbf{p}, \mathbf{q}) = \xi(1 - \boldsymbol{s}),$$

where

$$\mathcal{W}^{*p}_p(\bar{\mathbf{p}}, \bar{\mathbf{q}}) = \sum_{i=1}^{n_U+1} \sum_{j=1}^{n_P+1} \overline{\boldsymbol{C}}_{ij} \overline{\boldsymbol{T}}^*_{ij}$$

and $\overline{\boldsymbol{T}}^*$ is a solution of the regularised problem.

4. Numerical experiments

Partial Wasserstein in a PU Learning under SCAR Assumption (UCI data)

- One class that is positive, the other ones are negatives, drawn randomly.
- Average accuracy rates on various UCI datasets. **P-W 0** indicates no noise and **P-W 0.025** stands for a noise level.

DATASET	π	PU	PUSB	P-W 0	P-W 0.025
MUSHROOMS	0.518	91.1	90.8	96.3	**96.4**
SHUTTLE	0.786	90.8	90.3	**95.8**	94.0
PAGEBLOCKS	0.898	92.1	90.9	**92.2**	91.6
USPS	0.167	95.4	95.1	**98.3**	98.1
CONNECT-4	0.658	**65.6**	58.3	55.6	61.7
SPAMBASE	0.394	**84.3**	84.1	78.0	76.4

Partial Wasserstein in a PU Learning under SAR Assumption

- Following [Arjovsky et al., '19], we construct a colored version of MNIST: each digit is colored, either in green or red, with a probability of 90% to be colored in red.
- The **Unl** dataset is then mostly composed of red digits, while **Pos** dataset contains mostly green instances.

DATASET	π	PU	PUSB	P-W 0	P-W 0.025
ORIGINAL MNIST	0.1	97.9	97.8	**98.8**	98.6
COLORED MNIST	0.1	87.0	80.0	91.5	91.5

Take home message

- Consider partial Wasserstein distance to solve PU learning problem.
- Partial Wasserstein distance compete and sometimes outperforms the SOTA of PU learning methods.
- We also studied the case of partial Gromov-Wasserstein distance. Our approach for this setting is based on Franck-Wolf algorithm.
- An extension of this work can be tackle the case of partial sliced-OT that leads to lower the computational complexities of calculating an OT plans.

References

Bekker, J. and J. Davis (2018). Learning from positive and unlabeled data under the selected at random assumption. In *Proceedings of Machine Learning Research*, Volume 94, pp. 8–22.

Chapel L, Alaya MZ, Gasso G. Partial Optimal Tranport with applications on Positive-Unlabeled Learning. Adv Neural Inf Process Syst 2020;33:2903–13.

Courty, N., R. Flamary, D. Tuia, and A. Rakotomamonjy (2017). Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence 39*(9), 1853–1865.

Du Plessis, M., G. Niu, and M. Sugiyama (2014). Analysis of learning from positive and unlabeled data. In *Advances in neural information processing systems*, pp. 703–711.

Figalli, A. (2010). The optimal partial transport problem. *Archive for Rational Mechanics and Analysis 195*(2), 533–560.

Flamary, R. and N. Courty (2017). POT python optimal transport library.

Kato, M., T. Teshima, and J. Honda (2019). Learning from positive and unlabeled data with a selection bias. In *International Conference on Learning Representations*.

Thank you!